Protein Science (1996), *5*:2203–2216. Cambridge University Press. Printed in the USA.

Cleavage site analysis in picornaviral polyproteins: Discovering cellular targets by neural networks

NIKOLAJ BLOM,[1] JAN HANSEN,[1] DIETER BLAAS,[2] AND SØREN BRUNAK[1]

[1] Center for Biological Sequence Analysis, The Technical University of Denmark
[2] Institute of Biochemistry, University of Vienna, Dr. Bohr-Gasse 9/3, A-1030 Vienna, Austria

(RECEIVED May 28, 1996; ACCEPTED September 3, 1996)

Abstract

Picornaviral proteinases are responsible for maturation cleavages of the viral polyprotein, but also catalyze the degradation of cellular targets. Using graphical visualization techniques and neural network algorithms, we have investigated the sequence specificity of the two proteinases 2A^{pro} and 3C^{pro}. The cleavage of VP0 (giving rise to VP2 and VP4), which is carried out by a so-far unknown proteinase, was also examined. In combination with a novel surface exposure prediction algorithm, our neural network approach successfully distinguishes known cleavage sites from noncleavage sites and yields a more consistent definition of features common to these sites. The method is able to predict experimentally determined cleavage sites in cellular proteins. We present a list of mammalian and other proteins that are predicted to be possible targets for the viral proteinases. Whether these proteins are indeed cleaved awaits experimental verification. Additionally, we report several errors detected in the protein databases.

A computer server for prediction of cleavage sites by picornaviral proteinases is publicly available at the e-mail address NetPicoRNA@cbs.dtu.dk or via WWW at http://www.cbs.dtu.dk/services/NetPicoRNA/.

Keywords: cleavage site prediction; neural networks; picornavirus; proteinase; surface exposure

Members of the picornavirus family express their genomic RNA as a single polyprotein that is proteolytically processed to the mature polypeptides. At least three proteinases are required for the individual protein components to be released (reviewed in Krausslich & Wimmer, 1988; Hellen et al., 1989; Lawson & Semler, 1990). The primary cleavage, which severs the capsid precursor P1 from the nonstructural region P2–P3, is performed cotranslationally by the viral proteinase 2A^{pro} in enteroviruses and human rhinoviruses (HRVs; see Fig. 1). Most of the remaining cleavages are catalyzed by the viral proteinase, 3C^{pro}. In cardio-, hepato-, and aphthoviruses, which also belong to the picornavirus family, the L-proteinase performs functions similar to those of 2A^{pro}, resulting in a somewhat different cleavage scheme (see Fig. 1). Concomitantly with RNA encapsidation, VP0 is cleaved to VP4 and VP2; it is believed that the RNA itself exerts a catalytic function in this event (Arnold et al., 1987; Harber et al., 1991; Bishop & Anderson, 1993; Basavappa et al., 1994).

In addition to processing of the viral polyprotein, the proteinases also cleave cellular targets. When infected with poliovirus, at least nine acidic and five basic cellular proteins were shown to be degraded in two-dimensional gel electrophoresis (Urzanqui & Carrasco, 1989). The degradation of cellular proteins seems to be part of the viral attack mechanism, leading to *host cell shut-off*—a decrease in cellular transcription and translation that has no influence on viral replication. The best-studied event is the cleavage of the eukaryotic initiation factor 4G (eIF-4G), which is required for cap-dependent translation of cellular mRNA. This protein is degraded by 2A^{pro} in entero- and rhinovirus or by L-proteinase in aphthovirus, enabling cap-independent translation of viral RNA, while translation of capped cellular mRNA is suppressed. In addition to eIF-4G, other proteins were found to be cleaved by viral proteinases. For example, 3C^{pro} of foot-and-mouth disease virus (FMDV) cleaves histone H3 of baby hamster kidney cells (Tesar & Marquardt, 1990); 3C^{pro} of poliovirus cleaves and thereby in-

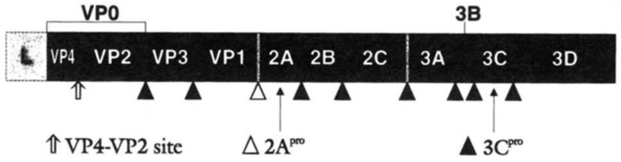

Fig. 1. Prototype picornavirus polyprotein and genetic map. P1 is the structural precursor; P2 and P3 are enzymatic precursors. The location of the proteinases 2A^{pro} and 3C^{pro} and their respective cleavage sites, as well as the VP4–VP2 autocatalytic site, are indicated by symbols. The L-proteinase, present in aphtho- and cardiovirus, is shown at the N terminus of the polyprotein.

Reprint requests to: Søren Brunak, Center for Biological Sequence Analysis, Department of Chemistry, Building 207, The Technical University of Denmark, DK-2800 Lyngby, Denmark; e-mail brunak@cbs.dtu.dk.

activates the transcription factors TFIIIC and TFIID (Clark & Dasgupta, 1990; Clark et al., 1991, 1993) and the microtubule-associated protein 4 (MAP-4) (Joachims & Etchison, 1992; Joachims et al., 1995) in HeLa cells.

The specificity of picornaviral proteinases was first studied in poliovirus and showed that $2A^{pro}$ preferentially cleaves between tyrosine–glycine (Y–G) pairs, whereas $3C^{pro}$ cleaves between glutamine–glycine (Q–G) pairs (reviewed in Krausslich & Wimmer, 1988). However, not all of the Y–G and Q–G dipeptides in the polyprotein are cleaved, and, in picornaviruses other than poliovirus, the Y–G and Q–G pairs are not always conserved at the cleavage sites. This suggested that substrate recognition and cleavage required at least more than two adjacent particular amino acids. The topology and accessibility of potential cleavage sites are certainly further determinants; e.g., it has been suggested that efficient cleavage pairs are located between beta-barrels (Ypma-Wong et al., 1988).

As mentioned above, the primary cleavage (by $2A^{pro}$) occurs cotranslationally, and thus most probably on a part of the protein that is not yet fully folded. However, in addition to this intramolecular cleavage (in *cis*), an additional intermolecular cleavage (in *trans*) in the P3 region has been noted in human rhinovirus serotype 1A and in poliovirus (McLean et al., 1976; Lee & Wimmer, 1988). The purpose of this alternative cleavage site is still unknown because it has been shown to be dispensable for viral replication (Hellen et al., 1992). On the other hand, the alternative cleavage of P3 may be similar to the cleavage of cellular targets, which have certainly acquired their final native structure when approached by the viral proteinases. However, because 3C, 3D, and the alternatively cleaved 3C′ and 3D′ are all rather stable proteins, it is also possible that the cleavage that occurs first leads to loss of accessibility of the second cleavage site.

Alignment of all known cleavage sites for either of the viral proteinases failed to yield a clear consensus sequence, although some common determinants were found. Because of the complexity of the cleavage site sequences, we decided to use neural network algorithms to attempt detection of specific features of the cleavage sites in question. The algorithm should distinguish cleavage sites from noncleavage sites and identify cleavage sites in cellular proteins. Identification of cellular target proteins might lead to a better understanding of mechanisms underlying takeover of the cellular synthesis machinery by these viruses.

Known cleavage sites and all sites predicted to be cleaved in proteins from the SwissProt database were also analyzed with a powerful neural network designed to predict if a given residue is exposed or buried (J. Hansen, O. Lund, H. Nielsen, S. Brunak, in prep.). We assume that cleavage sites in cellular target proteins are exposed. Therefore, proteins were ranked according to both the potential cleavage site probability and the probability of the site being accessible. In addition to the detailed description of proteinase cleavage sites and the prediction of possible targets, we report a number of errors in the SwissProt database entries used in this study.

Results

Properties of $2A^{pro}$ cleavage sites

In rhino- and enterovirus, $2A^{pro}$ cotranslationally cleaves at its own N terminus, thereby releasing the precursor protein P1 (which subsequently gives rise to the viral capsid proteins) from the P2–P3 region, whose mature proteins are involved in replication and are not associated with the virion. Subsequent processing of the P1 precursor is performed by $3C^{pro}$ (Fig. 1).

The separation of the structural region of the polyprotein from the enzymatic region is different in aphtho-, cardio-, and hepatovirus from the mechanism described above. In aphtho-, cardio-, and hepatovirus, the cleavage is mediated by proteinase 3C (Schultheiss et al., 1994), whereas the separation of 2A and 2B is mediated by an enzyme-independent autocatalytic mechanism involving the 2A polypeptide (Ryan & Drew, 1994). Because of these differences, only $2A^{pro}$ of rhino- and enterovirus was examined.

A sequence logo compiled from 22 unique $2A^{pro}$ cleavage sites of entero- and rhinoviruses is shown in Figure 2. The unique requirement for glycine at P1′ is evident and positions P4 and P2′ are most frequently occupied by hydrophobic amino acids. In addition, the preference for the turn-inducing proline at P2′ might suggest that this position is most likely structurally constrained. At position P1, Y is most abundant, correlating with all $2A^{pro}$ cleavages in poliovirus occurring between Y–G pairs (Lawson & Semler, 1990). However, Ala, Thr, or Val are found almost as frequently at position P1. Position P2 is occupied by neutral residues with a certain preference of threonine within positions P3–P1. Threonine at position P2 is the most conserved residue after P1′-glycine. From the logo, a higher conservation of the sequence C terminal to the cleavage site relative to the N-terminal side is also evident. Because the C-terminal (right) part of the cleavage region corresponds to the N terminus of $2A^{pro}$, this suggests a higher degree of conservation among the different 2A proteinases compared with the C terminus of the structural protein VP1 on the left side. The sequence logo thus visualizes very well the results of experimental work on $2A^{pro}$ of human rhinovirus 2 and coxsackievirus B4 (Sommergruber et al., 1992, 1994); based on cleavage experiments of oligopeptides in vitro, a consensus sequence of (Ile/Leu)-Xaa-Thr-Xaa*Gly (where Xaa is any amino acid) for the cleavage site was proposed. Similarly, studies on poliovirus $2A^{pro}$ stressed the importance of positions P2 and P1′, whereas P3, P1, and P2′ were less important for efficient cleavage (Hellen et al., 1992).

