Distilling Evidence of Long-Range Direction-Specific Causal Cross-Talk in Molecular Evolution of Retro-Viral Genomes

Ishanu Chattopadhyay and **Hod Lipson**
ic99@cornell.edu
Cornell University, Ithaca NY

Abstract

Rapid molecular evolution in retroviruses potentially pose a hurdle to effective vaccine design. While the coding sequence for viral surface proteins seemingly mutate randomly from point to point, the necessity of conserved function dictates the often suspected existence of hidden correlations and long-range dependencies between non-colocated sequence positions. In this initial report, we present a fundamentally new approach to infer the direction-specific causal dependencies that underlie the sequence changes driving viral evolution. Using no prior knowledge of viral genomes, or expectations of known patterns, we show that our algorithm distills the network of causality flows, identifying key regions of immunological vulnerabilities. Such computationally identified vulnerabilities may open the door to new vaccine designs that highly mutable retroviruses such as HIV fail to evade.

Motivation & Contribution

Design of an effective vaccine for the Human Immunodeficiency Virus has eluded researchers for the better part of last two decades. Sophisticated strategies such as carbohydrate cloaking and shape shifting identification molecules, endows the virus with an unmatched ability to evade the host immune response. Perhaps more important to such active evasive maneuvers is the rapid evolution of the viral genome, brought about by its intrinsically high per-base mutation rate. Evolving roughly 13 million times faster to the human genome, the HIV surface proteins present a constantly moving target for the host adaptive immune defense - which simply cannot keep up. Indeed, the genomic diversity of HIV within a single host is large enough to warrant treatment as a multi-species colony.

However, surface proteins are not redundant; they play a crucial role in viral assembly, and must therefore conserve function. It has long been suspected that hidden patterns and correlations are buried in the seemingly random alterations of the genomic sequences, and non-colocated mutations might have incipient statistical dependencies. Reported techniques that attempt to determine this hidden structure have investigated simple correlations in mutational frequencies. While correlation analyses have had some success, no notion of directional causality is obtainable via such symmetric approaches. In this preliminary report, we present a fundamentally new approach designed to go beyond simple correlations, and infer direction-specific causal dependencies in mutational dynamics at locations separated by possibly hundreds of bases in the coding sequences. For a specific surface protein, we show that our analyses reveals key vulnerabilities, which may be potentially exploited for vaccine design.

Relevant Work: Protein Sectors

Proteins display a hierarchy of structural features at primary, secondary, and tertiary levels; an organization that guides our current understanding of their functional properties. In (Halabi et al. 2009), the authors used statistical analysis of correlated evolution between amino acids to reveal a structural organization distinct from this traditional hierarchy. The analysis, applied to S1A serine proteases, indicated a decomposition into three quasi-independent groups of correlated amino acids, termed **"protein sectors."** Each sector is physically connected in the tertiary structure, has a distinct functional role, and constitutes an independent mode of sequence divergence in the protein family. Functionally relevant sectors are evident in other protein families as well, suggesting that they may be general features of proteins. The authors in (Halabi et al. 2009) proposed that sectors represent a structural organization of proteins that reflects their evolutionary histories.

From Correlated Evolution To Structure & Function

A standard measure of importance of protein residues is sequence conservation - the degree to which the frequency of amino acids at a given position deviates from random expectation in a well-sampled multiple sequence alignment of the protein family (Capra and Singh 2007; Ng and Henikoff 2006). The more unexpected the amino acid distribution at a position, the stronger the inference of evolutionary constraint and therefore of biological importance.

However, protein structure and function also depend on the cooperative action of amino acids, indicating that amino acid distributions at positions cannot be taken as independent of one another (Lockless and Ranganathan 1999). Indeed, analyses of correlations have contributed to the identi-

