DISCOVERING STATISTICAL VULNERABILITIES IN HIGHLY MUTABLE VIRUSES: A RANDOM MATRIX APPROACH

A. A. Quadeer[a], *R. H. Y. Louie*[a], *K. Shekhar*[b,c], *A. K. Chakraborty*[b,c,d,e,f], *I. Hsing*[g,h], *and M. R. McKay*[a,h]

[a]Department of Electronic and Computer Engineering, HKUST, Hong Kong

[b]Department of Chemical Engineering, MIT, Cambridge, MA 02139 USA

[c]Ragon Institute of Massachusetts General Hospital, MIT and Harvard, Boston, MA 02129 USA

[d]Department of Physics, Chemistry, and Biological Engineering, MIT, Cambridge, MA 02139 USA

[e]Institute for Medical Engineering and Science, MIT, Cambridge, MA 02139 USA

[f]Institute for Advanced Study, HKUST, Clear Water Bay, Kowloon, Hong Kong

[g]Department of Chemical and Biomolecular Engineering, HKUST, Hong Kong

[h]Division of Biomedical Engineering, HKUST, Hong Kong

ABSTRACT

The advancement in fast DNA sequencing technologies has opened up new opportunities to explore a diverse set of questions in biomedical research. In this paper, we review a general method which utilizes the available sequence data to determine potential weaknesses in highly mutable viruses, and which has shown promise in the design of vaccines. A key computational part of this method employs concepts from random matrix theory to obtain a robust estimate of a large covariance matrix. We apply this general method on hepatitis C virus as an example, and verify its usefulness by linking with the existing experimental and structural data.

Index Terms— Random matrices, estimation, hepatitis C virus.

1. INTRODUCTION

Since the completion of the Human Genome Project in 2003, the field of bioinformatics has developed at a rapid pace [1]. The advancement in high throughput sequencing technologies [2] has spawned an influx of sequence data in publicly available databases. This has opened up new opportunities and challenges in biomedical research for numerous scientific disciplines including computer science, statistical signal processing, and mathematics. Much effort has been focused on developing fast algorithms for sequence alignment [3], analyzing gene expression [4], and predicting drug outcome and resistance [5].

In this paper, we review an approach developed by the authors and their collaborators, first in [6] and subsequently in [7], which utilizes available viral sequence data[1] for identifying potential weaknesses in highly mutable viruses that has shown promise in the design of efficient vaccines. The procedure involves characterizing the correlations between substitutions at all pairs of sites[2] to determine groups of co-evolving sites[3], termed "sectors", that may correspond to certain structural and functional constraints essential for the survival of the virus. This approach is in contrast to the conventional vaccine design strategy which involves only identifying highly conserved parts of the virus, i.e., the parts comprising sites which show no (or very few) substitutions. Unfortunately, such parts are not present in highly mutable viruses, like human immunodeficiency virus (HIV), hepatitis C virus (HCV), and influenza. Moreover, targeting only the sites with high conservation overlooks the possibility of viral escape routes involving multiple sites. The presented method aims to alleviate these issues by incorporating the statistical information regarding simultaneous substitutions at multiple sites. This, in turn, can potentially serve as a useful guide for designing more efficient vaccines [6].

In this approach, when forming the sectors, a key computational challenge is to form a robust estimate of a covariance matrix of substitutions. With the currently available sequenced data, for many viruses, the number of sequences (observations) is comparable to the number of sites (variables) in the protein, with both being rather large. For covariance estimation under such scenarios, it is now well-known that the classical sample covariance estimator may fail remarkably. To address this issue and to obtain more robust estimators, methods based on random matrix theory (RMT) have proven particularly useful (see e.g., [8, 9]). The framework which we describe utilizes such RMT based covariance estimators, and was first employed by a sub-group of the authors and their collaborators in [6] to identify groups of co-evolving sites for HIV. In this paper, we apply this framework to HCV that causes an infectious liver disease called hepatitis C. A more complete account, including details on proposed vaccine design strategies stemming from these results, can be found in the extended journal contribution [7].

