Predicting evolutionary site variability from structure in viral proteins: buriedness, flexibility, and design

Amir Shahmoradi, [1,2] Daria K. Sydykova,[2] Stephanie J. Spielman,[2] Eleisha L. Jackson,[2] Eric T. Dawson,[2] Austin G. Meyer,[2] Claus O. Wilke[2]

[1] Department of Physics, The University of Texas at Austin, TX, 78712
[2] Institute for Cellular and Molecular Biology, The University of Texas at Austin, TX, 78712

Abstract

Several recent works have shown that site-specific evolutionary variation in proteins can be predicted from protein structure. Most prominently, sites that are buried and/or have many contacts with other sites in a structure evolve more slowly than surface sites with few contacts. Here, we present a comprehensive study of numerous different structural properties that may constrain sequence variation, including measures of buriedness (relative solvent accessibility, contact number), measures of structural fluctuations (B factors, root-mean-square fluctuations, variation in dihedral angles), and variability in designed structures. Structural fluctuation measures were obtained from molecular dynamics simulations performed on 9 non-homologous viral protein structures, and from variation in homologous variants of these proteins where available. Variability in designed structures was obtained from flexible-backbone design via Rosetta. We found that most of the structural properties correlate with site variation in the majority of structures, though the correlations are generally weak (correlation coefficients of 0.1 to 0.4). We further found that measures of buriedness tend to be better predictors of evolutionary variation than measures of structural fluctuations. Variability in designed structures was a weaker predictor of evolutionary variability than both measures of buriedness and measures of fluctuations. We conclude that simple measures of buriedness are better predictors of evolutionary variation than more complicated predictors obtained from dynamic simulations, ensembles of homologous structures, or computational protein design.

Introduction

Identification of the driving factors in protein evolution has been on of the important objectives in molecular biology and protein research (cite xx). It is already well-established and understood that highly conserved amino acid sites in the protein sequences often fall in hotspot regions responsible for the protein's biophysical function (cite xx) or happen to be pivotal in maintaining the protein's native conformation. Aside from biophysical constraints,

several structural determinants of protein dynamics and flexibility have been recently proposed to impose site-specific evolutionary pressure on the protein sequence. Examples include residue-level solvent exposure (cite xx), local protein density (cite xx), and measures of residue-level flexibility of the protein backbone (cite xx).

Among correlating variables, the Relative Solvent Accessibility (hereafter RSA) has gained special attention for its ability in predicting the general patterns of residue-level sequence variability and evolution in globular proteins. This variable is defined as a residue's site-specific Accessible Surface Area (ASA) to solvent molecules, normalized by the theoretically or experimentally determined maximum accessible area for the same residue (Tien, Rose). RSA was first introduced in the context of hydrophobicity scales derived by computational means from protein crystal structures (cite xx) and its association with sequence variability may be explained in terms of the residue hydrophobicity which correlates strongly with RSA (Cite xx). The core of globular proteins is generally thought as a region of near-zero solvent accessibility that is mainly occupied by tightly-packed hydrophobic amino acid side chains. It can be therefore expected that any mutations of these hydrophobic residues in the protein core to hydrophilic or bulkier side chains (cite xx) may result in significant changes the native conformation of the protein which might in turn adversely affect the biophysical functioning of the protein and hence exert selection pressure against such mutations. In other words, the positive association of RSA with sequence variability is not a causal relation but merely a reflection of the effects of geometrical (cite Mike xx) or chemical (cite xx) constraints on sequence variability.

Along with RSA, other measures of residue buriedness, such as residue contact number (cite xx) have been proposed and shown to correlate with sequence variability, or even argued to serve as better predictors than RSA (cite xx). Based on physical arguments and experimental evidence (xx what kind of arguments?), **?** argued that the local residue packing density is a direct proxy measure of residue and site-specific backbone flexibility, in particular the Debye-Waller factors, commonly known as B factors. Therefore, given the strong observational evidence for the significant positive correlation of residue density and packing with sequence evolution (**?**), one may also expect to observe a positive trend between local flexibility and sequence variability. Indeed, several authors have argued for the potential role of protein dynamics on sequence variability (e.g., Bahar et all. Cite etc XX).

Although multitude of structural variables have been shown or predicted to influence residue-level sequence variability, there is currently no consensus on which variable and to what extent has the dominant role in regulating sequence variability, independently of other structural determinants (see however, a recent work by Echave et al 2014, in press). So far, a comprehensive study of all potential structural determinants of protein sequence evolution has been missing in the literature, with the existing work mainly focusing on individual variables. In particular, measures of residue spatial fluctuations and protein dynamics have only received marginal attention and consideration as potential contributing factors to sequence evolution.

