

**Coevolving residues inform protein dynamics profiles and disease susceptibility
of nSNVs**

Brandon M. Butler^{1¶}, I. Can Kazan^{1¶}, Avishek Kumar^{1,2}, and S. Banu Ozkan^{1,*}

¹Department of Physics and Center for Biological Physics, Arizona State University,
Tempe, AZ, USA

²Harris School of Public Policy and Center for Data Science and Public Policy, University
of Chicago, Chicago, IL, USA

*Corresponding author

Email: Banu.Ozkan@asu.edu

[¶]These authors contributed equally to this work.

Abstract

The conformational dynamics of proteins is rarely used in methodologies used to predict the impact of genetic mutations due to the paucity of three-dimensional protein structures as compared to the vast number of available sequences. Until now a three-dimensional (3D) structure has been required to predict the conformational dynamics of a protein. We introduce an approach that estimates the conformational dynamics of a protein, without relying on structural information. This *de novo* approach utilizes coevolving residues identified from a multiple sequence alignment (MSA) using Potts models. These coevolving residues are used as contacts in a Gaussian network model (GNM) to obtain protein dynamics. B-factors calculated using sequence-based GNM (Seq-GNM) are in agreement with crystallographic B-factors as well as theoretical B-factors from the original GNM that utilizes the 3D structure. Moreover, we demonstrate the ability of the calculated B-factors from the Seq-GNM approach to discriminate genomic variants according to their phenotypes for a wide range of proteins. These results suggest that protein dynamics can be approximated based on sequence information alone, making it possible to assess the phenotypes of nSNVs in cases where a 3D structure is unknown. We hope this work will promote the use of dynamics information in genetic disease prediction at scale by circumventing the need for 3D structures.

Author Summary

Proteins are dynamic machines that undergo atomic fluctuations, side chain rotations, and collective domain movements that are required for biological function. There is, therefore, a need for quantitative metrics that capture the dynamic fluctuations per position to understand the critical role of protein dynamics in shaping biological functions. A limiting factor in incorporating structural dynamics information in the classification of non-synonymous single nucleotide variants (nSNVs) is the limited number of known 3D structures compared to the vast number of available sequences. We have developed a new sequence-based GNM method, termed Seq-GNM, which uses co-evolving amino acid positions based on the multiple sequence alignment of a given query sequence to estimate the thermal motions of C-alpha atoms. In this paper, we have demonstrated that the predicted thermal motions using Seq-GNM are in reasonable agreement with experimental B-factors as well as B-factors computed using 3D crystal structures. We also provide evidence that B-factors predicted by Seq-GNM are capable of distinguishing between disease-associated and neutral nSNVs.

Introduction

A 3D structure is still required to computationally obtain protein dynamics, drastically limiting the extent to which conformational dynamics can be incorporated into genomic analysis. The reason for this is that there are exponentially more sequences than experimental structures. Currently, UniProtKB contains more than 100 million sequence entries, whereas the PDB reports the number of known 3D structures to be

around 140,000 [1]. Furthermore, the number of known sequences is increasing at an exponential rate, compared to the much slower addition of new experimental PDB structures. This is due to the advent of high-throughput genomic sequencing, which is providing an unprecedented amount of data for genomic analysis. The vast amount of sequence data has driven the rapid classification of novel genetic variations through genome-wide association studies [2,3]. A large catalogue of non-synonymous single nucleotide variants (nSNVs) occurs in coding regions that can severely impact protein function, potentially leading to disease [4]. There are many *in silico* methods developed using evolutionary methodologies such as positional conservation and phylogeny and those that combine evolutionary approaches with biochemical and structural properties to diagnose neutral and disease associated nSNVs [5–11]. However, the accuracy of the majority of these *in silico* prediction methods is significantly lower for predicting the impact of nSNVs at highly evolving sites [12–16]. Protein dynamics can also be used to elucidate the functional impact of nSNVs and mechanisms of disease [5,17]. Our previous studies have evinced that a site-specific conformational dynamics analysis is capable of diagnosing nSNVs irrespective of evolutionary conservation [5,18,19] and recently has been incorporated as an additional feature for *in silico* prediction tools [20]. However, only a small fraction of the catalogued nSNVs in the coding regions (i.e. missense variants) have 3D experimental structures, [20], impeding broad application of protein dynamics in *in silico* tool predictions.

