LA-UR- 08-5513

Approved for public release;
distribution is unlimited.

Title:	Candidate Mosaic Proteins for a Pan-Filoviral Cytotoxic T-cell Lymphocyte Vaccine
Author(s):	P.Fenimore, Z#:181031, T-10/T-Division W. Fischer, Z#: 103477, T-10/T-Division C. Kuiken, Z#: 111147, T-10/T-Division B. Foley, Z#: 097029, T-10/T-Division J.R. Thurmond, Z#: 219839, T-10/T-Division K. Yusim, Z#: 169642, T-10/T-Division B.T. Korber, Z#: 108817, T-10/T-Division
Intended for:	J. Virology

Los Alamos
NATIONAL LABORATORY
EST. 1943

Form 836 (7/06)

Candidate Mosaic Proteins for a Pan-Filoviral Cytotoxic T-cell Lymphocyte Vaccine

P. W. Fenimore, W. M. Fischer, C. Kuiken, B. T. Foley, J. R. Thurmond, K. Yusim, H. Yoon, C. Calef, J. Dye, M. Parker, B. T. Korber

August 25, 2008

The extremely high fatality rates of many filovirus (FILV) strains[1], the recurrent but rarely identified origin of human epidemics[11], the only partly identified viral reservoirs[8], and the continuing non-human primate epizootics in Africa[9] make a broadly-protective filovirus vaccine highly desireable. Cytotoxic T-cells (CTL) have been shown to be protective in mice[10], guinea pigs[12] and non-human primates[4]. In murine models the cytotoxic T-cell epitopes that are protective against Ebola virus have been mapped[10] and in non-human primates CTL-mediated protection between viral strains (John Dye: specify) has been demonstrated using two filoviral proteins, nucleoprotein (NP) and glycoprotein (GP). These immunological results suggest that the CTL avenue of immunity deserves consideration for a vaccine. The poorly-understood viral reservoirs means that it is difficult to predict what strains are likely to cause epidemics. Thus, there is a premium on developing a pan-filoviral vaccine.

The genetic diversity of FILV is large, roughly the same scale as human immunodeficiency virus (HIV). This presents a serious challenge for the vaccine designer because a traditional vaccine aspiring to pan-filoviral coverage is likely to require the inclusion of many antigenic reagents. A recent method for optimizing cytotoxic T-cell lymphocyte epitope coverage with *mosaic* antigens was successful in improving potential CTL epitope coverage against HIV [5] and may be useful in the context of very different viruses, such as the filoviruses discussed here. Mosaic proteins are recombinants composed of fragments of wild-type proteins joined at locations resulting in exclusively natural k-mers, $9 \leq k \leq 15$, and having approximately the same length as the wild-type proteins. The use of mosaic antigens is motivated by three conjectures: (1) optimzing a mosaic protein to maximize coverage of k-mers found in a set of reference proteins will give better odds of including broadly-protective CTL epitopes in a vaccine than is possible with a wild-type protein, (2) reducing the number of low-prevalence k-mers minimizes the likelihood of undesirable immunodominance, and (3) excluding exogenous k-mers will result in mosaic proteins whose processing for presentation is close to what occurs with wild-type proteins.

The first and second applications of the mosaic method were to HIV and Hepatitis C Virus (HCV). HIV is the virus with the largest number of known sequences, and consequently a plethora of information for the CTL vaccine designer to incorporate into their mosaics. Experience with HIV and HCV mosaics supports the validity of the three conjectures above. The available FILV sequences are probably closer to the minimum amount of information needed to make a meaningful mosaic vaccine candidate. There were 532 protein sequences in the National Institutes of Health GenPept database in November 2007 when our reference set was downloaded. These sequences come from both Ebola and Marburg viruses (EBOV and MARV), representing transcripts of all 7

protein	length	ebola	length	marburg
NP	738-739	20	692-695	12
VP35	330-351	7	329	11
VP40	303	7	326-331	8
GP	681,	12	676-677	39
sGP				
VP30				
VP24				
L	2327-2331	14	2211-2213	15

Table 1: The number of unique NCBI GenPept proteins for each filoviral protein and the length of each protein. These sets define the reference set/training set used to generate mosaic proteins. GP and sGP are grouped as one protein because they are long and short products of gene 4. Rows in the table appear in their natural gene order.

genes. The coverage of viral diversity by the 7 genes is variable, with genes 1 (nucleoprotein, NP), 4 (glycoprotein, GP; soluble glycoprotein, sGP) and 7 (polymerase, L) giving the best coverage.