Proteins are intrinsically dynamic entities in vivo, far from their perceived rigidities in crystal structures and their dynamic behavior is expected to influence their sequence evolution (cite Marsh). However, contrary to RSA and residue contact number, an accurate determination of the protein's dynamical behavior and residue fluctuations—solely based

on the set of 3-dimensional atomic coordinates in crystal structures—remains a challenging task. B factors are generally considered as an attractive proxy to local flexibility, though the atomistic definition of B factor may not be appropriate for the study of side-chain flexibility and fluctuations. Experimental studies of protein dynamics in vivo has also proven extremely difficult if not impossible (cite xx).

Alternatively, Molecular Dynamics (hereafter, MD) simulations provide an ideally suited method of studying protein dynamics and its potential role in driving sequence evolution. Here in this work, we attempt to present a comprehensive study of several potential structural determinants of sequence variability from both protein crystallography and Molecular Dynamics perspectives. In addition to the factors already discussed in previous works, such as RSA and residue contact number, we also consider new dynamical measures of structural variability in the study, such as variance of the backbone and residue dihedral angles and residue spatial fluctuations from MD simulations and discuss their potential influence on sequence variability. The extent and breadth of such analysis however, limits diversity of the input data used in our work to highly evolving proteins that have multiple high-resolution homologous crystal structures in Protein Data Bank (cite xx). The availability of multiple structures is required for the calculation of site-specific spatial fluctuations and comparison of the results with the same variability measures based on MD simulations. In addition, the selected proteins should also have ample sequence data to ensure good statistic for sequence alignment and the calculation of sequence variability and evolutionary rates.

In the following sections, first we briefly present the methodology employed for data selection, sequence alignments, data analysis and the procedure for obtaining the relevant structural variables from Molecular Dynamics simulations and homologous structures, followed by the results and a discussion of the potential contributors to sequence variability or evolution. We show that measures of residue buriedness such as RSA and contact number outperform site-specific measures of residue fluctuations and discuss the potential underlying biases and reasons contributing to this observation.

Materials and Methods

Protein Family Selection and Sequence Alignment

All protein crystal structures in this work were taken from Data Bank (PDB) based on the availability of sequence data and homologous PDB structures. We focused our attention on viral proteins as they often have extensive sequence data available for alignment and for the calculation of evolutionary rates and sequence variability. Towards this, a total of 8 viral protein families were selected for analysis as tabulated in Table **??**.

All alignments were constructed using amino-acid sequences with MAFFT (**??**), specifying the "–auto" flag to select the optimal algorithm for the given data set, and then back-translated to a codon alignment using the original nucleotide sequence data. Evolutionary rates were calculated as described (**?**). In brief, we generated a phylogeny for each codon alignment in RAxML (**?**) using the GTRGAMMA model. Using the codon alignment and phylogeny, we inferred evolutionary rates with a Random Effects Likelihood (REL) model, using the HyPhy software (**?**). The REL model was a variant of the GY94

evolutionary model (**?**) with five *omega* rate categories as free parameters. We employed an Empirical Bayes approach (**?**) to infer dN/dS values for each position in the alignment. The amino-acid sequence of the seed protein structures were then mapped to the corresponding alignments for subsequent analyses.

As an alternative measure of site-specific sequence variability, we also calculated the Shannon entropy (H) defined as

$$H_i = -\sum_j P_{ij} \ln P_{ij} \tag{1}$$

for the i^{th} amino-acid site in each of the protein structures. Here P_{ij} is relative frequency of amino acid j at position i in the alignment.

Homologous Crystal Structures

Regarding Daria's work.

Molecular Dynamics Simulations

The computational expense of Molecular Dynamics (MD) simulations limited our analysis to only one representative PDB structure from each of the 8 protein families tabulated in Table **??**. However, to ensure the dynamic properties of homologous PDB structures do not differ significantly from each other, MD simulations were performed on two additional homologous PDB structures taken from Hepatitis C Protease. All simulations were performed on Lonestar's Dell Linux Cluster at Texas Advanced Computing Center (TACC) using the GPU implementation of *Amber12* Molecular Dynamics simulation package (cite xx) with the most recent release of Amber fixed-charge force field (ff12SB; c.f., AmberTools13 Manual).

Prior to MD production runs, all PDB structures were first energy minimized using the steepest descent method for 1000 steps followed by conjugate gradient for another 1000 steps. Then, the structures were constantly heated from $0K$ to $300K$ for $0.1ns$, followed by $0.1ns$ constant pressure simulations with positional harmonic restraints on all atoms to avoid instabilities during the equilibration process. The systems were then equilibrated for another $5ns$ without positional restraints, each followed by $15ns$ of production simulations in for subsequent post-processing and analyses. All equilibration and production simulations were run using SHAKE algorithm (cite xx) by which all bonds involving hydrogen are constrained to avoid instabilities for a choice of $2fs$ time step in simulations. Langevin dynamics were used for temperature control (cite for Langevin dynamics xx).