It is interesting to point out that the RMT based approach which we describe is closely analogous to an approach specified in finance [10], where it was employed to decouple intrinsic correlations in stock price fluctuations of companies from noisy time series data. This procedure was found to group companies into independent well-known "economic sectors". In biology, a similar approach was also proposed in [11] to predict groups of sites in a protein having distinct functions.

2. METHOD

In the following, we outline the framework that can be employed to obtain groups of co-evolving sites in a protein. This framework is general and can be applied to any virus.

[1]A virus consists of proteins and each protein comprises a collection (sequence) of amino acids.

[2]A site refers to a position in the protein.

[3]Co-evolving sites refers to sites with correlated substitutions.

978-1-4799-4975-5/14/$31.00 ©2014 IEEE

2.1. Constructing the covariance matrix

Amino acid sequences are first downloaded from a public database, and then aligned into an $M \times N$ matrix $\mathbf{A}$ (referred to as an amino acid multiple sequence alignment (MSA) matrix), where M denotes the number of sequences and N denotes the number of sites in the protein. Moreover, the (i,j)th element of $\mathbf{A}$ is given by A_{ij}, where A denotes the amino acid in sequence i at site j. The amino acid MSA matrix $\mathbf{A}$ is then converted into a binary MSA matrix $\mathbf{B}$, where the (i,j)th element of $\mathbf{B}$ is given by

$$B_{ij} = \begin{cases} 0 & \text{if } A_{ij} \text{ is a wild-type,} \\ 1 & \text{otherwise,} \end{cases} \tag{1}$$

where an amino acid A_{ij} is considered a wild-type if it is the most common amino acid at site j in $\mathbf{A}$. Thus a '1' at a specific site in $\mathbf{B}$ indicates a substitution. Subject to certain practical conditions as discussed in [7], this binary approximation will not significantly affect the results. A $N \times N$ sample covariance matrix $\mathbf{C}$ is then constructed based on the binary MSA matrix, with the (k,ℓ)th element given by

$$C_{k\ell} = \frac{f_{k\ell} - f_k f_\ell}{\sqrt{V_k V_\ell}}, \quad k,\ell = 1,2,...,N \tag{2}$$

where $f_{k\ell} = \frac{1}{M}\sum_{i=1}^{M} B_{ik}B_{i\ell}$ is the frequency of a simultaneous substitution at sites k and ℓ, $f_k = \frac{1}{M}\sum_{i=1}^{M} B_{ik}$ is the frequency of substitution at site k only, and $V_k = f_k(1-f_k)$ is the variance of substitution at site k.

2.2. Cleaning the covariance matrix

The covariance matrix constructed in (2) is corrupted by phylogeny[4] and finite sample noise (statistical noise) [6]. As the goal of this study is to identify sites that co-evolve due to only intrinsic structural and functional constraints, the sample covariance matrix needs to be "cleaned".

(a) Removing the effect of phylogeny: The covariance matrix can be written in terms of its eigenvalues as follows:

$$\mathbf{C} = \sum_{k=1}^{N} \lambda_k \mathbf{u}_k \mathbf{u}_k^{\mathsf{T}}, \tag{3}$$

where λ_k is the kth-largest eigenvalue and $\mathbf{u}_k$ is the corresponding eigenvector. A comparatively large value (λ_1) is observed among the eigenvalues due to phylogenetic correlations [6]. The effect of this common ancestry is removed by using a simple linear regression approach [10] that removes the maximum eigenvalue and its effect on the remaining data.

(b) Removing noise using RMT: As mentioned in Section 1, RMT is particularly useful in obtaining a robust covariance matrix estimate when both the number of observations and variables are large and comparable. Specifically as both quantities approach infinity in proportion, the Marčenko-Pastur (MP) law [12] indicates that the empirical eigenvalue distribution of a sample covariance matrix generated from independent and identically distributed (i.i.d.) elements converges to a deterministic distribution (with probability one). Thus by studying the eigenvalue distribution of $\mathbf{C}$, these RMT results can be used to distinguish between correlations due to statistical noise (which is generally i.i.d.) and the interesting correlations among sites. To determine which eigenvalues correspond to signal and which correspond to noise, rather than applying the MP law directly[5], the eigenvalue distribution of $\mathbf{C}$ is compared to that of an ensemble of "random matrices" obtained by randomly shuffling the entries along each column (site) of $\mathbf{B}$. The maximum eigenvalue of the randomized alignments, $\lambda_{\max}^{\text{rnd}}$, is considered as an upper bound for statistical noise and thus all the eigenvalues less than or equal to it are discarded [6]. The cleaned covariance matrix is thus given by

