Use of Mutual Information Arrays to Predict Coevolving Sites in the Full Length HIV gp120 Protein for Subtypes B and C

Bo Wei[#], Na Han[#], Hai-zhou Liu, Anthony Rayner and Simon Rayner[**]

(State Key Laboratory of Virology, Wuhan Institute of Virology, Chinese Academy of Sciences, Wuhan 430071, China)

Abstract: It is well established that different sites within a protein evolve at different rates according to their role within the protein; identification of these correlated mutations can aid in tasks such as ab initio protein structure, structure function analysis or sequence alignment. Mutual Information is a standard measure for coevolution between two sites but its application is limited by signal to noise ratio. In this work we report a preliminary study to investigate whether larger sequence sets could circumvent this problem by calculating mutual information arrays for two sets of drug naïve sequences from the HIV gp120 protein for the B and C subtypes. Our results suggest that while the larger sequences sets can improve the signal to noise ratio, the gain is offset by the high mutation rate of the HIV virus which makes it more difficult to achieve consistent alignments. Nevertheless, we were able to predict a number of coevolving sites that were supported by previous experimental studies as well as a region close to the C terminal of the protein that was highly variable in the C subtype but highly conserved in the B subtype.

Key words: Mutual information arrays; Predict coevolving sites; Protein evolve; HIV gp120 protein; B and C subtypes

INTRODUCTION

It is well established that different sites within a protein evolve at different rates according to their role within the protein [7,8]; sites that are involved in forming important components of the protein structure or are involved in key functions are generally highly

conserved, in contrast to other sites where mutations can be accommodated without loss of function. These sites are assumed to be less critical. A third possibility is that a site mutation may occur if there is a compensating mutation at a second site. Identification of these correlated mutations can aid in tasks such as *ab initio* protein structure [9, 19], structure function analysis [1, 14] or sequence alignment [4].

Entropy is a common measure of the variation in an alignment. The entropy $H(i)$ at a particular site i is defined in terms of the uncertainty of determining the

amino acid or nucleotide at that site; a site which is highly variable will have a high entropy value, whereas a site that is completely conserved will have an entropy of 0. In the same way, the entropy between two sites *i* & *j* can be defined as a measure of the certainty of predicting the sequence at site *j*, given the sequence at site *i*. These two metrics are combined to define the Mutual Information (MI) which is an explicit measure of the dependency between two sites.

There have been many reported studies that have used variations on MI to identify coevolving site pairs [2, 3, 18, 30]. However, several other approaches have also been developed for predicting coevolving site pairs. These include 1) identification of site pairs with frequencies that deviate from the expected distribution [16,21], 2) site substitution/correlation analysis [10, 26] and 3) statistical coupling analysis which calculates the degree to which the amino acid frequencies at one site are affected by a change in the frequency distribution at a second site[23, 27].

One limitation of all of these methods is that the detected signal is a combination of several components. These arise not only from structural and functional constraints, but also from background noise generated by stochastic events (for example, pairs of sites with high entropy values will have a strong MI signal) and phylogenetic effects (sets of closely related sequences can share two point mutations from a common ancestor which will be over represented in the analysis). Studies suggest that a minimum of 130 sequences are needed to distinguish true signal from background noise[24].

Current estimates indicate that globally, there are more than 33 million currently infected with AIDS[28]. Consequently, HIV is one of the most widely studied viruses with more than 350 000 nucleotide sequences submitted to the NCBI GenBank database as of Dec 2010. Within the genome, the gp120 protein is required for virus entry and plays a key role in the HIV life cycle by mediating binding to CD4 cell surface receptor on target cells [29] and is therefore one of the most frequently sequenced regions of the genome. Additionally, the structure of the protein has been solved in the both bound and unbound states [15, 22, 31, 32]. Thus, the gp120 protein seemed a good choice for a pilot study to investigate whether Mutual Information could be used to study an entire protein in order to identify coevolving sites.

In this study we first generated specific alignments for major subtypes or circulating recombinant forms and estimated the coevolution probability for all possible site pairs using mutual information as the metric. To determine the statistical significance of any site pair we bootstrapped the alignment and compared the Mutual Information between the random and real data sets. We then mapped the most probable coevolving sites on to the solved gp120 structure to interpret the significance of the predicted pair.

Our results suggest that, in the case of a virus such as a HIV with a high mutation rate, a larger sequence set is not necessarily beneficial as it becomes more difficult to achieve a reliable alignment in the highly variable regions of the sequences. Thus it seems that mutual information should still be combined with additional analysis techniques to achieve the best results. Nevertheless, using the larger dataset we were able to able to verify our approach and identify a number of sites that had been identified in previous experimental studies as well as locate sites close to the C terminus of the gp120 protein that only appeared to

be coevolving in the C subtype.

