Prime Scholars Library

Advance Journal of Virology, Epidemic and Pandemic Diseases

Full Length Research Paper

Vol. 1 (2), pp. 12 – 25 June, 2016

Available online at
https://primescholarslibrary.org/

Useful examination of Lassa infection glycoprotein from a recently recognized Lassa infection strain for conceivable use as antibody utilizing computational methods

L. E. Okoror and I. B. A. Momodu

.Department of Computer Sciences, Faculty of Natural Sciences, Ambrose Alli University, P. M. B. 14, Ekpoma, Nigeria.

Abstract

Lassa virus is the cause of morbidity and mortality in Ekpoma Nigeria. Recently, two new strains of the virus were identified. The genes were sequenced and deposited in the GenBank. We presume that genes of similar sequence will code for protein of similar function. Hence any vaccine produced using this protein could protect against any other similar virus due to conservation of active and functional regions. The gene codes for a glycoprotein which could be antigenic and stimulate the production of antibody. This is however if the protein could be made non virulence. It then becomes important to have a proper molecular knowledge and function of the protein. Determination of the function of the protein was first by global sequence alignment using blastp with different parameters. However, not all the parameters used produced hits even when e- values were adjusted. Pairwise alignment and multiple sequence alignment (MSA) of the two newly identified proteins were carried out but full analysis of only one of the protein (strain) was done. Other tools used in determining the function of the protein included hydrophobicity, leader sequence, transmembrane helices, Pfam using different tools from expasy and PHD tools. Prints and blocks databases were also searched, but the block database gave no hit. The hidden Markov's model structure was also done before searching for the 3D structure at PDB using phyre at expasy. The pairwise alignment and MSA were done with clustal W. This study gives a clear function of the Lassa virus glycoprotein, and also confirms the stability and antigenicity of the protein so long as multiple domain repeats are carried out before synthesis which will help increase the molecular weight while preserving protein function.

Keywords: Lassa virus, multiple sequence alignment, glycoprotein, protein structure, vaccine.

INTRODUCTION

Lassa virus belongs to a very large group of haemorrhagic fever viruses-arena viruses. It was classified in the old world arena viruses. Current evidence suggests that they have co-evolved along with their rodent hosts over the years a time scale of about 9 million years. Among the six arena viruses so far known, only one is known to cause illness in humans. Lassa virus is known to be localised in West Africa and is usually carried by the multimammate rat- *Mastomys natalensis* (Johnson et al., 1997). Among the arena viruses, Lassa virus affects by far the largest number of humans, encompassing perhaps 180 million people from Guinea to Eastern Nigeria. Over 200,000 infections are estimated to occur annually with several thousand deaths (Johnson et al., 1997).

Lassa fever begins insidiously after an incubation period of 7 days with fever, weakness, malaise, severe headache usually frontal and a very painful sore throat (Keelyside et al., 1983). More than 50% of the patients then develop joint and lumbar pain and 60% develop a non-productive cough. Many also develop severe restrosternal chest pain and about half will have nausea with vomiting or diarrhea and abdominal pain. On physical examination, the respiratory rate, temperature rate and pulse rate are elevated and blood pressure may be low

(Keelyside et al., 1983; McComicks, 1986). Lassa virus infections were said to be seasonal and usually occurs in epidemic scale within the dry season months of December to May, infections have also been seen during the months of February through April or May. The exact seasonality of Lassa virus infection is still not understood and therefore increase in the availability of the virus in village homes where the rodent reservoir is from may increase the risk of transfer to humans. Transmission from humans to humans has also been reported (Helmick et al., 1986; Buckley, 1990; Fisher -Hosch, 1989 and Frame, 1990). Lassa virus has been a major cause of mortality and morbidity every year in areas of outbreaks. Mortality and morbidity continue to increase with every year of outbreak despite circulating antibodies to Lassa virus. Increase in infection had been attributed to increase in bush burning due to increase in farming activities thereby forcing the rodents out of there natural habitat into the homes where they transfer the virus to humans (Okoror et al., 2005). Increase in the rate of reproduction of rodents was also suggested as a reason for the increase in infection rate of the virus. It is also been suggested that with all these there should be enough circulating antibodies in the population to curb or stem down the infection or stabilize the infection but contrary to this, there is usually increase in infection every year. Hence an understanding of the molecular evolution aspect of the virus may explain why the Lassa virus infection has remained unchecked for over a long period of time. Newer strains of the virus were usually isolated at every season of outbreak. It is expected that a new strain of the virus may produce a new protein. A good understanding of the protein sequence may bring about a good vaccine against the virus. Studies on prospect of Lassa virus vaccines began in the 1980s when Clegg expressed the nucleoprotein of Lassa virus in vaccinia and was able to show that the recombinant vaccine protected against lassa in guinea pigs and not in humans (Clegg, 1997). We suggest that the reason for this might be from the type of protein used, hence in this study we studied comprehensively the Lassa virus glycoprotein which has been the most frequently isolated from Lassa virus over the years. This also justifies the reason we analysed a most recently isolated Lassa virus glycoprotein. Analysis of protein diversity has become increasingly important due to their association with different diseases. Of the diseases oriented variations, we suggest that there may be deletion, substitution or alteration. These could well be classified as purifying selection, positive selection or neutral selection. All these selections may play a role in the development of newer strains of Lassa virus whose antibody do not exist in the population and hence an outbreak due to a new strain of the virus. The baseline of these selections is that they may lead to protein function/ protein structure alteration and or deletions which are usually the characteristic of purifying selection. The neutral selections are less important since they may not necessarily affect the function or structure of the protein while the positive selections are advantageous since they will give an in site to novel gene function as is the likely case in this new strain of Lassa virus. These selections give an understanding as to why there is constant development of new strains of the virus. A functional and structural prediction of these new glycoprotein could also gives an in site as to why despite circulating antibodies in the population, there is still an increase in infection rate. It may also explain the tenacity of Lassa virus seasonal outbreak.

Annotations of variation or mutational effects are rarely found in databases, mainly because mutagenesis experiments and functional assays are labour intensive and data accrual does not follow the pace of accumulation of rescriptive, mutation data. In many instances the scope of such functions are still not conclusive (Mathe et al., 2006). To overcome these limitations more and more computational methods are being developed to predict the function of these substitutions leading to change in acids residue and identify residues that have a significant effect of maintaining wild type function. Different approaches have been employed which include sequence based methods (Sunayev et al., 1999 and Yang et al., 2003) , structure based algorithms (Dambosky et al., 2001; Ferrer-Costa et al., 2002; Prokop et al., 2000; Stitziel et al., 2004 and Sunayev et al., 2000) and a combination of both (Ng and Henkoff, 2001, 2002; Ramensky et al., 2002). It is very important here to note that mutation has led to the development of the new strain of the virus, whether the mutation is purifying or positive selection is not yet understood. The use of multiple sequence alignment to align either closely related sequences, distant related sequences or both have highlighted two major trends that are unique to disease associated mutations than for neutral mutations (Miller and Kumar, 2001) . This trend exists because large biochemical changes between mutants and wild types are more likely to alter the structure and hence the function of the existing proteins explaining why changes are not normally tolerated during natural selection. Secondly, association with disease tends to occur at the residue positions that are conserved across species (Vitkup et al., 2003). One of the most important and widely studied problems in protein sequence analysis is identifying which residues in a protein are responsible for its function. Knowledge of a protein's functionally important sites has immediate relevance for predicting function, guiding experimental analysis, analyzing molecular mechanisms and understanding protein interactions. Many computational methods have been developed to predict functionally important residues given a protein sequence. John and Monah (2007) in their study, focused on one of the most common approaches: the analysis of a multiple sequence alignment (MSA) of the protein and homologous sequences in order to find columns that are preferentially conserved. These sites are presumed to be functionally or

or structurally important because they have accepted fewer mutations relative to the rest of the alignment (Higgins et al. 1992). Conservation analysis has proven to be a powerful indicator of functional importance and has been used to detect residues involved in ligand binding (Liang S 2006; Magliery and Regan, 2005), in protein-protein interaction interfaces (Caffery 2004; Guharoy and Chakrabarty, 2005 and Mintseris and Weng, 2005), in maintaining structure (Karlin and Brocchieri, 1996; Schueler-Furman and Baker, 2003 and Valdar and Thornton, 2001), and in determining protein functional specificity (Hannenhali and Russel, 2000; Kalinina et al., 2003 and Lichtarge et al., 1996). Conser-vation analysis has also been used in conjunction with structural information in many of these applications (Landau et al., 2005; Panchenko, 2004). Com-putational methods for identifying functional residues that do not use conservation exist, but they typically require structural information and are usually employed in the unusual case where there is an absence or paucity of sequence homologs. Such structural approaches (Jones and Thornton, 2004) work by either identifying local shared structural patterns (Fetrow and Skolnick, 1998; Stark and Russel, 2003 and Wallace, 1997) or by identifying residues in the protein structure with unusual electrostatic and ionization properties (Elock, 2001 and Ondrenchen, 2001) . Many recent methods have used conservation along with other predictors of function-al importance (e.g. solvent accessibility, secondary struc-ture, catalytic propensities of amino acids, etc.) in a sta-tistical learning framework (Bordner and Abagyan, 2005; Chung, 2006 and Guttridge et al., 2003). It has been found that conservation is the single most powerful attribute in predicting functional importance in these settings (Petrova and Wu, 2006). Hence disease causa-tion in Lassa virus will depend on amino acids that are conserved across the species and are more likely to have important structure and more functional roles. One popular method for measuring biochemical distances between pairs of amino acids is the Gratham Difference (Gratham 1974; Ng and Henikoff, 2001) which takes into account the composition, polarity and volume of mutants and wild type amino acids (Mathe et al., 2006). Biochemical nature of amino acids as determined by hydrophobicity, coils regions, signals transduction, molecular weight of individual amino acid residues and alpha helices give a better insight to the function of the protein because they involves all the secondary structure.