Broadlly-protective vaccine candidates for diverse viruses, such as HIV or Hepatitis C virus (HCV) have required pools of antigens. FILV is similar in this regard.

While we have designed CTL mosaic proteins using all 7 types of filoviral proteins, only NP, GP and L proteins are reported here. If it were important to include other proteins in a mosaic CTL vaccine, additional sequences would be required to cover the space of known viral diversity.

methods

The mosaic CTL vaccine candidates proposed here are optimized for 9-mer amino acid coverage of a reference set of protein sequences. Mosaic proteins are formed by recombination only in regions where two sequences share an identical amino acid 9-mer (or longer) and only when the length of the recombinant is not much longer than the longest reference-set protein. Maximizing the number of distinct 9-mers in the mosaic insures that short sequences are not generated during optimization.

Except for NP and the polymerase, there are no known amino acid 9-mers in common between Ebola and Marburg virus protein sequences. For the proteins with no 9-mers in common (*N. B.* GP and sGP), the lack of valid recombination sites means that optimization of two mosaic pools, one for Ebola and another for Marburg, gives average 9-mer coverage indistinguishable from a simulataneous optimization. For NP and L proteins, simultaneous optimization of Ebola and Marburg into a single set of mosaic antigens would mix otherwise phylogenetically divergent sequences and could potentially cause immunologically significant differences in the processing of mosaic proteins as compared with wild-type proteins. If MARV and EBOV were simultaneously optimized, one result might be the undesireable introduction of immunological preference for recoginition of one virus species at the expense of the other. Pools of mosaic proteins were made only for Ebola or Marburg.

The available protein sequences range from short fragments to full-length amino acid sequences. Based on maximum likelihood amino acid trees[6] and weighted-neighbor joined amino acid trees[2, 3], protein sequences were classified as either Marburg or Ebola, and into the 8 kinds of proteins. For most analysis GP and sGP were combined into a single set. textcolorred

The available information is inadequate to make good predictions about the likelihood of a par-

ticular filovirus strain causing a future outbreak because filovirus reservoir species are still incompletely identified [8]. The full-length protein sequences from the identified non-primate reserviors (bats) are absent, and because only genes 1 and 7 from non-primate species are sequenced. Furthermore, the number of sequence samples varies widely from one outbreaks to another. The available protein sequences were down-selected to retain only strictly non-redundant protein sequences. This proceedure is guaranteed to retain all possible amino acid 9-mers.

The uncertain nature of future filovirus strains is quite unlike the case for HIV[5] or HCV[13] where the prevalence of strains in the near future can be inferred from the known strains with very high confidence.

Proteins within these 14 less-biased sets (7 for EBOV and 7 for MARV) were recombined using a genetic algorithm to make the mosaic proteins. The current implementation of the optimizer is sensitive to the prevalence of 9-mers in the starting set, so winnowing redundant sequences helps to make 9-mer coverage more uniform across the diversity of sequences for a given filoviral protein.

Viral diversity was characterized by making phylogenetic trees using maximum likelihood, reconstruction independence and ?? (incorrectly called nearest-neighbor joining) on both full sequence alignments and gap-stripped alignments of both amino acids and nucleotides. To maintain both clarity and conciseness, we only report a subset of the self-consistent results of this analysis.