Structural flexibility and buriedness measures

Almost all structural determinants of sequence variability can be classified into either of the following two major categories:

1. **measures of local fluctuation/flexibility.** Among the most popular measures of local flexibility is the Root-Mean Square Fluctuation (RMSF) of the protein backbone with respects to a reference structure, often chosen to be the $3D$ coordinates in the seed

PDB file among homologous structures. In case of MD simulations, the reference can also be represented by the average structure over the entire MD trajectories, though it may not necessarily correspond to a realistic conformation of the protein. Based on MD trajectories, we calculated RMSF for the backbone C_α atoms of all protein structures with respect to the corresponding coordinates of the PDB crystal structures listed in Table **??**. In addition, the same RMSF measurements were carried out based on the structural alignments for four protein families that had adequate number of homologous structures in Protein Data Bank (how many specifically?? i.o.w. why did Daria end up with only these four? xx).

The variable RMSF, by its definition, can be potentially a biased measure of site-specific flexibility and fluctuations due to its dependence on the goodness of the structural fitting and alignments prior to RMSF calculations. We therefore considered also the variations in the backbone and residue dihedral angles (ϕ, ψ & χ_2) as independent measures of *local* residue and backbone flexibility which are less prone to systematic biases. The variances of ϕ & ψ can serve as approximate measures of the site-specific flexibility of the backbone atoms, while the variance of χ angle mainly represents the side-chain flexibility and local fluctuations. The χ_1 angle, however, is undefined for the two amino acids Alanine and Glycin and the corresponding sites in the proteins are excluded for the study of χ_1 and sequence variability relations.

2. **measures of residue buriedness and local density.** The site-specific Relative Solvent Accessibility (RSA) is likely the most prominent variable indicating the degree of amino acid exposure to solvent molecules. We used DSSP software (**?**) to calculate the Accessible Surface Area (ASA) for all protein crystal structures in this work, also for the ensemble of coordinates obtained from MD simulations. The ASA values were then normalized to the theoretical maximum ASA values of **?** to obtain the corresponding RSA values. Another indicator of residue solvent exposure, though indirectly, is the residue Contact Number (CN), defined as the total number of C_α atoms surrounding C_α atom of a desired site in a protein, within a spherical neighborhood of a predefined radius r_0. Following Franzosa (cite xx), we assume $r_0 = 13\mathring{A}$ for the calculation of the contact numbers for all structures and MD trajectories.

   Additionally, we also calculate a variant of Contact Number definition that was recently proposed by **?**, defined as the total number of contact C_α atoms at an amino acid site, weighted by the inverse square separation between the neighbor atoms and the site of interest. This Weighted Contact Number (WCN) avoids the problem of manually assigning a neighborhood radius in order to calculate the contact numbers. As it will be shown in the following sections, both definitions of contact numbers are negatively correlated with sequence variability, in contrast to other structural variability measures that often correlate positively. Thus, for the sake of consistency with other structural variables, we use the inverted values of the two contact number definitions and denote them by iCN and iWCN, respectively: iCN = 1/CN, iWCN = 1/WCN. Note that for Spearman correlations, which we use throughout here, replacing a variable by its inverse changes the sign of the correlation coefficient but not the magnitude.

Sequence Entropy from Designed Proteins

Design entropy was calculated as described (**?**). In brief, proteins were designed using RosettaDesign (Version 39284) (**?**) using a flexible backbone approach. This was done for all PDB structures in Table **??** as initial template structures. For each template, we created a backbone ensemble using the Backrub method (**?**). The temperature parameter in Backrub was set to 0.6, allowing for an intermediate amount of flexibility. For each of the 11 template structures we designed 100 proteins. *Is this correct? It shouldn't have been done that way.*The entropy of the designed sequences were subsequenctly calculated by first aligning the designed sequences using MAFFT

Availability of data and methods

All details of simulations, input/output files and scripts for subsequent analyses are available to view or download at `https://github.com/clauswilke/structural_prediction_of_ER`.

Results

Proteins are dynamic, flexible entities, and PDB crystal structures represent only a single snapshot of the infinite number of conformational variations a protein will sample over time. When we derive quantities such as RSA or contact numbers from PDB structures, we don't know *a priori* how representative these quantities are of the protein in solution. Therefore, we first assessed to what extent these quantities differed when we obtained them from crystal structures vs. from MD simulations, in the latter case averaged over 15ns of chemical time. We found that RSA, CN, and WCN from crystal structures were highly correlated with their averages over MD trajectories, for all protein structures we examined (Spearman rank correlation coefficients of > 0.9). Further, when we correlated these quantities with sequence variability (as measured by site-specific entropy H_i), we found that the correlation coefficients were virtually identical (Figure **??**A-C). Thus, in terms of predicting evolutionary variation, RSA and contact numbers obtained from the static structures perform as well as their dynamic equivalents averaged over short time scales.