MATERIALS AND METHODS

Sequence selection and Alignment

All full length drug naïve sequences which spanned the entire gp120 gene sequence were downloaded from the LANL HIV database as of 1/3/2009 for subtypes B & C and circulating recombinant form CRF07. A total of 1163, 1180 and 304 sequences were downloaded for these three sets respectively. Only 6 drug naïve samples were available for CRF08.

The sequences were aligned using T-COFFEE [25] and visually inspected and adjusted using SeaView [11] and BioEdit (http://www.mbio.ncsu.edu/bioedit/bioedit.html). Any sites containing more than 50% gaps were removed from the alignment as they introduced too much noise into the analysis. Sequences were initially aligned using nucleotide sequences and the final alignment was translated to protein sequence. In order to directly compare the results between sequence sets the sequences were realigned using sequence HXB2 as a reference .[17]

Calculation of MI values

Mutual information was calculated according to

$$M(i, j) = H(i) + H(j) - H(i, j)$$

Where H(i) and H(j) are the entropy of sites i and j respectively

$$H(i) = -\sum_{s_i} P(s_i) \log P(s_i)$$
$$H(j) = -\sum_{s_i} P(s_j) \log P(s_j)$$

and H(i,j) is the mutual entropy between sites i and j, $(i \neq j)$

$$H(i, j) = \sum P(s_i, s_j') \log P(s_i, s_j')$$

To estimate the statistical significance of the estimated MI values, each alignment was bootstrapped by randomly selecting and switching site pairs from two randomly selected sequences (i.e. site i in sequence k was switched with site j in sequence l). Each alignment was subjected to 5,000 bootstraps and sites that had MI values that were higher than the corresponding values in all 5,000 bootstrapped alignments were retained as most probable candidates for coevolving sites (i.e. $P < 2 \times 10^{-4}$).

Analysis of Significant MI Values

The statistically significant sites selected in the previous step were then mapped on to the solved structure for the gp120 trimer bound to the cell surface receptor CD4 [22] (PDB entry2NY7). This was done using code from the BioJava software package (www.biojava.org) to load the 3D structure and to thread the consensus sequence to obtain the most probable alignment. Based on this structure, the distance between the most probable site pairs could be estimated.

Software

The software used to analyze the alignments and generate the plots was written in java and can run on any platform that supports the java VM. The bootstrap analysis of the B subtype alignment took ~1 week on a standard desktop computer running Debian with a 2.6GHz AMD64 and required~1.6G of memory. Structures were viewed using pymol (www.pymol. org).

RESULTS

The alignment of the gp120 consensus sequence for subtype B and C to the HXB reference sequence is shown in Fig. 1. Key sites in the solved structure are shown in Table 1.