results

We present four phylogenetic trees: a gap-stripped genome maximum-likelihood phylogeny (12792 nucleotides), and gap-stripped protein (amino acid) maximum likelihood phylogenies for NP, GP/sGP and L proteins. The phylogeny of the four trees is consistent. Our first conclusion based on these trees is that FILV classification should add Ravn as a monophylitic clade, making 6 monophylitic clades in all. Our second conclusion is that a slighlty narrower set of taxa redefines a monophylitic Lake Victoria clade. Our six proposed clades are all well-supported: they are all distinguished from one another by branch lengths notably longer than intra-clade branch lengths, and have boot strap values no less than 48/50, and often 50/50, depending on the particular tree. The clades for Marburg virus are Ravn (boot strap 49 or 50/50, depending on tree) and Lake Victoria (49–50/50), for Ebola virus the monophylitic clades are Reston (50/50), Sudan (50/50), Ivory Coast (50/50) and Zaire (48–50/50). The small variation of Zaire clade's boot strap values away from unity results from the ocassional inclusion of Ivory Coast within the clade. This is most likely a result of the very small number of Ivory Coast sequences in GenPept, and the absence of a publicly-available genome sequence. An additional conclusion of examining the L-protein phylogeny is that fragmentary Yambio L-protein sequences currently listed as unclassified should belong to the Sudan clade `GEHATVRGSSFVTDLEKYNLAFRYEFTAPFIKYCNQCYG`.

The reconstructed trees include taxa covering all filoviral diversity available from the National Center for Biotechnology Information as of November 2007. The taxa cover the full range of outbreak dates, beginning in 1967, for which there are sequences in GenPept (i. e. the 2008 outbreak in Uganda is not included). Years on which sequences were collected are shown in the labels. All known host species are repsented in at least one tree. The human, cynomologus macaque, chimpanzee, gorilla, mouse, guinea pig, and two bat species are indicated by color in the trees. Not all clades are found in every tree because some proteins have no clade representative; in the case of Ivory Coast, there is no publicly available genome sequence. We have as complete a starting set for the three proteins (NP, GP/sGP and L) as is currently possible.

Mosaic pools sizes ranging from one to five proteins were tried for each protein for EBOV and MARV. Empricially it was found that roughly one mosaic protein was required for each Ebola clade

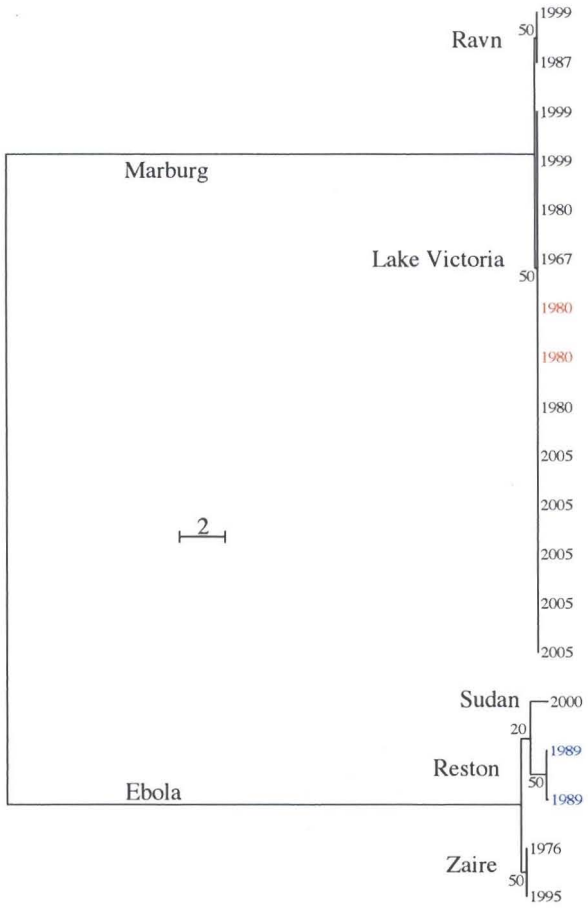

Figure 1: Maximum-likelihood tree for 19 filoviral genomes gap-stripped to 12792 nucleotides. The Ivory Coast clade is missing because there is no publicly-available whole-genome sequence for this clade. The five labelled monophylitic clades are supported by bootstrap values and intra-clade branch lengths that are consistently smaller than inter-clade lengths.

