Volume 218, number 2, 231–237

FEB 04835

June 1987

Prediction of antigenic determinants and secondary structures of the major AIDS virus proteins

Michael J.E. Sternberg, Geoffrey J. Barton, Marketa J. Zvelebil, John Cookson* and Anthony R.M. Coates⁺

Laboratory of Molecular Biology, Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX, *Hill Computer Centre and *Department of Medical Microbiology, London Hospital Medical College, Turner Street, London E1 2AD, England

Received 10 March 1987; revised version received 12 April 1987

Criteria for the design of peptide vaccines to prevent AIDS are presented. The best vaccine candidates contain both B and T lymphocyte-defined epitopes in regions conserved in sequence between viral isolates. We propose that attention should focus on proteins specified by the gag and, possibly, pol genes in addition to the env gene envelope glycoproteins being actively studied. The predictions of B- and T-epitopes are refined by consideration of secondary structure prediction and inter-isolate sequence variability to suggest peptides from env, gag and pol that would be the best vaccine candidates.

AIDS; Human immunodeficiency virus; Peptide vaccine; Epitope; Secondary structure prediction

1. INTRODUCTION

The dramatic spread of the human immunodeficiency virus (HIV-1) highlights the need for prevention of AIDS (acquired immune deficiency syndrome) by a vaccine (e.g. [1-6]). Towards this goal, the nucleotide sequences of several virus isolates have been determined [7-10] and they contain the three main retroviral genes (5' to 3') – gag, pol and env. gag encodes three proteins: p17, p24 (the major capsid protein) and p15 (a nucleic acid-binding protein); pol specifies a protease, a reverse transcriptase and an endonuclease and env determines the envelope glycoprotein and a transmembrane protein.

The search for vaccines has focussed on B lymphocyte-defined epitopes (B-epitopes) of env

Correspondence address: M.J.E. Sternberg, Laboratory of Molecular Biology, Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX, England proteins [1-5] encouraged by the use of retroviral envelope proteins to confer protection against viral challenge in animals [11]. Experimental identification of *env* B-epitopes by synthetic peptides [1,2] and by recombinant DNA technology [3] is often guided by predictions (e.g. [4]) that scan for local maxima in hydrophilicity [12]. As B-epitopes are often loops between the regular secondary structures [13], a further guide to their location would be to predict from the sequence which parts of the chain are not α -helices or β -sheets [14–17].

However, the importance of antibodies alone as the protective arm of the immune response is in doubt as, for instance, patients can die of AIDS with high levels of anti-*env* antibody in their blood [18]. In addition, only low levels of neutralising antibodies are present in HIV-1 infection and persist in AIDS [19,20]. Thus, cell-mediated immunity, particularly cytotoxic T-cells, may play an important role in resistance to HIV-1 as this type of cell induces protection in chronic viral infection [21]. The locations [22] of T lymphocyte-defined

Published by Elsevier Science Publishers B.V. (Biomedical Division) 00145793/87/\$3.50 © 1987 Federation of European Biochemical Societies epitopes (T-epitopes), which generally are on different parts of the protein from B-epitopes [12,13], can to some extent be predicted [22].

Any successful vaccine must be effective against a range of isolates. Thus, we consider that the best candidates for peptide vaccines should be one or several sections of the chain with both B- and Tepitopes that are conserved in sequence.

Recently, clinical progression to AIDS has been associated with a reduction in antibodies to the main capsid protein, gag p24 [23]. Thus, stimulation of these antibodies by a gag vaccine might prevent the onset of AIDS. In general, we propose that the search for a peptide vaccine should consider not only the *env* proteins but also those from gag and perhaps even *pol*. One strategy might be to include components from several proteins in the vaccine.

Here, algorithms to locate potential B- and Tepitopes [12,22] are refined by the identification of inter-strain sequence variability [14] and secondary structure prediction [14–17] and are applied to the *env, gag* and *pol* proteins. Interpretation of the secondary structure prediction also leads to the assignment of structural domains for the large proteins [24]. A preliminary account of this approach has been reported [6].

2. PROCEDURES

The studies considered the lymphadenopathyassociated virus (LAV) [7] with the gag, pol and env sequence files identified by the Protein Information Resource Databank [25] codes FOVWLV, GNVWLV and VCLJLV. In addition, analyses of sequence variation (section 2.1) considered three other isolates – HTLV-III (FOVWH3, GNVWH3 and VCLJH3); LV (FOVWVL, GNVWVL and VCLJVL) and ARV-2 (FOVWA2, GNVWA2 and VCLJA2). The following studies were performed and the results presented for LAV in figs 1–3.