Properties of $3C^{pro}$ cleavage sites

Proteinase 3C is present in all picornaviruses and therefore all $3C^{pro}$ sites were included in our initial analysis. However, using all available $3C^{pro}$ cleavage sites, the algorithm was unable to verify experimentally determined cleavage sites. This was taken to indicate that the specificities of cleavage by $3C^{pro}$ differ too much between the picornavirus genera to be considered unique. Experiments also suggest that rhino- and enterovirus $3C^{pro}$ have similar features, which differ from the remaining virus genera. For example, Joachims et al. (1995) showed that rhino- and poliovirus $3C^{pro}$ cleaved MAP-4, whereas cardiovirus $3C^{pro}$ did not.

For this reason, we decided to construct two separate $3C^{pro}$ algorithms, one for entero- and rhinovirus and one for aphthovirus.

The data for each of these neural networks were used to generate the sequence logos shown in Figures 3 and 4. In the sequence logo of 114 unique sites in the entero- and rhinovirus data set (Fig. 3), the preference for Q–G pairs at positions P1–P1′, as well as the importance of residues at positions P4 (mostly alanine) and P2′ (mostly proline or leucine), is readily seen. This observation clearly supports the experimental studies of HRV14 and poliovirus $3C^{pro}$ specificity (Cordingley et al., 1989, 1990; Long et al., 1989).

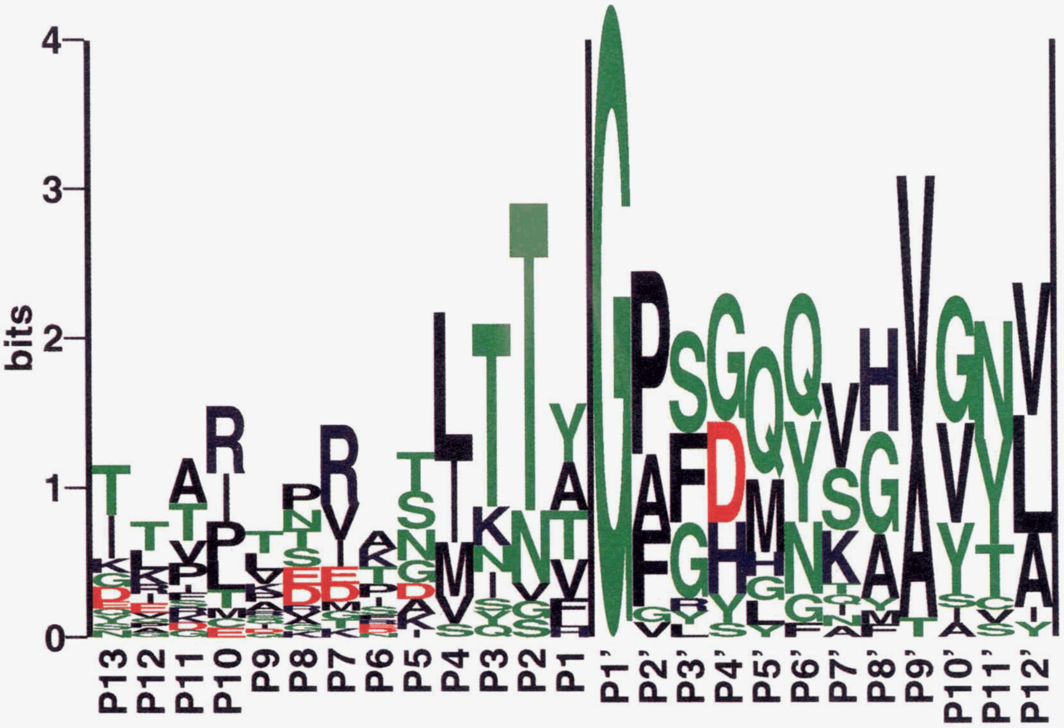

Fig. 2. Sequence logo (as described in Materials and methods) of 2A^{pro} cleavage site (22 unique sites from entero- and rhinovirus). Amino acids are color coded according to their physicochemical characteristics. Neutral and polar, green; basic, blue; acidic, red; neutral and hydrophobic, black. Amino acids are shown as one-letter standard code. Cleavage nomenclature according to Berger and Schechter (1970).

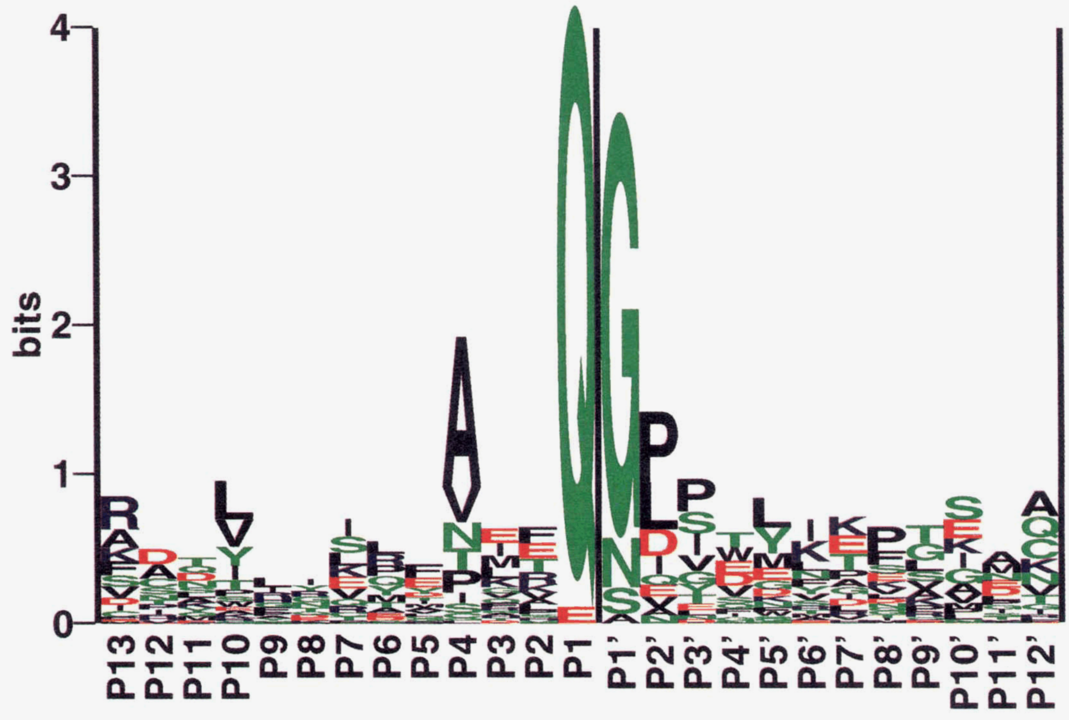

Fig. 3. Sequence logo of 3C^{pro} cleavage site (114 unique sites from entero- and rhinovirus).

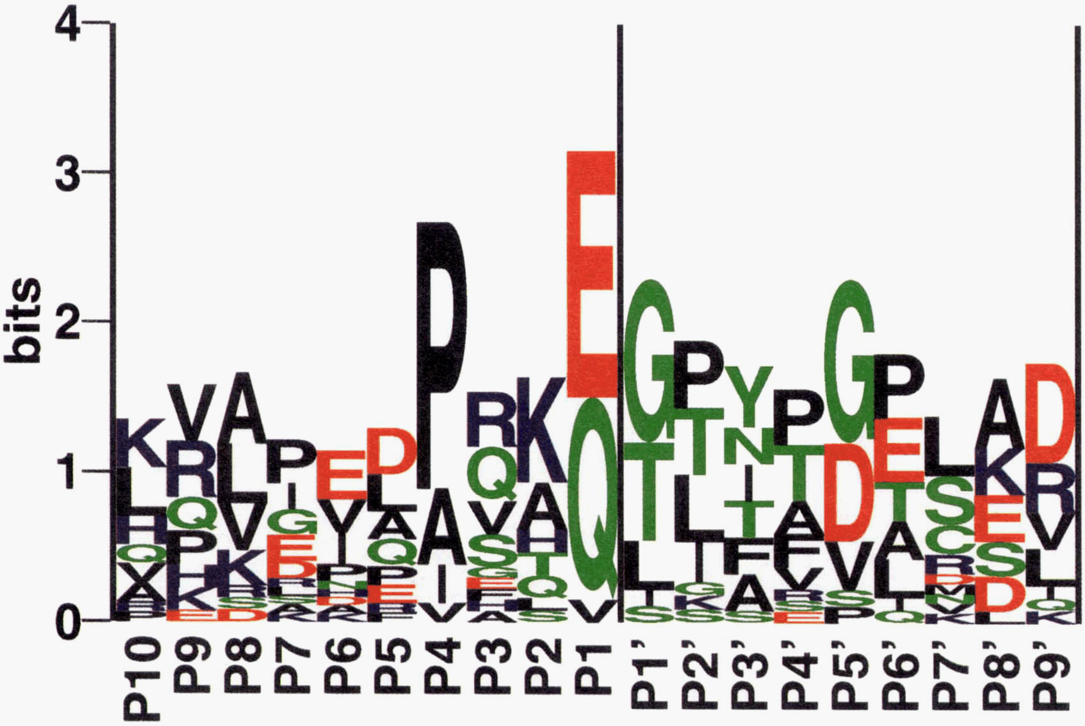

Fig. 4. Sequence logo of $3C^{pro}$ cleavage site (21 unique sites from aphthovirus).

The sequence logo for the 21 unique $3C^{pro}$ sites in the aphthovirus data set is shown in Figure 4. As seen for entero-/rhinovirus $3C^{pro}$, there is a preference for Q–G or E–G pairs and for hydrophobic residues at P4. This logo is based on our modification of certain VP1–2A cleavage sites in FMDV5 (ARQ*LL*NFDL) and FMDVO (VKQ*TL*NFDL). Our neural network method and multiple alignment indicated that the cleavage takes place at Q*L or Q*T, respectively, instead of the reported L*N cleavage site. The modified sites were used for further analysis (see also the Discussion).

Properties of VP4–VP2 (autocatalytic) cleavage sites

During maturation of the virion, the structural precursor protein VP0 is cleaved to yield the capsid proteins VP4 and VP2. This process is essential for viral infectivity and is assumed to be an autocatalytic process, not involving any of the known viral proteinases. In HRV14, cleavage occurs between residues N_{69} and S_{70}. Based on the three-dimensional structure of the protomer unit, it has been suggested that S_{10} of VP2 (S_{79} of VP0) acts as a nucleophile and catalyzes cleavage in concert with nucleotides of the viral RNA (Arnold et al., 1987; Lawson & Semler 1990). However, in a more recent study, S_{10} of VP2 was mutated to alanine or cysteine, which led to normal viral maturation, albeit at a somewhat reduced rate; the nature of this proteolytic activity is therefore still elusive (Harber et al., 1991).