The raw MI arrays calculated for the gp120_B and gp120_C alignments are shown in Fig. 2A and 2B

```
       . . . . .|. . . . | . . . .|. . . .| . . . .|. . . .| . . . .|. . . .| . . .Q|. . . .| . . . .|. . . .| . . . .|. . . .|
              10         20         30         40         50         60         70
C      MMRVKE---KY QHLWRW---- GTMLLGMLMI CSATEKLWVT VYYGVPVWKE ATTTLFCASD AKAYDTEVHN
CRF07  MMRVKE---KY QHLWRW---- GTMLLGMLMI CSATEKLWVT VYYGVPVWKE ATTTLFCASD AKAYDTEVHN
ASIA   MMRVKE---KY QHLWRW---- GTMLLGMLMI CSATEKLWVT VYYGVPVWKE ATTTLFCASD AKAYDTEVHN
B      MMRVKE---KY Q~~~~~~~~~ GTMLLGMLMS --AEKLWVT VYYGVPVWKE ATTTLFCASD AKAYDTEVHN
HXB2   MMRVKE---KY QHLWRWGWRW GTMLLGMLMI CSATEKLWVT VYYGVPVWKE ATTTLFCASD AKAYDTEVHN

       . . . . .|. . . . | . . . .|. . . .| . . . .|. . . .| . . . .|. . . .| . . . .|. . . .| . . . .|. . . .| . . . .|. . . .|
              80         90        100        110        120        130        140
C      MVWATHACVPT DPNPQEVVLN -VTENFNMWK NDMVEQMHED IISLWDQSLK PCVKLTPLCV SLKCTDL---
CRF07  MVWATHACVPT DPNPQEVVLN -VTENFNMWK NDMVEQMHED IISLWDQSLK PCVKLTPLCV SLKCTDL---
ASIA   MVWATHACVPT DPNPQEVVLV NVTENFNMWK NDMVEQMHED IISLWDQSLK PCVKLTPLCV SLKCTDTNTN
B      MVWATHACVPT DPNPQEVVLV NVTENFNMWK NDMVEQMHED IISLWDQSLK PCVKLTPLCV SLKCTDLKND
HXB2   MVWATHACVPT DPNPQEVVLV NVTENFNMWK NDMVEQMHED IISLWDQSLK PCVKLTPLCV SLKCTDLKND

       . . . . .|. . . . | . . . .|. . . .| . . . .|. . . .| . . . .|. . . .| . . . .|. . . .| . . . .|. . . .| . . . .|. . . .|
             150        160        170        180        190        200        210
C      M--------- -----EIKNC SFNISTSIRG KVQKEYAFFY KLDIIPIDND TTSYKLTSCN TSVITQACPK
CRF07  M--------- -----EIKNC SFNISTSIRG KVQKEYAFFY KLDIIPIDND TTSYKLTSCN TSVITQACPK
ASIA   MSSS~~~~GRM ~MEKGEIKNC SFNISTSIRG KVK-EYAFFY KLDIIPIDND TTSYKLTSCN TSVITQACPK
B      MTNTNSSSGRM ~MEKGEIKNC SFNISTSIRG KVQKEYAFFY KLDIIPIDND TTSYKLTSCN TSVITQACPK
HXB2   MTNTNSSSGRM IMEKGEIKNC SFNISTSIRG KVQKEYAFFY KLDIIPIDND TTSYKLTSCN TSVITQACPK

       . . . . .|. . . . | . . . .|. . . .| . . . .|. . . .| . . . .|. . . .| . . . .|. . . .| . . . .|. . . .| . . . .|. . . .|
             220        230        240        250        260        270        280
C      MVSFEPIPIHY CAPAGFAILK CNNKTFNGTG PCTNVSTVQC THGIRPVVST QLLLNGSLAE EEVVIRSVNF
CRF07  MVSFEPIPIHY CAPAGFAILK CNNKTFNGTG PCTNVSTVQC THGIRPVVST QLLLNGSLAE EEVVIRSVNF
ASIA   MVSFEPIPIHY CAPAGFAILK CNNKTFNGTG PCTNVSTVQC THGIRPVVST QLLLNGSLAE EEVVIRSVNF
B      MVSFEPIPIHY CAPAGFAILK CNNKTFNGTG PCTNVSTVQC THGIRPVVST QLLLNGSLAE EEVVIRSVNF
HXB2   MVSFEPIPIHY CAPAGFAILK CNNKTFNGTG PCTNVSTVQC THGIRPVVST QLLLNGSLAE EEVVIRSVNF

       . . . . .|. . . . | . . . .|. . . .| . . . .|. . . .| . . . .|. . . .| . . . .|. . . .| . . . .|. . . .| . . . .|. . . .|
             290        300        310        320        330        340        350
C      MTDNAKTIIVQ LNTSVEINCT RPNNNTRKRI RI--GPGRAF VTIGK-IGNM RQAHCNISRA KWNNTLKQIA
CRF07  MTDNAKTIIVQ LNTSVEINCT RPNNNTRKRI RI--GPGRAF VTIGK-IGNM RQAHCNISRA KWNNTLKQIA
ASIA   MTDNAKTIIVQ LNTSVEINCT RPNNNTRKRI RI--GPGRAF VTIGK-IGNM RQAHCNISRA KWNNTLKQIA
B      MTDNAKTIIVQ LNTSVEINCT RPNNNTRKRI RI--GPGRAF VTIGK-IGNM RQAHCNISRA KWNNTLKQIA
HXB2   MTDNAKTIIVQ LNTSVEINCT RPNNNTRKRI RIQRGPGRAF VTIGK-IGNM RQAHCNISRA KWNNTLKQIA

       . . . . .|. . . . | . . . .|. . . .| . . . .|. . . .| . . . .|. . . .| . . . .|. . . .| . . . .|. . . .| . . . .|. . . .|
             360        370        380        390        400        410        420
C      MSKLREQFGNN KTIIFKQSSG GDPEIVTHSF NCGGEFFYCN STQLF----- ---------- -------ITL
CRF07  MSKLREQFGNN KTIIFKQSSG GDPEIVTHSF NCGGEFFYCN STQLF----- ---------- -------ITL
ASIA   MSKLREQFG-N KTIIFKQSSG GDPEIVTHSF NCGGEFFYCN STQLFNSTWN S----~~~~~ ~TEGSDTITL
B      MSKLREQFG-- KTIIFKQSSG GDPEIVTHSF NCGGEFFYCN STQLFNSTWN ---------- -------ITL
HXB2   MSKLREQFGNN KTIIFKQSSG GDPEIVTHSF NCGGEFFYCN STQLFNSTWF NSTWSTEGSN NTEGSDTITL

       . . . . .|. . . . | . . . .|. . . .| . . . .|. . . .| . . . .|. . . .| . . . .|. . . .| . . . .|. . . .| . . . .|. . . .|
             430        440        450        460        470        480        490
C      MPCRIKQIINM WQKVGKAMYA PPISGQIRCS SNITGLLLTR DGGNSN-NES EIFRPGGGDM RDNWRSELYK
CRF07  MPCRIKQIINM WQKVGKAMYA PPISGQIRCS SNITGLLLTR DGGNSN-NES EIFRPGGGDM RDNWRSELYK
ASIA   MPCRIKQIINM WQKVGKAMYA PPISGQIRCS SNITGLLLTR DGGNS~~~ES EIFRPGGGDM RDNWRSELYK
B      MPCRIKQIINM WQKVGKAMYA PPISGQIRCS SNITGLLLTR DGGNSN--ES EIFRPGGGDM RDNWRSELYK
HXB2   MPCRIKQIINM WQKVGKAMYA PPISGQIRCS SNITGLLLTR DGGNSN-NES EIFRPGGGDM RDNWRSELYK

       . . . . .|. . . . | . . . .|. . . .| . . . .|. . . .|.
             500        510
C      MYKVVKIEPLG VAPTKAKRRV VQREKR
CRF07  MYKVVKIEPLG VAPTKAKRRV VQREKR
ASIA   MYKVVKIEPLG VAPTKAKRRV VQREKR
B      MYKVVKIEPLG VAPTKAKRRV VQREKR
HXB2   MYKVVKIEPLG VAPTKAKRRV VQREKR
```