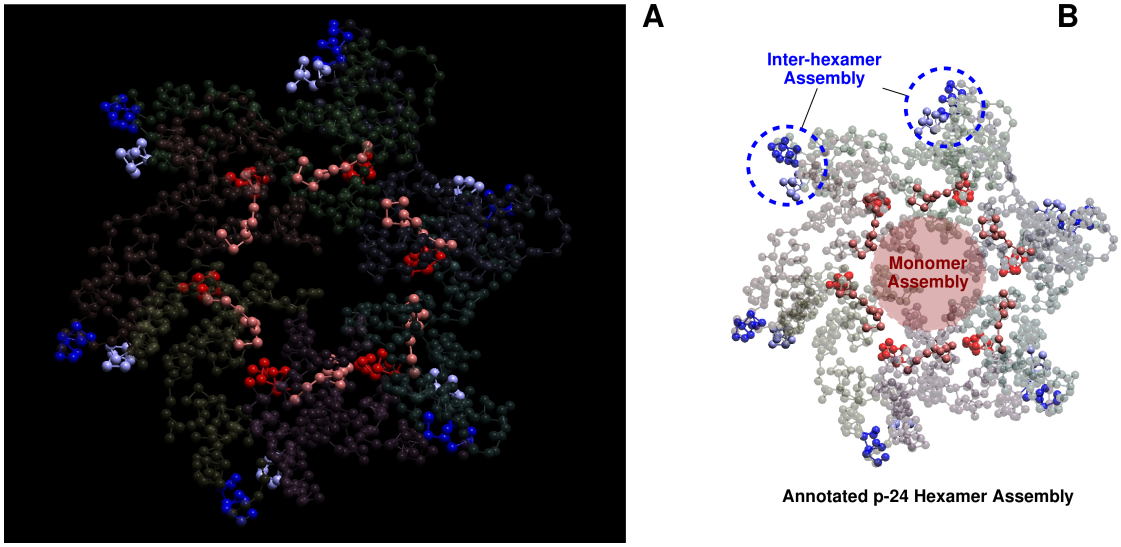

Figure 1: p24 hexamer (3GV2 PDB Database), with neighborhoods of the loci identified above highlighted. Note that the loci $45, 57$ correspond to the regions involved in the assembly of the six p24 monomers to form the hexamer. Theses hexamers then need to assemble to generate the capsid, and this assembly requires molecular interactions around the $201, 217$ region.

fication of allosteric mechanisms in proteins (Ferguson et al. 2007), and even new vaccine design strategies against HIV (Dahirel et al. 2011).

Far greater insight into protein structure and function may be obtained if we can infer the causal directions of positional correlations. Put simply, such inferred directed causality flow will potentially reveal the fundamental constraints, feedbacks and emergent high-level dynamical relationships on evolutionary change through time. In the context of rapidly evolving viral proteins, analysis of such an inferred causality network can reveal regions of immunological vulnerabilities, and will provide immediate insights into vaccine design strategies.

In this initial report, we delineate a theoretically sound solution to this inference problem. At a conceptual level, we implement Granger's notion of causality (Granger 1969):

> Entity X causally drives the evolution of entity Y, if we can predict the dynamics of Y better from a knowledge of X than the other way around.

Granger's notion, however, is more intuitive than constructive. Additionally, the dynamics of the position-specific variations of the amino-acid residues are not known a priori, and hence any assumption on the statistics of such dynamics ($e.g.$ Gaussian, independence or otherwise) could be biasing our conclusions; and perhaps more importantly, make it impossible to correctly capture the emergent statistical patterns.

Inferring Causality Networks

To find a non-parametric implementation of Granger's notion, which is independent of a priori knowledge of the dynamics of the entities involved, we need a more refined approach to the one pursued in (Halabi et al. 2009); namely computing the simple correlation matrix, followed by a cluster analysis would not suffice. Our approach consists of the following conceptual building blocks:

- *Representation of position-specific variational data with symbol streams:* Given a sufficiently large set of sequences representing variations of a protein of interest, we generate position-specific symbol sequences that capture the residue variation. As a simple example (See Fig. 2A), a neighborhood length is chosen, and the sequence fragments from within this neighborhood (of the specific position) are concatenated. We replace the consensus (or the most frequent) residue at each position by a 0, and a variation or mutation by a 1; resulting in a binary sequence s_i for each position i.
- *Inference of probabilistic transducers:* We use a generalized version of the algorithm reported in (Chattopadhyay and Lipson 2013) to infer finite state probabilistic transducers G_{ij} for each pair of binary streams s_i, s_j. These probabilistic transducers are directional, $i.e.$, $G_{ij} \neq G_{ji}$. Intuitively, G_{ij} represents the "transfer function" from $i \rightarrow j$, $i.e.$, given the sequence s_i, there exists a operation $\otimes$, such that:
$$G_{ij} \otimes s_i = s'_j \tag{1}$$
where $s'_j \simeq s_j$ in a well-defined statistical sense. Thus, the degree of causal connection in the direction $i \rightarrow j$ may be assumed to be inversely related to: $\Theta(s_j, G_{ij} \otimes s_i)$, where $\Theta(\cdot, \cdot)$ is a distance function defined on the space of admissible sequences. Thus, smaller the distance between s_j and $G_{ij} \otimes s_i$, better s_i is in predicting s_j via the transfer function G_{ij}. Thus, dynamics at position i "drives" that in position j, if and only if we have the inequality:
$$\Theta(s_j, G_{ij} \otimes s_i) \lneqq \Theta(s_i, G_{ji} \otimes s_j) \tag{2}$$
- *Quantification of causal similarity between sequences:* To

complete the above formalization, we need an appropriate distance metric $\Theta(\cdot,\cdot)$ on sequence space; particularly one that captures a well-defined distance between the hidden generating processes from the observed sequences alone. Essentially we need to be able to estimate a distance between two stochastic processes from finite sample paths. Fortunately, a notion of such causal similarity has been reported in (Chattopadhyay and Lipson 2014), where the only assumptions on the processes are that of ergodicity and stationarity.