As depicted in the sequence logo of the VP4–VP2 cleavage site (Fig. 5), only D or S are present at the P1′ position, whereas at P1 several residues with different physicochemical properties are found. However, in the 26 cases examined, L at P2 and E at P5′ are completely conserved. Serine at position P10′, which has been claimed to be involved in the cleavage process, is not present in all cases. VP2 residues on the C-terminal side of the cleavage site (positions P5′–P12′) appear to be more conserved, possibly reflecting the requirement for particular structural features in this region of the VP0 protein. Until autocatalytic cleavage has been demonstrated unequivocally, it cannot be excluded that a cellular proteinase is responsible for this processing step.

Neural network prediction

Performance of neural networks

The performance of the four types of networks optimized for each of the cleavage categories is summarized in Table 1. A very high percentage of both cleavage and noncleavage sites in the test data sets were predicted correctly.

Prediction of $2A^{pro}$ cleavage sites

For poliovirus and for HRV1A, an additional cleavage of $2A^{pro}$ in the 3D region has been described previously. In studies with poliovirus, this cleavage was shown to be nonessential for viral replication to occur (Lee & Wimmer, 1988). In order to avoid possible cleavage sites in the training set used for noncleavage, polyprotein sequences were only used up to 3B with 3C and 3D being excluded (amino acids 1–1760). As compared with earlier training runs, this change considerably improved the performance of the neural network in accurately recognizing cleavage sites, without the necessity to use hidden units. This observation implies that the potential second cleavage site within the 3D region, previously assigned as a noncleavage site, added substantially to the complexity of the problem. The neural network trained on several subsets of the picornavirus cleavage sites was then tested for its ability to correctly predict the $2A^{pro}$ cleavage sites in human and rabbit initiation factor eIF-4G (Lamphear et al., 1995). The cleavage site in human eIF-4G gave a high score, whereas prediction of the rabbit

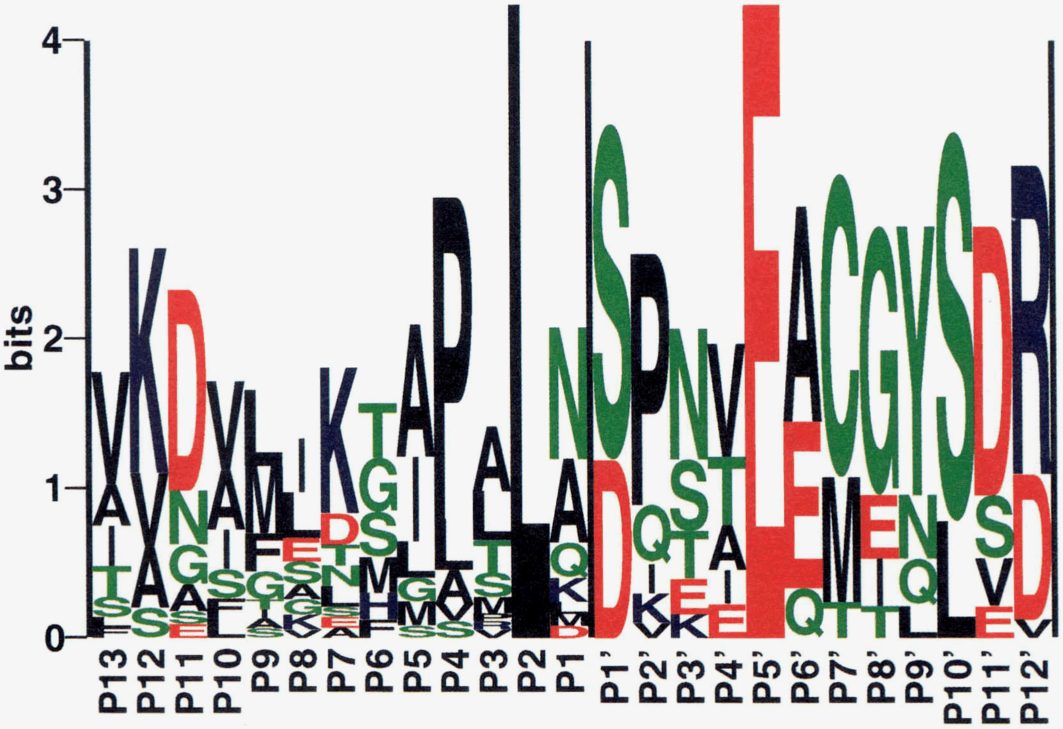

Fig. 5. Sequence logo of autocatalytic site (26 unique sites from all picornaviruses).

eIF-4G cleavage site gave a score just above the threshold of 0.5. Correct prediction of the latter site was highly dependent on the presence of the HRV14 (POLG_HRV14) sequence in the training set (see Table 2 and Fig. 6).

Multiple sequence alignments of the rhinovirus polyproteins identify HRV14 as evolutionarily more distant to all other known HRVs (see Palmenberg alignments at http://www.bocklabs.wisc.edu/seq.html). Moreover, based on similarity with poliovirus, the $2A^{pro}$ cleavage site in HRV14 was assigned originally between Y_{856} and G_{857} (Stanway et al., 1984). However, when this assignment was used in the training set, the cleavage site in rabbit eIF-4G could not be predicted correctly. When cleavage was assigned to L_{858}, as later suggested by the Palmenberg alignments, the site was predicted correctly, although with low score. In the absence of experimental data on the determination of the $2A^{pro}$ cleavage site in HRV14, we suggest that cleavage takes place between L_{858} and G_{859}.

Scanning SwissProt by neural networks

All 50,000+ sequence entries in the SwissProt database (rel. 33), comprising more than 18 million amino acid residues, were scanned with each of the three cleavage site neural networks. The highest cleavage scores along with the corresponding surface prediction scores of each cleavage category are shown in Tables 3, 4, and 5

Table 1. *Neural network performance for each kind of cleavage category*[a]

Type	Window	Hidden	C/NC	Train	Test
$2A^{pro}$	15	0	C	21/21 (100%)	6/6 (100%)
			NC	2,072/2,073 (99.9%)	370/370 (100%)
$3C^{pro}$(ER)	9	2	C	97/97 (100%)	17/17 (100%)
			NC	601/601 (100%)	412/418 (98.6%)
$3C^{pro}$(FMDV)	9	2	C	18/18 (100%)	4/4 (100%)
			NC	1,008/1,018 (99.0%)	337/347 (97.1%)
Autocatalytic	15	0	C	35/35 (100%)	11/11 (100%)
			NC	277/277 (100%)	93/93 (100%)

[a] ER, entero-/rhinovirus network; FMDV, aphthovirus network; Window, optimal window size; Hidden, number of hidden units; C/NC, cleavage/noncleavage sites in data set; Train, performance on training set and number of correctly predicted/total number of sites; Test, performance on test set.

Table 2. *Cleavage sites in 2A^{pro} data set*[a]

Entry	Sequence	Position
POLG_BOVEV	SNRASLTSY*GPFGQQQG	840
POLG_COXA2	TKVDSITTF*GFGHQNKA	879
POLG_COXA9	GDMSTLNTH*GAFGQQSG	867
POLG_COXB3	QSITTMTNT*GAFGQQSG	851
POLG_POL1S	LSTKDLTTY*GFGHQNKA	881
POLG_SVDVU	TDITTMKTT*GAFGQQSG	851
HRV1A_MOD	VRRNTITTA*GPSDLYVH	832
POLG_HRV1B	VPRASMKTV*GPSDLYVH	857
POLG_HRV2	VTRPIITTA*GPSDMYVH	850
POLG_HRV89	PDVFTVTNV*GPSSMFVH	866
POLG_EC11G	TVKPDLSNY*GAFGYQSG	40
POLG_POL2W	LPEKGLITY*GFGHQNKA	879
POLG_POL3L	LSEKGLTTY*GFGHQNKA	878
HRV85	KERASLTTA*GPSDMYVH	[b]
POLG_COXB1	TTRSNITTT*GAFGQQSG	848
POLG_COXB4	AERASLITT*GPYGHQSG	849
POLG_COXB5	TEITAMQTT*GVLGQQTG	851
POLG_HUEV7	PNDINLTTA*GPGYGGAF	867
HRV15	ENVRAIVNV*GPSDMYVH	[b]
HRV16	RPRTNLTTV*GPSDMYVH	[b]
HRV50	ATRPKITVA*GPSDMYVH	[b]
POLG_COXA4	IAVENINTF*GGFGHQNM	885
POLG_POL2L	LPGKGLTTY*GFGHQNKA	879
HRV9	DNVRAVKNV*GPSDMYVH	[b]
POLG_HRV14	KGDIKSYGL*GPRYGGIY	858
IF4G_HUMAN[c]	LGRTTLSTR*GPPRGGPG	485
IF4G_RABIT[c]	LGRPALSSR*GPPRGGPG	486

[a] SwissProt entry name, cleavage sequence, and position in database file are indicated.

[b] Sequences do not exist in the SwissProt database.

[c] Initiation factor-4G sequences were used in the validation of the trained network and were not included in the training procedure.

for human proteins. Proteins of nonhuman origin, which are discussed in the text, are also shown.

The 2A^{pro} neural network

Sites in human proteins with the highest cleavage prediction scores and their corresponding surface exposure scores are shown in Table 3. Predicted sites that were topologically inaccessible to the proteinase (e.g., extracellular domains) were removed from the list. A graphical representation of both prediction coordinates (cleavage and surface exposure scores) is shown in Figure 6.

Somewhat surprisingly, the surface exposure scores for the known viral cleavage sites cluster around 0.5, indicating that these sites are neither fully exposed nor fully buried (Fig. 6). This may reflect the nature of the intramolecular *cis* cleavage versus the *trans* cleavage, which certainly requires exposure of the target site. The cleavage scores of the viral cleavage sites lie between 0.8 and 0.9, in contrast to those of the only two experimentally proven nonviral target proteins eIF-4G from rabbit and human, respectively, which have scores of between 0.5 and 0.7. The lower cleavage scores in these proteins thus seem to be compensated by the higher surface exposure scores. The difference in the observed scores may reflect a feature of the algorithm used, a difference in the accessibility of the proteins inside the cell, or a combination of both.