Fig. 1. Realignment of gp120 sequences sets for subtype B, subtype C, CRF07 to reference sequence HXB-2. " ~ " indicates sites that contain less than three amino acids and cannot be translated.

Table 1.   Key structure elements within the gp120 sequences.

Site		Feature	Site		Feature	Site		Feature
HXB	2NY7		HXB	2NY7		HXB	2NY7	
96	99	B1	268	270	B9	391	396	B17
100	103	B1_TO_A1	271	277	LOOP_C	397	401	A4
104	121	A1	278	280	B10	402	405	B18
122	123	A1_TO_B2	281	290	LOOP_D	406	411	LOOP_V4
124	128	B2	291	296	B11	412	423	B19
129	204	V1V2_LOOP	297	298	B11_TO_B12	424	429	B20
205	208	B3	299	304	B12	430	431	B20_TO_B21
209	219	B3_TO_B4	305	338	V3_LOOP	432	438	B21
220	226	B4	339	343	B13	439	445	LOOP_F
227	228	B4_TO_B5	344	356	A2	446	452	B22
229	236	B5	357	367	LOOP_E	453	461	B23
237	241	LOOP_A	368	371	B14	462	468	LOOP_V5
242	246	B6	372	374	B14_TO_B15	469	476	B24
247	247	B6_TO_B7	374	378	B15	477	480	B24_TO_A5
248	253	B7	379	383	A3	481	489	A5
254	258	B8	384	387	B16	490	498	B25
259	267	LOOP_B	388	390	B16_TO_B17			

gp120 sequences realigned to reference sequence HXB2 (left column). Second column is the corresponding position in solved structure (PDB entry 2NY7). Sequences were aligned using custom biojava code to thread the HXB2 sequence on to the solved structure.

respectively. The figures also show the location of the points with respect to the regions of the solved structure and the specific secondary structure features that have been identified within the structure. In both arrays the V1/V2 and V3 loops show the greatest variation. Although there appears to be more variation in the V1/V2 loop, this is because many of the sites within the V3 loop contained large numbers of insertions or deletions and couldn't be included in the analysis. For the raw data, the most notable difference occurs close to the amino terminal of the protein. The C subtype array appears to contain a region of coevolving sites which is absent in the B subtype array.

The statistically significant points ($P < 0.0002$) that remain after bootstrapping are shown in Fig. 2C and 2D. For subtype B, there are two clearly defined regions that remain and which are distinct from the three regions identified in the subtype C array. The figure also shows the range of sites that define

hotspots, but it seems likely that some of the signal is due to noise and not all of the sites are coevolving. Region A occurs around aa140 and corresponds to the N terminal of the V1 loop. This is not present in the solved structure, but the loop is thought to be located at the top of the trimer away from where the protein binds to the CD4 receptor (arrow A in Fig. 3). This prediction is supported by a fusogenicity study which found that a mutation at position aa140 had a negative effect on the fusion ability of the virus [6]. A second region is predicted to occur between positions 140 and 339 (A and B on Fig. 3). In this case, the positions are on opposite sides of the protein and even in the trimer are not in close proximity; therefore it seems unlikely that these sites are coevolving and the signal is probably produced by random co-mutations that the bootstrapping failed to remove.

