Prediction of the Virulencies of Some Enveloped Viruses from the Structure of the Cleavage Recognition Site of Viral Glycoprotein Essential for Infectivity. II. Deviation Analysis

Yukio Kiho, Keizo Miyata, and Yoshio Okada*

Central Research Laboratory, Ishihara Sangyo kaisha, Ltd., Kasatsu, Shigaken, Japan 525
* Institute for Molecular and Cellular Biology, Osaka University, 1-3, Yamadaoka, Suita, Osaka 565, Japan

ABSTRACT. Analysis of the amino acid sequence in protein (deviation analysis) suggests that the binding between trypsin and the enveloped virus is the first step of their interaction, which occurs in a specified configuration. It is possible that the distance between their active sites is important for the viral sensitivity to trypsin, which is related to the virulence of the enveloped virus.

In a previous report (1), the amino acid sequence distribution in protein was analysed. It was found that the frequency of occurrence of each amino acid in a defined sequence with functional significance more or less deviates from uniform distribution. The deviation found in this defined sequence seems to parallel the function of this sequence. On the basis of this finding, peptide-peptide interaction was analysed. Suppose that the receptor protein ABC binds with ligand peptide E at site B, forming complex ABEC. As no more E can bind with B in ABEC, the binding function of B is lost and the deviation of the B site at ABC will decrease at ABEC. This expectation was realized in the interaction of trypsin with its inhibitor (1).

In this study, the deviation of the viral glycoprotein region containing the cleavage recognition site was examined and the effect of the incorporation of the active site of trypsin was observed. When the deviation of the viral site decreased, we assigned it as the binding site for trypsin.

MATERIALS AND METHODS

Deviation was formulated (1) as:

$$dev = \sum_{z=A}^{V} Dz = \sum_{z=A}^{V} [N(z) - xF(z)]^2,$$

Abbreviations used: SV5, simian virus 5; ReSV, respiratory syncytial virus; VISV, Visna virus; NDV (vir), Newcastle disease virus, virulent strain; NDV (avir), Newcastle disease virus, avirulent strain; RSV, Rous sarcoma virus; MMTV, mouse mammary tumor virus; HIV, human immunodeficiency virus type I; FPV, fowl plague virus; HTLV-1, human T lymphotropic virus; WSN, influenza virus A/WSN/33; PR8, influenza virus PR8; HVJ (Z), Sendai virus, Z strain; HVJ (TR2), Sendai virus, TR2 strain. The amino acid abbreviations follow standard conventions.
where \( z \) = amino acid, A, R, N, D, C, Q, E, G, H, L, I, K, M, F, P, S, T, W, Y, V, \( x \) = size of the restricted sequence for calculation, \( F(z) \) = frequency of the amino acid \( z \) in envelope protein, \( N(z) \) = number of amino acid \( z \) in the restricted sequence.

In this report, a seven-consecutive-amino-acid sequence (restricted sequence, \( x = 7 \)) in the viral region around the viral recognition cleavage site was examined. Then a three-amino acid consecutive sequence around the active site of trypsin was added and the entire sequence was examined again (\( x = 10 \)). In each case, specific deviation (sp-dev) was calculated using the program “SP-DEV” (Appendix).

\[
\text{sp-dev} = \frac{\text{dev}}{x}
\]

When sp-dev at \( x = 10 \) was less than that at \( x = 7 \), the following DD (deviation decrease) was calculated:

\[
\text{DD} = \left(1 - \frac{\text{sp-dev at } x = 10}{\text{sp-dep at } x = 7}\right) \times 100.
\]

The interaction energy between trypsin and virus was calculated as reported in the previous paper, I, of this series (2), in which viral virulency was analysed in terms of the interaction energy. Experimental findings on the viral virulency were also reviewed in ref. 2.

**RESULTS AND DISCUSSION**

The deviation profile of trypsin is shown in Fig. 1. The deviation pattern around the region containing the active site (171-D) is fairly simple but is overlapped with neighbours. Details around this region are shown in Fig. 2. As it is possible that the neighbours are functionally related to the active site region, calculation was carried out over a fairly wide range (CALC region), 164-GYLEGGKDSCQGD.