However, the same was not true when we considered backbone variability as measured by root mean fluctuations (RMSF). RMSF cannot be obtained from a single crystal structure, but we can calculate it from an alignment of multiple crystal structures were available. When we compared RMSF from MD to RMSF from crystal structures, we found that they were generally quite different. In particular, the strength of the correlation between site entropy and RMSF from MD was independent of the strength of the correlation between site entropy and RMSF from crystal structures (Figure **??**D). In fact, for the structure for which RMSF from MD had the highest explanatory power for site entropy (Which structure is this?), the RMSF from crystal structures had the least explanatory power for site entropy (Figure **??**D). The reverse was also true. (Which is the structure for which RMSF from crystal structure works best?)

To further investigate the relationship between backbone fluctuations and sequence variability, we also considered the correlation between sequence entropy and B factors of the

C_α atoms in the protein backbone. We found that these correlations were generally different from the ones found for either the MD RMSF or the crystal structure RMSF (Figure **??**). Thus, B factors, MD RMSF, and crystal RMSF, though all measures of backbone fluctuations, contained distinct information about sequence variability in our data set.

Continue editing from here

Focusing on structural fluctuations first, we compared six different measures of local spatial fluctuations in predicting sequence entropy. The results are illustrated in Figure **??**, where the Spearman's correlation coefficients of entropy with individual structural variables are compared against each other, for all 9 protein families considered in the study. Given our data and results, the *backbone* dihedral angles seem to be the least explanatory variables of sequence entropy, while the variance of χ_1 angle generally exhibits stronger and more significant correlation with entropy. Despite the observed inconsistency between B factor and the two measures of RMSF at the protein family level (Figures **??** & **??**) in the previous section, it is evident that these three measures of local fluctuations explain the same amounts of sequence entropy at about the same significance levels on average.

Next, we compared the three most contributing fluctuation factors to sequence entropy with measures of buriedness and local density (RSA & iWCN). The Spearman correlation strength of the variables with sequence entropy are compared to each other in Figure **??**. A trend can be seen in the explanatory power of the variables, from left to the right of the plot. The variables RSA and the weighted contact number (iWCN) exhibit similar correlation strengths with sequence entropy with average $\rho \sim 0.26$ & $\rho \sim 0.22$ respectively, over all protein structures. I contrast, the three fluctuation measures VAR(χ_1), MD RMSF & B factor exhibit lower levels of associations with sequence entropy corresponding to average $\rho \sim 0.17$, $\rho \sim 0.10$ & $\rho \sim 0.13$ respectively, over all protein structures.

Sequence Entropy from Designed and Natural Proteins

Discussion of the PCA analysis and potential RMSF-designed entropy bias.

Discussion

References

Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. 11:725–736.

Halle B. 2002. Flexibility and packing in proteins. Proc. Natl. Acad. Sci. USA 99:1274–1279.

Jackson E L, Ollikainen N, III A W C, Kortemme T, Wilke C O. 2013. Amino-acid site variability among natural and designed proteins. PeerJ 1:e211.

Katoh K, Kuma K I, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucl. Acids Res. 33:511–518.

Katoh K, Misawa K, Kuma K I, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucl. Acids Res. 30:3059–3066.

Kosakovsky Pond S L, Frost S D W, Muse S V. 2005. HyPhy: hypothesis testing using phylogenetics. Bioinformatics 21:676–679.

Leaver-Fay A, Tyka M, Lewis S M, Lange O F, Thompson J, Jacak R, Kaufman K, Renfrew P D, Smith C A, Sheffler W, Davis I W, Cooper S, Treuille A, Mandell D J, Richter F, Ban Y E A, Fleishman S J, Corn J E, Kim D E, Lyskov S, Berrondo M, Mentzer S, Popovi Z, Havranek J J, Karanicolas J, Das R, Meiler J, Kortemme T, Gray J J, Kuhlman B, Baker D, Bradley P. 2011. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods in enzymology 487:545–574. ISSN 1557-7988. doi:10.1016/B978-0-12-381270-4.00019-6. PMID: 21187238.

Smith C A, Kortemme T. 2008. Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. Journal of molecular biology 380(4):742–756. ISSN 1089-8638. doi:10.1016/j.jmb.2008.05.023. PMID: 18547585 PMCID: PMC2603262.