Structural back bone of the protein is also a good tool in predicting the structure of the protein. Other pattern databases search also help in predicting the structure of the protein. Such database includes blocks and print databases. A major function predicting tool is the motif like prosite or proscan at expasy. The 3D structure is also a good tool at predicting the function of the protein; this will display the structural occurrence of the protein in space. All these put together gives protein function prediction a good credibility as well as improves the accuracy of prediction and thereby its reliability.

In this study we combine the sequence alignment algorithm with that of the structural algorithm as well as biochemical nature to get a proper and reliable analysis. We carried out a global alignment using different matrices in order to get both closely and distantly related proteins. From here we selected both closely and distantly related protein and constructed a multiple sequence alignment (MSA) and combine a conservation score with a measure of biochemical difference between mutants and a wild type Lassa virus with respect to the alignment (GD). This extension of Gratham difference is called Align-GVGD which has previously been applied to p35 protein and contributed to the clinical categorization of eight previously unclassified missense mutations (Vitkup et al., 2003). The result obtained was compared with a well known prediction method-SIFT which normalises probabilities that specific substitutions would be tolerated at a given position and assign the mutation effect from a specific probability cut-off value (Ferrer-Costa et al., 2002 and Prokop et al., 2000). Hence in this study, confirmation of our prediction were carried out using other profiles like TM Pred, PHDhtm which further predicts both hydrophobicity of the protein and the probability of assigning a helix at different positions. And other alignment algorithm like mahom alignment which takes into consideration the biochemical composition of the protein, prodom which runs alignment based on the protein family domain and predicts alignment with closely related protein. The PSI blast which was also helpful in comparing results from other alignments was also used.

MATERIALS AND METHODS

MSA

The protein sequence being analysed is the Lassa virus glycoprotein recently isolated in Nigeria (Omilabu et al., 2005) with Gen Bank ac-cession ID: AAY67753.1 and from Lassa virus strain Nig04-010 with Gen Bank DQ010031. The glycoprotein multiple sequence align-ment was constructed with clustal W in biology work bench. This was done after carrying out global alignment with blastp at 100PAM and blosum 62, however all other default parameters were used. After which sequences for MSA were selected which included lym-phocystic chorionmeningitis virus glycoprotein (DQ286931.1), Lassa virus glycoprotein (strain Nig04-02), and mopeia virus nucleoprotein (DQ328875).

Other alignment algorithms used included PSI-blast, mahom alignment and prodom (protein family domain alignment). A blast was also constructed in the block database as well as prints database and the signal peptide.

Secondary structure prediction

Secondary structures of the protein were predicted using hydrophobicity profile, transmembrane helices, alpha helices and profile of individual amino acids were predicted using the protscale at expasy server and confirmed using TMpred, PHDhtm and prof prediction. The windows for predicting hydrophobicity was set at 14 while that for alpha helix was set at 9. Protein structural back bone was predicted at expasy using the Ramanchandran tool which uses an

Table 1. PSI-blast alignment of the Lassa virus glycol protein.

Protein ID	LSEQ2	IDE	SIM	LALI	LGAP	B SCORE	B EXPECT	PROTEIN
trembl|Q6GWR4_9VIRU|Q6GWR4_9VIRU	490	85	91	84	0	139	3e-32	Glycoprotein
swiss|P17332|VGLY_LASSG	490	85	91	84	0	136	1e-31	Glycoprotein polyprotein
trembl|Q27YE4_9VIRU|Q27YE4_9VIRU	495	75	88	84	0	136	2e-31	Glycoprotein
trembl|Q91B79_9VIRU|Q91B79_9VIRU	490	80	91	84	0	136	2e-31	Glycoprotein
trembl|Q2A069_9VIRU|Q2A069_9VIRU	491	77	84	84	0	134	6e-31	Glycoprotein
trembl|Q6GWR6_9VIRU|Q6GWR6_9VIRU	491	87	94	84	1	134	8e-31	Glycoprotein
swiss|P19240|VGLY_MOPEI	489	72	88	84	0	134	9e-31	Glycoprotein polyprotein
trembl|Q5S582_MOPEI|Q5S582_MOPEI	489	71	86	84	0	134	1e-30	Glycoprotein
trembl|Q6GWS4_9VIRU|Q6GWS4_9VIRU	491	85	94	84	1	134	1e-30	Glycoprotein
trembl|Q27YF0_MOPEI|Q27YF0_MOPEI	489	75	89	84	0	133	2e-30	Glycoprotein
trembl|Q8AZ52_9VIRU|Q8AZ52_9VIRU	507	41	64	84	1	132	2e-30	Glycoprotein
trembl|Q9DQX8_9VIRU|Q9DQX8_9VIRU	491	81	94	84	1	132	3e-30	Glycoprotein
swiss|P08669|VGLY_LASSJ	491	84	94	84	1	132	4e-30	Glycoprotein polyprotein
trembl|Q6GWS0_9VIRU|Q6GWS0_9VIRU	491	84	94	84	1	132	4e-30	Glycoprotein
trembl|Q5S586_9VIRU|Q5S586_9VIRU	491	84	94	84	1	132	4e-30	Glycoprotein
trembl|Q6Y627_9VIRU|Q6Y627_9VIRU	491	84	92	84	1	131	7e-30	Glycoprotein
trembl|Q90037_9VIRU|Q90037_9VIRU	488	48	64	84	0	127	6e-29	Glycoprotein
trembl|Q8B114_9VIRU|Q8B114_9VIRU	483	42	67	83	0	126	2e-28	Glycoprotein
trembl|O90423_PIARV|O90423_PIARV	508	41	62	84	1	126	2e-28	Glycoprotein
trembl|O11998_PIARV|O11998_PIARV	508	43	63	84	1	125	3e-28	Glycoprotein
trembl|Q4VZZ3_9VIRU|Q4VZZ3_9VIRU	232	85	94	84	1	125	3e-28	Envelope glycoprotein
trembl|Q4VZY9_9VIRU|Q4VZY9_9VIRU	164	85	91	84	0	125	3e-28	Envelope glycoprotein
swiss|P03540|VGLY_PIARV	503	43	63	84	1	125	4e-28	Glycoprotein polyprotein
trembl|Q9YTW8_PIARV|Q9YTW8_PIARV	508	40	61	84	1	125	4e-28	Glycoprotein
trembl|Q9YTW9_PIARV|Q9YTW9_PIARV	508	40	61	84	1	125	5e-28	Glycoprotein
trembl|Q84168_9VIRU|Q84168_9VIRU	518	45	62	84	2	124	5e-28	Glycoprotein
trembl|Q8B118_9VIRU|Q8B118_9VIRU	480	44	66	83	0	124	8e-28	Glycoprotein
trembl|Q9YTX1_PIARV|Q9YTX1_PIARV	508	38	60	84	1	124	8e-28	Glycoprotein
trembl|Q911P0_9VIRU|Q911P0_9VIRU	480	44	66	83	0	124	9e-28	Glycoprotein G1+G2
trembl|O11997_PIARV|O11997_PIARV	508	43	63	84	1	124	9e-28	Glycoprotein
trembl|O11999_PIARV|O11999_PIARV	509	43	63	84	1	124	1e-27	Glycoprotein
trembl|Q9DK03_9VIRU|Q9DK03_9VIRU	507	42	60	84	1	124	1e-27	Glycoprotein
trembl|Q9DK06_9VIRU|Q9DK06_9VIRU	507	43	62	84	1	124	1e-27	Glycoprotein
trembl|Q9IMI8_9VIRU|Q9IMI8_9VIRU	490	96	98	84	0	124	1e-27	Glycoprotein
trembl|Q8B121_9VIRU|Q8B121_9VIRU	515	41	59	84	2	123	1e-27	Glycoprotein
trembl|Q8B116_9VIRU|Q8B116_9VIRU	515	41	59	84	2	123	1e-27	Glycoprotein
trembl|Q8AYY5_9VIRU|Q8AYY5_9VIRU	485	40	66	83	0	123	2e-27	Glycoprotein
trembl|Q4VZZ1_9VIRU|Q4VZZ1_9VIRU	188	84	94	84	1	122	3e-27	Envelope glycoprotein
trembl|Q4VZZ2_9VIRU|Q4VZZ2_9VIRU	183	76	86	84	0	121	5e-27	Envelope glycoprotein
trembl|Q8B119_9VIRU|Q8B119_9VIRU	510	36	54	84	2	121	6e-27	Glycoprotein
trembl|Q995C5_9VIRU|Q995C5_9VIRU	509	36	54	84	2	120	9e-27	Glycoprotein
trembl|Q8B117_9VIRU|Q8B117_9VIRU	507	37	64	84	1	120	9e-27	Glycoprotein
trembl|Q8B120_9VIRU|Q8B120_9VIRU	507	37	64	84	1	120	1e-26	Glycoprotein
trembl|O10429_JUNIN|O10429_JUNIN	485	42	57	82	3	118	4e-26	Glycoprotein
swiss|P26313|VGLY_JUNIN	485	42	57	82	3	118	5e-26	Glycoprotein polyprotein
trembl|O10428_JUNIN|O10428_JUNIN	485	42	57	82	3	118	5e-26	Glycoprotein
trembl|Q6UY73_JUNIN|Q6UY73_JUNIN	485	43	57	82	3	118	5e-26	Glycoprotein
trembl|O10430_JUNIN|O10430_JUNIN	485	42	57	82	3	118	5e-26	Glycoprotein
trembl|Q642U8_JUNIN|Q642U8_JUNIN	485	43	57	82	3	118	6e-26	Glycoprotein
trembl|Q6IVU3_JUNIN|Q6IVU3_JUNIN	485	43	57	82	3	118	6e-26	Glycoprotein
trembl|Q6GWR0_9VIRU|Q6GWR0_9VIRU	490	84	91	84	0	117	8e-26	Glycoprotein
trembl|Q9IMJ0_9VIRU|Q9IMJ0_9VIRU	490	84	91	84	0	117	8e-26	Glycoprotein
trembl|Q8B115_9VIRU|Q8B115_9VIRU	480	39	62	82	0	117	1e-25	Glycoprotein
trembl|Q8AYW1_9VIRU|Q8AYW1_9VIRU	479	38	59	83	0	116	2e-25	Glycoprotein
swiss|P31840|VGLY_TACVT	483	42	62	80	3	113	2e-24	Glycoprotein polyprotein
swiss|P31841|VGLY_TACV5	483	42	62	80	3	113	2e-24	Glycoprotein polyprotein
swiss|P31842|VGLY_TACV7	482	42	62	80	3	113	2e-24	Glycoprotein polyprotein
swiss|P18141|VGLY_TACV	495	42	62	80	3	112	4e-24	Glycoprotein polyprotein
trembl|Q4VZY8_9VIRU|Q4VZY8_9VIRU	188	72	82	84	0	110	9e-24	Envelope glycoprotein
trembl|Q98VU0_9VIRU|Q98VU0_9VIRU	85	87	94	84	1	110	1e-23	Putative glycoprotein
trembl|Q1L746_9VIRU|Q1L746_9VIRU	84	100	100	84	0	109	3e-23	Glycoprotein (Fragment)

Table 1. contd.