For gp120 subtype C there are three regions that remain after bootstrapping. The first of these occurs at
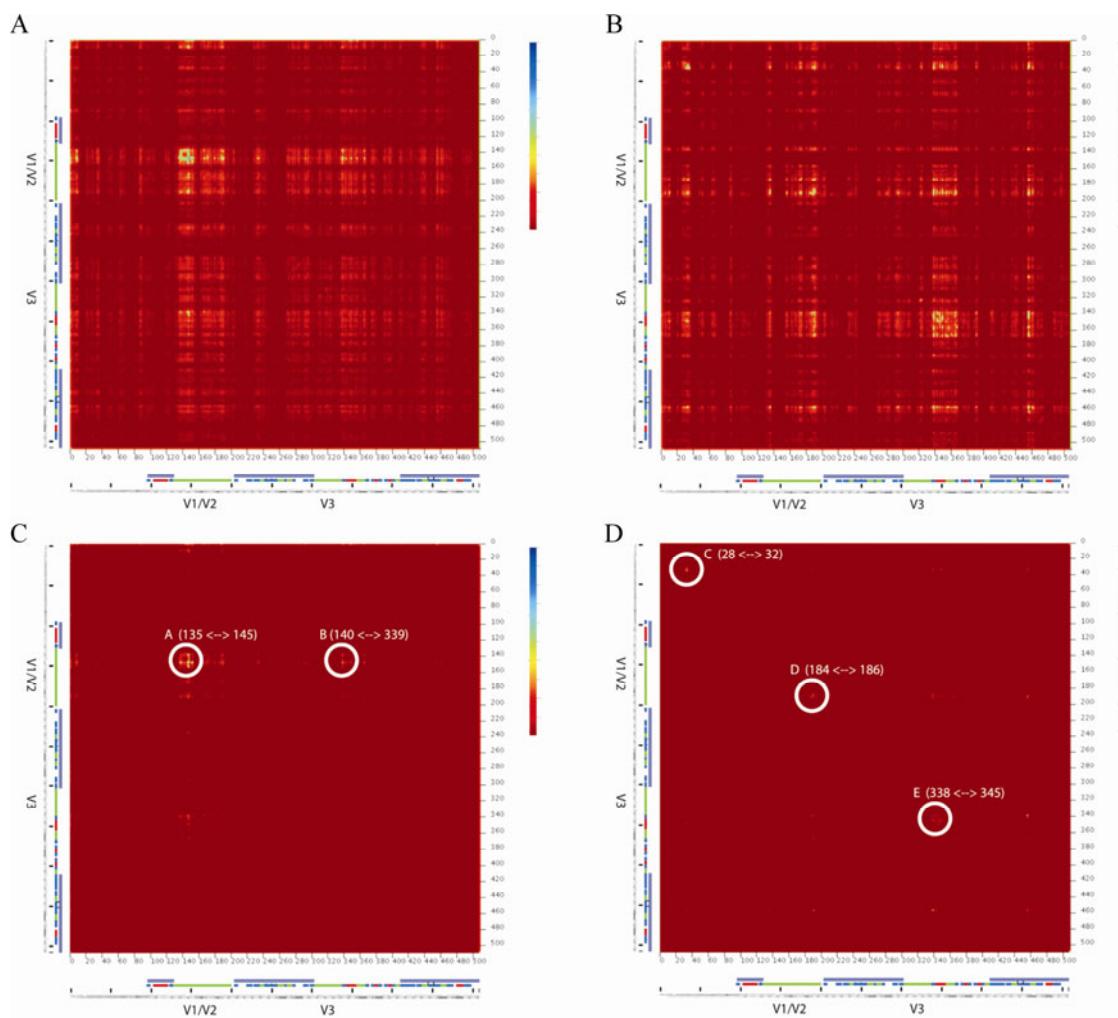
Fig. 2. Calculated mutual information arrays for gp120 subtype B and subtype C. (A) and (B) raw mutual information for subtype B and C respectively. (C) and (D), mutual information arrays after removing sites that were not statistically significant at 5000 bootstraps. Circled regions highlight regions that appear to contain sites of interest. Letters correspond to the marked positions on the structure in Fig. 3.