Spielman S J, Wilke C O. 2013. Membrane environment imposes unique selection pressures on transmembrane domains of G protein-coupled receptors. J. Mol. Evol. 76:172–182.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690.

Yang Z. 2000. Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. J. Mol. Evol. 51:423–432.

Yeh S W, Liu J W, Yu S H, Shih C H, Hwang J K, Echave J. 2014. Site-specific structural constraints on protein sequence evolutionary divergence: Local packing density versus solvent exposure. Mol. Biol. Evol. 31:135–139.

Tables

Table 1: PDB structures considered in this study

Viral Protein	PDB Structure	PDB Chain	Number of Residues	Number of Sequences
Hemagglutinin Precursor	1RD8	AB	503	1039
Dengue Protease Helicase	2JLY	A	451	2362
West Nile Protease	2FP7	B	147	237
Japanese Encephalitis Helicase	2Z83	A	426	145
Hepatitis C Protease	3GOL	A	557	1021
Hepatitis C Protease	3GSZ	A	558	1021
Hepatitis C Protease	3I5K	A	566	1021
Rift Valley Fever Nucleoprotein	3LYF	A	244	95
Crimean Congo Nucleocapsid	4AQF	B	474	69
Marburg RNA Binding Domain	4GHA	A	122	42
Influenza Nucleoprotein	4IRY	A	404	943

