Online Supporting Information A

The benchmark data for predicting the cleavage sites in proteins by HIV-1 and HIV-2 proteases (from K.C. Chou, Prediction of HIV protease cleavage sites in proteins, Analytical Biochemistry, 1996, 233, 1-14)

(1A) $\mathbb{S}_1^+$: list of 64 cleavable octapeptides by HIV-1 protease

CANLSTFA
VVIATVIV
TQIMFETF
GQVNYEEF
PFIFEEEP
SFNFPQIT
DTVLEEMS
ARVLAEAM
AEELAEIF
SLNLRETN
ATIMMQRG
AECFRIFD
DQILIEIC
DDLFFEAD
YEEFVQMM
PIVGAETF
TLNFPISP
REAFRVFD
AETFYVDK
AQTFYVNL
PTLLTEAP
SFIGMESA
DAINTEFK
QITLWQRP
ELEFPEGG
ANLAEEA
SQNYPIVQ
PGNFLQSR
KLVFFAE
GDALLERN
KELYPLTS
RQANFLGK
SRSLYASS
AEAMSQVT
RKILFLDG
GSHLVEAL
GGVYATRS
FRSGVETT
VEVAEEEE
LPVNGEFS
ETTALVCD
HLVEALYL
HYGFPTYG
DSADAEED
GWILGEHG

GWILAEHG
QAIYLALQ
EKVYLAWV
VEICTEME
TQDFWEVQ
LWMGYELH
GDAYFSVP
ELELAENR
SKDLIAEI
LEVNIVTD
GGNYPVQH
ARLMAEAL
PFAAAQQR
PRNFPVAQ
GLAAPQFS
SLNLPVAK
AETFYTDG
RQVLFLEK
QMIFEEHG

(1B) $\mathbb{S}_1^-$: list of 239 non-cleavable octapeptides by HIV-1 protease

KVFGRCEL
VFGRCELA
FGRCELAA
GRCELAAA
RCELAAAM
CELAAAMK
ELAAAMKR
LAAAMKRH
AAAMKRHG
AAMKRHGL
AMKRHGLD
MKRHGLDN
KRHGLDNY
RHGLDNYR
HGLDNYRG
GLDNYRGY
LDNYRGYS
DNYRGYSL
NYRGYSLG
YRGYSLGN
RGYSLGNW
GYSLGNWV
YSLGNWVC
SLGNWVCA
LGNWVCAA
GNWVCAAK
NWVCAAKF
WVCAAKFE
VCAAKFES
CAAKFESN
AAKFESNF

AKFESNFN
KFESNFNT
FESNFNTQ
ESNFNTQA
SNFNTQAT
NFNTQATN
FNTQATNR
NTQATNRN
TQATNRNT
QATNRNTD
ATNRNTDG
TNRNTDGS
NRNTDGST
RNTDGSTD
NTDGSTDY
TDGSTDYG
DGSTDYGI
GSTDYGIL
STDYGILQ
TDYGILQI
DYGILQIN
YGILQINS
GILQINSR
ILQINSRW
LQINSRWW
QINSRWWC
INSRWWCN
NSRWWCND
SRWWCNDG
RWWCNDGR
WWCNDGRT
WCNDGRTP
CNDGRTPG
NDGRTPGS
DGRTPGSR
GRTPGSRN
RTPGSRNL
TPGSRNLC
PGSRNLCN
GSRNLCNI
SRNLCNIP
RNLCNIPC
NLCNIPCS
LCNIPCSA
CNIPCSAL
NIPCSALL
IPCSALLS
PCSALLSS
CSALLSSD
SALLSSDI
ALLSSDIT
LLSSDITA
LSSDITAS
SSDITASV

SDITASVN
DITASVNC
ITASVNCA
TASVNCAK
ASVNCAKK
SVNCAKKI
VNCAKKIV
NCAKKIVS
CAKKIVSD
AKKIVSDG
KKIVSDGN
KIVSDGNG
IVSDGNGM
VSDGNGMN
SDGNGMNA
DGNGMNAW
GNGMNAWV
NGMNAWVA
GMNAWVAW
MNAWVAWR
NAWVAWRN
AWVAWRNR
WVAWRNRC
VAWRNRCK
AWRNRCKG
WRNRCKGT
RNRCKGTD
NRCKGTDV
RCKGTDVQ
CKGTDVQA
KGTDVQAW
GTDVQAWI
TDVQAWIR
DVQAWIRG
VQAWIRGC
QAWIRGCR
AWIRGCRL
KETAAAKF
ETAAAKFE
TAAAKFER
AAAKFERQ
AAKFERQH
AKFERQHM
KFERQHMD
FERQHMDS
ERQHMDSS
RQHMDSST
QHMDSSTS
HMDSSTSA
MDSSTSAA
DSSTSAAS
SSTSAASS
STSAASSS
TSAASSSN