2.1. Sequence variation

A multiple alignment of the sequences of the four isolates was obtained [14]. In the row denoted VARIABLE of the figures, the letter V indicates that there is sequence variability at this position whilst a G denotes that a gap was introduced into the alignment.

2.2. Secondary structure prediction

The algorithms of Zvelebil et al. [14], Chou and Fasman [15], Lim [16] and Rose [17] were used to predict α -helices, β -sheets and bends in the proteins. There is considerable variation in the results between the different algorithms (see fig.3) and to obtain a final prediction of the secondary structure (denoted SS-PRED in figs 1–3), it is necessary to interpret the individual predictions using an understanding of the main features of protein architecture. Of particular importance is that large proteins tend to form domains linked by sections of the polypeptide chain that are hydrophilic. These domains tend to belong to one of four structural classes [24]: α/α , β/β , α/β and $\alpha + \beta$.

2.3. B-epitopes

The algorithm of Hopp and Woods [12] searches for a local maximum in a hydrophilicity profile smoothed over an average of 6 residues. The most hydrophilic peak was selected first and then others taken in order of hydrophilicity until on average there was one site per 30 residues of the polypeptide chain. In figs 1–3, the numbers indicate the rank order of the peaks. The best candidates for vaccines, denoted by a + sign, were selected so that the 6 residues, and one residue before and after, were in a sequence-conserved region and only one of the 6 residues overlapped with a predicted secondary structure. The other peaks are denoted by a - sign.

2.4. T-epitopes

The approach of DeLisi and Berzofsky [22] was applied. The amphipathicity of sections of 7 residues was calculated and regions with a periodicity corresponding to that of an α -helix were identified. The letter T denotes the best vaccine candidates where the section, and one residue before and after, is in a sequence-conserved region, otherwise a / is used.

3. RESULTS AND DISCUSSION

We must emphasise that the description below refers to predicted secondary structures, domain links and epitopes, being only suggestions for experimental verification. The *env* gene (fig.1) begins with a hydrophobic signal peptide (residues