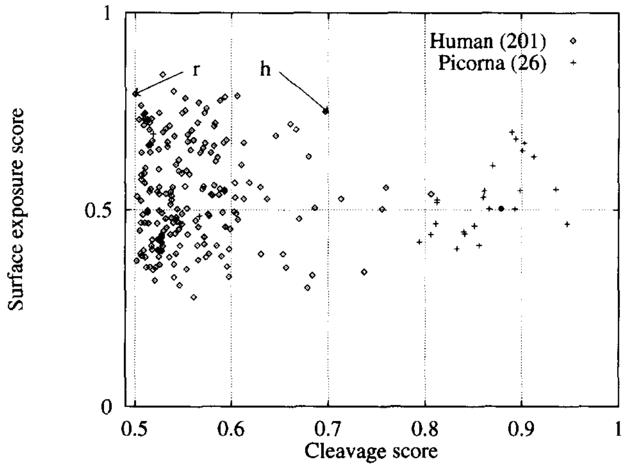

Fig. 6. Human proteins (a total of 201) from SwissProt with a 2A^{pro} cleavage score greater than 0.5 and their respective surface exposure scores. Also included are the cleavage and exposure scores for 26 cleavage sites in the picornavirus data set. Individual points correspond to a single Xaa*Glycine site. The position of coordinates for human(h) and rabbit(r) eIF4G is indicated by arrows.

Table 3. *Predicted human target proteins for rhino-/enterovirus 2A^{pro}*[a]

Entry	Position	Cleavage	Surface	Sequence
IF4G_RABIT	486	0.501	0.794	PALSSR*GPPRG
VATD_YEAST	125	0.848	0.465	FRLTGL*GRGGQ
YMC1_YEAST	73	0.747	0.745	KLLANE*GPRGF
DSC2_HUMAN	761	0.737	0.343	FTTQTV*GASAQ
IF4G_HUMAN	485	0.697	0.749	TTLSTR*GPPRG
PUR8_HUMAN	130	0.684	0.334	ASLPTL*GFTHF
DMD_HUMAN	588	0.680	0.635	NKIHTT*GFKDQ
RCC_HUMAN	196	0.679	0.302	GDLYTL*GCGEQ
TNRL_HUMAN	425	0.661	0.718	ATPSNR*GPRNQ
CIK6_HUMAN	423	0.654	0.388	VTMTTV*GYGDM
GLI3_HUMAN	446	0.646	0.688	EDLPSP*GARGQ
RL1X_HUMAN	103	0.630	0.558	RDLTTA*GAVTQ
HNFA_HUMAN	206	0.613	0.670	RNRFKW*GPASQ
ABP_HUMAN	170	0.611	0.614	FFLNTT*GFSFQ
SYEP_HUMAN	1,297	0.607	0.475	PIRLEV*GPRDM
AHR_HUMAN	318	0.606	0.789	AELCTR*GSGYQ
AC13_HUMAN	14	0.605	0.491	PHLLVY*GPSGA
AC14_HUMAN	75	0.604	0.568	PNIIIA*GPPGT
HO1_HUMAN	97	0.604	0.531	DLAFWY*GPRWQ

[a] Human proteins with highest scores are shown as well as proteins discussed in the text (shown in bold). Proteins are indicated by their entry codes (SwissProt), cleavage position, cleavage score (0.000–1.000), surface score (0.000–1.000), and target sequence.

The only cellular protein identified experimentally as target for 2A^{pro} is the translational initiation factor eIF-4G; cleavage sites in the human as well as in the rabbit protein have been determined after in vitro cleavage by bacterially expressed recombinant 2A^{pro} from HRV2 and from coxsackievirus B4 (Lamphear et al., 1993, 1995; Liebig et al., 1993). The 2A^{pro} network predicts these sites correctly for human eIF-4G (IF4G_HUMAN), being cleaved at R_{485} (cleavage score 0.697 and surface score of 0.749), and rabbit eIF-4G (IF4G_RABIT), being cleaved at R_{486} (cleavage score 0.501 and surface score 0.794) (see Fig. 6).

Table 4. *Predicted human target proteins for rhino-/enterovirus 3C^{pro}*[a]

Entry	Position	Cleavage	Surface	Sequence
TFC3_HUM (U02619)[b]	732	0.644	0.606	QPPVPQ*GEAEE
TF2D_HUMAN	18	0.716	0.732	GLASPQ*GAMTP
MAP4_HUMAN	188	0.638	0.652	TAVVPQ*GWSVE
NU21_HUMAN	692	0.959	0.655	AVAEKQ*GHQWK
LYN_HUMAN	63	0.959	0.615	KDPEEQ*GDIVV
RIR1_HUMAN	550	0.956	0.698	DLAKEQ*GPYET
HS1_HUMAN	17	0.955	0.580	VSVETQ*GDDWD
KGPB_HUMAN	266	0.952	0.569	EYIIRQ*GARGD
T2FB_HUMAN	49	0.951	0.603	RIAKTQ*GRTEV
KAP0_HUMAN	179	0.950	0.502	FYVIDQ*GETDV
PIP5_HUMAN	881	0.949	0.680	LEPKEQ*GDPPV
PTN1_HUMAN	85	0.948	0.644	SYILTQ*GPLPN
SYV_HUMAN	62	0.947	0.682	LPALEQ*GPGGL
PTPM_HUMAN	972	0.947	0.648	HYIATQ*GPMQE
IRE1_HUMAN	471	0.947	0.555	VEAITQ*GDLVA
PRGB_HUMAN	29	0.946	0.629	DYVNTQ*GPSLF
OC3A_HUMAN	39	0.945	0.732	TWLSFQ*GPPGG
PTPB_HUMAN	1,776	0.945	0.582	EYIVTQ*GPLPG
OC3N_HUMAN	246	0.943	0.742	PPPPPQ*GPPGH
DMD_HUMAN	2,313	0.943	0.605	ALPEKQ*GEIEA
PLAK_HUMAN	68	0.943	0.540	GVPPSQ*GDLEY
5H1A_HUMAN	283	0.943	0.533	NGAVRQ*GDDGA
PRPC_HUMAN	47	0.942	0.687	IDEERQ*GPPLG
OCT2_HUMAN	21	0.940	0.715	LEAEKQ*GLDSP
GLI3_HUMAN	1,358	0.940	0.637	HINIYQ*GPESC

[a] Human proteins with highest scores are shown as well as proteins discussed in the text (shown in bold). Proteins are indicated by their entry codes (SwissProt), cleavage position, cleavage score (0.000–1.000), surface score (0.000–1.000), and target sequence.

[b] Denotes GenBank accession number; protein is not present in SwissProt.

Upon scanning all proteins in SwissProt, 24 of the top 100 proteins having the highest cleavage scores were of viral origin (data not shown); 5 were from DNA viruses (alphaherpes- and mastadenovirus), and 19 were from RNA viruses (calici-, poty-, como-, corona-, and orthomyxoviruses). Calici-, poty-, and comoviruses are related to picornaviruses, all having a polyprotein precursor that is proteolytically processed to the mature polypeptides. This suggests the presence of cleavage patterns beyond the picornavirus family and may reflect a conservation of ancestral viral sequences (see also Hellen et al., 1989).

Expression of 2A^{pro} in yeast cells resulted in a marked decrease in protein synthesis and a change in morphology (Barco & Carrasco, 1995; Klump et al., 1996). However, the yeast homologue of mammalian eIF-4G was found to be intact, indicating that the effects on yeast cells have to be attributed to cleavage of so-far unknown proteins. In the list of probable cleavage sites resulting from the scanning of SwissProt, several yeast proteins appear (data not shown). Most interesting is the protein with the highest scoring site, the probable vacuolar ATP synthase subunit d (VATD_YEAST), with a cleavage score of 0.848 and surface score of 0.465 (see Table 3). Because the morphology change observed in yeast is associated with an increased number of vacuoles, cleavage of this protein might be related to these changes. YMC1_YEAST, with a cleavage score of 0.747 and a surface score of 0.745, was also considered a possible candidate (Table 3). However, the comment lines in the database identify this polypeptide as an integral membrane protein of the inner mitochondrial membrane, making it highly unlikely to be accessible for the proteinase. Any of the proteins with a predicted cleavage site must thus be examined for its cellular location and cleavage must be tested experimentally before any conclusions can be drawn; nevertheless, the method described here provides a way of limiting the number of choices before conducting the experiment.

The 3C^{pro} neural network

The 3C^{pro} neural network trained on entero- and rhinovirus cleavage sites was used to scan all human protein sequences in SwissProt (rel. 33). Sites in human proteins with the highest cleavage prediction scores and their corresponding surface exposure scores are shown in Table 4.

As also seen with the 2A^{pro} network, many viral sequences were predicted as potential cleavage sites if all proteins of SwissProt were scanned (data not shown). Of the top 100 cleavage scores, 11 were of viral origin. In contrast to 2A^{pro}, the majority (eight) were from DNA viruses and three were from RNA viruses. The latter were all retroviruses.