With these key notions in place, we can infer the causality network given a sufficiently large set of aligned variations of the candidate protein. Eliminating weaker connections via a chosen cut-off value, we get a clear picture of how residue variations propagate across the primary structure. Importantly, the causality network captures the directional constraints arising from high-level protein function; and hence includes but are not limited to those arising from secondary and tertiary structure formation, and even constraints arising from required interaction with functional molecules external to the residue sequence under consideration.

Mechanistic Insights From Inferred Causality Flow

The inferred causality network is significantly more informative compared to the symmetric correlation matrix obtained in (Halabi et al. 2009) and (Dahirel et al. 2011). We, for example, can immediately compute the global effect of specific perturbations: how does the effect of the variation or mutation at a position propagate throughout the structure; which positions incur necessary variations as a result, and which one must necessarily not change to preserve high-level function.

Additionally, the post-analysis now admits standard tools from network theory. In the example (See next section) we note how immunologically vulnerable positions on the p24 capsid protein for HIV-1A is identifiable via the simple task of finding nodes with maximal degrees; this is computationally far easier, concrete, and parameter-free compared to finding approximate clusters.

Analyzing The HIV Capsid Protein p24

To illustrate the applicability of the inferred causality flow in inferring functional constraints, we analyze the HIV capsid protein p24. The genome of HIV, which is composed of two strands of RNA, is packaged inside a distinctive cone-shaped capsid, which protects the RNA and delivers it to target cells. The capsid is built from a single protein, known as CA or p24, which folds to form two domains connected by a flexible linker. The larger domain associates with 6 other copies of the protein to form a hexamer. The smaller domain then links these rings together to form the larger structure.

Such a complex assembly process places functional constraints on mutational variations of the primary sequence of p24. However, the constraints are subtle enough to not manifest simply as conserved regions, but do so in the form of statistical signatures of correlated mutations across the primary structure. Host cellular immune control of HIV is mediated, in part, by induction of single amino acid mutations that attempt to reduce viral fitness, but compensatory mutations limit this effect. The authors in (Dahirel et al. 2011) used the protein sector analysis described above to conclude that higher order constraints on viral evolution do indeed exist, and that some coordinately linked combinations of mutations may severely hurt viability.

Network Analysis Results

The sector-based analysis in (Dahirel et al. 2011) identified two key regions of interest from the viewpoint of immunological vulnerability: the region that interact to form the p24-hexamer using monomeric building blocks, and the one that is involved in the inter-hexamer binding to assemble the capsid. However, no justification is given as to why this particular sector houses the vulnerable regions; the sector-based analysis in itself only finds approximate clusters and cannot provide such information. Some additional spectral analysis was employed to reach the conclusion, which still does not elucidate why this high-level constraint shows up in this particular sector.

We analyzed the p24 variants as described in the previous section. Very rough alignment suffices, as our inferred probabilistic transducers accommodate for small relative shifts automatically. The inferred network is shown in Fig. 2B.

Note the nodes with maximal in-degree has a unambiguous interpretation: these are the positions, which along with their small neighborhoods, are maximally constrained by the rest of the structure. Looking back to the assembly (See Fig. 1), we find that these regions indeed correspond to the two binding regions described above: the monomeric binding regions to yield the hexamer, and the inter-hexamer binding regions to assemble the capsid. In contrast to (Dahirel et al. 2011), the the network analysis provides a clear explanation of the observed immunological vulnerabilities at these regions.

In addition, the information-rich network may be used to derive other functional insights, and may even provide a way to simulate the effects of specific mutations, and delineate additional vulnerabilities. Targeting such regions with higher order evolutionary constraints provides a novel approach to immunogen design for a vaccine against HIV and other rapidly mutating viruses.

References

Capra, J. A., and Singh, M. 2007. Predicting functionally important residues from sequence conservation. *Bioinformatics* 23(15):1875–1882.

Chattopadhyay, I., and Lipson, H. 2013. Abductive learning of quantized stochastic processes with probabilistic finite automata. *Philos Trans A* 371(1984):20110543.