VARIABLE SS-PRED B-EPITOPE T-EPITOPE	MRVKEN V GV	10 (YOHLWRN 7000000 HHHH 5 //////	20 GMKWGTMLI VVVV HHHHHHHH	30 LGILMICSATEK V ӨННННН	40 LWVIVYYGVPVN HHH	50 KEATTTLFCA MIRTHHIM TTTTT	60 SDAKAYDTEV V ННННН 26	70 НИУЖАТНАСУР? НИНИНИНИН	80 90 IDPNPQEVVLVNVTEN V SSSSSSSS 21	100 1 FNNSK NDMVEQMHED) V V HHHHHHHH 2215- //////	10 120 1 3LWDQSLKPCV #HHHHHH TTTTTTT
VARIABLE SS-PRED B-EPITOPE T-EPITOPE	KLTPLO	130 CVSLKCTD V V SSSSS	140 LGNATNTN: VVV 14-	150 SSNTNSSSGEMM GGGGGVVVVV	160 MEKGEIKNCSFN VG 5	170 IISTSIRGKV(V V V HIGHHHIGHHI 10	180 PKEYAFFYKLD VVV IHHH SSSSS	190 IIPIDNDTTSY VV GV V SS 28	200 210 FLTSCNTSVITQACPR GGVV V SSSSSSSS TTFTTT	220 VSFEPIPIHYCAPAG V	230 240 FAILKCNNKTFNG SSSSSS
VARIABLE SS-PRED R-EPITOPE T-EPITOPE	TGPCTT V	250 NVSTVQCT SSSSSS /	260 HGIRPVVS V SSS	270 TQLLLNGSLAEE SSSSS 17	280 EVVIRSANFTDA V V SSSSSSSS	290 IAKTIIVQLN(SSSSSSS	300 SVEINCTRPN 7 V SSSSSSSS ++2	310 NNTRKSIRIOR V GG 4+16++	320 330 GPGRAFVTIGKIGNMG V V VGGVV SSSSS	340 RQAHCNISRAKWNATL VVV HHHHH	350 360 KOIASKLREQFGN V VV HHHHHHHHH 11
VARIABLE SS-PRED B-EPITOPE T-EPITOPE	NKTIII V SSS	370 FKQSSGGE V SSS 23	380 PEIVTHSF V SSSSSSS	390 NCGGEFFYCNST V V SSSSS	400 Clfnstwfnstv VVVV(410 NSTECSNNTEX GCVGGGV	420 SSDTITLPCRI VV +27	430 KQFINMMQEVG V IHHHHHHHHHHHHHHH TTTTT	440 450 KAMYAPPISGQIRCS: V V H TT	460 SNITGLLLTRDGGNNN VGG SSSSSSS 25	470 480 NGSEIFRPGGGDM VV V +
VAR I ABLE SS-PRED B-EP I TOPE T-EP I TOPE	RD NWR +12++	490 SELYKYKV SSSSSS	500 VKIEPLGV V V SSSS	510 APTKAKRRVVQF +++4++ +++8	520 REKRAVGIGALFI GG V H HIDDHHDD }++	530 LGFLGAAGSTI MHHMMM	540 MGARSMILIVÇ V V SSSSSS	550 DARQLLSGIVQQ SSS TTTTTT	560 570 ջ NNLLRA I EAQQHL Lչ HENENHHHHHHHH	580 QLTVNGIKQLQARILA V HHHHHHHHHHHHHH	590 600 VERYLKDQQLLGI V HHHHHH 9
VARIABLE SS-PRED B-EPITOPE T-EPITOPE	WGCSG	610 KLICTTAN SSSSSSS	620 /PWNASWSN 5 H	630 KSLEQIWNNMTN V V HHHHHHBBBBBBB ////////////////////////	640 Memoreinnyt VVV (ннныннн	650 BLIHSLIEES VV VV V 20	660 QNQQEKNEQEI ~- +++2++	670 Lleldkwaslwa V Homenichinger	680 690 MFNITNHLWYIKIFII VV V HH HHH TTTTT	700 MIVGGLVGLRIVFAVL HINDDHHIMHHIMHH ////	710 720 SIVNRVRQGYSPL V HHHH
VARIABLE SS-PRED B-EPITOPE T-EPITOPE	SFQTH V	730 ILPTPRGPI V 7-	740 DRPEGIEEE V	750 XIGERDRDRSIRI V V SSS: 3++1++	760 LVNGSLALIWDD VVVV SS SSSSSS 29	770 LRSLCLFSYH V	780 RLRDLLLIVIT VV HH IITHHII HHI	790 RIVELLGRRGME VVV IHHHHH	800 810 АLКҮНИКLQҮНSQEI V V НННННННННННН	820 LKNSAVSLLNATAIAV V HHHHHHHHHHH	830 840 AEGTDRVIEVVQG V V V IH НННЯЛНН 13 ////////////////////////////////
VARIABLE SS-PRED B-EPITOPE T-EPITOPE	ACRAI V HHHHH	850 RHIPRRII V V IHH HI 19	860 ROGLERILI V KHHHHHHH 18	4							

Fig.1. Analysis of the *env* gene. The LAV sequence [7] is given. SS-PRED gives the result of the final interpretation of the secondary structure prediction. H denotes an α -helix and S, a β -sheet.

10-28). After this until 130 there is a region predicted to be mainly α -helices. Residues 130–165 are devoid of structure and may well be a domain link. The next region is predicted to be rich in β sheets which is typical of the influenza virus glycoproteins, neuraminidase and haemagglutinin [26,27]. Residues 391-425 form another domain link to a region that ends at 516. This Lys/Arg-rich region is considered to be a proteolytic cleavage site [7] before the hydrophobic transmembrane α helical section (517-532). There is another highly hydrophobic section (689-708) which would be suitable for another transmembrane α -helix. All the rest of the polypeptide chain (i.e. 533-688 and 709-C-terminus) is not particularly hydrophobic and so we propose that only these two α -helical sections are buried in the membrane. Indeed, peptides from HTLV-III which correspond to residues 501-529, 584-604 and 732-751 of LAV can raise antisera that recognise the virus [1-3] suggesting that these regions are not buried within the membrane. The first and third of these regions are predicted B-epitopes in our analysis.