Protein ID	LSEQ2	IDE	SIM	LALI	LGAP	B SCORE	B EXPECT	PROTEIN
trembl|Q1L747_9VIRU|Q1L747_9VIRU	84	98	100	84	0	109	3e-23	Glycoprotein (Fragment)
trembl|Q91B94_9VIRU|Q91B94_9VIRU	498	56	76	82	7	104	8e-22	Glycoprotein
trembl|Q91B92_9VIRU|Q91B92_9VIRU	498	56	75	82	7	103	1e-21	Glycoprotein
trembl|Q9IFT1_JUNIN|Q9IFT1_JUNIN	130	43	57	82	3	103	2e-21	Glycoprotein 1 (Fragment)
trembl|Q9IFT0_JUNIN|Q9IFT0_JUNIN	130	43	57	82	3	103	2e-21	Glycoprotein 1 (Fragment)
trembl|Q9IFT3_JUNIN|Q9IFT3_JUNIN	130	43	57	82	3	103	2e-21	Glycoprotein 1 (Fragment)
trembl|Q9IFT2_JUNIN|Q9IFT2_JUNIN	130	43	57	82	3	103	2e-21	Glycoprotein 1 (Fragment)
trembl|Q9IFQ6_JUNIN|Q9IFQ6_JUNIN	130	43	57	82	3	103	2e-21	Glycoprotein 1 (Fragment)
trembl|Q9IFQ7_JUNIN|Q9IFQ7_JUNIN	130	43	57	82	3	102	3e-21	Glycoprotein 1 (Fragment)
trembl|Q9IFR5_JUNIN|Q9IFR5_JUNIN	130	42	57	82	3	102	3e-21	Glycoprotein 1 (Fragment)
trembl|Q9IFR3_JUNIN|Q9IFR3_JUNIN	130	42	57	82	3	102	3e-21	Glycoprotein 1 (Fragment)
trembl|Q9IFR4_JUNIN|Q9IFR4_JUNIN	130	42	57	82	3	102	3e-21	Glycoprotein 1 (Fragment)
trembl|Q9IFR6_JUNIN|Q9IFR6_JUNIN	130	42	57	82	3	102	3e-21	Glycoprotein 1 (Fragment)
trembl|Q9IFS2_JUNIN|Q9IFS2_JUNIN	130	43	57	82	3	102	3e-21	Glycoprotein 1 (Fragment)
trembl|Q9IFS7_JUNIN|Q9IFS7_JUNIN	130	43	57	82	3	102	3e-21	Glycoprotein 1 (Fragment)
trembl|Q9IFR0_JUNIN|Q9IFR0_JUNIN	130	41	56	82	3	102	3e-21	Glycoprotein 1 (Fragment)
trembl|Q9IFQ2_JUNIN|Q9IFQ2_JUNIN	130	43	57	82	3	102	4e-21	Glycoprotein 1 (Fragment)
trembl|Q9IFQ3_JUNIN|Q9IFQ3_JUNIN	130	43	57	82	3	102	4e-21	Glycoprotein 1 (Fragment)
trembl|Q9IFQ8_JUNIN|Q9IFQ8_JUNIN	130	43	57	82	3	102	4e-21	Glycoprotein 1 (Fragment)
trembl|Q9IFR7_JUNIN|Q9IFR7_JUNIN	130	42	58	82	3	102	4e-21	Glycoprotein 1 (Fragment)
trembl|Q9IFS1_JUNIN|Q9IFS1_JUNIN	130	42	57	82	3	102	4e-21	Glycoprotein 1 (Fragment)
trembl|Q9IFR8_JUNIN|Q9IFR8_JUNIN	130	42	57	82	3	102	4e-21	Glycoprotein 1 (Fragment)
trembl|Q9IFR2_JUNIN|Q9IFR2_JUNIN	130	42	57	82	3	102	4e-21	Glycoprotein 1 (Fragment)
trembl|Q9IFS0_JUNIN|Q9IFS0_JUNIN	130	42	57	82	3	102	4e-21	Glycoprotein 1 (Fragment)
trembl|Q9IFR9_JUNIN|Q9IFR9_JUNIN	130	42	57	82	3	102	4e-21	Glycoprotein 1 (Fragment)
trembl|Q9IFS3_JUNIN|Q9IFS3_JUNIN	130	43	58	82	3	102	4e-21	Glycoprotein 1 (Fragment)
trembl|Q9IFS6_JUNIN|Q9IFS6_JUNIN	130	43	57	82	3	102	4e-21	Glycoprotein 1 (Fragment)
trembl|Q9IFQ0_JUNIN|Q9IFQ0_JUNIN	130	42	57	82	3	102	5e-21	Glycoprotein 1 (Fragment)
trembl|Q9IFQ5_JUNIN|Q9IFQ5_JUNIN	130	43	57	82	3	101	5e-21	Glycoprotein 1 (Fragment)
trembl|Q9IFS4_JUNIN|Q9IFS4_JUNIN	130	43	57	82	3	101	5e-21	Glycoprotein 1 (Fragment)
trembl|Q9IFS8_JUNIN|Q9IFS8_JUNIN	130	42	58	82	3	101	5e-21	Glycoprotein 1 (Fragment)
trembl|Q9IFS5_JUNIN|Q9IFS5_JUNIN	130	43	57	82	3	101	5e-21	Glycoprotein 1 (Fragment)
trembl|Q9IFQ4_JUNIN|Q9IFQ4_JUNIN	130	43	57	82	3	101	5e-21	Glycoprotein 1 (Fragment)
trembl|Q9IFQ1_JUNIN|Q9IFQ1_JUNIN	130	43	57	82	3	101	5e-21	Glycoprotein 1 (Fragment)
trembl|Q9IFS9_JUNIN|Q9IFS9_JUNIN	130	43	57	82	3	101	5e-21	Glycoprotein 1 (Fragment)
trembl|Q9IFQ9_JUNIN|Q9IFQ9_JUNIN	130	43	57	82	3	101	5e-21	Glycoprotein 1 (Fragment)
trembl|Q9IFR1_JUNIN|Q9IFR1_JUNIN	130	42	57	82	3	101	5e-21	Glycoprotein 1 (Fragment)
trembl|Q9IFP9_JUNIN|Q9IFP9_JUNIN	130	43	58	82	3	101	6e-21	Glycoprotein 1 (Fragment)
trembl|Q6IUF7_MACHU|Q6IUF7_MACHU	496	45	63	81	3	100	1e-20	Glycoprotein
trembl|Q6PXP4_MACHU|Q6PXP4_MACHU	496	42	60	81	3	100	2e-20	Glycoprotein
trembl|Q8AZ57_MACHU|Q8AZ57_MACHU	496	45	63	81	3	100	2e-20	Glycoprotein
trembl|Q9IFT5_JUNIN|Q9IFT5_JUNIN	130	43	57	82	3	100	2e-20	Glycoprotein 1 (Fragment)
trembl|Q9IFT4_JUNIN|Q9IFT4_JUNIN	130	43	57	82	3	100	2e-20	Glycoprotein 1 (Fragment)
trembl|Q9IFT6_JUNIN|Q9IFT6_JUNIN	130	43	57	82	3	100	2e-20	Glycoprotein 1 (Fragment)
trembl|Q6IVT5_MACHU|Q6IVT5_MACHU	496	45	63	81	3	99	2e-20	Glycoprotein
trembl|Q9IFT7_JUNIN|Q9IFT7_JUNIN	130	43	57	82	3	99	3e-20	Glycoprotein 1 (Fragment)
trembl|Q6PXT5_MACHU|Q6PXT5_MACHU	257	45	63	81	3	98	4e-20	Glycoprotein 1 (Fragment)
trembl|Q6PXT6_MACHU|Q6PXT6_MACHU	257	45	63	81	3	98	5e-20	Glycoprotein 1 (Fragment)
trembl|Q6PXT7_MACHU|Q6PXT7_MACHU	257	44	63	81	3	98	6e-20	Glycoprotein 1 (Fragment)
trembl|Q6PXS5_MACHU|Q6PXS5_MACHU	257	45	63	81	3	98	7e-20	Glycoprotein 1 (Fragment)
trembl|Q6PXT8_MACHU|Q6PXT8_MACHU	257	45	63	81	3	97	7e-20	Glycoprotein 1 (Fragment)

Table 1. contd.