<table>
<thead>
<tr>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>60</th>
</tr>
</thead>
<tbody>
<tr>
<td>IGGYTCGQN</td>
<td>TYPYQVSLNS</td>
<td>GYHPCGQLS</td>
<td>NSQYVSAAH</td>
<td>CYKSIQVYRL</td>
<td>GEDNINVEG</td>
</tr>
<tr>
<td>70</td>
<td>60</td>
<td>90</td>
<td>100</td>
<td>110</td>
<td>120</td>
</tr>
<tr>
<td>NEQPISAKS</td>
<td>IVHPINPVNT</td>
<td>LANDMILNL</td>
<td>KMSASLNGY</td>
<td>ASISLPTSCA</td>
<td>SAGTMILSG</td>
</tr>
<tr>
<td>130</td>
<td>140</td>
<td>150</td>
<td>160</td>
<td>170</td>
<td>180</td>
</tr>
<tr>
<td>WGNKFGSGTS</td>
<td>YPDYKCLKA</td>
<td>PILSDDSCRS</td>
<td>AYQQTINSMK</td>
<td>FCAYLEGGKI</td>
<td>DCSCGDSGGP</td>
</tr>
<tr>
<td>190</td>
<td>200</td>
<td>210</td>
<td>220</td>
<td></td>
<td></td>
</tr>
<tr>
<td>YVCGRLEQGI</td>
<td>YSWSGCAQK</td>
<td>NRPGVVTKVC</td>
<td>NYSNKWQTI</td>
<td>ASN</td>
<td></td>
</tr>
</tbody>
</table>

**Total amino acid residues = 223**

\( A = 14 \quad R = 2 \quad N = 16 \quad D = 6 \quad C = 11 \)

\( Q = 9 \quad E = 4 \quad G = 26 \quad H = 3 \quad I = 15 \)

\( L = 14 \quad K = 14 \quad M = 3 \quad F = 3 \quad P = 9 \)

\( S = 34 \quad T = 10 \quad W = 4 \quad Y = 10 \quad V = 17 \)

![Graph showing deviation](image)

**Fig. 1.** Deviation was calculated (\( x = 10 \)) and shown in histograms. Active site (171-D) is indicated by an arrow.
Deviation profiles around the recognition cleavage site of the envelope protein of the NDV virulent strain and avirulent strain are shown in Fig. 3. Sp-dev of the restricted sequence (x=7) was calculated consecutively, as is shown in the first column of Table 1. To each sequence, three amino acids from the trypsin CALC region (top row in Table 1) were added consecutively and again sp-dev was calculated. When the sp-dev of the restricted region was decreased by the addition of three amino acids of trypsin, DD was calculated as shown in Table 1. These DD values are plotted along the sequence of the CALC region and are shown in Fig. 4(a). GYL and LEG had a similar pattern and a peak at 90-R, indicating that these peptides are near 90-R. Sp-dev and DD values are assigned to the central amino acid in the restricted region and/or in the three amino acids added. Following the hypothesis (1), "deviation (sp-dev) parallels the function," we estimated the binding activity by DD value. YLE showed fairly large DD over the entire CALC region, in which 84-S is nearest. These data suggest that Y and E are close to 90-R and L is positioned close to 84-S, but also not far from other points in the entire region.

Fig. 3. Deviation around the recognition cleavage site of NDV, (a) virulent strain, (b) avirulent strain.
TABLE 1. **Effect of trypsin portion on the deviation of the CALC region of NDV (virulent strain).**