The analysis of the *pol* gene (fig.2) uses the proposition of Johnson et al. [28] that the order of the proteins is: protease – reverse transcriptase – ribonuclease – endonuclease. The protease ends between residues 160 and 180 and mainly consists of β -sheet structure. Residues 160–180 are a proline-rich region without α - or β -structure. The reverse transcriptase ends around residue 430 and is an α/β -structure. Sequence comparisons [28] have identified a consensus polymerase sequence around two conserved aspartate residues (340 and 341). In addition, there is a conserved triplet LPQ (304–306) and a conserved doublet SP (311–312)

Volume 218, number 2

VAR IABLE SS-PRED B-EPITOPE	10 FFREDLAFLQGKA GGGVVV HHHHHHHH 	20 REFSSEQTRANSP IMH 19	30 40 TTRRELQVWGRDNNSLSI GC VV V -18	50 EAGADROGTVSI S8555 - 25	60 FNFPQITLNQRE 355	70 PLVTIKIGGOLK V SSSSSSS	80 EALLDTGADDT ++24++	90 100 VLEEMSLPGROM(PM V V	110 MIGGICGFIKVRG SSSSS	120 YDQILI VV SSS	
VARIABLE SS-PRED	130 EICCHKAIGTVLV SS SSSSSSS	140 GPTPVN I IGRNLI S	150 160 LTQIGCTLNFPISPIET 3855555555	170 VPVKLKPCHEDG	180 PKVKQMPLTEE HERSED	190 200 XIKALVEICTEMEXEGKISKIC		210 220 ENPYWIPVFAIKKI 55555	230 (DSTKHRKLVDFRE S\$SSSSS	230 240 TKHRKLVDFRELNKRTQ SSSSSS S	
T-EPITOPE	250 DFWEVQLGIPHPA	260 Glkk kksvtvld v	270 280 JGDAYFSVPLDEDFRKY V	290 TAFTIPSINNE	300 TPGIRYQYWVLI	310 POGNKGSPAIFY	320 SSMIX ILEPFI	330 340 RKQNPDIVIYQYMDX	350 DLYVGSDLEIGQHI	TTT 360 RTKIEEL	
SS-PRED B-EPITOPE T-EPITOPE	370 370	588888 9	5 H AMADABABA M 2	410	\$ 555\$ \$555	+ 18445	**************************************	5555555 -16	H2	480	
VARIABLE SS-PRED B-EPITOPE T-EPITOPE	RQHLLRWGLTTPDKKHQKEPP V HH HEBBH E +++6++		GVELHPOKWIVQPIVLP V +	PERDSHTVNDIQKLVGKLINNAS HHUUREEFEEDEEFEH ++ 31++ TTTTTTTT		IYPGIKURQLCKLLRGTKALT V V HEREFEREFEREFE		/IPLTEEAELELAEI ++15++	NREILKEPVHGVYY V ++21++	/DPSKDL HHHHH 29-	
VARIABLE SS-PRED B-EPITOPE T-EPITOPE	490 IAEIQKQGQGQWI V HHHARHH SS -	500 YQIYQEPFKNLK SSSSS	510 520 IGKYARTRGAHINDVKQ V H	530 LTEAVQKITTE VV PD#8000000000000000000000000000000000000	530 540 VQK ITTES IV IMGKTPKFKI VV V I DIGINGSEEEE I ////		. 550 560 57 LPIQKETNETNHTEYHQATNIPEN V V SSSSS		590 WYQLEKEPIVCAE SSSS 32	600 FFYVDCA	
VARIABLE SS-PRED B-EPITOPE T-EPITOPE	610 ASRETKLGKAGYV V V 22	620 TNRCRQKVVTLT VV VVV	630 640 Ditmorteloaihlalo V V SSSSSSSS	650 DSGLEVN IVTD SSSSS	660 Sqyalgi iqaq S3355555	670 PDKSESELVNQ HJOHHAGA 1 ///	680 I IEQLIKKEKV 19999999999 20- /////	690 700 YLAMVPAHKGIGGN 	710 EQVDKLVSÅGIRK HH RTTGREFTERT -26 TITIT	720 VLFLDGI V V HH	
VAR I ABLE SS-PRED B-EPITOPE T-EPITOPE	730 DKAQDEHEKYHSN V	740 Imramasdfnlpp I	750 760 VVAKEIVASCDKCQLKC Hudhhudhugggggggg	770 Seanngqvdcsf	780 GINQLDCTHLE S55 55555 HH	790 SKVILVAVHVA V HTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT	800 SGYIEAEVIPA HERRINERI	810 820 ETGQETAYFLLKLA SSSSSSSSS	830 GRNPVKTIHTDNG	840 SNFTSTT V S	
VAR IABLE SS-PRED R-EP ITOPE T-EP ITOPE	850 VKAACHHAGIKQI SSSSSSSSSS	860 EFGIPYNPQSQGV SSSS	870 880 VESMNKELKKIIGQVRI V HHODHESUGDUGGG 122: ///////////////////////////////////	890 Do aehlktavor H HHHHHH H	900 IAVFIHNFKRKG IAUHOCHOCH 33-	910 GIGGYSAGERI HHH	920 VDI IATDIQTK HHPPPPIGHUH 28- TTT	930 940 ELQKQITKIQNFRV H HRB HHHHHH -	950 YYRDSRDPLWKGP VVV	960 Akllmikg S	
VAR IABLE SS-PRED B-EPITOPE T-EPITOPE	970 EGAVVIQDNSDI SSSSSSS 27+	980 KVVPRRKAK I IRD ++8++ TIT	990 1000 YGKQMAGDDCVASRQDI +++4+*		•						