$$\hat{\mathbf{C}} = \sum_{\lambda_{\max}^{\text{rnd}} < \lambda_k < \lambda_1} \lambda_k \mathbf{u}_k \mathbf{u}_k^{\mathsf{T}}. \tag{4}$$

There are alternative RMT-based approaches present in the literature, for example [9, 13], that can be employed for cleaning statistical noise.

2.3. Forming the sectors

Once the covariance matrix is cleaned of phylogeny and noise, the sites in the protein are grouped into sectors by studying the elements of the eigenvectors corresponding to the eigenvalues in the range $\lambda_{\max}^{\text{rnd}} < \lambda_k < \lambda_1$. Let α be the total number of eigenvalues in this range. These eigenvalues are arranged in descending order, i.e., $\lambda_1' > \lambda_2' > \cdots > \lambda_\alpha'$, and the corresponding eigenvectors are numbered accordingly. The sectors are then formed based on the eigenvectors as follows:

$$\text{Sector } k := \{n : \left|\mathbf{u}_k'(n)\right| > \epsilon \text{ and } n \in \{1,2,...,N\}\} \tag{5}$$

where $k = 1,2,...,\alpha$, $|\mathbf{u}_k'(n)|$ is the absolute value of the nth element of $\mathbf{u}_k'$, and ϵ is a small positive constant value that serves as a threshold. A small value of ϵ is chosen to include many sites in the sectors, followed by a pruning procedure discussed below. The resulting sectors may have overlapping sites[6]. This issue is resolved by assigning each overlapping site to the sector with the sites of which it has the highest mean absolute correlation coefficient (based on $\hat{\mathbf{C}}$) [6].

Within the binary approximation (1), a pair of sites can either be positively correlated, negatively correlated or uncorrelated. A positive correlation implies that a simultaneous substitution at both sites occurs more frequently than if the sites mutated independently. Thus a positive correlation between sites can correspond to "compensatory interactions" between the sites. In contrast, a negative correlation means that a simultaneous substitution at both sites occurs less frequently than if the sites mutated independently, suggesting that such a double mutant would be deleterious for viral fitness. The reason why this is important for vaccine design is because a vaccine induces an immune response which would force the virus to mutate in order to survive. Thus, one would like to design a vaccine which forces multiple substitutions at sites where the virus does not tolerate simultaneous substitutions (i.e., the negatively correlated sites) and avoid the sites where multiple simultaneous substitutions are more probable (i.e., the positively correlated sites) [6, 7]. In order to identify statistically significant positively and negatively correlated pairs of sites, the following convention is used:

$$\text{Sites } i \text{ and } j \text{ are } \begin{cases} \text{Positively correlated if } \hat{C}_{ij} \geq \delta^+, \\ \text{Negatively correlated if } \hat{C}_{ij} \leq \delta^-, \\ \text{Uncorrelated if } \delta^- < \hat{C}_{ij} < \delta^+, \end{cases} \tag{6}$$

where $\hat{C}_{ij}$ denotes the (i,j)th element of $\hat{\mathbf{C}}$ and the thresholds δ^+ and δ^- are obtained from the ensemble of the covariance matrices

[4] The inherent correlation among sequences due to evolution from a single ancestor. This concept of phylogeny is analogous to the "market mode" in finance as discussed in [10].

[5] To account for finite size effects, the MP law is not applied directly here.

[6] Sites which fall into multiple sectors.