<table>
<thead>
<tr>
<th>NDV seq</th>
<th>sp-dev</th>
<th>TRYPSIN</th>
<th>GYL</th>
<th>YLE</th>
<th>LEG</th>
<th>EGG</th>
<th>GGK</th>
<th>GKD</th>
<th>DSC</th>
<th>SCQ</th>
<th>CQG</th>
<th>QGD</th>
</tr>
</thead>
<tbody>
<tr>
<td>81-VTTSGGR</td>
<td>0.9861</td>
<td></td>
<td>2</td>
<td>35</td>
<td>4</td>
<td>8</td>
<td>5</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TTSGGR</td>
<td>1.3369</td>
<td></td>
<td>8</td>
<td>32</td>
<td>9</td>
<td>12</td>
<td>10</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TSGGR</td>
<td>1.1596</td>
<td></td>
<td>1</td>
<td>30</td>
<td>4</td>
<td>7</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SGGRQR</td>
<td>1.8656</td>
<td></td>
<td>10</td>
<td>28</td>
<td>11</td>
<td>2</td>
<td>13</td>
<td>12</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GGRQR</td>
<td>2.8312</td>
<td></td>
<td>16</td>
<td>27</td>
<td>17</td>
<td>10</td>
<td>25</td>
<td>24</td>
<td>18</td>
<td>3</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>GRRQRF</td>
<td>2.6713</td>
<td></td>
<td>21</td>
<td>26</td>
<td>22</td>
<td>4</td>
<td>3</td>
<td>15</td>
<td>23</td>
<td>22</td>
<td>15</td>
<td>7</td>
</tr>
<tr>
<td>RRQRF</td>
<td>2.6756</td>
<td></td>
<td>28</td>
<td>26</td>
<td>29</td>
<td>19</td>
<td>18</td>
<td>22</td>
<td>23</td>
<td>22</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>RQRF</td>
<td>1.7273</td>
<td></td>
<td>17</td>
<td>25</td>
<td>19</td>
<td>8</td>
<td>21</td>
<td>19</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 4. Decrease of the deviation (DD) due to the insertion of trypsin portion is plotted (a). Possible structure model based on DD is shown (b).

Fig. 5. Same as Fig. 4, but for NDV, avirulent strain.
Similar examination with GKD, KDS and DSC suggests that 171-D is close to 88-R. From these data, we can construct the structure model shown in Fig. 4(b). The active site 171-D of trypsin is positioned near four amino acids upstream of the cleavage site (between 91-R and 92-F). Similar analysis was done with the avirulent strain of NDV (Fig. 5). Generally, DD values are smaller than that of the virulent strain, except in the case of YLE. Furthermore, 171-D is nearest to 84-S and is far from the cleavage site. Thus, in the case of the avirulent strain, 171-D binds weakly to a position distant from the cleavage site, which may explain the weak virulency of this virus.

Deviation analysis was done with two strains of HVJ. They are 1) the Z strain, which is sensitive to trypsin, and 2) the TR-2 strain, which is resistant to trypsin. Figure 6 shows the DD analysis of both strains interacted with the trypsin portion. On GYLE binding, similar patterns were observed with both strains, and they are also similar to NDV. With KDS, as with the avirulent strain of NDV, binding oc-

![Graphs showing deviation analysis for two strains of HVJ.](image-url)
Y. Kiho, K. Miyata and Y. Okada

An analysis of the trypsin-virus interaction was extended to other enveloped viruses. The results are shown in Table 2. Binding of trypsin-YLE and -KDS to the virus was examined. It was found that 1) in the case of a virulent virus, trypsin-KDS binds near the cleavage site and trypsin-YLE bindings upstream as in the case of the virulent strain of NDV, and 2) in the case of a less virulent virus, trypsin-KDS binding took place to a lesser extent (low DD) and far from the cleavage site of the virus. In some cases, for example, PR8, WSN or HTLV-1, trypsin-YLE binds downstream of the binding site of trypsin-KDS. Although the meaning of the binding of trypsin-YLE to the virus is not clear at present, it is possible that a somewhat different interaction occurs in these cases, which might also explain the lesser sensitivity to trypsin.

The binding site deduced from the deviation analysis so far are more or less distant from the cleavage sites. We infer that this binding takes place as the first step of virus-trypsin interaction, followed by the interaction of active sites and the cleavage site.

![Diagram](image)

Fig. 7. Acrophilicity profiles of the region around 108-Q of HVJ are shown. Circle represents amino acid, its radius is equal to van der Waals radius, and the center corresponds to its acrophilicity value. Acrophilicity profile of the active site of trypsin (KDS) is positioned as close as possible to those of HVJ. Distances between the amino acids in the respective profile were determined on pairs less than 50 A apart. Interaction energy was then calculated.
reaction. By the above analysis, we could not show the latter step, probably because of the unstable nature of the binding, which was already discussed in the previous report (1). Summing up the results obtained, we can say that the trypsin sensitivity of the enveloped virus seems to depend on 1) the distance between the active site (171-D) of trypsin and the cleavage site of the virus; 2) the mutual configuration in the first step of their interaction, the binding which is derived from the deviation analysis; and 3) the amount of DD. In addition to the binding of 171-D, trypsin-YLE, several amino acids upstream from 171-D, also bindings near the cleavage site of the virus.