Figures

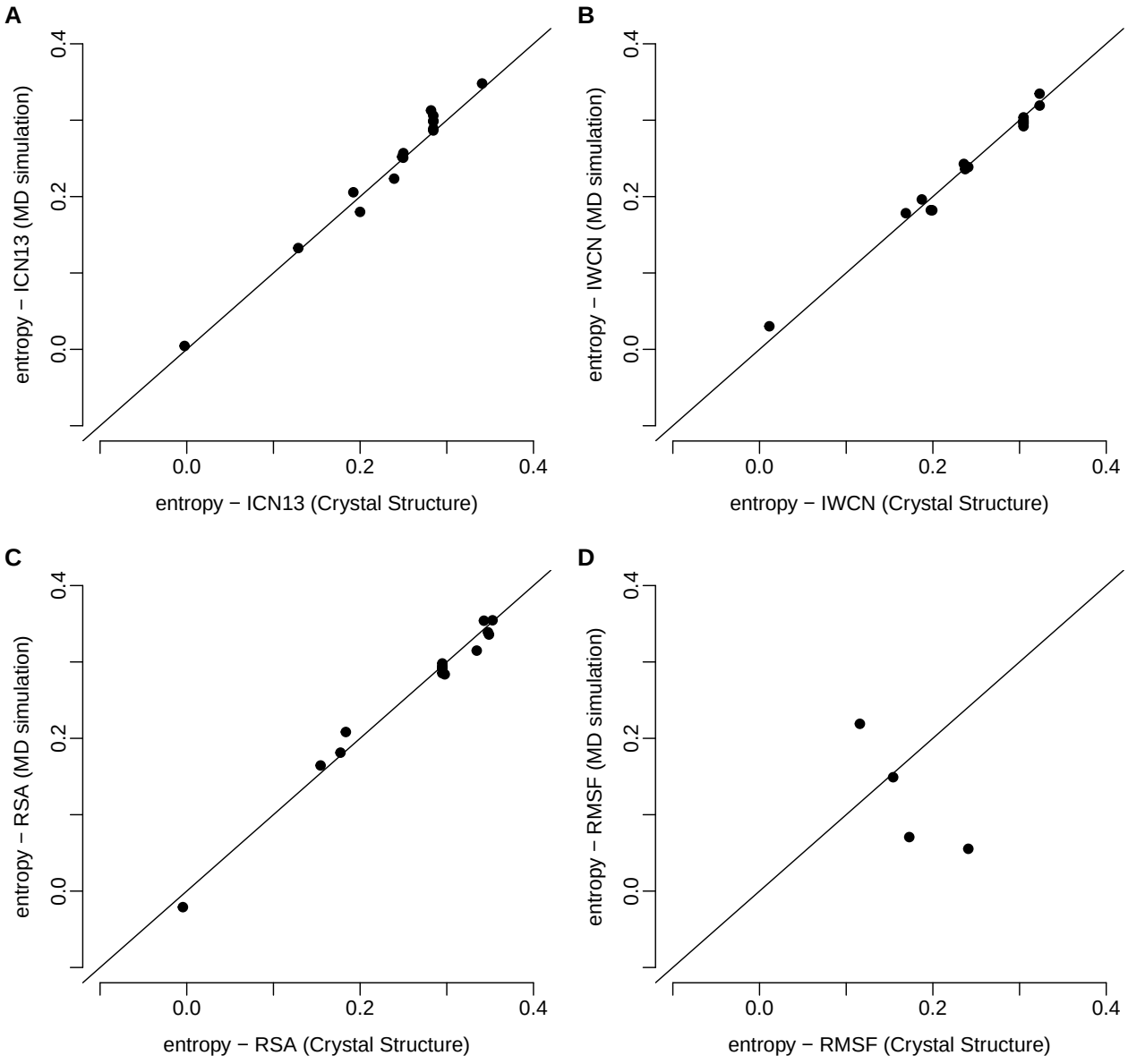

Figure 1: **Molecular dynamics vs. protein crystal structures in predicting sequence entropy.** The vertical axes in all plots represent the Spearman's rank correlation coefficient of sequence entropy with one structural variable obtained from $15ns$ of Molecular Dynamics (MD) simulations. The horizontal axes represent the Spearman's rank correlation coefficient of sequence entropy with the same structural variable as in the vertical axes, but measured from protein crystal structures. Each black dot in the plots represents one protein structure provided in Table **??**.

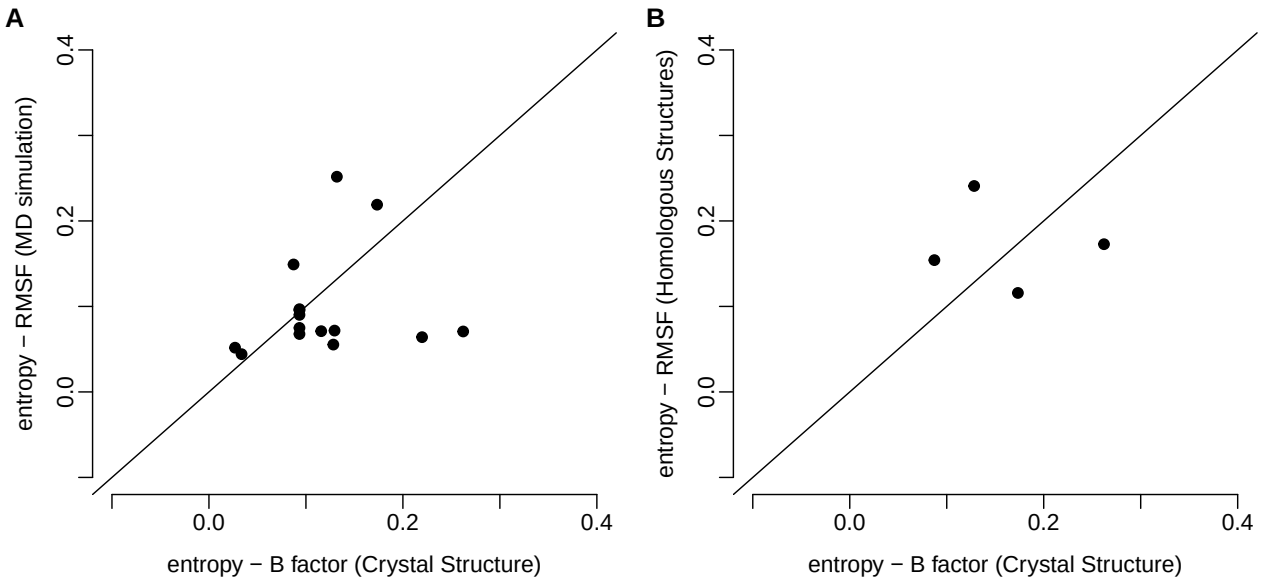

Figure 2: **RMSF vs. B factor in sequence entropy prediction.** **(A)** Comparison of the Spearman's rank correlation coefficients of entropy–RMSF with entropy-B factor. Each black dot represents one protein structure. **(B)** Comparison of the Spearman's rank correlation coefficients of entropy–RMSF with entropy–B factor. Each black dot represents one protein structure for which there were adequate number of homologous structures in the Protein Data Bank, in order to calculate RMSF. As evidenced from the data in the plots, the three different measures of site-specific structural flexibility – MD RMSF, CS RMSF, and B factor – are inconsistent with each other.