~aa30 at the N terminus of the protein. Again, this region falls outside the solved structure but it is likely this region is located at the base of the trimer where the structure binds to the CD4 cell surface receptor (Fig. 2). Because the alignments for both subtypes are realigned according to reference sequence HXB2 it is possible to compare this region in the two alignments. Interestingly, between positions aa28 to aa35 in subtype C this region does appear to show greater variation than subtype B. Fig. 4 shows an entropy plot for this region. While the entropy is relatively well conserved across the region, the average entropy between aa28 to aa35 for subtype C is an order of magnitude higher than the value for subtype B. This finding is further supported by visual inspection of the alignment (data not shown).

Region D corresponds to the C terminal of the V2 protein at ~aa180. Again, because this region is not part of the solved structure it is difficult to interpret the significance of these sites. However, once again, the prediction is supported by the same fusogenicity study which found that a mutation at aa180 also
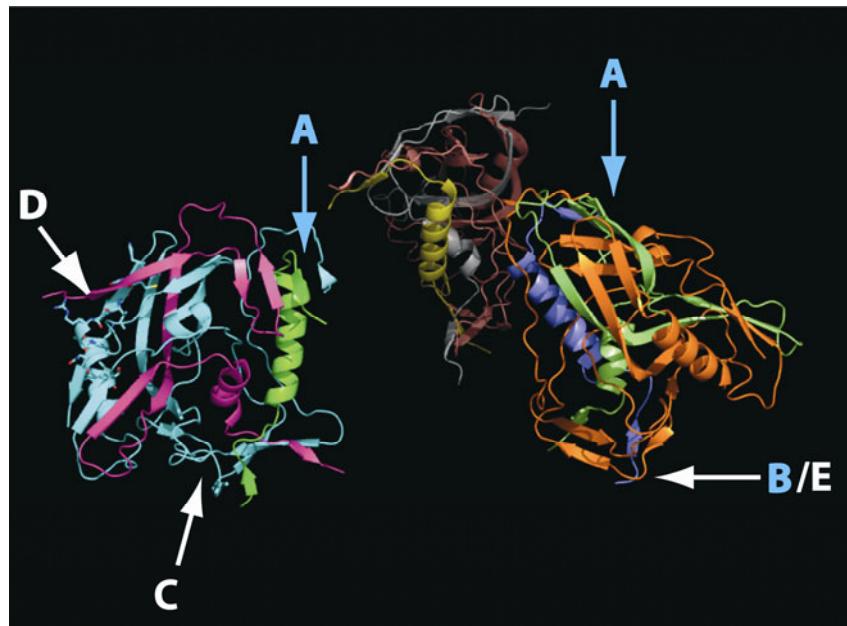
Fig. 3. Location of predicted coevolving sites on the Solved structure for the gp120 trimer (PDB structure 2NY7). The base of the trimer binds to the CD4 cell surface receptor. Letters refer to the highlighted regions in Fig. 2. A and D correspond to the N terminus of the V1 and C terminus of the V2 loop respectively, B/E corresponds to the V3 loop in the CPGR region. C corresponds to a region close to the N terminus of the gp120 protein. See text for details
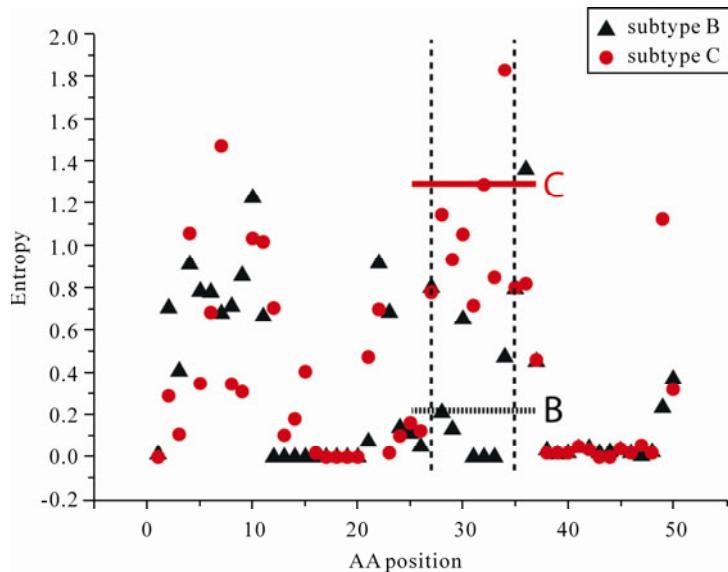
Fig. 4. Entropy vs amino acid position in the area around region C in Fig. 2 and Fig. 3 for subtype B and C. Subtype B is shown as black triangle, subtype C is shown as red circles. In this region, the mutual information array predicts a number of coevolving sites for subtype C, but not for subtype B. The entropy plot indicates that the subtype B alignment is highly conserved whereas the subtype C sequences show greater variation. The vertical dashed lines indicate the range of region C marked in Fig. 2D, The horizontal lines show the average entropy across the region for each subtype: subtype B dashed black line; subtype C red solid line. Outside this region the entropy is approximately the same for both subtypes.

produced a decrease in the fusion ability of the virus [6]. Region E in Fig. 1D corresponds to the V3 loop, close to the CPGR motif [12, 13, 20]. There are many coevolving sites predicted in this region which is consistent with one of the earliest mutual information studies on HIV which examined the variation in V3 loop within a broad sequence set [18].