Studies by Clark and colleagues (Clark & Dasgupta, 1990; Clark et al., 1991, 1993) have shown that poliovirus infection of HeLa cells causes an inhibition of all three classes of host cell RNA polymerases. The inhibition of RNA polymerase III-mediated transcription in HeLa cells on poliovirus infection or expression of 3C^{pro} seemed to be caused by proteolysis of the TFIIIC complex, but the cleavage site was not determined. Predictions from our

Table 5. *Proteins from SwissProt containing a predicted VP4–VP2 "autocatalytic" site*[a]

Entry	Position	Cleavage	Surface	Sequence
DPO1_BACCA	304	0.758	0.511	MAFTLA*DRVTE
MX1_ANAPL	313	0.731	0.395	TKPDLV*DIGTE
MEP_SCHCO	265	0.728	0.460	KAPELA*SGDAE
MANA_YEAST	166	0.725	0.619	RIPELR*NIVGE
MYSN_CHICK	1,110	0.719	0.691	KIRELE*SQITE
LCB2_YEAST	411	0.710	0.460	TISSLQ*TISGE
CYB_STRPU	314	0.702	0.544	LMPLLN*TSKNE
HS27_HUMAN	173	0.698	0.352	PMPKLA*TQSNE
ATPA_SCHPO	501	0.695	0.595	FIPYLR*SSGAE
ODB2_HUMAN	326	0.689	0.575	QFPILN*ASVDE
APC_HUMAN	1,714	0.685	0.656	TIPELD*DNKAE
ANFK_RHOCA	325	0.684	0.405	ANPDLA*IGLTE
SFCA_ECOLI	267	0.682	0.540	AMPLLN*RYRNE
ENDR_BOVIN	260	0.681	0.721	TSPSLN*GRCTE
PAL1_PHAVU	31	0.681	0.398	QAFELA*NINSE
CLH_DROME	864	0.676	0.697	LLPWLE*SRVHE
CHI1_TOBAC	19	0.675	0.351	FSLLLL*SASAE
YEM6_YEAST	555	0.673	0.507	TIDSLA*DAINE
RAD4_YEAST	545	0.670	0.379	YIPPLA*SASGE
HVC2_HETFR	383	0.667	0.344	FIYSLL*SIAAE
ARSA_HUMAN	183	0.664	0.452	PIPLLA*NLSVE
PR05_YEAST	343	0.662	0.436	PTRELA*LQIHE
MNS1_YEAST	347	0.661	0.461	MGGLLA*SGSTE
ASRC_SALTY	86	0.659	0.635	LEPFLR*EIEIE
YO11_MOUSE	358	0.659	0.558	TNPALQ*RIITE
ZN35_HUMAN	114	0.658	0.376	KNLQLL*VPKTE
DPOL_CHVN2	862	0.657	0.750	FEPLLD*DPETE
CHIC_LYCES	18	0.657	0.319	FSVLLL*SASAE
CSP_PLABE	304	0.655	0.740	EDLTLE*DIDTE

[a]Proteins discussed in the text are shown in bold. Cleavage, cleavage score (0.000–1.000); Surface, surface score (0.000–1.000).

neural network method suggest that human TFIIIC (TFC3_HUM) may be cleaved between Q_{732}–G (see Table 4), with a cleavage score of 0.644. This prediction supports the results presented in a recent report (Shen et al., 1996).

The effect of $3C^{pro}$ on the activity of RNA polymerase II-mediated transcription was taken to indicate that the TATA-binding protein (TBP) of the TFIID complex was the target (Clark et al., 1993). The neural network method predicts a cleavage site in SwissProt entry TF2D_HUMAN at Q_{18}–G (cleavage score 0.716, surface score 0.732, see Table 4). This potential site has the right combination of a high cleavage score and a high surface exposure score and may be a candidate for further experiments. Furthermore, the $3C^{pro}$ network predicts a cleavage site in transcription initiation factor IIF (T2FB_HUMAN, Table 4) at Q_{49}–G, with a high score of 0.951. This factor may also be involved in the poliovirus-induced decrease in RNA polymerase II-mediated transcription.

Structural proteins have also been shown to be cleaved by $3C^{pro}$. In HeLa cells, expression of poliovirus or HRV14 $3C^{pro}$ results in cleavage of the microtubule-associated protein 4 (MAP-4) (Joachims et al., 1995). The exact cleavage site has not yet been determined, but is suggested to lie at Q_{188}–G. Our neural network method confirms that this site is a likely candidate (cleavage score 0.638, surface score 0.652, Table 4).

Histone H3 in baby hamster kidney cells was reported to be cleaved by $3C^{pro}$ from FMDV (Tesar & Marquardt, 1990). The $3C^{pro}$ neural network trained on aphthovirus sequences was used to try to verify this finding. The actual cleavage site was located experimentally at L_{20}-A of histone H31_HUMAN (sequence APRKQ$_{19}$L$_{20}$A$_{21}$TKAA) (Falk et al., 1990), but our results strongly favor cleavage between Q_{19}–L, with a cleavage score of 0.723 (cleavage score for L_{20}–A is 0.169). Furthermore, multiple alignment with known FMDV $3C^{pro}$ cleavage sites also suggests that Q_{19}–L is the target site (data not shown).

The "autocatalytic site" neural network

The cleavage sites with highest scores are shown in Table 5. It is interesting to note that, of the top 100 proteins, 23 are described as precursor proteins. In some proteins, the predicted autocatalytic site coincides with the cleavage site of a potential signal peptide, e.g., CHIC_LYCES and CHI1_TOBAC (Table 5). The autocatalytic conversion from VP0 to VP4 and VP2, and the cleavage of a signal peptide, are both a conversion from a precursor to a mature product. This might suggest that a common or at least similar mechanism might operate in the maturation of other proteins.

Discussion

We have examined the performance of neural networks in the prediction of sites present in cellular proteins that might be cleaved by viral proteinases when the host cells are infected with picornaviruses. The neural networks were trained to distinguish cleaved from noncleaved sites, making use of the known sites in the viral polyproteins. Based on a test data set independent from the training data set, the performance of the networks was examined and approached 100% accuracy. Because cleavage sites must be accessible, an additional analysis was performed by a network that had been trained to recognize surface-exposed regions in proteins. In cases of alternative predicted cleavage sites in a specific protein, the surface score may be used as an additional criterion (see Tables 3, 4, 5).

The eukaryotic translation initiation factor eIF-4G is the only known cellular target for $2A^{pro}$; the site cleaved by the proteinase has been determined experimentally for the human as well as for the rabbit protein. The combination of the two neural networks predicted cleavage within human eIF-4G with high confidence, whereas prediction of cleavage of the rabbit protein was lower. In the experiments that had led to identification of the cleavage sites by amino acid sequencing, no comparison with regard to the efficiency of cleavage of these two proteins was made; it therefore cannot be excluded that cleavage of rabbit eIF-4G by the purified proteinase in vitro is also less efficient.

Several other proteins that might be targets for $2A^{pro}$ were found in the SwissProt database. Inspection for their function in the cellular life cycle did not yield any hint as to their possible role in viral infection. It is clear that proteins other than eIF-4G are processed by $2A^{pro}$ because, e.g., yeast cells expressing this proteinase substantially change their morphology, although the yeast homologue of eIF-4G is not cleaved. Whether the proteins identified by the network are real targets remains to be shown by experiment.

Whereas $2A^{pro}$ in entero- and rhinoviruses performs one essential cleavage between VP1 and $2A^{pro}$ (with a second dispensable cleavage in 3D), $3C^{pro}$ is responsible for eight cleavages within the polyprotein. These cleavages exhibit different kinetics and processing of P1 even requires 3CD rather than its maturation product

$3C^{pro}$ for efficient cleavage. These differences have not been taken into account in the construction of the present $3C^{pro}$ networks; input sequences were only classified as cleaved or noncleaved. Nevertheless, the essential features of potential cleavage sites of entero- and rhinoviruses are clearly apparent from the sequence logo shown in Figure 3. Glutamine at P1, glycine at P1′, and small hydrophobic amino acids at P4 are most important for interaction with the substrate binding pocket of $3C^{pro}$, as shown recently by modeling the trypsin Bowman–Birk inhibitor complex onto the three-dimensional structure of HRV14 $3C^{pro}$ (Matthews et al., 1994) and by studies of cleavage site mutants occurring with the introduction of two VPg coding units into the poliovirus genome (Cao & Wimmer, 1996).

Initial attempts to combine $3C^{pro}$ data from entero-, rhino-, and aphthoviruses in the training set resulted in poor prediction of proteins that had been demonstrated experimentally to be cleaved by the respective proteinases. Therefore, two networks were constructed using either entero- and rhinovirus $3C^{pro}$ sites, or only aphthovirus cleavage sites, respectively. The entero- and rhinovirus $3C^{pro}$ network predicted cleavage sites in all proteins that have been identified experimentally as targets for these proteinases; our predictions agree with those suggested by the authors.

Training of the FMDV $3C^{pro}$ network using the known cleavage sites within the FMDV polyproteins questions some of the tentative assignments in the literature. Learning the sites ARQLL*NFDL (FMDV5) and VKQTL*NFDL (FMDVO) proved most difficult, whereas the alternative sites ARQ*LLNFDL and VKQ*TLNFDL, respectively, were accepted easily. Sequence alignments were also found to favor this alternative possibility; although there is currently no experimental proof for these assignments, these sites were consequently used in the training set. The difference in the preferred cleavage sites of the two groups of 3C proteinases is reflected clearly in their respective sequence logos (cf. Figs. 3, 4) and in the correct identification of cellular proteins either cleaved by entero- and rhino $3C^{pro}$ or FMDV $3C^{pro}$. The latter network predicted cleavage of histone H3 to occur at APRKQ*LATKAA instead of APRKQL*ATKAA, as reported by Falk et al. (1990). In the absence of additional data, we have to assume that cleavage at L*A occurs due to structural constraints in the folded protein that force the proteinase to cleave downstream of its preferred site. Alternatively, a cellular aminopeptidase might cleave off one or two aminoterminal residues immediately after processing by $3C^{pro}$ has taken place.

The site between VP2 and VP4, which is thought to be cleaved autocatalytically, was also presented to a neural network. The network learned to recognize this site in the viral polyproteins and predicted possible cleavage sites in proteins within the SwissProt database. The significance of this is not clear at the present time, and it will be necessary to examine experimentally the identified proteins for possible autoproteolysis. In addition, if the process really is autocatalytic, the site need not be exposed and the surface exposure prediction becomes obsolete in this case.

In order to optimize the three different neural networks for recognition of the respective sites, the window sizes were varied. The best performance was seen with window sizes of 15 for $2A^{pro}$, of 9 for $3C^{pro}$, and with 15 for the "autocatalytic" site. Nine amino acids is the minimum length required for a synthetic peptide to be cleaved by $2A^{pro}$ in vitro, but longer peptides are processed more efficiently (Sommergruber et al., 1992). In the absence of structural information on the 2A proteinase, it is not currently clear whether all 15 amino acids, which contribute to the quality of prediction of the network, are also involved in physical interactions with the enzyme. It is also possible that they take part in folding the vicinity of the scissile bond.

From modeling of the structure of a peptide bound to HRV14 $3C^{pro}$, it can be inferred that about eight amino acids are contacting the enzyme; in this case, the network seems to take into account all amino acids involved in interactions. How the 15 amino acids of the "autocatalytic site" are involved in the cleavage reaction is totally unknown at the present time.

In summary, we have attempted to identify possible cleavage sites within known cellular proteins by a neural network approach. Whether these proteins are real targets for the viral proteinases will be seen in the future. As became clear for YMC1_YEAST, the location and accessibility in the cellular context also has to be taken into account. This protein might be cleaved when solubilized and purified, but, when associated with the inner mitochondrial membrane, it is certainly not. It should also be kept in mind that proteins that are cleaved in the host cell might not be represented in the database, and thus escape identification. If such proteins are being discovered, they can then be subjected to site detection analysis and the validity of the neural network approach may be tested.