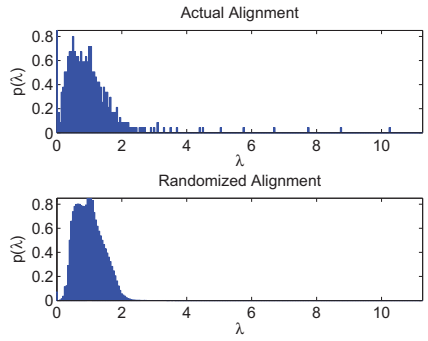


Fig. 1: Eigenvalue distribution of NS3 computed from the covariance matrix resulting from the actual alignment (upper) and 1000 randomized alignments (lower).

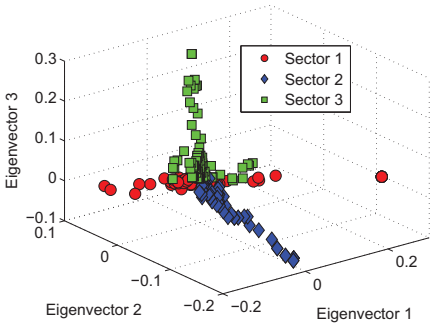
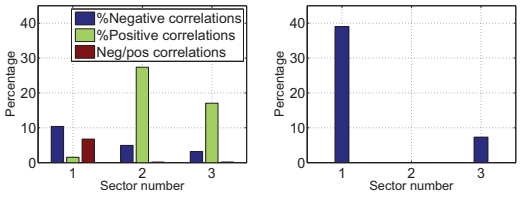


Fig. 2: 3-D scatter plot of the loadings of the eigenvectors showing the three distinct sectors.



(a) (b)

Fig. 3: (a) Percentage of negatively correlated sites, percentage of positively correlated sites, and ratio of negative to positive correlations in the three sectors of NS3. (b) Percentage of sites in the parts targeted by HCV controllers present in the three sectors of NS3.

constructed by randomizing the mutants in the binary MSA matrix (i.e., randomly shuffling the columns). Specifically, these thresholds are chosen such that the correlations with magnitude larger than δ^+ and smaller than δ^- arise with a very low probability, $P = 10^{-2}$, in the randomized case. The sites in each sector that do not show any correlation larger (smaller) than the positive (negative) threshold $\delta^+(\delta^-)$ with the sites in all the sectors are discarded. In addition to this approach, other methods can also be used to form sectors (see [11] and [14]). One can also potentially use sophisticated eigenvector estimation methods that utilize the inherent sparsity to form sectors. These methods are being investigated and are not further discussed here.

3. APPLICATION TO NS3 PROTEIN OF HCV

In this section, we apply the general framework presented in Section 2 to HCV. HCV causes an infectious disease called hepatitis C that affects mainly the liver. It infects more than 170 million people globally. Although there is a treatment available for HCV [15], it is prolonged, quite expensive, and has extensive side-effects. Much efforts are being carried out to design a functioning vaccine for HCV; see ([16] and the references therein). The major problem related to the design of a HCV vaccine is the virus's extreme mutability which helps it to evade the immune system [17]. Interestingly however, there are rare individuals who are able to clear or control the virus without any treatment (referred as "HCV controllers"). Thus, it is important to understand how the immune systems of these individuals are able to control HCV. Experiments indicate that the NS3 protein is one of the important targets of the immune systems of HCV controllers [18]. Thus, the major focus of this section is to identify the groups of co-evolving sites in NS3. The amino acid MSA matrix of NS3 HCV subtype 1a (that predominates in North America) was downloaded from the Los Alamos database (`http://hcv.lanl.gov`) during May 2013. The data consisted of $M = 2815$ sequences and $N = 631$ sites.

The eigenvalue distribution of $\mathbf{C}$ is plotted in Fig. 1 (Upper). The corresponding eigenvalue distribution of 1000 randomized alignments is also shown in Fig. 1 (Lower). Following the procedure discussed in Section 2, three distinct sectors comprising co-evolving sites are determined in the NS3 protein. Fig. 2 shows a 3-D scatter plot of the elements of the eigenvectors in which the three sectors are clearly visible. Note that the qualitative results remained the same if statistical noise was cleaned using alternate RMT-based methods (mentioned in Section 2.2) [7]. Moreover, the sectors formed by using the alternate methods (mentioned in Section 2.3) were also found to be approximately the same [7].