Protein ID	LSEQ2	IDE	SIM	LALI	LGAP	B SCORE	B EXPECT	PROTEIN
trembl|Q6PXT4_MACHU|Q6PXT4_MACHU	257	45	63	81	3	97	8e-20	Glycoprotein 1 (Fragment)
trembl|Q6PXU2_MACHU|Q6PXU2_MACHU	257	44	63	81	3	97	8e-20	Glycoprotein 1 (Fragment)
trembl|Q6PXU1_MACHU|Q6PXU1_MACHU	257	44	63	81	3	97	8e-20	Glycoprotein 1 (Fragment)
trembl|Q6PXU9_MACHU|Q6PXU9_MACHU	257	44	63	81	3	97	8e-20	Glycoprotein 1 (Fragment)
trembl|Q6PXU3_MACHU|Q6PXU3_MACHU	257	44	63	81	3	97	8e-20	Glycoprotein 1 (Fragment)
trembl|Q6PXU5_MACHU|Q6PXU5_MACHU	257	44	63	81	3	97	8e-20	Glycoprotein 1 (Fragment)
trembl|Q6PXU0_MACHU|Q6PXU0_MACHU	257	44	63	81	3	97	8e-20	Glycoprotein 1 (Fragment)
trembl|Q6PXU4_MACHU|Q6PXU4_MACHU	257	44	63	81	3	97	8e-20	Glycoprotein 1 (Fragment)
trembl|Q6PXU8_MACHU|Q6PXU8_MACHU	257	44	63	81	3	97	8e-20	Glycoprotein 1 (Fragment)
trembl|Q6PXU6_MACHU|Q6PXU6_MACHU	257	44	63	81	3	97	8e-20	Glycoprotein 1 (Fragment)
trembl|Q6PXU7_MACHU|Q6PXU7_MACHU	257	44	63	81	3	97	8e-20	Glycoprotein 1 (Fragment)
trembl|Q6PXS7_MACHU|Q6PXS7_MACHU	257	45	63	81	3	97	8e-20	Glycoprotein 1 (Fragment)
trembl|Q6PXS9_MACHU|Q6PXS9_MACHU	257	45	63	81	3	97	8e-20	Glycoprotein 1 (Fragment)
trembl|Q6PXS6_MACHU|Q6PXS6_MACHU	257	45	63	81	3	97	8e-20	Glycoprotein 1 (Fragment)
trembl|Q6PXS8_MACHU|Q6PXS8_MACHU	257	45	63	81	3	97	8e-20	Glycoprotein 1 (Fragment)
trembl|Q6PXT3_MACHU|Q6PXT3_MACHU	257	45	63	81	3	97	8e-20	Glycoprotein 1 (Fragment)
trembl|Q6PXT2_MACHU|Q6PXT2_MACHU	257	45	63	81	3	97	8e-20	Glycoprotein 1 (Fragment)
trembl|Q6PXT9_MACHU|Q6PXT9_MACHU	257	45	63	81	3	97	9e-20	Glycoprotein 1 (Fragment)
trembl|Q6PXS3_MACHU|Q6PXS3_MACHU	257	42	60	81	3	97	9e-20	Glycoprotein 1 (Fragment)
trembl|Q6PXT0_MACHU|Q6PXT0_MACHU	257	45	63	81	3	97	9e-20	Glycoprotein 1 (Fragment)
trembl|Q6PXT1_MACHU|Q6PXT1_MACHU	257	45	63	81	3	97	9e-20	Glycoprotein 1 (Fragment)
trembl|Q8B122_9VIRU|Q8B122_9VIRU	481	32	60	82	0	97	1e-19	Glycoprotein
trembl|Q8B113_9VIRU|Q8B113_9VIRU	481	32	60	82	0	97	1e-19	Glycoprotein
trembl|Q6PXS4_MACHU|Q6PXS4_MACHU	257	45	63	81	3	96	2e-19	Glycoprotein 1 (Fragment)
trembl|Q27V72_9VIRU|Q27V72_9VIRU	498	52	75	82	7	91	8e-18	Glycoprotein
trembl|Q82997_9VIRU|Q82997_9VIRU	494	55	75	82	7	90	9e-18	Envelope glycoprotein C
trembl|Q45R56_9VIRU|Q45R56_9VIRU	498	52	75	82	7	90	9e-18	Envelope glycoprotein
swiss|P09991|VGLY_LYCVA	498	52	75	82	7	90	9e-18	Glycoprotein polyprotein
trembl|Q49K87_9VIRU|Q49K87_9VIRU	498	52	75	82	7	90	1e-17	Glycoprotein
trembl|Q82996_9VIRU|Q82996_9VIRU	498	53	75	82	7	89	3e-17	Envelope glycoprotein C
trembl|Q9WA75_LYCVW|Q9WA75_LYCVW	309	51	74	82	7	89	4e-17	Glycoprotein 1 (Fragment)
trembl|Q9WA77_LYCVW|Q9WA77_LYCVW	309	51	74	82	7	89	4e-17	Glycoprotein 1 (Fragment)
swiss|P07399|VGLY_LYCVW	498	51	73	82	7	89	4e-17	Glycoprotein polyprotein
trembl|Q9WA34_LYCVW|Q9WA34_LYCVW	309	51	74	82	7	89	4e-17	Glycoprotein 1 (Fragment)
trembl|Q77AV9_LYCVW|Q77AV9_LYCVW	309	51	74	82	7	89	4e-17	Glycoprotein 1 (Fragment)
trembl|Q9QDK7_9VIRU|Q9QDK7_9VIRU	498	51	73	82	7	89	4e-17	Glycoprotein LCMVGP
trembl|Q9ICW1_9VIRU|Q9ICW1_9VIRU	498	51	73	82	7	89	4e-17	Glycoprotein C
trembl|Q9WA79_LYCVW|Q9WA79_LYCVW	309	51	74	82	7	89	4e-17	Glycoprotein 1 (Fragment)
trembl|Q9WA78_LYCVW|Q9WA78_LYCVW	309	51	74	82	7	89	4e-17	Glycoprotein 1 (Fragment)
trembl|Q9WA81_LYCVW|Q9WA81_LYCVW	309	50	73	82	7	88	6e-17	Glycoprotein 1 (Fragment)
trembl|Q9WA82_LYCVW|Q9WA82_LYCVW	309	50	73	82	7	88	6e-17	Glycoprotein 1 (Fragment)
trembl|Q9WA76_LYCVW|Q9WA76_LYCVW	309	51	74	82	7	88	6e-17	Glycoprotein 1 (Fragment)
trembl|Q9W855_LYCVW|Q9W855_LYCVW	293	51	73	82	7	88	7e-17	Glycoprotein 1 (Fragment)
trembl|Q77AU8_LYCVW|Q77AU8_LYCVW	293	51	73	82	7	88	7e-17	Glycoprotein 1 (Fragment)
trembl|Q9W941_LYCVW|Q9W941_LYCVW	309	50	73	82	7	88	7e-17	Glycoprotein 1 (Fragment)
trembl|Q77AV8_LYCVW|Q77AV8_LYCVW	309	50	73	82	7	88	7e-17	Glycoprotein 1 (Fragment)
trembl|Q9WA86_LYCVW|Q9WA86_LYCVW	293	51	73	82	7	87	8e-17	Glycoprotein 1 (Fragment)
trembl|Q77AU6_LYCVW|Q77AU6_LYCVW	293	51	73	82	7	87	8e-17	Glycoprotein 1 (Fragment)
trembl|Q9W997_LYCVW|Q9W997_LYCVW	293	51	73	82	7	87	8e-17	Glycoprotein 1 (Fragment)
trembl|Q77AU7_LYCVW|Q77AU7_LYCVW	293	51	73	82	7	87	8e-17	Glycoprotein 1 (Fragment)

Table 1. contd.

Protein ID	LSEQ2	IDE	SIM	LALI	LGAP	B SCORE	B EXPECT	PROTEIN
trembl|Q77AU5_LYCVW|Q77AU5_LYCVW	293	51	73	82	7	87	8e-17	Glycoprotein 1 (Fragment)
trembl|Q77AV7_LYCVW|Q77AV7_LYCVW	309	50	73	82	7	87	8e-17	Glycoprotein 1 (Fragment)
trembl|Q9W8A6_LYCVW|Q9W8A6_LYCVW	309	50	73	82	7	87	8e-17	Glycoprotein 1 (Fragment)
trembl|Q77AV6_LYCVW|Q77AV6_LYCVW	309	50	73	82	7	87	8e-17	Glycoprotein 1 (Fragment)
trembl|Q9WA80_LYCVW|Q9WA80_LYCVW	309	50	73	82	7	87	8e-17	Glycoprotein 1 (Fragment)
trembl|Q9WA83_LYCVW|Q9WA83_LYCVW	309	50	73	82	7	87	8e-17	Glycoprotein 1 (Fragment)
trembl|Q9QNP0_9VIRU|Q9QNP0_9VIRU	304	50	73	82	7	87	1e-16	Glycoprotein 1 (Fragment)
trembl|Q9QNP3_9VIRU|Q9QNP3_9VIRU	304	50	73	82	7	87	1e-16	Glycoprotein 1 (Fragment)
trembl|Q9QNP2_9VIRU|Q9QNP2_9VIRU	304	50	73	82	7	87	1e-16	Glycoprotein 1 (Fragment)
trembl|Q9QNP1_9VIRU|Q9QNP1_9VIRU	304	50	73	82	7	87	1e-16	Glycoprotein 1 (Fragment)
trembl|Q9QNN4_9VIRU|Q9QNN4_9VIRU	304	50	73	82	7	87	1e-16	Glycoprotein 1 (Fragment)
trembl|Q9QNN5_9VIRU|Q9QNN5_9VIRU	304	50	73	82	7	87	1e-16	Glycoprotein 1 (Fragment)
trembl|Q9QNN6_9VIRU|Q9QNN6_9VIRU	304	50	73	82	7	87	1e-16	Glycoprotein 1 (Fragment)
trembl|Q9QNP4_9VIRU|Q9QNP4_9VIRU	304	50	73	82	7	87	1e-16	Glycoprotein 1 (Fragment)
trembl|Q9QNN9_9VIRU|Q9QNN9_9VIRU	304	50	73	82	7	87	1e-16	Glycoprotein 1 (Fragment)
trembl|Q9QNN7_9VIRU|Q9QNN7_9VIRU	304	50	73	82	7	87	1e-16	Glycoprotein 1 (Fragment)
trembl|Q9QNN8_9VIRU|Q9QNN8_9VIRU	304	50	73	82	7	87	1e-16	Glycoprotein 1 (Fragment)
trembl|Q1EPV7_9VIRU|Q1EPV7_9VIRU	498	50	71	82	7	87	1e-16	Envelope glycoprotein
trembl|Q9WA84_LYCVW|Q9WA84_LYCVW	222	50	73	82	7	86	2e-16	Glycoprotein 1 (Fragment)
trembl|Q9WKU1_9VIRU|Q9WKU1_9VIRU	140	52	75	82	7	80	2e-14	Envelope glycoprotein
trembl|Q9QSP5_9VIRU|Q9QSP5_9VIRU	140	50	73	82	7	79	2e-14	Envelope glycoprotein
trembl|Q82995_9VIRU|Q82995_9VIRU	487	53	74	75	7	78	6e-14	Envelope glycoprotein C
trembl|Q4VZZ0_9VIRU|Q4VZZ0_9VIRU	207	66	77	65	1	69	4e-11	Envelope glycoprotein