Materials and methods

Extraction of PICORNA sequence data from SwissProt

Entries containing the string "PICORNA" in the "OC" (organism classification) description field (a total of 53) were extracted from the SwissProt database (rel. 33, Feb 1996) and examined in detail. Two entries were discarded because the sequence failed to encompass any cleavage site. In addition, two rhinovirus sequences, HRV9 and HRV85, were obtained prior to publication from G. Leckie and J.W. Almond, and from G. Stanway, respectively. The resulting data set of 53 entries contained 4 small fragments (~200 residues), 9 partial fragments (800–1,300 residues), and 40 full-length polypeptides (~2,200 residues). A list of the sequences used in this study is shown in Table 6; 8 aphtho-, 20 entero-, and 7 rhinoviruses were included, whereas 18 cardio- and hepatoviruses were excluded. The amino acid composition was compared with the total SwissProt database and showed no obvious difference between picornaviral proteins and the whole database; only a slightly lower content of charged amino acids was found to be present in picornaviral polyproteins (22.7% versus 24.8% in SwissProt).

Data set for a $2A^{pro}$ neural network

The data used for training of a $2A^{pro}$ neural network consisted of sequences belonging to the entero- and rhinovirus genera only. Because a second cleavage site has been reported in the 3D region of HRV1A and poliovirus (see above), which might also be present in the other viruses, this part of the polyprotein sequence was omitted from the noncleavage data set to avoid conflicting data. In addition, due to their high similarity with other entries (in parentheses), POLH_POL1M and POLG_POL1M (POLG_POL1S), POLG_POL32 (POLG_POL3L), and POLG_SVDVH (POLG_SVDVU) were removed from the data set. From the sequence alignment data by A. Palmenberg (http://www.bocklabs.wisc.edu/seq.html), HRV15, HRV16, and HRV50 $2A^{pro}$ cleavage sites were available and were included in the training set; for HRV1A, the cleavage site as given by the Palmenberg alignment was used

Table 6. *Description of the data set, showing length (in residues) and name of entry, accession number, and cleavage type*[a]

Length	Name	Accession	3C										2A	Auto	L
Aphtho															
2,333	POLG_FMDV1	P03306	504	725	935	1107	1425	1578	1601	1625	1649	1863	953	286	201
230	POLG_FMDV5	P03307		4	214										
2,332	POLG_FMDVA	P03308	503	724	935	1107	1425	1578	1601	1625	1649	1862	953	285	200
216	POLG_FMDVC	P03309			200										
234	POLG_FMDVI	P03310		10	218										
2,332	POLG_FMDVO	P03305	504	724	935	1107	1425	1578	1601	1625	1649	1862	953	286	201
861	POLG_FMDVS	P03311		46	252	391									
1,011	POLG_FMDVT	P15072	504	723	930								948	286	201
Entero															
205	POLG_COXA3	P08490							22						
2,175	POLG_BOVEV	P12915	317	559	990	1089	1419	1508	1531	1714			840	69	
2,206	POLG_COXA2	P22055	341	581	1028	1125	1453	1540	1562	1745			879	69	
2,214	POLG_COXA4	P36290	340	580	1035	1132	1461	1548	1570	1753			885	69	
2,201	POLG_COXA9	P21404	330	568	1017	1116	1445	1534	1556	1739			867	69	
2,182	POLG_COXB1	P08291	332	570	998	1097	1426	1515	1537	1720			848	69	
2,185	POLG_COXB3	P03313	332	570	1001	1100	1429	1518	1540	1723			851	69	
2,183	POLG_COXB4	P08292	330	568	999	1098	1427	1516	1538	1721			849	69	
2,185	POLG_COXB5	Q03053	330	568	1001	1100	1429	1518	1540	1723			851	69	
1,374	POLG_EC11G	P29813			190	289	618	707	729	912			40		
2,194	POLG_HUEV7	P32537	319	561	1014	1113	1443	1532	1554	1737			867	69	
2,207	POLG_POL1M	P03299	340	578	1029	1126	1455	1542	1564	1746			880	69	
2,209	POLG_POL1S	P03301	341	579	1030	1127	1456	1543	1565	1748			881	69	
2,207	POLG_POL2L	P06210	340	578	1028	1125	1454	1541	1563	1746			879	69	
2,205	POLG_POL2W	P23069	340	578	1028	1125	1454	1541	1563	1746			879	69	
2,206	POLG_POL32	P06209	340	578	1027	1124	1453	1540	1562	1745			878	69	
2,206	POLG_POL3L	P03302	340	578	1027	1124	1453	1540	1562	1745			878	69	
2,185	POLG_SVDVH	P16604	330	568	1001	1100	1429	1518	1540	1723			851	69	
2,185	POLG_SVDVU	P13900	330	568	1001	1100	1429	1518	1540	1723			851	69	
2,209	POLH_POL1M	P03300	341	579	1030	1127	1456	1543	1565	1748			881	69	
Rhino															
2,179	POLG_HRV14	P03303	331	567	1002	1099	1429	1514	1537	1719			858	69	
832	POLG_HRV1A	P23008	307	545									832	44	
2,157	POLG_HRV1B	P12916	332	570	999	1094	1416	1493	1514	1697			857	69	
2,150	POLG_HRV2	P04936	330	567	992	1087	1409	1486	1507	1690			850	69	
2,164	POLG_HRV89	P07210	336	574	1008	1103	1424	1500	1521	1704			866	69	
2,159	HRV85	(NSP)	335	573	1001	1096	1418	1495	1516	1699			859	69	
2,157	HRV9	(NSP)	331	569	1000	1095	1417	1494	1515	1698			858	69	

[a] Auto, VP4–VP2 cleavage; L, aphthovirus L-proteinase cleavage. Cleavage site assignments used in this analysis are indicated by their positions in the polyprotein. Corrected assignments that are erroneous in the SwissProt entries are shown by underline. NSP, not in SwissProt.

because it is not indicated in the SwissProt entry (Table 2). As validation set, the sequences of the human and rabbit eukaryotic translation initiation factor eIF-4G (entries IF4G_HUMAN and IF4G_RABIT, respectively), which have been shown to be cleaved by recombinant HRV2 and coxsackievirus B4 (CoxB4) $2A^{pro}$ in vitro (Sommergruber et al., 1992, 1994; Lamphear et al., 1993, 1995; see Table 2), were used.

For all data sets used for training of neural networks, care was taken to eliminate highly similar cleavage sites because these would have a tendency to skew the prediction from the network.

Data set for $3C^{pro}$ and autocatalytic site neural networks

The prediction of $3C^{pro}$ target sites was split into two: rhino-/enterovirus and aphthovirus. The known cleavage sites used were as described in Table 6. Approximately 80% of the sites were used as training examples, and the remaining 20% were used for testing the neural network. After optimization of window size and the number of hidden units, the neural network was used to scan the SwissProt database.

The data used for training of an autocatalytic (cleavage of VP0 to VP2 and VP4) site neural network were as described in Table 6. The only modifications to the SwissProt data were as described below.

Cleavage sites—Missing or erroneous assignments

Where errors were obvious, the data were modified as follows. Of the 53 picornavirus sequences present in SwissProt, the $3C^{pro}$ cleavage sites have been determined experimentally only for poliovirus (Kitamura et al., 1981). For all other viruses, the cleavage sites had been assigned by similarity. In some cases, cleavage sites were either identical or very similar to those present in poliovirus and therefore straightforward to assign. In other cases, some doubt about the correct assignment of the cleavage site is expressed in the original paper. This uncertainty is not reflected in the SwissProt entry, where only one of the alternatives is indicated. Based on the data available, some of the suggested cleavage sites appear highly unlikely, or else annotated incorrectly, in the SwissProt entry [for errors in the database, see also Korning et al. (1996)]. When using data-driven methods, such as neural networks, and the data set is limited, the quality of the data is extremely important. In the present case, the data set was small enough for a complete visual inspection and therefore errors could be found easily.

Another, more objective way to verify cleavage assignments is to study how a neural network learns to recognize different cleavage patterns (Brunak et al., 1990a, 1990b). Sites that are "difficult" to learn, i.e., they require exorbitantly many learning cycles, are indicative either of an unusual sequence or of an erroneous assignment in the database.

When training the neural net, three kinds of errors were encountered: (1) Missing assignments; (2) erroneous assignments; and (3) erroneous annotation in the database.

Missing assignments

Due to the nature of the neural network algorithm, which uses training on both cleavage and noncleavage sites, a missing assignment may be as serious to the algorithm as an incorrect assignment. Where missing assignments in SwissProt sequence entries were encountered (see Table 7), they were added only when there was no doubt (almost 100% sequence identity with known sites) about the cleavage site position.

Erroneous assignments

Based on multiple alignments and the inability of the neural network algorithm to learn certain sites, it was possible to identify several sequence entries that were misassigned. Examples of obvious wrong cleavage site assignments are shown in Table 8, e.g., the $2A^{pro}$ site in POLG_HRV2. However, the correct cleavage site was used in a recent paper (Sommergruber et al., 1994), but the SwissProt entry obviously has not been updated. The GenBank entries corresponding to the relevant SwissProt entries were also examined; in some cases, the cleavage sites deduced from translation of the nucleotide sequence were assigned erroneously (see also Table 8).

Table 7. *Missing cleavage assignments in SwissProt entries and the proposed cleavage sites based on homology to other members of that virus group*

Site	Sequence	Entry	Position
$2A^{pro}$			
—	SNLG*PFF	POLG_FMDV1	953
—	SNPG*PFF	POLG_FMDVO	953
—	SNPG*PFF	POLG_FMDVT	948
$3C^{pro}$			
—	AEKQ*LKA	POLG_FMDV1	1107
—	IFKQ*ISI	POLG_FMDV1	1425
—	PQQE*GPY	POLG_FMDVO	1601
—	VVKE*GPY	POLG_FMDVO	1625

Erroneous annotation in database

Some of the errors listed in Table 8 were clearly caused by typographical errors that occurred when data were being entered from the original report into SwissProt (indicated by "ok" in the Reference field) or from conversion of the corresponding GenBank DNA entry (indicated by "ok" in the GenBank field).