SAASSSNY
AASSSNYC
ASSSNYCN
SSSNYCNQ
SSNYCNQM
SNYCNQMM
NYCNQMMK
YCNQMMKS
CNQMMKSR
NQMMKSRN
QMMKSRNL
MMKSRNLT
MKSRNLTK
KSRNLTKD
SRNLTKDR
RNLTKDRC
NLTKDRCK
LTKDRCKP
TKDRCKPV
KDRCKPVN
DRCKPVNT
RCKPVNTF
CKPVNTFV
KPVNTFVH
PVNTFVHE
VNTFVHES
NTFVHESL
TFVHESLA
FVHESLAD
VHESLADV
HESLADVQ
ESLADVQA
SLADVQAV
LADVQAVC
ADVQAVCS
DVQAVCSQ
VQAVCSQK
QAVCSQKN
AVCSQKNV
VCSQKNVA
CSQKNVAC
SQKNVACK
QKNVACKN
KNVACKNG
NVACKNGQ
VACKNGQT
ACKNGQTN
CKNGQTNC
KNGQTNCY
NGQTNCYQ
GQTNCYQS
QTNCYQSY
TNCYQSYS
NCYQSYST

CYQSYSTM
YQSYSTMS
QSYSTMSI
SYSTMSIT
YSTMSITD
STMSITDC
TMSITDCR
MSITDCRE
SITDCRET
ITDCRETG
TDCRETGS
DCRETGSS
CRETGSSK
RETGSSKY
ETGSSKYP
TGSSKYPN
GSSKYPNC
SSKYPNCA
SKYPNCAY
KYPNCAYK
YPNCAYKT
PNCAYKTT
NCAYKTTQ
CAYKTTQA
AYKTTQAN
YKTTQANK
KTTQANKH
TTQANKHI
TQANKHII
QANKHIIV
ANKHIIVA
NKHIIVAC
KHIIVACE
HIIVACEG
IIVACEGN
IVACEGNP
VACEGNPY
ACEGNPYV
CEGNPYVP
EGNPYVPV
GNPYVPVH
NPYVPVHF
PYVPVHFD
YVPVHFDA
VPVHFDAS
PVHFDASV

(2A) $\mathbb{S}_2^+$: list of 22 cleavable octapeptides by HIV-2 protease

SQNYPIVQ
EEELAECF
TQIMFETP
GQVNYEEF

GGNYPVQH
PRNFPVAQ
AEELAEIF
PFAAAQQR
RQVLFLEK
ATIMMQRG
SLNLPVAK
 ANLAEEA
PTLLTEAP
SFIGMESA
YEEFVQMM
RHVMTNLG
YISAAELR
GLAAPQFS
DGNGTIDF
GDALLERN
NPTEAELQ
RQAGFLGL

(2B) $\mathbb{S}_2^-$: list of 127 non-cleavable octapeptides by HIV-2 protease

KVFGRCEL
VFGRCELA
FGRCELAA
GRCELAAA
RCELAAAM
CELAAAMK
ELAAAMKR
LAAAMKRH
AAAMKRHG
AAMKRHGL
AMKRHGLD
MKRHGLDN
KRHGLDNY
RHGLDNYR
HGLDNYRG
GLDNYRGY
LDNYRGYS
DNYRGYSL
NYRGYSLG
YRGYSLGN
RGYSLGNW
GYSLGNWV
YSLGNWVC
SLGNWVCA
LGNWVCAA
GNWVCAAK
NWVCAAKF
WVCAAKFE
VCAAKFES
CAAKFESN
AAKFESNF
AKFESNFN

KFESNFNT
FESNFNTQ
ESNFNTQA
SNFNTQAT
NFNTQATN
FNTQATNR
NTQATNRN
TQATNRNT
QATNRNTD
ATNRNTDG
TNRNTDGS
NRNTDGST
RNTDGSTD
NTDGSTDY
TDGSTDYG
DGSTDYGI
GSTDYGIL
STDYGILQ
TDYGILQI
DYGILQIN
YGILQINS
GILQINSR
ILQINSRW
LQINSRWW
QINSRWWC
INSRWWCN
NSRWWCND
SRWWCNDG
RWWCNDGR
WWCNDGRT
WCNDGRTP
CNDGRTPG
NDGRTPGS
DGRTPGSR
GRTPGSRN
RTPGSRNL
TPGSRNLC
PGSRNLCN
GSRNLCNI
SRNLCNIP
RNLCNIPC
NLCNIPCS
LCNIPCSA
CNIPCSAL
NIPCSALL
IPCSALLS
PCSALLSS
CSALLSSD
SALLSSDI
ALLSSDIT
LLSSDITA
LSSDITAS
SSDITASV
SDITASVN

DITASVNC
ITASVNCA
TASVNCAK
ASVNCAKK
SVNCAKKI
VNCAKKIV
NCAKKIVS
CAKKIVSD
AKKIVSDG
KKIVSDGN
KIVSDGNG
IVSDGNGM
VSDGNGMN
SDGNGMNA
DGNGMNAW
GNGMNAWV
NGMNAWVA
GMNAWVAW
MNAWVAWR
NAWVAWRN
AWVAWRNR
WVAWRNRC
VAWRNRCK
AWRNRCKG
WRNRCKGT
RNRCKGTD
NRCKGTDV
RCKGTDVQ
CKGTDVQA
KGTDVQAW
GTDVQAWI
TDVQAWIR
DVQAWIRG
VQAWIRGC
QAWIRGCR
AWIRGCRL
SQNYYIVQ
SQNYFIVQ
SQNYLIVQ
SQNYMIVQ
SQNYVIVQ