Some comments on the nature of the above methods should be made. Deviation was defined as the composition of amino acid within the restricted portion of the sequence relative to the mean (1). Further details on the deviation are cannot be determined, although it may include hydrophilicity, hydrophobicity, acrophilicity and so on. Thus, the deviation analysis gives only the functional site along the sequence. Our hypothesis is that the deviation parallels the function. Since the deviation is defined on the basis of the compositional difference from the mean, still other sequences besides the natural one may be more or less functional. Actually, even the so-called consensus sequence has some variety. Such examples may be found in the future by improving the assay method for function and by introducing mutation.
REFERENCES


(Received for publication, July 14, 1989 and in revised form, September 21, 1989)
Appendix I Program; SP-DEV

1000 'Specific Deviation in a Specified Region of Protein
1010 'SP-DEV
1020 
1030 DIM N$(6000):SP$=""
1040 CLS 3:CONSOLE 0,25,0,1:SCREEN 3,0,0
1050 FILES 2:PRINT:PRINT
1060 INPUT "File Name ----- ",SN$
1070 OPEN "2:"+SN$ FOR INPUT AS #1
1080 INPUT #1,F$,CMT$ 
1090 NSEQ=0
1100 WHILE NOT EOF(1)
1110 INPUT #1,A$
1120 NSEQ=NSEQ+1:NS(NSEQ)=AS
1130 WEND:CLOSE #1
2000 '
2010 'Display
2020 '
2030 CLS 3
2040 COLOR 4:PRINT CMT$
2050 PRINT
2060 COLOR 7:PRINT "Total amino acids ";NSEQ
2070 PRINT:PRINT
2080 CL=60:CB=10
2090 P$="###########"
2100 PL=1-CL
2110 PL=PL+CL:PR=PL+CL-1
2120 IF PR>NSEQ THEN PR=NSEQ
2130 FOR I=PL TO PR
2140 IF I MOD CB=0 THEN PRINT USING P$;1;
2150 NEXT I:PRINT
2160 PRINT SPS;
2170 FOR I=PL TO PR
2180 PRINT N$(1);
2190 IF I MOD CB=0 THEN PRINT SPS;
2200 NEXT I
2210 PRINT:IF PR=NSEQ THEN LOCATE 0,15:GOTO 2300
2220 GOTO 2110
2300 LOCATE 0,24:PRINT "DO YOU PRINT THE SEQUENCE(Y/N)?";ANS$
2310 IF ANS$="Y" THEN PRINT USING P$;1;
2320 NEXT I:PRINT
2330 IF PR=NSEQ THEN PRINT SPS;
2340 FOR I=PL TO PR
2350 IF I MOD CB=0 THEN PRINT USING P$;1;
2360 NEXT I
2370 PRINT SPS;
2380 FOR I=PL TO PR
2390 IF I MOD CB=0 THEN PRINT USING P$;1;
2400 NEXT I:PRINT
2410 LPRINT SPS;
2420 FOR I=PL TO PR
2430 LPRINT N$(I);
2440 IF I MOD CB=0 THEN LPRINT SPS;
2450 NEXT I
2460 LPRINT:IF PR=NSEQ THEN GOTO 2480
2470 GOTO 2350
2480 'calculation for the amino acid composition
2490 2500 NA=0:NR=0:NN=0:ND=0:NC=0:NQ=0:NE=0:NG=0:NH=0:N1=0:NL=0:NK=0:NM=0:NF=0:NP=0:
NS=0:NT=0:NW=0:NY=0:NV=0
2510 FOR I=1 TO NSEQ
2520 IF NS$(I)="A" THEN NA=NA+1
2530 IF NS$(I)="R" THEN NR=NR+1
2540 IF NS$(I)="N" THEN NN=NN+1
2550 IF NS$(I)="D" THEN ND=ND+1
2560 IF NS$(I)="C" THEN NC=NC+1
2570 IF NS$(I)="Q" THEN NQ=NQ+1
2580 IF NS$(I)="E" THEN NE=NE+1
2590 IF NS$(I)="G" THEN NG=NG+1
2600 IF NS$(I)="H" THEN NH=NH+1
2610 IF NS$(I)="I" THEN NI=NI+1
2620 IF NS$(I)="L" THEN NL=NL+1
2630 IF NS$(I)="K" THEN NK=NK+1
2640 IF NS$(I)="M" THEN NM=NM+1
2650 IF NS$(I)="F" THEN NF=NF+1
2660 IF NS$(I)="P" THEN NP=NP+1
2670 IF NS$(I)="S" THEN NS=NS+1
2680 IF NS$(I)="T" THEN NT=NT+1
2690 IF NS$(I)="W" THEN NW=NW+1
2700 IF NS$(I)="Y" THEN NY=NY+1
2710 NEXT I
2720 IF I=1 TO NSEQ
Y. Kiho, K. Miyata and Y. Okada