Fig.2. Analysis of the pol gene, see legend to fig.1.

before this consensus section. The secondary structure prediction gives a $\beta\alpha\beta$ unit and all the conserved sequences would lie at the C-end of the two strands and at the N-terminus of the α -helix. This is reminiscent of a binding motif using the positive N-terminus of the α -helix to bind negatively charged phosphate groups such as in nucleotides [29]. Residues 370–405 form a proline-rich section devoid of α - and β -structure. Johnson et al. [28] propose that residues 430-590 form a tether region between the reverse transcriptase and the ribonuclease. This section has a structure with α helices and β -strands arranged in a similar fashion to that of the reverse transcriptase. The ribonuclease (590–718) has an $\alpha + \beta$ -structure. The endonuclease (719-C-terminus) is predicted to adopt an α/β -structure.

The results for the gag polyproteins (fig.3) have been presented elsewhere [6]. In outline, p17 (1-132) and p24 (133 to about 373) belong to the α/α class of proteins. The last 40 residues of p17 and residues 220-260 of p24 are probably exposed loops that would be ideal candidates for stimulating a B-cell antigenic response. The Cterminal section, p15, is predicted not to have regular secondary structure.

Individually secondary structure algorithms have accuracies of between 50% and 65% for prediction of α - and β -structures. However, regions where there is agreement from several methods have been shown to be predicted more accurately [30]. To provide an indication of the consistency of the individual predictions, the results of the individual algorithms are given in fig.3.