a. The psi-blast showed proteins with very high scores, which goes to say that there could be antigenic relationship between Lassa virus glycoprotein and other distantly related viral glycoproteins. This also suggests the possibilities of antigenically cross reaction between strains of Lassa virus and other viruses alike especially the arena virus family to which Lassa virus belong. However, because of such high scoring protein it could be difficult to suggest if Lassa virus glycoprotein is responsible for the high antigenic variation in Lassa virus. LSEQ2 = length of aligned sequence, IDE = % of pairwise sequence identity, SIM = % of similarity, LALI-Number of residue aligned, LGAP = Number of residues in all indels, BSCORE = Blast score bits), BEXPECT = Blast expectation value, PROTEIN = online description of aligned protein, ID = Identifier of aligned (homologous) protein.

approximation of the protein main chain and its movement in space to determine the stability of the protein. Such movement is in the form of rotation which is permitted around the N-C and C -C single bonds of all residues (with one exception: proline). The angles and around these bonds, and the angle of rotation around the peptide bond, , define the conformation of a residue. The peptide bond itself tends to be planar, with two allowed states: trans, 180° (usually) and cis, 0° (rarely, and in most cases at a proline residue). The sequence of , and angles of all residues in a protein defines the backbone conforma-tion.The 3D structural prediction of the protein was carried out with phyre at expasy server. The hidden Markov's model struc-ture-CM (HMMSTR-CM) was used to confirm the structure of the protein. It predicts 2D contact map from the sequence alone. A contact potential energy map (Eij) is calculated using HMMSTR contact potentials. Energies for each possible ij contact are displayed using a red-to-blue colour scale. A threshhold is chosen and all ij contacts with energy that cutoff are predicted to be in contact. In practice, a contact map predicted using a simple thres-hold is usually not accurate or even physically possible. Such energies confirm the stability of the protein structure. The stability of the protein depends on the energy level. Pfam and interpro were finally used to confirm the protein family structure and consequently the function.

Functional motif prediction

Functional motif was predicted at expasy server using the proscan, and was also confirmed with prosite motif search at the predict protein server.

RESULTS

Sequence alignment

The blastp result reveals 100 proteins with similarity between 44-85% (Table 1). All the proteins were of viral origin with only one bacterium with 44% similarity. All the proteins were high scoring with the highest score by another Lassa virus strain which have probably co-evolved with the virus under study. The result of multiple sequence alignment reveals some functionally conserved residues which were done using heuristic alignment with no gap penalty (Figure 1) some of the residues were strongly conserved while others were weakly conserved. The distance from their ancestor was

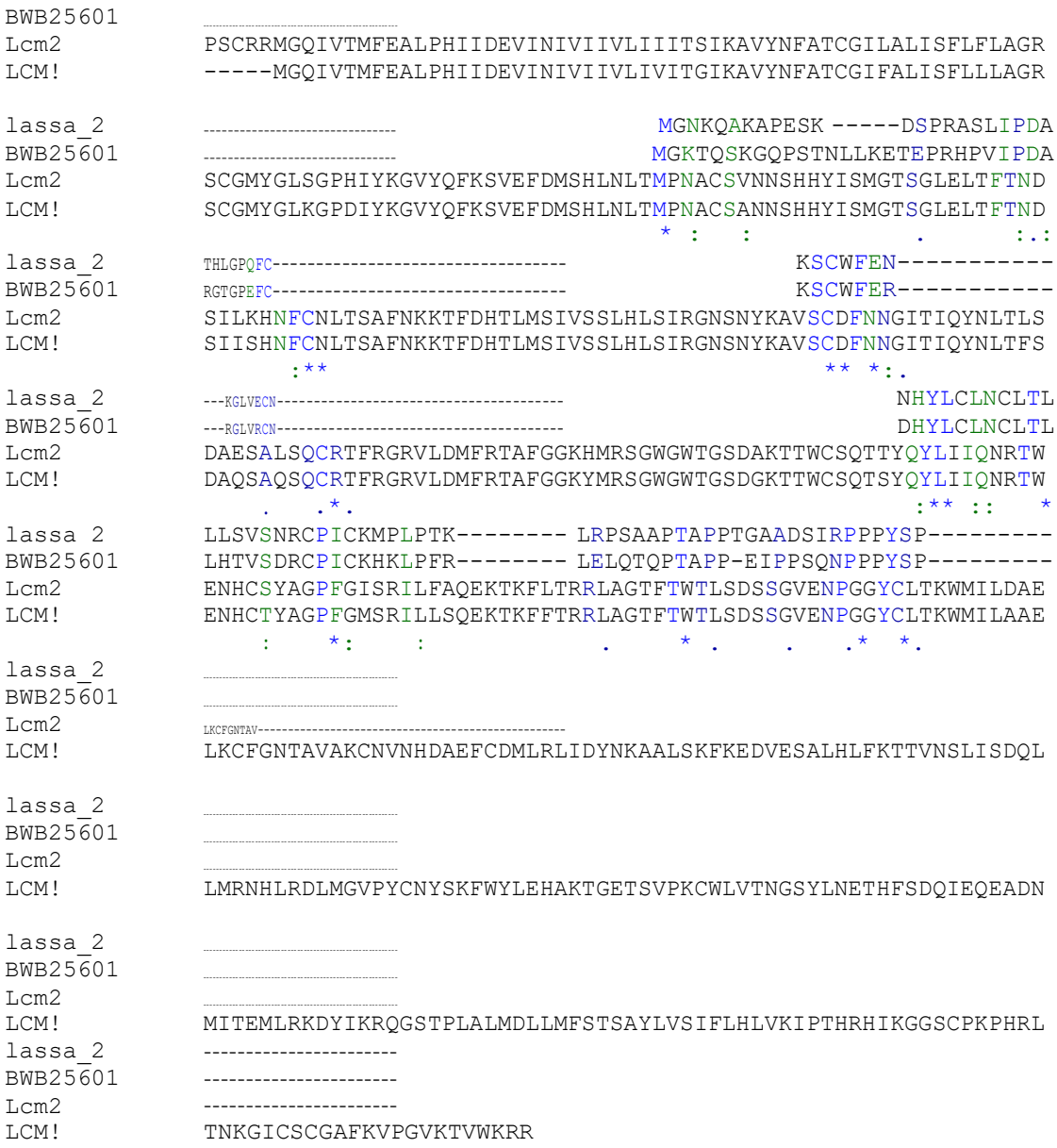

Figure 1. Multiple Sequence Alignment of the Lassa virus glycoprotein sequence.
a. The MSA shows that the Lassa glycoproteins are fully conserved across species and with other strains of the Lassa virus but there are also high amount of amino acids that are not conserved and this may be responsible for the high strain variation in Lassa virus. It could also be suggested that such variation could be responsible for little cross immunity which seems to be non protective to newer strains during epidemic outbreaks protective. **Key:** * - single, fully conserved residue; - conservation of strong groups; - conservation of weak groups; no consensus

16450. There were both deletions and substitutions along the lines of alignments. The 16450 was the longest dis-distance while the shortest distance was to the parent was 134.5 without gap penalty. MSA also predicted 14 amino acid residues that were fully singly conserved while 12 amino acids were strongly conserved and 11 were weakly conserved (Figure 1). The pairwise alignment score of the Lassa virus glycoprotein with the LCM2 virus was 2, while that for LCM1 was 3 and for the closely related other Lassa virus strain was 55 (CI = 95%).

Psi-blast, Mahom alignment and ProDom

Other alignment algorithms used were the psi-blast with best score of 136 and lowest of 65 and they were both

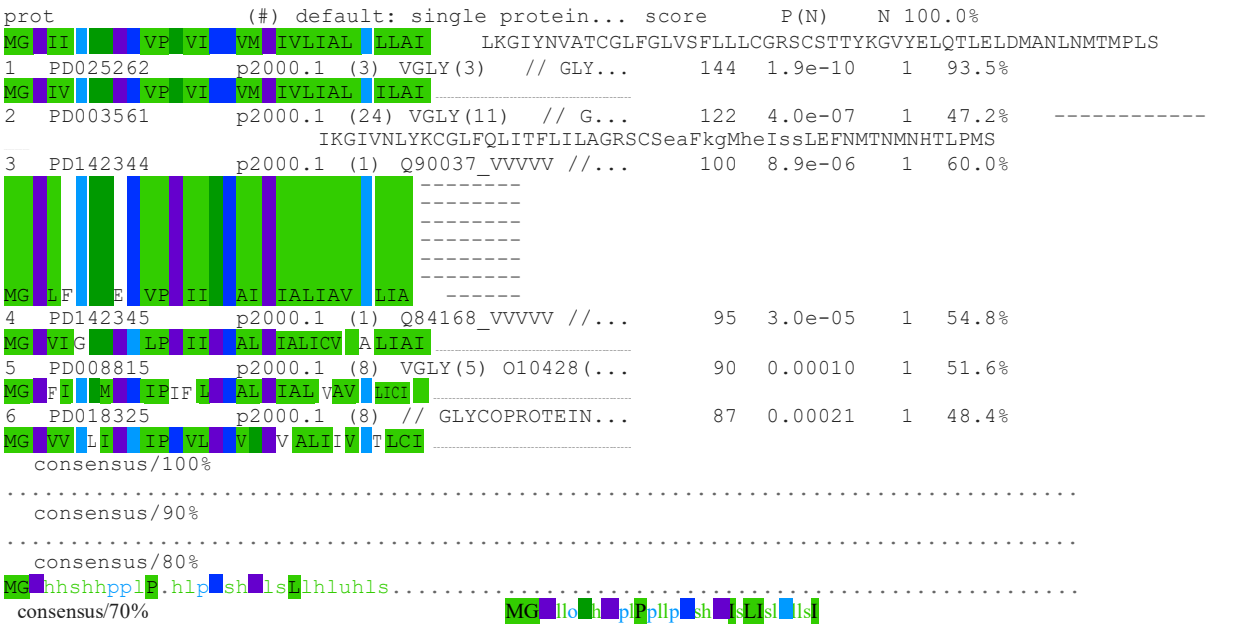

Figure 2. PRODOM alignment for protein domain using the Predict Protein server. There was high level of consensus between the families and domains. There was high consensus among the highly related domains. Coloured by consensus.