DISCUSSION

Identification of coevolving sites can provide insight into protein structure and function and can provide guidance in tasks such as sequence analysis [4] and protein structure and function prediction [1, 9, 14, 19]. Many methods have been developed to identify these sites [10, 16, 21, 23, 26, 27]. Calculation of the mutual information between sites is a relatively simple method but has the advantage of being a more intuitive quantity. The downside of the method is it generates a high level of background signal; our data would probably benefit from additional bootstrapping. The use of mutual information to identify coevolving sites is not particularly novel but our goal here was to investigate whether larger datasets could improve the prediction ability of the method.

We should emphasize that our results are preliminary and should be interpreted with caution. In particular, we attempted to identify coevolving sites in the gp120 B and C subtypes by using a single dataset for each subtype. A better approach might be to first estimate the phylogenetic relationship between the sequences within a particular subtype and then analyze subgroups that are defined according to highly supported sub-clades. This would produce better alignments which would reduce the background signal and help to identify less frequent co-mutations that may

nevertheless be significant.

Another improvement worth investigating would be the separation of the alignment into highly variable and less variable regions. Currently, our approach is best suited to identification of sites in highly variable regions – any covariation in the more conserved regions may be lost in the noise generated by the highly variable regions. This occurs because, during bootstrapping, this randomness becomes somewhat dispersed through the alignment as the sequence is shuffled and this may be sufficient to mask some of the low, but nevertheless significant, signal. By dividing the array into high variable/highly variable, more conserved/more conserved and highly variable/more conserved sections and analyzing these regions separately, it may be possible to identify additional sites of statistical significance.

Finally, in this study we only used the Mutual Information metric to identify coevolving sites. It has already been reported that the principal drawback with this method is the high background signal. Our goal was to see whether larger datasets would lead to improved signal to noise ratio and we deliberately chose HIV as this is one of the most sequenced viruses to date. However, it seems that there is a trade off between the benefit associated with the larger number of sequences and the high mutation rate of the virus which makes it difficult to obtain good alignments and which increases the background signal. It seems therefore that using Mutual Information as the only measure for identifying coevolution is not sufficient and it would be worthwhile to estimate additional quantities based on analysis of site pair frequency [16, 21] distributions, site substitution/correlation analysis [10, 26] and statistical coupling analysis [23, 27].

We are in the process of incorporating these methods into our software; our latest software uses the Average Product Correction to estimate the background signal in the mutual information [5].

Nevertheless, in spite of all these concerns, we were able to make predictions for (i) the V1/V2 loop that were supported by experimental studies, (ii) the V3 loop that were supported both by results from previous analysis and experimental studies and (iii) a novel prediction at the C terminal of the protein. Future versions of the software may be for useful functional analysis of well studied proteins with known or partially solved structures

References

1. **Atchley W R, Terhalle W, Dress A**. 1999. Positional dependence, cliques, and predictive motifs in the bHLH protein domain. **J Mol Evol**, 48 (5): 501-516.

2. **Clarke N D**. 1995. Covariation of residues in the homeodomain sequence family. **Protein Sci**, 4 (11): 2269-2278.

3. **Daub C O, Steuer R, Selbig J, et al.** 2004. Estimating mutual information using B-spline functions--an improved similarity measure for analysing gene expression data. **BMC Bioinformatics**, 5: 118.

4. **Dickson R J, Wahl L M, Fernandes A D, et al.** 2010. Identifying and seeing beyond multiple sequence alignment errors using intra-molecular protein covariation. **PLoS ONE**, 5 (6): e11082.

5. **Dunn S D, Wahl L M, Gloor G B**. 2008. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. **Bioinformatics**, 24 (3): 333-340.

6. **Etemad-Moghadam B, Sun Y, Nicholson E K, et al.** 2000. Envelope glycoprotein determinants of increased fusogenicity in a pathogenic simian-human immunodeficiency virus (SHIV-KB9) passaged *in vivo*. **J Virol**, 74 (9): 4433-4440.

7. **Excoffier L, Hofer T, Foll M**. 2009. Detecting loci under selection in a hierarchically structured population. **Heredity**, 103 (4): 285-298.