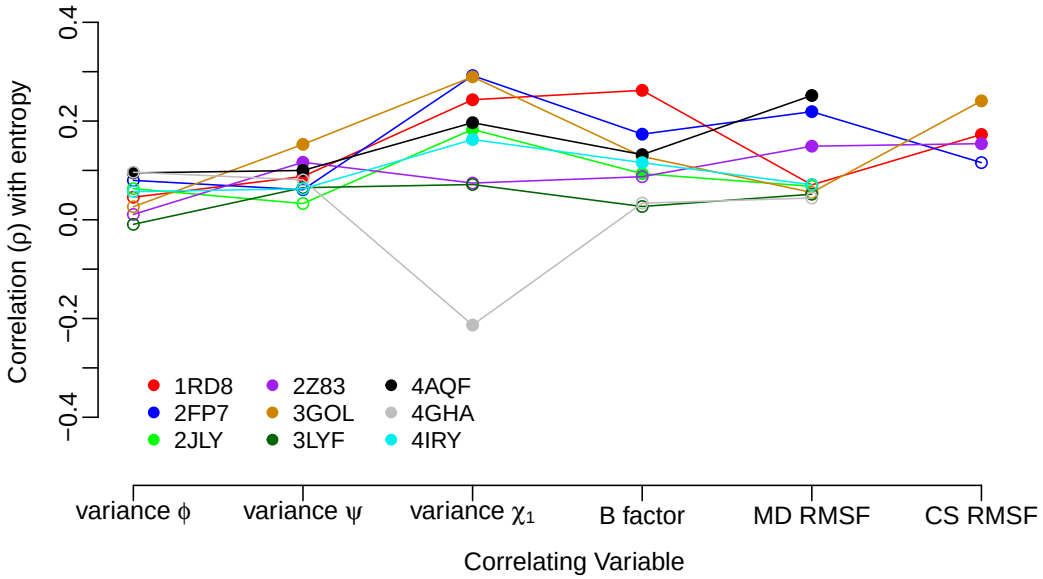

Figure 3: Comparison of the Spearman's rank correlation coefficients of different structural variability measures with sequence entropy. Different colors represent data for different protein PDB structures. The correlation coefficients are represented by solid dots where significant (i.e., p-value ≤ 0.05) or empty dots otherwise (where p-value ≥ 0.05). The variables ϕ, ψ, and χ_1 represent the protein's backbone dihedral angles, measured from MD simulations. The Root-Mean-Square Fluctuations for the proteins' C_α atoms are calculated from both MD simulations (variable MD RMSF) and Crystal Structures (CS RMSF). Almost all structural measures of variability correlate weakly, but significantly with sequence entropy.

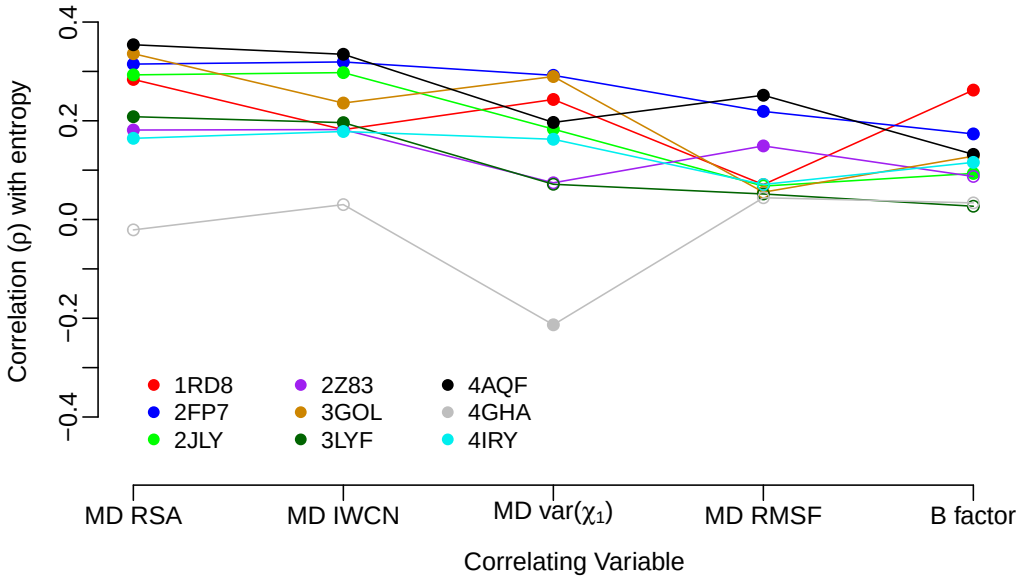

Figure 4: Comparison of the Spearman's rank correlation coefficients of sequence entropy with structural variability and buriedness measures. Different colors represent data for different protein structures. The correlation coefficients are represented by solid dots where significant (i.e., p-value ≤ 0.05) or empty dots otherwise (where p-value ≥ 0.05). The variables: B factor, RMSF and the backbone dihedral angle χ_1 represent different measures of structural variability and fluctuations, while the RSA and IWCN assess the amino acids solvent accessibility and the local packing density respectively. Compared to fluctuation measures, the buriedness variables consistently show stronger correlations with sequence entropy among all viral proteins considered in this analysis.