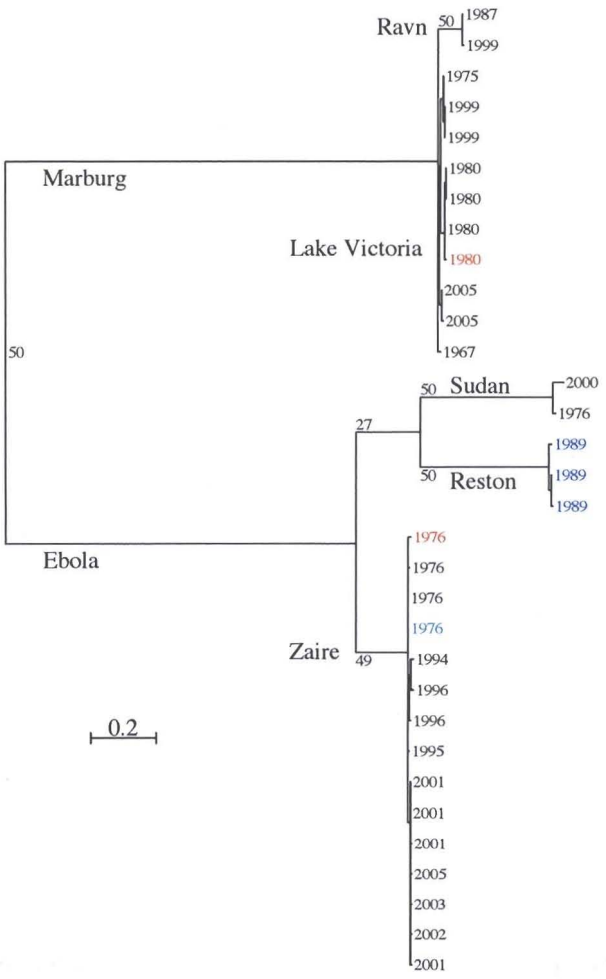

Figure 2: Maximum-likelihood tree for the gap-stripped, non-redundant amino acid sequences of the filoviral nucleoprotein. 450 amino acids are inncluded in the gap-stripped alignment. Sequences from human, macaque, mouse host species are included in the tree. Boot strap values ...

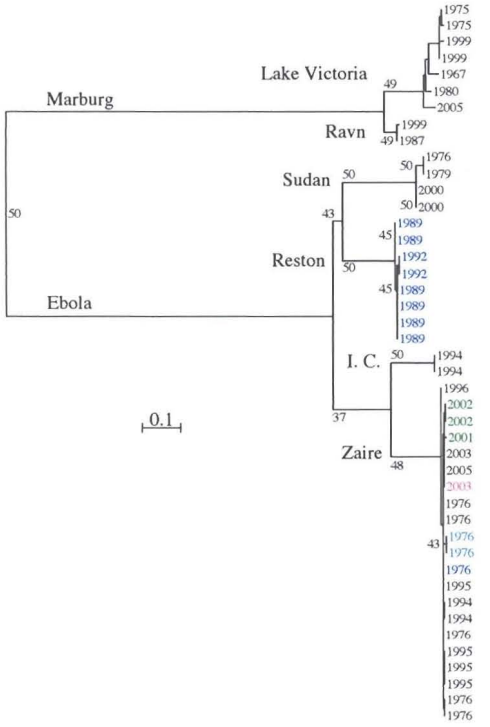

Figure 3: Maximum-likelihood tree for the gap-stripped, non-redundant amino acid sequences of the filoviral glycoprotein (GP) and soluble glycoprotein (sGP). 264 amino acids are included in the gap-stripped alignment. Boot strap values ...

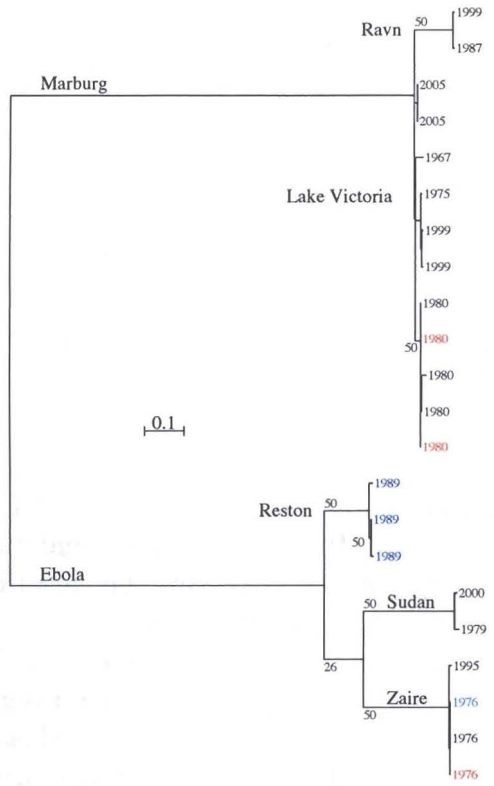

Figure 4: Maximum-likelihood tree for the gap-stripped, non-redundant sequences of the filoviral polymerase (L) protein. 2193 amino acids are included in the gap-stripped alignment. Boot strap values ...