2710 IF N$(I)="V" THEN NV=NV+1
2720 NEXT I
2730 PRINT USING "R=### P=### D=### F=###";NR,NP,ND,NF
2740 PRINT USING "C=### L=### E=### T=###";NC,NL,NE,NT
2750 PRINT USING "G=### K=### A=### I=###";NG,NK,NA,NI
2760 PRINT USING "Y=### N=### Q=### V=###";NY,NN,NQ,NV
2770 PRINT USING "M=### W=### H=### S=###";NM,NW,NH,NS
2780 PRINT:PRINT
2785 INPUT "DO YOU PRINT THE COMPOSITION(Y/N)?";ANS$"N"
2790 IF ANS$="N" THEN GOTO 3010
2795 IF ANS$="Y" THEN GOTO 2800
2800 LPRINT:IF N$(I)="V" THEN GOTO 3000
2810 LPRINT USING "A=### R=### N=### D=### C=###";NA,NR,NN,ND,NC
2820 LPRINT USING "Q=### E=### G=### H=### I=###";NQ,NE,NG,NH,NI
2830 LPRINT USING "L=### K=### M=### F=### P=###";NL,NK,NM,NF,FP
2840 LPRINT USING "S=### T=### W=### Y=### V=###";NS,NT,NW,NY,NV
2850 LPRINT:IF N$(I)="V" THEN GOTO 3000
2860 'DEVIATION CALCULATION IN SPECIFIED REGION
2870 'CLS 3
2880 PRINT:INPUT "SIZE OF THE CALC REGION";RNA
2890 INPUT "A=";PA
2900 INPUT "R=";PR
2910 INPUT "N=";PN
2920 INPUT "D=";PD
2930 INPUT "C=";PC
2940 INPUT "Q=";PQ
2950 INPUT "E=";PE
2960 INPUT "G=";PG
2970 INPUT "H=";PH
2980 INPUT "L=";PL
2990 INPUT "I=";PI
3000 INPUT "K=";PK
3010 INPUT "M=";PM
3020 INPUT "F=";PF
3030 INPUT "P=";PP
3040 INPUT "S=";PS
3050 INPUT "T=";PT
3060 INPUT "W=";PW
3070 INPUT "Y=";PY
3080 DA=(PA-RNA*NA/NSEQ)-2
3090 DR=(PR-RNA*NR/NSEQ)-2
3100 DN=(PN-RNA*NN/NSEQ)-2
3110 DD=(PD-RNA*ND/NSEQ)-2
3120 DC=(PC-RNA*NC/NSEQ)-2
3130 DQ=(PQ-RNA*NQ/NSEQ)-2
3140 DE=(PE-RNA*NE/NSEQ)-2
3150 DG=(PG-RNA*NG/NSEQ)-2
3160 DH=(PH-RNA*NH/NSEQ)-2
3170 DL=(PL-RNA*NL/NSEQ)-2
3180 DI=(PI-RNA*N1/NSEQ)-2
3190 DK=(PK-RNA*NK/NSEQ)-2
3200 DM=(PM-RNA*NM/NSEQ)-2
3210 DS=(PS-RNA*NS/NSEQ)-2
3220 DT=(PT-RNA*NT/NSEQ)-2
3230 DW=(PW-RNA*NW/NSEQ)-2
3240 DY=(PY-RNA*NV/NSEQ)-2
3250 DEV=DA+DR+DN+DD+DC+DQ+DE+DH+DL+DI+DK+DM+DS+DT+DW+DY+DY
3260 SPDEV=DEV/RNA
3270 PRINT "DEV=DEV/RNA"
3280 PRINT "DEVIATION = ":DEV
3290 PRINT "SPECIFIC DEVIATION = ":SPDEV
3300 INPUT "TRY AGAIN(Y/N)?";TRY$"N"
3310 IF TRY$="N" THEN END
3320 IF TRY$="Y" THEN 3010