8. **Excoffier L, Yang Z**. 1999. Substitution rate variation among sites in mitochondrial hypervariable region I of humans and chimpanzees. **Mol Biol Evol**, 16 (10): 1357-1368.

9. **Fariselli P, Olmea O, Valencia A, et al.** 2001. Progress in predicting inter-residue contacts of proteins with neural networks and correlated mutations. **Proteins**, Suppl 5: 157-162.

10. **Gobel U, Sander C, Schneider R, et al.** 1994. Correlated mutations and residue contacts in proteins. **Proteins**, 18 (4): 309-317.

11. **Gouy M, Guindon S, Gascuel O**. 2010. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. **Mol Biol Evol**, 27 (2): 221-224.

12. **Handley M A, Steigbigel R T, Morrison S A**. 1996. A role for urokinase-type plasminogen activator in human immunodeficiency virus type 1 infection of macrophages. **J Virol**, 70 (7): 4451-4456.

13. **Hartley O, Klasse P J, Sattentau Q J, et al.** 2005. V3: HIV's switch-hitter. **AIDS Res Hum Retroviruses**, 21 (2): 171-189.

14. **Hemmerich C, Kim S**. 2007. A study of residue correlation within protein sequences and its application to sequence classification. **EURASIP J Bioinform Syst Biol**, 2007: doi:10.1155/2007/87356.

15. **Huang C C, Tang M, Zhang M Y, et al.** 2005. Structure of a V3-containing HIV-1 gp120 core. **Science**, 310 (5750): 1025-1028.

16. **Kass I, Horovitz A**. 2002. Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations. **Proteins**, 48 (4): 611-617.

17. **Korber B T F B T, Kuiken C L, Pillai S K, et al.** 1998. Numbering Positions in HIV Relative to HXB2CG. In: **Human Retroviruses and AIDS 1998: A Compilation and Analysis of Nucleic Acid and Amino Acid Sequences.**: Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM. 102-111.

18. **Korber B T, Farber R M, Wolpert D H, et al.** 1993. Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. **Proc Natl Acad Sci U S A**,

90 (15): 7176-7180.

19. **Kundrotas P J, Alexov E G**. 2006. Predicting residue contacts using pragmatic correlated mutations method: reducing the false positives. **BMC Bioinformatics**, 7: 503.

20. **Laakso M M, Lee F H, Haggarty B*, et al.*** 2007. V3 loop truncations in HIV-1 envelope impart resistance to coreceptor inhibitors and enhanced sensitivity to neutralizing antibodies. **PLoS Pathog**, 3 (8): e117.

21. **Larson S M, Di Nardo A A, Davidson A R**. 2000. Analysis of covariation in an SH3 domain sequence alignment: applications in tertiary contact prediction and the design of compensating hydrophobic core substitutions. **J Mol Biol**, 303 (3): 433-446.

22. **Liu J, Bartesaghi A, Borgnia M J*, et al.*** 2008. Molecular architecture of native HIV-1 gp120 trimers. **Nature**, 455 (7209): 109-113.

23. **Lockless S W, Ranganathan R**. 1999. Evolutionarily conserved pathways of energetic connectivity in protein families. **Science**, 286 (5438): 295-299.

24. **Martin L C, Gloor G B, Dunn S D*, et al.*** 2005. Using information theory to search for co-evolving residues in proteins. **Bioinformatics**, 21 (22): 4116-4124.

25. **Notredame C, Higgins D G, Heringa J**. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. **J Mol Biol**, 302 (1): 205-217.

26. **Olmea O, Rost B, Valencia A**. 1999. Effective use of sequence correlation and conservation in fold recognition. **J Mol Biol**, 293 (5): 1221-1239.

27. **Suel G M, Lockless S W, Wall M A*, et al.*** 2003. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. **Nat Struct Biol**, 10 (1): 59-69.

28. **UNaids**. 2011. UNAIDS Report on the global AIDS epidemic 2010. **http://www.unaids.org/globalreport/ Global_report.htm**.

29. **Wu H, Kwong P D, Hendrickson W A**. 1997. Dimeric association and segmental variability in the structure of human CD4. **Nature**, 387 (6632): 527-530.

30. **Wu T D, Schiffer C A, Gonzales M J*, et al.*** 2003. Mutation patterns and structural correlates in human immunodeficiency virus type 1 protease following different protease inhibitor treatments. **J Virol**, 77 (8): 4836-4847.

31. **Wyatt R, Kwong P D, Desjardins E*, et al.*** 1998. The antigenic structure of the HIV gp120 envelope glycoprotein. **Nature**, 393 (6686): 705-711.

32. **Zhou T, Xu L, Dey B*, et al.*** 2007. Structural definition of a conserved neutralization epitope on HIV-1 gp120. **Nature**, 445 (7129): 732-737.