Chattopadhyay, I., and Lipson, H. 2014. Data Smashing. *ArXiv 1401.0742*.

Dahirel, V.; Shekhar, K.; Pereyra, F.; Miura, T.; Artyomov, M.; Talsania, S.; Allen, T. M.; Altfeld, M.; Carrington, M.; Irvine, D. J.; Walker, B. D.; and Chakraborty, A. K. 2011. Coordinate linkage of HIV evolution reveals regions of immunological vulnerability. *Proc. Natl. Acad. Sci. U.S.A.* 108(28):11530–11535.

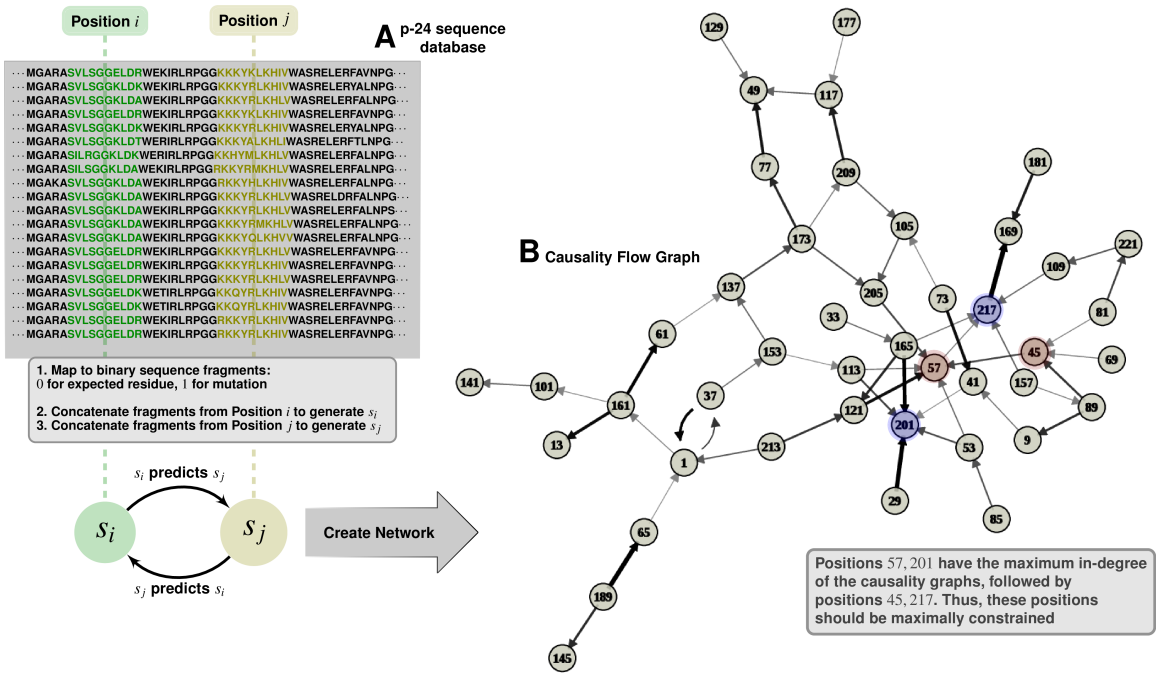

Figure 2: Variants of the primary structure (amino-acid sequences) for the p24 capsid protein in HIV-1A (Source: LANL HIV Database) are used to generate the causality graph shown in plate B. The sequences are analyzed for each position pair. Using a neighborhood size of 4 residues, we generate sequence fragments around the positions, map them to binary sequences (where 1 represents a mutation from the average residue at that position, and 0 otherwise), and finally procure two binary streams for each position pair. We then ask if one can predict the other, which induces a directional causality graph as shown. Note that the indicated positions $(201, 57)$ have the largest in-degree, signifying the most-constrained loci. These loci have a clear structural interpretation, as shown in Figure1.

Ferguson, A. D.; Amezcua, C. A.; Halabi, N. M.; Chelliah, Y.; Rosen, M. K.; Ranganathan, R.; and Deisenhofer, J. 2007. Signal transduction pathway of TonB-dependent transporters. *Proc. Natl. Acad. Sci. U.S.A.* 104(2):513–518.

Granger, C. W. J. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* 37(3):424–438.

Halabi, N.; Rivoire, O.; Leibler, S.; and Ranganathan, R. 2009. Protein sectors: evolutionary units of three-dimensional structure. *Cell* 138(4):774–786.

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

Ng, P. C., and Henikoff, S. 2006. Predicting the effects of amino acid substitutions on protein function. *Annu Rev Genomics Hum Genet* 7:61–80.