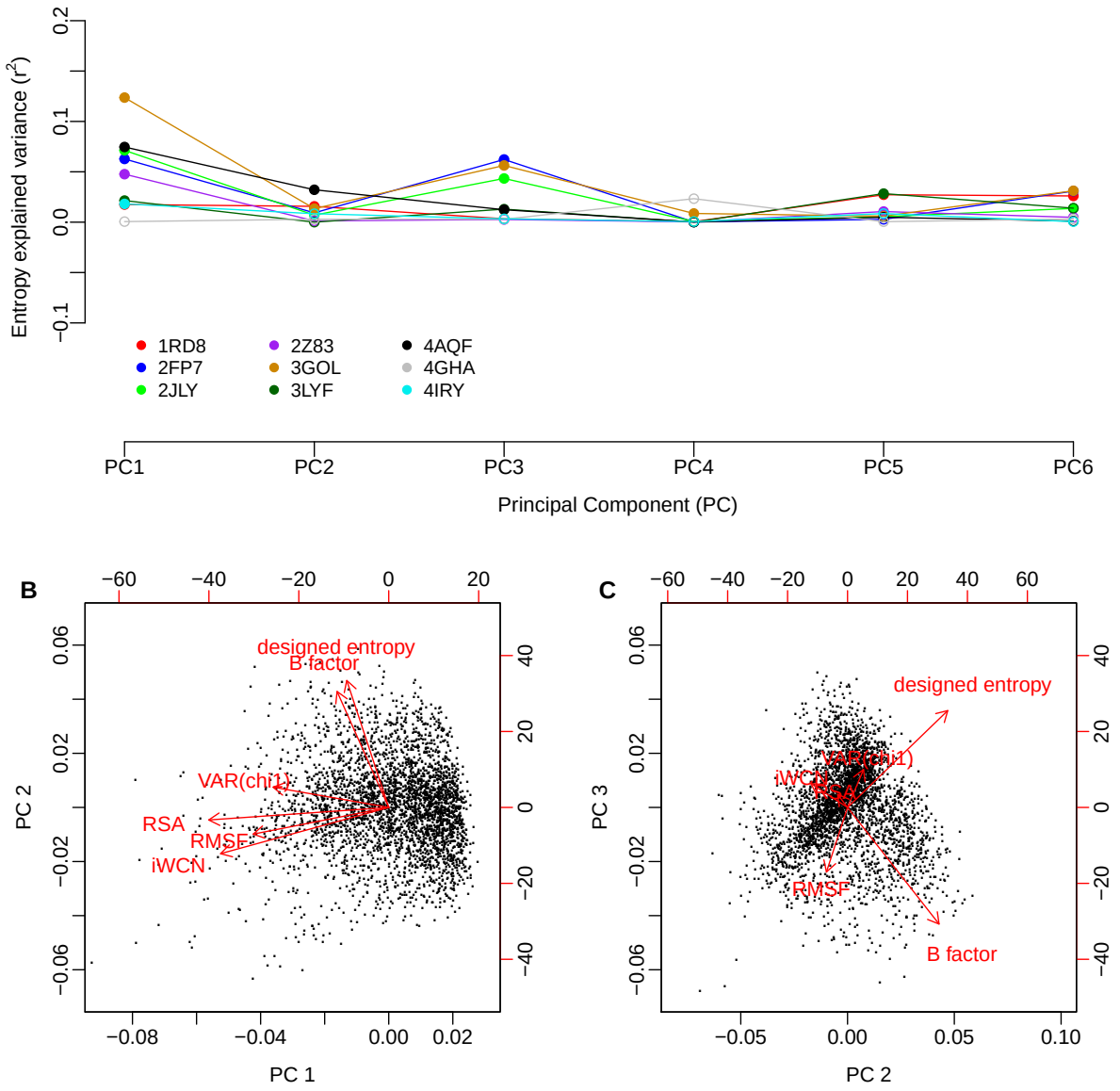

Figure 5: Principal Component (PC) Regression of sequence entropy given the structural variables: **(A)** An illustration of the Spearman's rank correlation coefficients of sequence entropy with different components of PC analysis on 6 structural variability measures: RSA, iWCN, MD RMSF, B factor & sequence entropy from designed proteins. Different colors represent data for different protein structures. The correlation coefficients are represented by solid dots where significant (i.e., p-value ≤ 0.05) or empty dots otherwise (where p-value ≥ 0.05). **(B)** Plot of the second vs. first principal components depicting the major contribution of buriedness measures to the first component and therefore, to the explained variance of sequence entropy in plot **(A)**. Red vectors represent the contributions of each of the structural variables to principal components and the black dots represent the amino acid sites on all 9 protein structures considered. **(C)** Plot of the third vs. first principal components depicting the major contribution of – xx needs further discussion with Claus. I doubt if the original conclusion about these two plots is credible. xx