glycoproteins. However, most sequences occurring in the psi-blast revealed some high scoring sequences that were not high scoring in the blastp result (Table 1). The protein domain alignment using PRODOM (Figure 2) predicted 6 major domains with highest score of 144 and lowest score of 87, there were however consensus sequences. The maxhom alignment also predicted 12 related proteins with 4 consensus sequences (Figure 3). PFAM and INTERPRO predicted a structure common to all glycoproteins and confirm that the glycoprotein is in the family of arena virus glycoprotein.

Biochemical structure prediction

Biochemical structural prediction, with the profile of the 20 amino acids reveals a highly hydrophobic phenyl-alanine (scale 1.920) and this amino acid was highly con-served. Asperagine was lowest in the hydrophobicity profile with a scale of -1.310 and was not conserved (Table 2). The alpha helix residue of the individual amino acids which are determinants of the nature of coiling on the protein and its structure were shown in Table 3.

There were 8 coil domains with probability > 0.5. The PHDhtm predicted 2 transmembrane helices with the first score of 0.7049 with best transmembrane helices prediction of 0.9102, a c -N range of 20 - 37. And a second transmembrane helix of score 0.8321 and the best was 0.7968 and c-N range of 42 - 59. There were 2 outside region of transmembrane helices predicted, one at position 1 - 19 and 60 - 84; there were also 1 inside region at position 38 - 41.

Secondary structure prediction

Predicted secondary structure by Profsec included 58.33% probability for assigning a helix and 19.05 probability for assigning a strand while the probability for assigning either of them is 22.62.The structural back bone of the protein sequence with Ramanchandran (Figure 4) regions predicted high areas of phi and psi regions including dense areas of helix as well as cis -peptide bonds. The HMMSTR -CM (Figure 5) reveals high amount of non polar proteins interwinged by few polar proteins with uncharged proteins forming the structural back-bones.

Prints, blocks databases and signal P

Other profile search in the prints database predicted two channels which included CLChannels 7 PR011185/5 CLC 7 mouse 070496 chloride channels with a score of 28 at 2.2E and CLChannel PR00762 7/7 Q21791 Q21791 R 07B7 1 also with a score 28 and e-value of 2.2 but this time was a human protein. There was however, no statistically significant score at the block database. There were two signal peptides using signal p.

Functional motifs and 3D structure

The motif search predicted 3 functional motifs which included N-glycosylation site with pattern ID: ASN_GLY-COSYLATION, protein kinase C phosphorylation site with

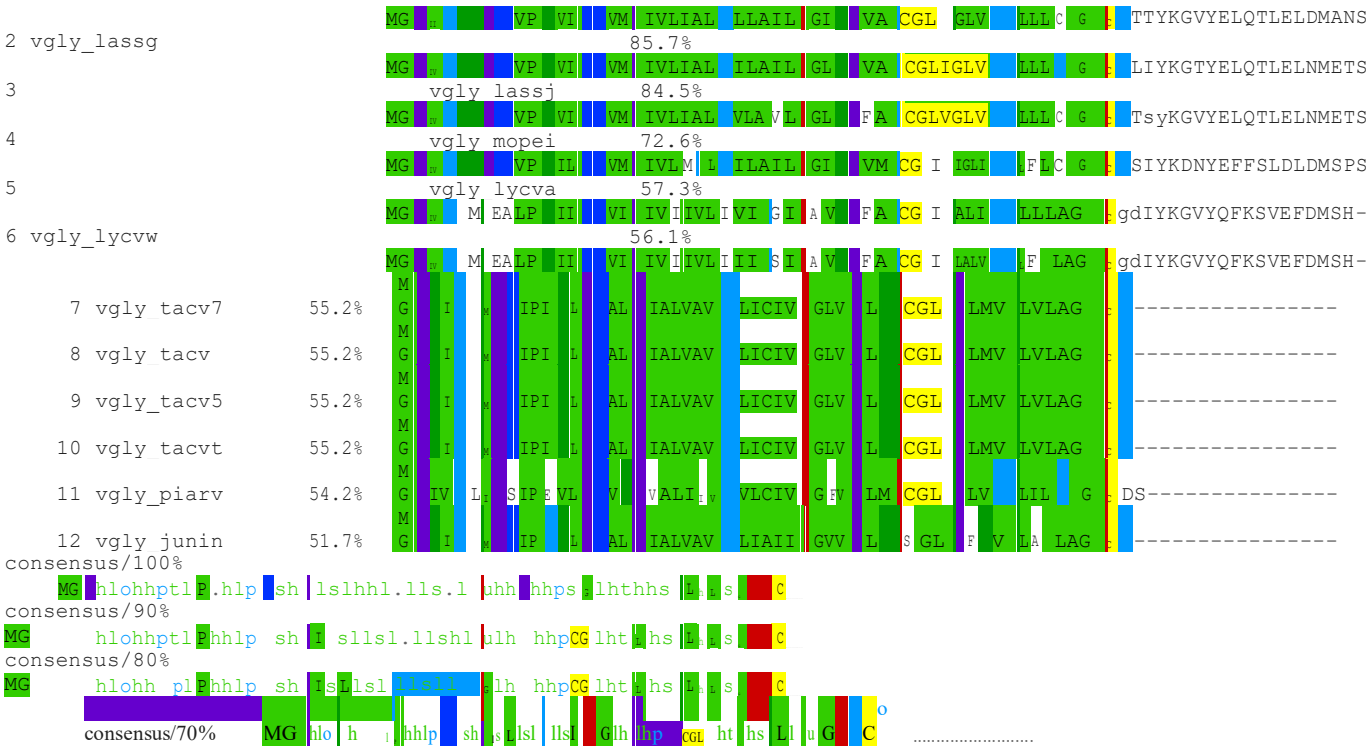

Figure 3. Maxhom alignment from Predict Preotein server. Mahom alignment further highly conservative assessment of the protein family. The amino acids were highly conserved along the families of glycoprotein even with non Lassa viruses like Junin virus. Coloured by consensus and properties.

Table 2. Using the scale hydrophobicity OMH/Sweet et al., the individual values for the 20 amino acids are:

Ala: -0.400 Arg: -0.590 Asn: -0.920 Asp: -1.310 Cys: 0.170
Gln: -0.910 Glu: -1.220 Gly: -0.670 His: -0.640 Ile: 1.250 Leu: 1.220
Lys: -0.670 Met: 1.020 Phe: 1.920 Pro: -0.490 Ser: -0.550 Thr: -0.280
Trp: 0.500 Tyr: 1.670 Val: 0.910 Asx: -1.115 Glx: -1.065 Xaa: 0.000

The amino acids with the highest probability of hydrophobicity like phenalanine determine the hydrophobicity of the protein. And the determination of such factor like the hydrophobicity is imperative during vaccine production. It is also important in determining the administration of the eventual vaccine.

Table 3. Using the scale alpha-helix/Levitt, the individual values for the 20 amino acids are:

Ala: 1.290 Arg: 0.960 Asn: 0.900 Asp: 1.040 Cys: 1.110
Gln: 1.270 Glu: 1.440 Gly: 0.560 His: 1.220 Ile: 0.970 Leu: 1.300 Lys:1.230 Met: 1.470 Phe: 1.070 Pro: 0.520 Ser: 0.820 Thr: 0.820 Trp: 0.990

The alpha –helix defines the probability of assigning a helix to the individual amino acids. And most of the amino acids gave high probabilities. The helical structure of the protein, and knowing the helical nature of individual amino acids could help determine the stability of the vaccine to be produced from the protein especially when domain repeats is to be used to increase the molecular weight of the of the protein to make it more antigenic.

pattern ID: PKC_PHOSPHO_SITE and N-myristoylation site with pattern ID: MYRISTYL. The glycosylation site occurred at the site 78 along sequence with motif NMTM while the phosphorylation site was at the 60th TYK as the motif and the myristoylation site started at the amino acid residue 2 with motif as GQIITF. The 3D structure (Figure 6) of the protein was just a little above the twilight zone which was predicted at 26% similarity. The twilight zone is usually taken as the minimum accepted zone of similarity which is usually 25%.