3.1. Statistical analysis of the NS3 sectors

Fig. 3a presents the statistical analysis of the three sectors of NS3. We see that Sector 1 is comprised of around 10% sites that are negatively correlated, and 5% and 3% for Sector 2 and Sector 3 respectively. Further taking into account the percentage of positive correlations and the ratio of negative to positive correlations, it is evident that Sector 1 consists of a large proportion of sites that do not mutate simultaneously. Note that these statistics of the sectors remained the same qualitatively even if the thresholds in (6) were changed [7].

3.2. Experimental and structural validation

In [18], it was determined experimentally that the immune systems of most of the HCV controllers recognize three specific parts of NS3. Figure 3b shows the percentage of sites in these parts that fall within the three NS3 sectors. This percentage is calculated as $N_{\rm ps}/N_{\rm p} \times 100$, where $N_{\rm ps}$ is the number of sites in the parts (targeted by HCV controllers) that are present in the sector and $N_{\rm p}$ is the total number of sites in these parts. It can be observed that 39% of these parts is present in Sector 1 while 7% is in Sector 3. This result for Sector 1 was found to be statistically significant [7]. Moreover, using the covariance matrix without cleaning the statistical noise (i.e., using all the eigenvalues except the phylogenetic one) does not yield any meaningful results. This further exemplifies the importance of RMT approaches in obtaining robust covariance statistics.

In addition, the sites present in Sector 1 are found to have structural significance. Specifically, the sites in the protein interface region, formed between the two NS3 molecules (dimer), were reported to be important for the unpacking of viral genes and viral replication [19]. A large proportion of the sites in the interface of this dimer

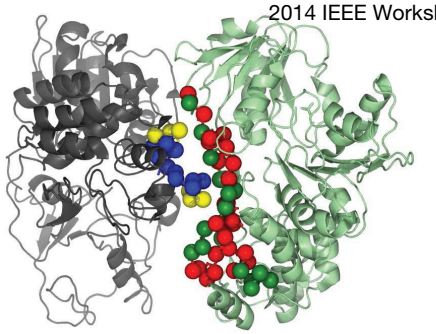


Fig. 4: The sites at the interface between the two NS3 helicase chains A (pale green) and B (grey) in the dimer structure (PDB code 2F55) are shown as dark green and blue spheres respectively. The interface sites of chain A and B present in Sector 1 are shown as red and yellow spheres respectively.

structure are present in Sector 1 (20 out of 46 ($\sim$44%)). Fig. 4 shows the Sector 1 sites present at the interface region of the NS3 dimeric structure.

4. CONCLUSION

In this paper, a general framework was presented for identifying groups of co-evolving sites in a viral protein by studying the correlations of the substitutions at sites using ideas from RMT. The application of this procedure on the NS3 protein of HCV led to the discovery of three groups of co-evolving sites. One of these groups, Sector 1, was found to comprise a large percentage of sites for which HCV resists simultaneous substitutions most likely due to their deleterious effect on viral fitness, and therefore serves as a useful target for vaccine design [7]. The significance of Sector 1 was corroborated by validation against clinically-reported experiments and structural data. Moreover, this study signifies the importance of statistical signal processing tools (in particular RMT and robust covariance estimation) for providing a better understanding of harmful viruses. This can potentially help to design effective vaccines against catastrophic diseases.

5. ACKNOWLEDGMENTS

This work was supported by the Hong Kong PhD Fellowship Scheme, Research Grants Council, Hong Kong (A.A.Q.), HKUST research grant IGN13EG02 (R.H.Y.L.), the Ragon Institute of MGH, MIT, & Harvard (K.S. and A.K.C.), and the Hari Harilela endowment fund, R8002 (M.R.M.).

6. REFERENCES

[1] D. W. Mount, *Sequence and Genome Analysis*. Bioinformatics: Cold Spring Harbour Laboratory Press: Cold Spring Harbour 2, 2004.

[2] A. Motahari, G. Bresler, and D. Tse, "Information theory of DNA shotgun sequencing," *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6273–6289, Sep. 2013.