		10	20	30	40	50	60	70	80	90)	100	110	120
	MGARASV	LSCGELD	RWEKIRLRP	GGKKKYKLK	HIVWASRELEF	FAVNPGLLE	TSEGCROIL	GLOPSLOTGS	EELRSLYNT	VATLYCVH	RIEIKDI	KEALDK IEEE	ONKSKKKAO	OAAA
VARTABLE		,	1									v		G
ZVET FRII.	нынын	ннн	анны	нинин	нининини	(H)H	5555		нинин з	SSSSSSHH	0000000	000 0000 0000		нннй
CE HIBENT		BBBBHHH	-	ваннынын		HHHBBBB	BBBB	BBBBBBBB	ннннн	HHD	*****	*****	HRRRRHHHH	ныны
		00001101		CC.	8000		CCCC	190	CCCCCCCC		teedd			
Cr SHELL				33	00000			755 RJUU	ULLEULLE					
LIM					22222		กษณฑาการณ	8991 DB				-	000	
ROSE BEND		801	000	88888							DL	10 QOQ	BDD	
SS-PRED		ннн	ннннн	НН	ннининини	1111	ннннн	ากกก	HARMAN	HHHHH				
B-EPITOPE			16-+	+14++-9	6						17	b101++	3+++2++	
T-EPITOPE							TITTITT		TTTTTTT	TTT		//////		
VARIABLE ZVELEBIL CF H/BEND CF SHEET LIM	DTCHSSQ G V BBBB	130 VSONYPI SSSS BBBB S H	140 VONIOCOMV SSS SSS HUHHHH SSSSSSSSS HHHHHHHHHH	150 HQAISPRTL SSSS HHHHBBBBH SSSS S HHHH	160 NAWVKVVEEK HIIIIIIIIII IIIIIIIIIIIIIII IIIIIIIIII	170 AFSPEVIPM HHHHHHH HBBBBHHH SSS	180 rsalsegatp инн иннин в	190 DLNTMLNTVG SSS555 BBB BBB HHRHHHHHH	200 GHQAAMQMI HIBHHHHH BBHHHHHH SSSS HIBH SSSS	21(KETINEEA UNIONINI UNIONINI UNIONINI SS) AEMORVHE INHIMH (HIMHMH	220 PVHAGPIAPGQ BBBB BBBB	230 MREPRGSDI SSS BBBB	240 AGTT
ROSE BEND			BBB	BBBBB	BBBB	B	BBBBB	E	388	BBBBB			BBBBB	
SS-PRED		HH	нннанананан	инник н		-TH		HHHDDDDD	0 00000000 000	*******	ннн			
B-EPITOPE					13	-							++12++	
T-FPITOPE						TTTT.	TITTT	TTTTT	TITI	TTT				
1 11 11010														
VARIABLE 2VELEBIL CF H/BEND CF SHEET LIM ROSE BEND SS-PRED S-EPITOPE T-EPITOPE VARIABLE	STLQEQ I SSSSS HHITITI SSSSSS BBBB	250 GMMTNNP BBBB BBBB 370 SQVTNSA	260 PIPVGE1YK SSS5555 HHH BBB HHHHH BBBE HHHHHH 380 TIMMQRCNE V SSS55	270 (RWIILGLNK SSSSSS S CHAHAH HAH SSSSSS SSS AH HH HAH B BBE HHHHHHHHH S BBE CRNQRK IVK(V S 955	280 (IVRMYSPTSI) 1985858 1984 BBBH 1984 BBB 1994 BBB 1994 BB 1994 BB 1	290 LDIRQGPKEI SSS BBBB +15+1 TTTTT RNCRAPRKK	300 PFRDYVDRFY BBBB BBBB HHHHH BB 1++ TTT TTTTTT GCWKCGKEGH V V	310 KTLRAEQASQE U U U U U U U U U U U U U U U U U U U	320 SVKNMTETI 7 800810000 80081000000 955555 0000100000 955555 0000100000 955555 0000100000 95555 0000000000	33: LLVQNANPDO HITH BBB SSSSS SSSSS BBB HITH HITH 45: YKGRPGNFL	0 CKTILKAI SSSSSS SSSSSS HINHIN BB TTTTTTT 0 QSRPEPT GGVVVV	340 GPAATLEEM HIGHAHIGHA 3 SS HIGHAGHAHI BBBB HOHAGHAHI FTTT TTT 460 APPEESFRSGV VVGGVVVVVV	350 TACQGVGGF HRAN BE SSSSS SSSSS HRANGTAR HBAN TTTTTT 470 ETTTPSQKQ GGCGCGGVVV	360 CHKA HH BBBHH H BBBB H H 480 DEPID
CF H/BEND	ннннн	ннн	BBBB	BBBB	BBBB B	BBB BBBB	BBBB HHH	ннн ниннни	ннн ввв	B BBBB	BBBB	BBBB BBBB	BBBB H	нанны
CF SHEET		SSSSS	SSSS	SSSS	55			SSS	855					
TIN	нычны	ннн	55555	нннннн	нынынын нын	HIHHH S	55555	HOHOM	нинини					
DOGE SEMT				RABBR	BBBB	BBBBB	BBBB	BBBBB	1	BBBBB	BBB	BBBBB	BBE	BBBBB
COSE DENE	บบบบบบ	มแม		20020										
B.FDITODE						+++4+	+	+++7++						8
TO EDITOPE		ALALALALAL.					11111111							
1-EPILOPE														
VARIABLE ZVELEBIL CF H/BENI CF SHEET LIM ROSE BENI	KELYPL VVVVVV HHH SSSS	490 TSLRSLFG VGGGGGGGG S SSSSS SSSSS	500 SNDPSSQ GGGGGGG BBBB BBBB											
SS-PRED														
B-EPITOPI	5 3													

Fig.3. Analysis of the gag gene. The results of the individual secondary structure predictions are given with B denoting bend (i.e. turn) regions. CF indicates the approach of Chou and Fasman [15] and all potential α -helices and β -sheets are given for later selection using the results from the other predictions.