DISCUSSION

Previous studies have found a strong correlation between highly conserved residues and intolerance of mutations

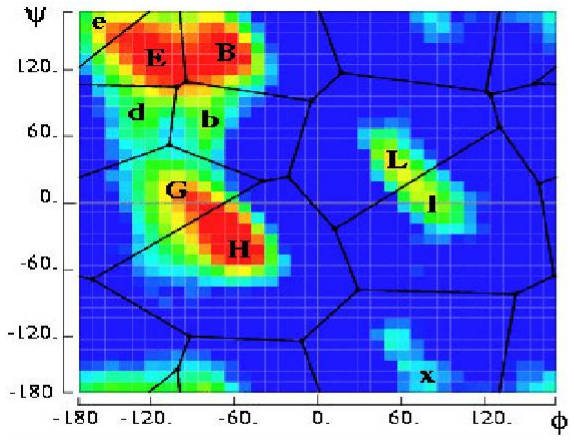

Figure 4. The Ramanchandran backbone of the Lassa virus glycoprotein from expasy server. Note the clustering of residues in areas with red colour labeled E, H and B and that most of the exceptions occur in Glycine residues (labeled G). H represents the areas with helix while the L represents the probability of assigning a strand and E is either a strand or a helix. The allowed regions generate standard conformations. A stretch of consecutive residues in the H conformation (typically 6–20 in native states of globular proteins) generates an -helix. Repeating the L conformation generates an extended -strand. Helices are 'standard' or 'prefabricated' structural pieces that form components of the conformations of most proteins. They are stabilized by relatively weak interactions, *hydrogen bonds,* between mainchain atoms.

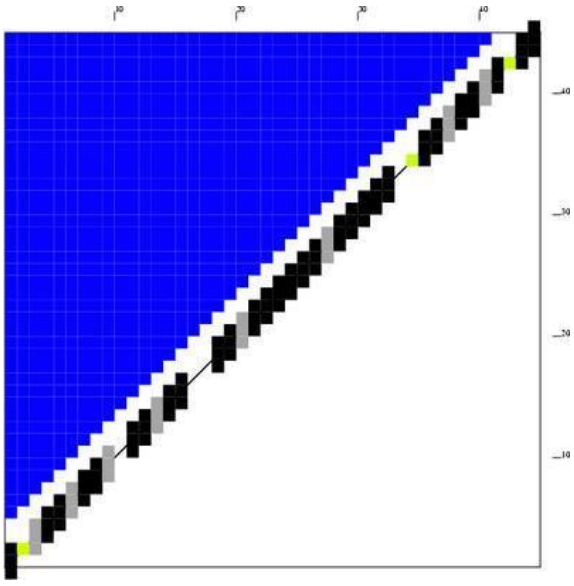

Figure 5. The hidden Markov's model structure of the protein. The bars along the diagonal represent the amino acid type: black bars for non-polar, grey bars for uncharged polar, no bar for charged side chains, yellow-green bar for glycine. The upper triangle shows the contact potential for each pair of positions, colour red for low energy to blue for high energy. Hence the Lassa virus glycolprotein has high potential for energy.

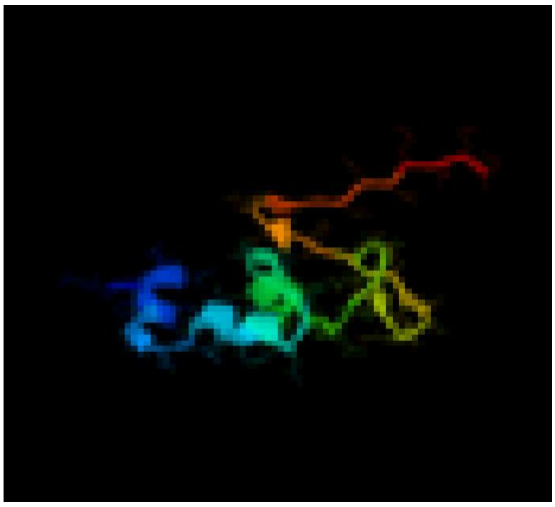

Figure 6. The 3D structure of the Lassa virus glycolprotein using phyre at the expasy server. The red colour shows area of alpha helix and are at most 60% conserved. The blue colour shows area of beta strands which are least conserved while others are the coil re- gion. They are also conserved. The coil regions are also hydrophobic.

which are likely to cause a change in the structure and function of the protein (Sunayev et al., 2000 and Ng and Henikoff, 2001) . The majority of mutation associated with diseases show larger biochemical differences between mutants and wild type amino acids than between amino acids observed in MSA for a given position (Ng and Henikoff, 2001). These observations provide the basics for structural biochemical analysis of proteins which takes into accounts different biochemical profiles which includes hydrophobicity profile and alpha helices. Mathe et al. (2006) had used the align-GVGD to study the biochemical nature of alignments but this does not give detail analysis of the biochemical nature of the protein and hence less predictive. No one method is sufficiently enough for prediction of function of a protein and in this study we have combine both the alignment algorithms to the detailed biochemical structures to determine the function of the protein. These combinations of different methods are more reliable and hence more predictive. Results obtained from each of the methods can be compared before making a final inference on the function of the protein under study.

Alignments

It is also very important to do a global alignment in the form of a blast like in this study the blast was done with blastp where both closely and distantly related protein were identified before being used in construction of the MSA. The blastp result which gave a similarity between 85 - 44% similarity, those that were below 50% similarity were treated as distantly related, however this does not

mean that they are of different ancestors or their amino acids are not conserved. We only assumed that for the construction of MSA. It is also assumed that the probability of having homologous sequences with scores above 50% is very high. Hence the protein under study is very homologous to known sequences but this does not rule out the probability of having orthologues and paralogues. The fact that the Lassa virus protein sequence is homologous to known sequences was seen in the MSA carried out with clustal W using another recently identified Lassa virus glycoprotein from another Lassa virus strain, LCM 1, Mopeia virus and LCM2 virus serving as distantly related. Because of the highly conserved regions in the MSA, it could be conclusively said that these viruses may have a common ancestor and hence explains the possibility of cross immunity among strains in infected patients. The high level conservation of amino acids between LCM and Lassa virus infer that they may have evolved from the same ancestor and hence antibody cross reaction in population between the two viruses. The MSA also inferred that there have been deletions, alterations and substitution and a high level of purifying and positive selection. There have been considerable amount of non synonymous substitutions leading to the formation of a new protein from a probably new strain of the virus and also explains why there could be new outbreaks by new strains. The psi- blast which showed very high scoring proteins may lead to a suggestion that there could be cross immunity between strains of Lassa virus and other related viruses, but since hospital reports have shown that individuals are susceptible to re-infection it means that such antigenic cross reaction does not confer protection. It could also be that such immunity due to a previous infection wane off with time and hence exposes the individual to re-infection. Resnick and colleagues (Valdar and Thornton, 2001) have found that mutations occur in these areas where such substitution have taken place resulting in mutant proteins with increase in transactivation activities (called supermutants). Such selection could be counter-selected. Previous studies also suggested that using both closely related and distantly related sequences is most suitable for accurate construction of MSA (Ramensky et al., 2002 and Hannanhali and Russell, 2000). This is because relying on closely related sequences will result in apparent lack of sequence variation due to little divergence between the individual sequences. On the other hand, using sequences that are too divergent increases the risk of including sequences that codes for protein of different functions. This explains why we carried out the global alignment before going on to construct MSA, the global alignment help to identify closely related and distantly related sequences. It also explains why sequences from mopeia virus were included in the MSA and the distant was considered not enough for high risk divergence and in order to reduce the risk of

divergence their glycoproteins were used. The PRODOM alignment predicted that the protein has been conserved within the protein family of glycoproteins even from other viruses with consensus as high as 90% down to 48% and this is a pointer to high family conservation even with viruses outside the arena viruses family. Hence the protein if used in vaccine preparation, could offer a wide range of protection. The maxhom alignment also agrees that the biochemical nature of the individual amino acids have been conserved. The strong conservation was seen in the prediction of only 12 strongly related proteins which were homologous. The importance of conservation based analysis cannot be overemphasized; however, no one method is completely adequate for protein function prediction. This explains why we combine both the conservation algorithm with secondary structure prediction. The result of the MSA was further validated by the biochemical structure analysis. This study suggests the high number of non-conserved amino acids could be responsible for strain variation and this could also be the cause of recurring high rates of infection. The multiple sequence alignment also suggests that there is possible cross immunity between strains, however, hospital reports have it that the same patient have reported for the same illness. Whether or not there is protection from re-infection of new outbreak was not covered in this study. The variation generated by the Lassa virus epidemics was detected in the Lassa virus glycoprotein and may be responsible for the antigenic escape of new strains especially from the results of the global alignment and multiple sequence alignment. However, there could be other proteins that could determine strain variation like the Lassa virus nucleoprotein. The strength of such variation still needs to be studied. There were however, high regions of sequence conservation over time but those areas where amino acids were not conserved were enough to cause strain variation which could be responsible for the yearly epidemics. This high conservation makes the protein ideal for use as vaccine.

Biochemical structure analysis

This included the hydrophobicity profile of individual amino acids with phenylanine having the highest score and occurring along the sequence much more times and this was to be confirmed at PHDhtm where there were a high percentage of 2 transmembrane helices, this goes to predict that the Lassa virus glycoprotein is hydrophobic. The hydrophobicity of the protein becomes important in the preparation of the vaccine and also its administration and storage as these are determinants of a good vaccine. Hence hydrophobicity is a very important factor in protein function determination. It is also important in the administration of the vaccine and also its pharmacokinetics. The Ramanchandran back bone which deals with the stability of the protein predicts that the protein is relatively stable considering its movement in space from the ψ region to the ϕ within the 180^{o} and -180^{o}. The typical globular protein like in this case, there were several helices and/or sheet

regions connected by turns. Usually the ends of the helix or strand regions on the surface of the domain of a protein structure. In the Lassa virus glycoprotein, the high probability to assign a helix or a strand which are connected to each other by weak hydrogen bonds contributes to its stability and hence makes its use as a good vaccine plausible. Since one of the properties of a good vaccine is stability over time. HMMSTR -CM was able to estimate the connection between the different amino acids in the protein sequence, the nature of the individual amino acids was also estimated and all these confirms the structure of the of the protein and hence its stability. The 2D structure as seen in the HMMSTR shows that the protein is of high energy and the probability of changing from one state to another state is limited. This is also true for the probability of loosing energy and going to a lower energy level. The protein backbone further supported all other predictions by revealing the position of those compounds involved in hydrophobicity, hydrophilicity as well as the coil structure and the region of the formation of the helix. The prediction with the alpha helix at prot scale shows that the protein could be a helical protein with a high probability for assigning a helix at Prof Sec. This also contributes to the stability of the protein.