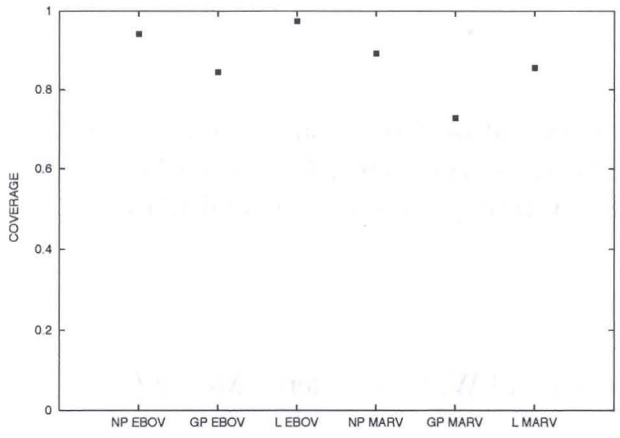

Figure 5: The fraction of covered amino acid 9-mers for each protein and viral grouping.

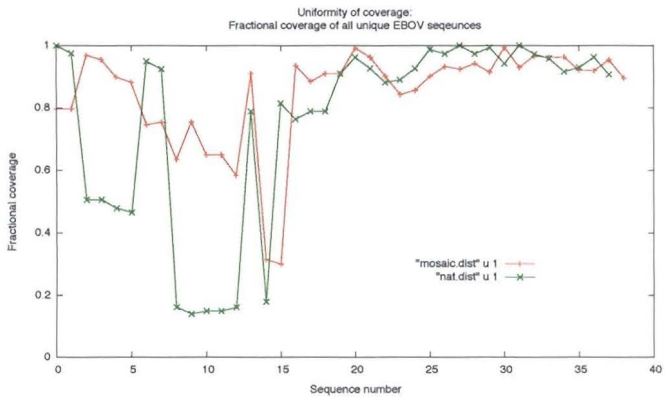

Figure 6: The fraction of 9-mers in each Ebola GP or sGP protein covered by the mosaic pool or the set of best natural proteins.

to achieve good coverage of known 9-mers. Marburg coverage with one mosaic was better than would be expected based on the experience with Ebola: two mosaic protein were not essential to achieving good coverage of known 9-mers. Mosaic pools are superior to the same-sized optimal pool of reference-set (i. e. non-mosaic) proteins by as much as 5%. Especially noteworthy, the variation in coverage from one reference strain to the next is reduced by using mosaic proteins.

Several known murine CTL epitopes[10] against reference-set proteins are also present in the mosaic proteins. The murine epitopes either occur in exactly one mosaic protein or are absent. When present, the epitope aligns with the murine epitopes in the reference set protein. Mosaic proteins lacking known murine epitopes differ by at least 3 amino acids from the aligned murine epitope. The presence a few murine epitopes in the mosaic proteins provides a positive control for the antigenicity of specific mosaic proteins.

Lower coverage variability by 9-mer of the wild-type sequences and a reduced number of "unique" 9-mers in the mosaic proteins is especially significant because it reduces the occurence of immunodominance of one viral strain over others or immunodominance of a 9-mer found in only a few, or perhaps a single, viral strain[7].

There are probably an insufficient number of distinctly sampled filoviral strains to determine what constitutes a unique filoviral 9-mer.

acknowledgments

We would like to acknowledge illumination discussions with T. K. Leitner and B. K. Gaschen. This work was performed under D. O. E. contract DE-AC52-06NA25396. Funding was provided by the Defense Threat Reduction Agency through TA2F06062 and LDRD ....