The algorithm to locate B-epitopes tends to overpredict with sections of chain located that, as yet, are not known to be antigenic [12]. Accordingly, our prediction of B-epitopes used an additional constraint that the chain section is not in a secondary structure. This approach is however a simplification designed to suggest the most likely B-epitopes but after these candidates have been tried, attention should focus on predicted Bepitopes that overlap with a predicted secondary structure. Indeed, in foot and mouth disease virus, a major B-epitope lies within residues 141–160 of the capsid polypeptide VP1 but this region is predicted by the approach of Chou and Fasman [15] to be α -helical [31].

The difficulty of predicting B-epitopes can be seen from a comparison of our results and those of Robson and co-workers [5] on the *env* protein. We locate 5 good candidates for B-epitopes (denoted +) whilst Robson et al. [5] suggest 7 sections. Residues 480–485, 503–515 and 659–665 in our predictions correspond to sites 6, 7 and 1 of Robson et al. [5]. Sites 2 and 4 of Robson et al. [5] (residues 75–84 and 367–375) are predicted potential B-epitopes of rank 21 and 23 but we do not consider them as good candidates due to overlaps with a predicted secondary structure and a region of sequence variability, respectively. Sites 3 and 5 of Robson et al. [5] are not predicted as B-epitopes by the algorithm of Hopp and Woods [12]. Our site around residues 303–311 is probably not predicted by Robson et al. [5] as they exclude potential glycosylation sites and residues 305–307 have the sequence NNT. The most hydrophilic peak of Hopp and Woods' algorithm [12] is 745–749 and forms one of our sites but is not predicted by Robson et al. [5], possibly due to sequence variation nearby.

T-epitope predictions are still in the early stage of development and a different approach is being developed (Rothbard, J. and Taylor, W.R. personal communication). This gives predictions with substantial overlap with the algorithm of DeLisi and Berzofsky [22]. Thus, both the B- and Tepitope predictions must be considered as suggestions that should be refined as improved algorithms are developed.

4. CONCLUSION

Studies on other proteins, such as influenza haemagglutinin and VP1 of polio [32], have used B-epitope predictions to design synthetic peptide vaccines that yield some protection in animals. We have located B- and T-epitopes in sequenceconserved regions in all three HIV-1 gene products. These epitopes in env, gag and possibly pol should be considered as candidates for vaccines. This paper proposes that the best vaccine candidates are those sections with both B- and Tepitopes in sequence-conserved regions. Because of the variability of the env sequences, no 15-30 residue section meets this condition and this suggests that larger sections of the molecule should be used. However, in gag, which has a more conserved sequence than env, a short peptide containing residues 288-304 (GPKEPFRDY-VDRFYKTL) meets this condition and may be effective as a synthetic vaccine. Although there is no evidence yet that the immune response to any *pol* proteins is important clinically, this analysis shows that, as these proteins have highly conserved sequences, there are several short regions which meet the condition and might well prove successful as peptide vaccines. These considerations are affecting our decisions as to which peptides to synthesise for in vitro testing as vaccines.

ACKNOWLEDGEMENTS

We thank Professor T.L. Blundell and Drs J. Thornton, W. Taylor and J. Rothbard for helpful discussion. The SERC and MRC provided financial support.