Signal Peptide

The occurrence of 2 signal peptide predicted with signal p reveals that the protein could occur in the cell. This goes to point to a possible protein-protein interaction. The protein could actually be responsible for disease causation through interaction with the protein inside the cell; this however, depends on the kind of interaction. It is possible that such an interaction, if modified could lead to antibody production and hence immunity.

Coils

There were 8 coils regions predicted, which also confirm that the protein could also be a helical protein with that nature of coiling. These coils also predicted not the nature of the protein alone but also the nature of the virus which is a segmented virus. This also goes to postulate that when the analysis of structural prediction of a protein is known, it is also possible to predict the nature of the corresponding virus from the predicted protein structure.

Prints

The prints database which predicted two channels one in mouse and one in humans goes to suggest that the protein could have interactions both in humans and in mouse. The rodent is its natural reservoir where the virus is transmitted to man.

Functional motifs

However the fact that the protein is said to be a glycoprotein does not say its function which justifies this research; but one could expect the protein to behave like every other glycoprotein with some distinctive functions. The 3 predicted functional motifs gave a deeper look into the function of the protein. It is predicted that the protein could add phosphorus to the cell during interaction; it could also glycosylate the cell and also add myristoy to the cell, the glycosylation site of the cell is antigenic, while the myristoylation site could be the site for virulence which has being reported to be responsible for viral replication (Strecker et al., 2006). Hence removing the myristoylation site could remove viral virulence and carrying out multiple domain repeats will increase its molecular weight and then antigenicity. Molecular weight of protein should be improved upon computationally after removal of myristoylation site before being synthesized and used as vaccine. This prediction gives a deeper look into the function of the protein and its likely effect in disease causation and antigenicity.

REFERENCES

Bordner A, Abagyan R (2005). Statistical analysis and prediction of protein-protein interfaces. Proteins 60: 353-366.

Buckley SM (1990). Lassa fever, a new virus disease of man fro West Africa. J. Trop. Hyg. Med. 19: 680-91.

Caffrey D (2004). Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? Protein Sci. 13: 190-202.

Chung J (2006). Exploiting sequence and structure homologs to identify protein-protein binding sites. Proteins 62: 630–640.

Clegg JC (1997). Vaccinia recombinant exposing Lassa virus internal nucleocapsid protein project guinea pigs against Lassa fever. Lancet 2(8552): 1870-187.

Dambosky J, Prokop M, Koca J (2001). TRITON: graphic software for rational engineering of enzymes. Trends. Biochem. Sci. 26: 71–73.

Elock A (2001). Prediction of funtionally important residues based solely on the computed energetics of protein structure. J. Mol. Biol. 312: 885-896.

Ferrer-Costa C, Orozco M, de la Cruz X (2002). Characterization of disease Associated single amino acid polymorphisms in terms of Sequence and structure properties. J. Mol. Biol. 315: 771-786.

Fetrow J, Skolnick J (1998). Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/ thioredoxins and T1 ribonucleases. J. Mol. Biol. 281: 949-968.

Fisher-Hoch SP (1989). Protection of Rhesus monkey from fatal fever by vaccination with a recombinant vaccine virus containing the Lassa virus glycoprotein gene. Puas 85: 317-321.

Frame JD (1990). Lassa fever, a new virus disease of man from West Africa-clinical description and pathological findings. Am. J. Trop. Med. Hyg. 19: 670-6.

Grantham R (1974). Amino acid difference formula to help explain protein evolution. Science 185: 862-864.

Guharoy M, Chakrabarti P (2005). Conservation and relative importance of residues across protein-protein interfaces. Proc. Natl Acad. Sci. USA, 102: 15447-15452.

Gutteridge A (2003) Using a neural network and spatial clustering to predict the location of active sites in enzymes. J. Mol. Biol. 330: 719-734.

Hannenhalli S, Russell R (2000). Analaysis and prediction of functional subtypes from protein sequence alignments. J. Mol. Biol. 303: 61-76.

Helmick CG, Webb PA, Scribner CL, Krebs JW, McCormick JB (1986).

No evidence for increased risk of Lassa fever infection in hospital staff. Lancet 2: 1202-1205.
Higgins DG, Bleasby AJ, Fuchs R (1992). CLUSTAL V: improved software for multiple sequence alignment. in the Computer Applications Biosciences (CABIOS), 8(2): 189-191.
John AC, Monah S (2007). Predicting functionally important residue from sequence conservation. Bioinformatics 23(15): 1875-1882.
Johnson KM, Webb PA, Kuns ML, Valurde L (1997). On the mode of transmission of Bolivian hemorrhagic fever. J. Med. Sci. Biol. 20: 153-159.
Jones S, Thornton J (2004). Searching for functional sites in protein structures. Curr. Opin. Chem. Biol. 8: 3-7.
Kalinina O (2003). Automated selection of positions determining functional specificity of proteins by comparative analysis of orthologous groups in protein families. Protein Sci. 13: 443-456.
Karlin S, Brocchieri L (1996) Evolutionary conservation of reca genes in relation to protein structure and function. J. Bacteriol. 178: 1881-1894.
Keelyside RA, McCormick JB, Webb PA, Smith E, Elliot L, Johnson KM (1983). Case –control study of *Mastomys natalensis* and humans in Lassa virus-infections in households in Sierra Leone. Am. J. Trop. Med. Hyg. 32: 829-837.
Landau M (2005). Consurf 2005: the projection of evolutionary conservation scores of residues on protein structures. Nucleic Acids Res. 33: W299-W302.
Liang S (2006). Protein binding site prediction using and empirical scoring function. Nucleic Acids Res. 34: 3698-3707.
Lichtarge O (1996). An evolutionary trace method defines binding surfaces common to protein families. J. Mol. Biol. 257: 342–358.
Magliery T, Regan L (2005). Sequence variation in ligand binding sites in proteins. BMC Bioinformatics pp.6-240.
Mathe E, Olivier M, Kato S, Ishioka C, Hainaut Pand Tavtigian SV (2006). Computational approaches for predicting the biological effect of p35 missense mutations: a comparison of three sequence analysis methods. Nucleic Aids Res. 34(5): 1317-1325.
McCormick JB (1986). Lassa fever. Effective Therapy with Ribavirin. N. Engl. Med. 314(1): 20-26.
Miller MP, Kumar S (2001). Understanding human disease mutations through the use of interspecific genetic variation. Hum. Mol. Genet. 10: 2319-2328.
Mintseris J, Weng Z (2005). Structure, function, and evolution of transient and obligate protein-protein interactions. Proc. Natl Acad. Sci. USA 102: 10930-10935.
Ng PC, Henikoff S (2001). Predicting deleterious amino acid substitutions. Genome Res. 11: 863-874.
Ng PC, Henikoff S (2002). Accounting for human polymorphisms predicted to affect protein function. Genome Res. 12: 436-446.
Okoror LE, Agbonlahor DE, Esumeh FI, Umolu PI (2005). Lassa virus: Seroepidemiological survey of rodents caught in Ekpoma. Tropical Doctor.
Omilabu SA, Gunther S, Asogun A, Okokhere P (2005). Lassa fever, Nigeria, 2003 and 2004. Emerging infectious Diseases.
Ondrechen M (2001). Thematics: a simple computational predictor of enzyme function from structure. Proc. Natl Acad. Sci. USA 98: 12473-12478.
Panchenko A (2004) Prediction of functional sites by analysis of sequence and structure conservation. Protein Sci. 13: 884-892.
Petrova N, Wu C (2006). Prediction of catalytic residues using support vector machines with selected protein sequence and structural properties. BMC Bioinformatics pp.7- 312.
Prokop M, Damborsky J, Koca J (2000). TRITON: in silico construction of protein mutants and prediction of their activities. Bioinformatics, 16: 845-846.
Ramensky V, Bork P, Sunyaev S (2002). Human non-synonymous SNPs: server and survey. Nucleic Acids Res. 30: 3894-3900.
Rost B, Yachdav G, Liu J (2004). The Predict Protein Server. Nucleic Acids Res. 32(Web Server issue): W321-W326**.**
Schueler-Furman O, Baker D (2003). Conserved residue clustering and protein structure prediction. Proteins 52: 225-235.
Stark A, Russell R (2003). Annotation in three dimensions. PINTS: patterns in non-homologous tertiary structures. Nucleic Acids Res. 31: 3314-3344.
Stitziel NO, Binkowski TA, Tseng YY, Kasif S, Liang J (2004). Toposnp: a topographic database of non-synonymous single nucleotide polymorphisms with and without known disease association. Nucleic Acids Res. 32: D520-D522.
Strecker T, Maisa A, Daffis S, Eichler R, Lenz O, Garten W (2006). The role of myristoylation in the membrane association of the Lassa virus matrix protein Z. Virol. J. 3: 93, Bio Medical Central. Doi: 10.1186/1743-422X-3-93.
Sunayev S, Hanke J, Aydin A, Wirkner U, Zastrow I, Reich J, Bork P (1999). Prediction of nonsynonymous single nucleotide polymorphisms in human disease-associated genes. J. Mol. Med. 77: 754-760.
Sunayev S, Ramensky V, Bork P (2000). Towards a structural basis of human non-synonymous single nucleotide polymorphisms. Trends Genet. 16: 198-200.
Valdar W, Thornton J (2001). Conservation helps to identify biologically relevant crystal contacts. J. Mol. Biol. 313: 399-416.
Vitkup D, Sander C, Church GM (2003). The amino-acid mutational spectrum of human genetic disease. Genome Biol. 4-R72.
Wallace A (1997). Tess: a geometric hasing algorithm for deriving 3d co-ordinate templates for searching structural databases. Protein Sci. 6: 2308-2323.
Yang Z, Ro S, Rannala B (2003). Likelihood models of somatic mutation and codon substitution in cancer genes. Genetics 165: 695-705.