Quantification of sequence information content

When a large set of sequences is aligned, the Shannon information measure (Shannon, 1948) can be used to quantify the randomness in each column. The information content was computed by the formula:

$$I(i) = \log_2 20 - \sum_{L=1}^{20} p_i^L \log_2 p_i^L, \qquad (1)$$

where p_i^L is the probability of occurrence for a particular amino acid L at position i. The unit of information is bits/amino acid. A completely conserved position will have an information content of 4.32 bits [$\log_2(20)$]. Information content may be displayed in the form of sequence logos (Schneider & Stephens, 1990). Instead of histograms or curves showing the variation of the information content, amino acid symbols themselves are used to represent the value of I at a given position. The sum of the height of the letters indicates the value of I and the height of each letter represents its frequency at the position. This visualization approach is used here and makes it much easier to comprehend the numerical variation than more conventional methods.

The neural network algorithm

The neural networks used were of the standard feed-forward type (Minsky & Papert, 1969; Hertz et al., 1991). They were equipped with an input layer scanning the sequence of amino acids and two layers of processing neurons delivering cleavage site classifications of the sequence fragments in the window. The architecture has been reviewed in many recent papers, where details may be found (Qian & Sejnowski, 1988; MacGregor et al., 1989; Brunak et al., 1991; Rost & Sander, 1993).

Conventional sparse encoding (Qian & Sejnowski, 1988; Hertz et al., 1991) was used to convert the amino acids into numerical form. Adding an extra symbol for handling incomplete windows in the initial and terminal parts of the polyprotein chains, the input consisted of a block of 21 binary values for each amino acid: alanine was represented by (100000000000000000000), cysteine by (010000000000000000000), etc.

The training algorithm was of the gradient descent type, where the adjustable network weights iteratively are modified in order to make the network produce the correct output category on presentations of the sequence fragments. The cleavage (e.g., $3C^{pro}$, $2A^{pro}$, or autocatalytic site) category of the central amino acid decides the output category for each fragment.

When training the networks, a slightly more powerful error function suggested by McClelland was used

$$E = -\sum_{\alpha} \log(1 - (O^{\alpha} - T^{\alpha})^2),$$

replacing the conventional error function (Rumelhart et al., 1986)

$$E = \sum_{\alpha} (O^{\alpha} - T^{\alpha})^2,$$

Table 8. *Entries with erroneous assignments in SwissProt*[a]

Cleavage	Sequence	Entry	Old	New	GenBank	Reference
L^{pro}						
—	RKLK*GAG	POLG_FMDVT	216	201	ERR (X00130)	ERR (Beck et al., 1983)
$2A^{pro}$						
—	LSNY*GAF	POLG_EC11G	43	40	ok (X80059)	ERR (Auvinen & Hyypia, 1990)
—	LTTA*GPG	POLG_HUEV7	871	867	ERR (D00820)	ERR (Ryan et al., 1990)
—	INTF*GGF	POLG_COXA4	888	885	n.a. (D90457)	ok (Supanaranond et al., 1992)
—	LNTH*GAF	POLG_COXA9	870	867	ERR (D00627)	ok (Chang et al., 1989)
—	ITTT*GAF	POLG_COXB1	851	848	ok (M16560)	ok (Iizuka et al., 1987)
—	LITT*GPY	POLG_COXB4	852	849	ERR (X05690)	ERR (Jenkins et al., 1987)
—	SYGL*GPR	POLG_HRV14	856	858	ERR (K02121)	ERR (Stanway et al., 1984)
—	ITTA*GPS	POLG_HRV2	856	850	ERR (X02316)[b]	ERR (Skern et al., 1985)
$3C^{pro}$						
—	ARAQ*GIE	POLG_COXA4	577	580	n.a. (D90457)	ok (Supanaranond et al., 1992)
—	TQSQ*GEI	POLG_POL1S	1747	1748	ok (V01150)	ok (Nomoto et al., 1982)
—	TQSQ*GEI	POLH_POL1M	1747	1748	ok (V01148)	ok (Kitamura et al., 1981)
—	PRSQ*TTA	POLG_FMDVA	723	724	ok (M10975)	ok (Robertson et al., 1985)

[a] The cleavage type and corrected cleavage sequence and entry name are shown, as well as position of erroneous site (Old) and corrected (New). The status of the corresponding GenBank entry is also indicated if available (plus accession number). Original reference and the error status is indicated. n.a., not applicable; ERR, error; ok, no error.

[b] Entry HRV2 (X02316) was recently corrected in the EMBL database.

where T^{α} is the training target value and O^{α} the actual value of the output unit of the network. This logarithmic error function reduced the convergence time considerably, and also has the property of making a given network architecture learn more complex tasks (compared with the standard error measure) without increasing the network size.

Neural network surface exposure prediction

Profile searches and similar pattern matching algorithms have been used previously in attempts to discover cellular targets of viral proteinases (Sommergruber et al., 1994). However, these searches did not take into account that matching sequences might be buried in the folded protein. Therefore, we decided to combine the cleavage site prediction method with an algorithm for the prediction of the surface-forming potential of the site in question.

A neural network was trained to recognize the relative exposure of a selected amino acid in the context of a given protein sequence. Training was performed on data derived from three-dimensional structures of proteins, as present in the Protein Data Bank. The Connolly surface assignment method was used to assign a measure of exposure to each residue (Conolly, 1983; J. Hansen, O. Lund, H. Nielsen, S. Brunak, in prep.). Based on the assumption that cleavage sites must be accessible on the protein surface, we expected that the combined method would reduce the number of possible proteinase targets, as predicted by the cleavage network.

Output from neural networks

The output from the cleavage site prediction networks is a score between 0.0 and 1.0, where scores above the threshold of 0.5 indicate a cleavage site. The surface prediction neural network also produces an output between 0.0 and 1.0, where scores above the threshold of 0.5 indicate an exposed residue and scores below 0.5 a buried residue. Thus, sequences with high cleavage and high surface score are probable targets for the proteinase. Sites scoring low in the surface prediction might well be cleavable when contained within short synthetic peptides, but are probably inaccessible within the context of the native protein.

Cleavage site prediction terminology

Picornaviral amino acid sequences were presented to a neural network in symmetric windows comprising an odd number of amino acids. The central amino acid was designated as either cleavage or noncleavage, and the actual cleavage site was located between the central residue and the following (C-terminal) residue, e.g., the cleavage site V_{858} refers to the cleavage between valine-858 and residue 859. All references to cleavage sites herein refer to the P1 residue according to cleavage terminology, where cleavage takes place between P1 and P1′ (Berger & Schechter, 1970).

Training based on sequence logos

Training of neural networks was performed for each kind of cleavage ($2A^{pro}$, $3C^{pro}$, or autocatalytic). In case of $2A^{pro}$, only cleavages in entero- and rhinovirus polyproteins were examined. In order to cut down on the number of negative training examples, only those sequences that contained specific amino acids at given positions were used. Based on the sequence logos (see Fig. 2), the $2A^{pro}$ network would thus only consider sequences with G at position P1′. Likewise, the $3C^{pro}$ network for rhino-/enterovirus would accept only sequences with Q or E at position P1 (see Fig. 3). The $3C^{pro}$ network for aphthovirus was not limited to specific residues (Fig. 4). The "autocatalytic" (VP4-VP2) site network was limited to sequences containing L at position P2 and E at position P5′ (see Fig. 5).

Electronic prediction server publicly available

A computer server for prediction of cleavage sites by picornaviral proteinases is made publicly available via e-mail or via WWW at http://www.cbs.dtu.dk/services/NetPicoRNA/. To use the e-mail gateway, sequences in single-letter amino acid code of 80 characters per line should be put into a file and mailed to the Internet address NetPicoRNA@cbs.dtu.dk. Mail the word "help" to receive information about the sequence format. The prediction from the networks will be returned promptly.

Acknowledgments

We thank Kristoffer Rapacki, Karsten Dalsgaard, and Hans Henrik Stærfeldt for competent computer assistance. This work was supported by the Danish National Research Foundation.

References

Arnold E, Luo M, Vriend G, Rossmann MG, Palmenberg C, Parks GD, Nicklin MJH, Wimmer E. 1987. Implications of the picornavirus capsid structure for polyprotein processing. *Proc Natl Acad Sci USA 84*:21–25.

Auvinen P, Hyypia T. 1990. Echoviruses include genetically distinct serotypes. *J Gen Virol 71*:2133–2139.

Barco A, Carrasco L. 1995. Poliovirus 2Apro expression inhibits growth of yeast cells. *FEBS Lett 371*:4–8.

Basavappa R, Syed R, Icenogle JP, Filman DJ, Hogle JM. 1994. Role and mechanism of the maturation cleavage of VP0 in poliovirus assembly: Structure of the empty capsid assembly intermediate at 2.9 Å resolution. *Protein Sci 3*:1651–1669.

Beck E, Forss S, Strebel K, Cattaneo R, Feil G. 1983. Structure of the FMDV translation initiation site and of the structural proteins. *Nucleic Acids Res 11*:7873–7885.

Berger A, Schechter I. 1970. Mapping the active site of papain with the aid of peptide substrates and inhibitors. *Phil Trans R Soc Lond Biol 257*:249–264.

Bishop NE, Anderson DA. 1993. RNA-dependent cleavage of VP0 capsid protein in provirions of hepatitis A virus. *Virology 197*:616–623.

Brunak S, Engelbrecht J, Knudsen S. 1990a. Cleaning up gene databases. *Nature 343*:123.

Brunak S, Engelbrecht J, Knudsen S. 1990b. Neural network detects errors in the assignment of pre-mRNA splice site. *Nucleic Acids Res 18*:4797–4801.

Brunak S, Engelbrecht J, Knudsen S. 1991. Prediction of human mRNA donor and acceptor sites from the DNA sequence. *J Mol Biol 220*:49–65.

Cao X, Wimmer E. 1996. Genetic variation of the poliovirus genome with two VPg coding units. *EMBO J 15*:23–33.

Chang KH, Auvinen P, Hyypia T, Stanway G. 1989. The nucleotide sequence of coxsackievirus A9; implications for receptor binding and enterovirus classification. *J Gen Virol 70*:3269–3280.

Clark ME, Dasgupta A. 1990. A transcriptionally active form of TFIIIC is modified in poliovirus-infected HeLa cells. *Mol Cell Biol 10*:5106–5113.