REFERENCES

- Chanh, T.C., Dreesman, G.R., Kanda, P., Linette, G.P., Sparrow, J.T., Ho, D.D. and Kennedy, R.C. (1986) EMBO J. 5, 3605-3671.
- [2] Wang, J.J.G., Steel, S., Wisniewolski, R. and Wang, C.Y. (1986) Proc. Natl. Acad. Sci. USA 83, 6159-6163.
- [3] Crowl, R., Ganguly, K., Gordon, M., Conroy, R., Schaber, M., Kramer, R., Shaw, G., Wong-Stahl, F. and Reddy, E.P. (1985) Cell 41, 979-986.
- [4] Pauletti, D., Simmonds, R., Dreesman, G.R. and Kennedy, R.C. (1985) Anal. Biochem. 151, 540-546.
- [5] Robson, B., Fishleigh, R.V. and Morrison, C.A. (1987) Nature 325, 395.
- [6] Coates, R.M., Cookson, J., Barton, G.J., Zvelebil, M.J. and Sternberg, M.J.E. (1987) Nature, in press.
- [7] Wain-Hobson, S., Sanigo, P., Danos, O., Cole, S. and Alizon, A. (1985) Cell 40, 9–17.
- [8] Ratner, L., Haseltine, W., Patarca, R., Livak, K.J., Starcich, B., Josephs, S.F., Doran, E.R., Rafalski, J.A., Whitehorn, E.A., Baumeister, K., Ivanoff, L., Petteway, S.R., Pearson, M.L., Lautenberger, J.A., Papas, T.S., Ghrayeb, J., Chang, N.T., Gallo, R.C. and Wong-Staal, F. (1985) Nature 313, 277-284.
- [9] Sanchez-Pescador, R., Power, M.D., Barr, P.J., Steiner, K.S., Stempien, M.M., Brown-Shimer, S.L., Gee, W.W., Renard, A., Randolf, A., Levy, J.A., Dina, D. and Luciw, P.A. (1985) Science 227, 484-492.
- [10] Muesing, M.A., Smith, D.H., Cabradilla, C.D., Benton, C.V., Lasky, L.A. and Capon, D.J. (1985) Nature 313, 450-458.
- [11] Hunsmann, G., Schneider, J. and Schulz, A. (1981) Virology 113, 602–612.
- [12] Hopp, T.P. and Woods, K.R. (1981) Proc. Natl. Acad. Sci. USA 78, 3824–3828.
- [13] Thornton, J.M., Edwards, M.S., Taylor, W.R. and Barlow, D.J. (1986) EMBO J. 5, 409-413.

Volume 218, number 2

- [14] Zvelebil, M.J., Barton, G.J., Taylor, W.R. and Sternberg, M.J.E. (1987) J. Mol. Biol., in press.
- [15] Chou, P.Y. and Fasman, G.D. (1978) Adv. Enzymol. 47, 45-148.
- [16] Lim, V.I. (1974) J. Mol. Biol. 88, 873-894.
- [17] Rose, G.D. (1978) Nature 272, 586-591.
- [18] Schupbach, J., Haller, O., Vogt, M., Luthy, R., Joller, H., Oelz, O., Popovic, M., Sarngadharan, M.G. and Gallo, R.C. (1985) N. Engl. J. Med. 312, 265-270.
- [19] Weiss, R.A., Clapham, P.R., Cheingsong-Popov, R., Dalgleish, A.G., Carne, C.A., Weller, I.V.D. and Teder, R.S. (1985) Nature 316, 69-72.
- [20] Robert-Guroff, M., Brown, M. and Gallo, R.C. (1985) Nature 316, 72-74.
- [21] Oldstone, M.B.A., Blount, P., Southern, R.J. and Lampert, P.W. (1986) Nature 321, 239-243.
- [22] DeLisi, C. and Berzofsky, J.A. (1985) Proc. Natl. Acad. Sci. USA 82, 7048-7052.
- [23] Weber, J.N., Clapham, P.R., Weiss, R.A., Parker, D., Roberts, C., Duncan, J., Weller, I., Carne, C., Tedder, R.S., Pinching, A.J. and Cheingsong-Popov, R. (1987) Lancet 1, 119-122.

- [24] Levitt, M. and Chothia, C. (1977) Nature 261, 552-558.
- [25] George, D.G., Barker, W.C. and Hunt, L.T. (1986) Nucleic Acids Res. 14, 11-17.
- [27] Wilson, I.A., Skehel, J.J. and Wiley, D.C. (1981) Nature 289, 366–373.
- [28] Varghese, J.N., Laver, W.G. and Colman, P.M. (1983) Nature 305, 35-40.
- [29] Johnson, M.S., McClure, M.A., Feng, D.F., Gray,
 J. and Doolittle, R.F. (1986) Proc. Natl. Acad. Sci.
 USA 83, 7648-7652.
- [30] Hol, W.G.J., Van Duijnen, P.T. and Berendsen, H.J.C. (1978) Nature 273, 443-446.
- [31] Nishikawa, K. and Ooi, T. (1986) Biochim. Biophys. Acta 871, 45-54.
- [32] Rowlands, D.J., Clarke, B.E., Carroll, A.R., Brown, F., Nicholson, B.H., Bittle, J.L., Houghten, R.A. and Lerner, R.A. (1983) Nature 306, 694-697.
- [33] Zanetti, M., Sercarz, E. and Salk, J. (1987) Immunol. Today 8, 18-25.