[3] L. Weiguo, B. Schmidt, G. Voss, and W. Muller-Wittig, "Streaming algorithms for biological sequence alignment on GPUs," *IEEE Trans. Parallel Distrib. Syst.*, vol. 18, no. 9, pp. 1270–1281, Sep. 2007.

[4] Y. Lai, "The analysis of ordered changes of gene expression and gene-gene co-expression patterns," *IEEE Int. Conf. Comp. Adv. in Bio and Medical Sciences*, pp. 117–122, Feb. 2011.

[5] R. W. Harrison and I. T. Weber, "HIV drug resistance prediction using multiple regression," *IEEE Int. Conf. Comp. Adv. in Bio and Medical Sciences*, pp. 1–2, Jun. 2013.

[6] V. Dahirel et.al., "Coordinate linkage of HIV evolution reveals regions of immunological vulnerability," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 28, pp. 11 530–11 535, Jul. 2011.

[7] A. A. Quadeer, R. H. Y. Louie, K. Shekhar, A. K. Chakraborty, I. Hsing, and M. R. McKay, "Statistical Linkage Analysis of Substitutions in Patient-Derived Sequences of Genotype 1a Hepatitis C Virus Non-Structural Protein 3 Exposes Targets for Immunogen Design," *accepted in Journal of Virology*, 2014.

[8] R. Couillet and M. Debbah, *Random Matrix Methods for Wireless Communications*. Cambridge: Cambridge University Press, 2011.

[9] J.-P. Bouchaud and M. Potters, "Financial applications of random matrix theory: A short review," in *The Oxford Handbook of Random Matrix Theory*, G. Akemann, J. Baik, and P. D. Francesco, Eds. Oxford University Press, 2011.

[10] L. Laloux, P. Cizeau, M. Potters, and J.-P. Bouchaud, "Random matrix theory and financial correlations," *International Journal of Theoretical and Applied Finance*, vol. 3, pp. 391–397, Jul. 2000.

[11] N. Halabi, O. Rivoire, S. Leibler, and R. Ranganathan, "Protein sectors: Evolutionary units of three-dimensional structure," *Cell*, vol. 138, no. 4, pp. 774–86, Aug. 2009.

[12] V. A. Marčenko and L. A. Pastur, "Distribution of eigenvalues for some sets of random matrices," *Mathematics of the USSR-Sbornik*, vol. 1, no. 4, pp. 457–483, 1967.

[13] R. Nadakuditi and A. Edelman, "Sample eigenvalue based detection of high-dimensional signals in white noise using relatively few samples," *IEEE Trans. Signal Process.*, vol. 56, no. 7, pp. 2625–2638, Jul. 2008.

[14] R. Ranganathan and O. Rivoire, "Note 109: A summary of SCA calculations," 2012. [Online]. Available: http://systems.swmed.edu/rr_lab/sca.html

[15] "National Institutes of Health Consensus Development Conference Statement: Management of hepatitis C: 2002–June 10-12, 2002," *Hepatology*, vol. 36, no. 5 Suppl. 1, pp. S3–S20, Nov. 2002. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12407572

[16] J. Halliday, P. Klenerman, E. Barnes, T. G. Unit, and J. R. Hospital, "Vaccination for hepatitis C virus: Closing in on an evasive target," *Expert Review of Vaccines*, vol. 10, no. 5, pp. 659–672, May 2011.

[17] D. Moradpour, F. Penin, and C. M. Rice, "Replication of hepatitis C virus," *Nature Reviews. Microbiology*, vol. 5, no. 6, pp. 453–463, Jul. 2007.

[18] J. Schulze et.al., "Broad repertoire of the CD4+ Th cell response in spontaneously controlled hepatitis C virus infection includes dominant and highly promiscuous epitopes," *The Journal of Immunology*, vol. 175, pp. 3603–3613, Jun. 2005.

[19] S. G. Mackintosh et.al., "Structural and biological identification of residues on the surface of NS3 helicase required for optimal replication of the hepatitis C virus," *The Journal of Biological Chemistry*, vol. 281, no. 6, pp. 3528–3535, Feb. 2006.