Clark ME, Hammerle T, Wimmer E, Dasgupta A. 1991. Poliovirus proteinase 3C converts an active form of transcription factor IIIC to an inactive form: A mechanism for inhibition of host cell polymerase III transcription by poliovirus. *EMBO J 10*:2941–2947.

Clark ME, Lieberman PM, Berk AJ, Dasgupta A. 1993. Direct cleavage of human TATA-binding protein by poliovirus protease 3C in vivo and in vitro. *Mol Cell Biol 13*:1232–1237.

Conolly ML. 1983. Solvent-accessible surfaces of proteins and nucleic acids. *Science 221*:709–713.

Cordingley MG, Callahan PL, Sardana VV, Garsky VM, Colonno RJ. 1990. Substrate requirements of human rhinovirus 3C protease for peptide cleavage in vitro. *J Biol Chem 265*:9062–9065.

Cordingley MG, Register RB, Callahan PL, Garsky VM, Colonno RJ. 1989. Cleavage of small peptides in vitro by human rhinovirus 14 3C protease expressed in *Escherichia coli*. *J Virol 63*:5037–5045.

Falk MM, Grigera PR, Bergmann IE, Zibert A, Multhaup G, Beck E. 1990. Foot-and-mouth disease virus protease 3C induces specific proteolytic cleavage of host cell histone H3. *J Virol 64*:748–756.

Harber JJ, Bradley J, Anderson CW, Wimmer E. 1991. Catalysis of poliovirus VP0 maturation cleavage is not mediated by serine 10 of VP2. *J Virol 65*:326–334.

Hellen CU, Lee CK, Wimmer E. 1992. Determinants of substrate recognition by poliovirus 2A proteinase. *J Virol 66*:3330–3338.

Hellen CUT, Krausslich HG, Wimmer E. 1989. Proteolytic processing of polyproteins in the replication of RNA viruses. *Biochemistry 28*:9881–9890.

Hertz J, Krogh A, Palmer R. 1991. *Introduction to the theory of neural computation*. Redwood City, California: Addison-Wesley.

Iizuka N, Kuge S, Nomoto A. 1987. Complete nucleotide sequence of the genome of coxsackievirus B1. *Virology 156*:64–73.

Jenkins O, Booth JD, Minor PD, Almond JW. 1987. The complete nucleotide sequence of coxsackievirus B4 and its comparison to other members of the Picornaviridae. *J Gen Virol 68*:1835–1848.

Joachims M, Etchison D. 1992. Poliovirus infection results in structural alteration of a microtubule-associated protein. *J Virol 66*:5797–5804.

Joachims M, Harris KS, Etchison D. 1995. Poliovirus protease 3C mediates cleavage of microtubule-associated protein 4. *Virology 211*:451–461.

Kitamura N, Semler BL, Rothberg PG, Larsen GR, Adler CJ, Dorner AJ, Emini EA, Hanecak R, Lee JJ, vander Werf S, Anderson CW, Wimmer E. 1981. Primary structure, gene organization and polypeptide expression of poliovirus RNA. *Nature 291*:547–553.

Klump H, Auer H, Liebig HD, Kuechler E, Skern T. 1996. Proteolytically active 2A proteinase of human rhinovirus 2 is toxic for *Saccharomyces cerevisiae* but does not cleave the homologues of eIF-4 gamma in vivo or in vitro. *Virology 220*:109–118.

Korning PG, Hebsgaard SM, Rouzé P, Brunak S. 1996. Cleaning the GenBank *Arabidopsis thaliana* data set. *Nucleic Acids Res 24*:316–320.

Krausslich HG, Wimmer E. 1988. Viral proteinases. *Annu Rev Biochem 57*:701–754.

Lamphear BJ, Kirchweger R, Skern T, Rhoads RE. 1995. Mapping of functional domains in eukaryotic protein synthesis initiation factor 4G (eIF4G) with picornaviral proteases. Implications for cap-dependent and cap-independent translational initiation. *J Biol Chem 270*:21975–21983.

Lamphear BJ, Yan R, Yang F, Waters D, Liebig HD, Klump H, Kuechler E, Skern T, Rhoads RE. 1993. Mapping the cleavage site in protein synthesis initiation factor eIF-4 gamma of the 2A proteases from human coxsackievirus and rhinovirus. *J Biol Chem 268*:19200–19203.

Lawson MA, Semler BL. 1990. Picornavirus protein processing—Enzymes, substrates, and genetic regulation. *Curr Topics Microbiol Immunol 161*:49–88.

Lee CK, Wimmer E. 1988. Proteolytic processing of poliovirus polyprotein: Elimination of 2Apro-mediated, alternative cleavage of polypeptide 3CD by in vitro mutagenesis. *Virology 166*:405–414.

Liebig HD, Ziegler E, Yan R, Hartmuth K, Klump H, Kowalski H, Blaas D, Sommergruber W, Frasel L, Lamphear B, Rhoades R, Kuechler E, Skern T. 1993. Purification of two picornaviral 2A proteinases: Interaction with eIF-4 gamma and influence on in vitro translation. *Biochemistry 32*:7581–7588.

Long AC, Orr DC, Cameron JM, Dunn BM, Kay J. 1989. A consensus sequence for substrate hydrolysis by rhinovirus 3C proteinase. *FEBS Lett 258*:75–78.

MacGregor MJ, Flores TP, Sternberg MJE. 1989. Prediction of beta-turns in proteins using neural networks. *Protein Eng 2*:521–526.

Matthews DA, Smith WW, Ferre RA, Condon B, Budahazi G, Sisson W, Villafranca JE, Janson CA, McElroy HE, Gribskov CL, Worland S. 1994. Structure of human rhinovirus 3C protease reveals a trypsin-like polypeptide fold, RNA-binding site, and means for cleaving precursor polyprotein. *Cell 77*:761–771.

McLean C, Matthews TJ, Rueckert RR. 1976. Evidence of ambiguous processing and selective degradation in the noncapsid proteins of rhinovirus 1A. *J Virol 19*:903–914.

Minsky M, Papert S. 1969. *Perceptrons*. Cambridge, Massachusetts: MIT Press.

Nomoto A, Omata T, Toyoda H, Kuge S, Horie H, Kataoka Y, Genba Y, Nakano Y, Imura N. 1982. Complete nucleotide sequence of the attenuated poliovirus Sabin 1 strain genome. *Proc Natl Acad Sci USA 79*:5793–5797.

Qian N, Sejnowski TJ. 1988. Predicting the secondary structure of globular proteins using neural network models. *J Mol Biol 202*:865–884.

Robertson BH, Grubman MJ, Weddell GN, Moore DM, Welsh JD, Fischer T, Dowbenko DJ, Yansura DG, Small B, Kleid DG. 1985. Nucleotide and amino acid sequence coding for polypeptides of foot-and-mouth disease virus type A12. *J Virol 54*:651–660.

Rost B, Sander C. 1993. Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol 232*:584–599.

Rumelhart DE, Hinton GE, Williams RJ. 1986. Learning internal representations by error propagation. In: Rumelhart D, McClelland J, PDP Research Group, eds. *Parallel distributed processing: Explorations in the microstructure of cognition, vol 1. Foundations*. Cambridge, Massachusetts. MIT Press. pp 318–362.

Ryan MD, Drew J. 1994. Foot-and-mouth disease virus 2A oligopeptide mediated cleavage of an artificial polyprotein. *EMBO J 13*:928–933.

Ryan MD, Jenkins O, Hughes PJ, Brown A, Knowles NJ, Booth D, Minor PD, Almond JW. 1990. The complete nucleotide sequence of enterovirus type 70: Relationships with other members of the picornaviridae. *J Gen Virol 71*:2291–2299.

Schneider TD, Stephens RM. 1990. Sequence logos: A new way to display consensus sequences. *Nucleic Acids Res 18*:6097–6100.

Schultheiss T, Kusov YY, Gauss-Muller V. 1994. Proteinase 3C of hepatitis A virus (HAV) cleaves the HAV polyprotein P2-P3 at all sites including VP1/2A and 2A/2B. *Virology 198*:275–281.

Shannon CE. 1948. A mathematical theory of communication. *Bell System Tech J 27*:379–423/623–656.

Shen Y, Igo M, Yalamanchili P, Berk AJ, Dasgupta A. 1996. DNA binding domain and subunit interactions of transcription factor IIIC revealed by dissection with poliovirus 3C protease. *Mol Cell Biol 16*:4163–4171.

Skern T, Sommergruber W, Blaas D, Gruendler P, Fraundorfer F, Pieler C, Fogy I, Kuechler E. 1985. Human rhinovirus 2: Complete nucleotide sequence and proteolytic processing signals in the capsid protein region. *Nucleic Acids Res 13*:2111–2126.

Sommergruber W, Ahorn H, Klump H, Seipelt J, Zophel A, Fessl F, Krystek E, Blaas D, Kuechler E, Liebig HD, Skern T. 1994. 2A proteinases of coxsackie- and rhinovirus cleave peptides derived from eIF-4gamma via a common recognition motif. *Virology 198*:741–745.

Sommergruber W, Ahorn H, Zophel A, Maurer-Fogy I, Fessl F, Schnorrenberg G, Liebig HD, Blaas D, Kuechler E, Skern T. 1992. Cleavage specificity on synthetic peptide substrates of human rhinovirus 2 proteinase 2A. *J Biol Chem 267*:22639–22644.

Stanway G, Hughes PJ, Mountford RC, Minor PD, Almond JW. 1984. The complete nucleotide sequence of a common cold virus: Human rhinovirus 14. *Nucleic Acids Res 12*:7859–7875.

Supanaranond K, Takeda N, Yamazaki S. 1992. The complete nucleotide sequence of a variant of Coxsackievirus A24, an agent causing acute hemorrhagic conjunctivitis. *Virus Genes 6*:149–158.

Tesar M, Marquardt O. 1990. Foot-and-mouth disease virus protease 3C inhibits cellular transcription and mediates cleavage of histone H3. *Virology 174*:364–374.

Urzanqui A, Carrasco L. 1989. Degradation of cellular proteins during poliovirus infection: Studies by two-dimensional gel electrophoresis. *J Virol 63*:4729–4735.

Ypma-Wong MF, Filman DJ, Hogle JM, Semler BL. 1988. Structural domains of the poliovirus polyprotein are major determinants for proteolytic cleavage at Gln-Gly pairs. *J Biol Chem 263*:17846–17856.