References

[1] *Infectious Diseases.* J. Cohen and W. G. Powderly, Mosby (2004) pp. 2106–2107.

[2] W. J. Bruno, *Modeling residue usage in aligned protein sequences via maximum likelihood,* Molecular Biology and Evolution, 13 (1996), pp. 1368–1374. rind.

[3] W. J. Bruno and N. D. S. and. A. L. Halpern, *Weighted neighbor joining: A likelihood-based approach to distance-based phylogeny reconstruction,* Mol. Bio. Evol., (2000), p. 189.

	35 40 45 | | |
GP Mosaic 1/1	FSIPLGVIHNSTLQV
GP Mosaic 2/3	..M....VT....E.
GP Mosaic 3/3	I.M...IVT....KA
Zaire GP	.....
	145 150 155 | | |
GP Mosaic 1/3	AQ.....P..Y....
GP Mosaic 2/3	VSGTGPCAGDFAFHK
GP Mosaic 3/3	.Q.....P..L....
gp 141-155	.....
	290 295 300 | | |
GP mosaic 1/3	.....TSLEK
GP mosaic 2/3	.....TSLGK
GP mosaic 3/3	GEWAFWETKKNFSQQ
gp 286-300 CD4 VRP Ad	.....LTRK
	245 250 255 | | |
GP mosaic 1/3	F.L.....F.
GP mosaic 2/3	....ESRF.....L.
GP mosaic 3/3	YVQLDRPHTPQFLVQ
Sudan Boniface gp 241-255	....ETRF.....L.

Table 2: The alignment between known CTL epitope-containing motifs in wild-type GP and mosaic GP proteins. Motifs 1 and 2 are exact matches, motif 3

	45 50 | |
NP 44-52	VYQVNNLEEIC
EBOV NP 1/3	..V.SD..G..
EBOV NP 2/3	..L.D...AM.
EBOV NP 3/3	.....

Table 3: The alignment between known CTL epitope-containing motifs in wild-type NP and mosaic NP proteins.

[4] J. Dye, *unpublished data.*

[5] W. Fischer, S. Perkins, J. Theiler, T. Bhattacharya, K. Yusim, R. Funkhouser, C. Kuiken, B. Haynes, N. L. Letvin, B. D. Walker, B. H. Hahn, and B. T. Korber, *Polyvalent vaccines for optimal coverage of potential t-cell epitopes in global hiv-1 variants*, Nature Medicine, 13 (2007), pp. 100–106.

[6] S. Guindon and O. Gascuel, *A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood*, Systematic Biology, 52 (2003), pp. 696–704.

[7] B. T. Korber.

[8] E. M. Leroy, B. Kumulungui, X. Pourrut, P. Rouquet, A. Hassanin, P. Yaba, A. Délicat, J. T. Paweska, J.-P. Gonzalez, and R. Swanepoel, *Fruit bats as reservoirs of ebola virus*, Nature, 438 (2005), pp. 575–576.

[9] E. M. Leroy, P. Rouquet, P. Formenty, S. Souquière, A. Kilbourne, J.-M. Froment, M. Bermejo, S. Smit, W. Karesh, R. Swanepoel, S. R. Zaki, and P. E. Rollin, *Multiple ebola virus transmission events and rapid decline of central african wildlife*, Science, 303 (2004), pp. 387–390.

[10] G. G. Olinger, M. A. Bailey, J. M. Dye, R. Bakken, A. Kuehne, J. Kondig, J. Wilson, R. J. Hogan, and M. K. Hart, *Protective cytotoxic t-cell responses induced by venezuelan equine encephalitis virus replicons expressing ebola virus proteins*, Journal of Virology, 79 (2005), pp. 14189–14196.

[11] WHO I. Commission to Sudan, *Ebola haemorrhagic fever in zaire, 1976*, Bulletin of the WHO, (1978), p. 271.

[12] L. Xu, A. Sanchez, Z.-Y. Yang, S. R. Zaki, E. G. Nabel, S. T. Nichol, and G. J. Nabel, *Immunization for ebola virus infection*, Nature Medicine, (1998), p. 37.

[13] K. Yusim, *A genotype 1 and a global hepatitis c t-cell vaccine design optimized to address genetic diversity*, (in preparation).