Coevolution, on the other hand, has become a valuable tool for its ability to predict structural contacts of 3D structures, particularly using global information through Potts

models [21–27]. Coevolving residues are inferred from a multiple sequence alignment (MSA) of a given protein family, whereby if two given amino acids exhibit concordant patterns of evolution throughout the MSA then they are assumed to be in close spatial proximity in the folded 3D structure. This evolutionary principle can be leveraged so that sequence information can be used to describe protein topology, making *de novo* structure prediction possible [24,27]. It has been reported that only one correct contact for every 12 residues in a protein is necessary for accurate topology-level modeling [28]. In addition to structure prediction, coevolution analysis has also been used to identify critical interactions between protein complexes [22] important functional sites [24] and allosteric response [29]. The use of coevolution for structure prediction is largely possible for two reasons. First, the amount of sequence data for different protein families is sufficient to be leveraged by this technique to make predictions. Second, the methods for inferring coevolving residues from an MSA are becoming increasingly robust [30–34].

Inferring evolutionary couplings from an MSA are based on two primary approaches categorized as local [35–37] and global approaches [37–39]. The global approaches detangle direct evolutionary couplings from indirect couplings which enables them to capture spatial contacts [40]. Regardless of the method, the accuracy of detecting coevolving residues that correspond to structural contacts is fundamentally limited by the number of sequence homologs in the MSA. While most of the current methods use only the sequence homologs of the protein family belonging to target sequence, integrating multiple orthology protein families (i.e. families that share similar phylogeny and retain similar functions) was used to increase the number of homologs to produce a more

accurate statistical inference [41]. RaptorX, leverages this joint family methodology; it uses an ultra-deep neural network combining coevolution information with sequence conservation information to infer 3D contacts and has produced higher accuracy than other methods [42–44].

In this paper, we will demonstrate the efficacy of our novel sequence-based GNM approach, called Seq-GNM, to estimate the dynamics profile of a protein with no *a priori* knowledge of its 3D structure. This *de novo* approach based on a Gaussian network model (GNM) enables the prediction of the magnitude of mean-square fluctuations of residues, which are proportional to the B-factors determined by X-ray crystallography experiments. However, instead of using a cutoff distance to determine 3D contacts as does the original structure-based GNM, we use coevolving residues (evolutionary couplings) in our model. We show that the theoretical predictions from our Seq-GNM are in reasonable agreement with experimental crystallographic B-factors as well as the values obtained from the structure GNM models that use spatial contacts. We also extend this analysis to determine the capacity of our model to assess the functional impact of nSNVs. We will demonstrate that the dynamics predicted by Seq-GNM can adequately classify disease and benign nSNVs across the proteome.

Results

B-factor Correlations: Sequence, Structure, and Experimental

We considered a high-resolution protein (2.25 Å) that is involved in amino acid catabolism, acyl-CoA dehydrogenase (1JQI), as an example case to examine the B-factor profiles and predicted contact maps using Seq-GNM. Coevolution analysis using direct coupling analysis (DCA) has been shown to recapitulate accurate structural contact maps for a wide range of proteins [21,23,24,27,31,45]. As expected, the contact maps of Seq-GNM and structural GNM are similar (Fig 1). In a comparison of their B-factor profiles, both Seq-GNM and structural GNM exhibit good agreement with observed B-factors, capturing flexible and rigid positions. Using evolutionary coupling (EC) values obtained from RaptorX, the correlation between the Seq-GNM and observed B-factors is 0.77, whereas the correlation between the structural GNM and observed B-factors is 0.57 (Fig 1a). Similarly, using EC values obtained by EVcouplings produced a correlation of 0.60 between the Seq-GNM and observed B-factors (Fig 1b). The scores obtained from EVcouplings are still reasonable, yet relatively lower correlations compared to those obtained by the RaptorX. This is likely due to the relatively noisy contact map predictions by EVcouplings compared to the more reliable contact maps produced by RaptorX (we think this is due to their inclusion of multiple orthology protein families) [42].

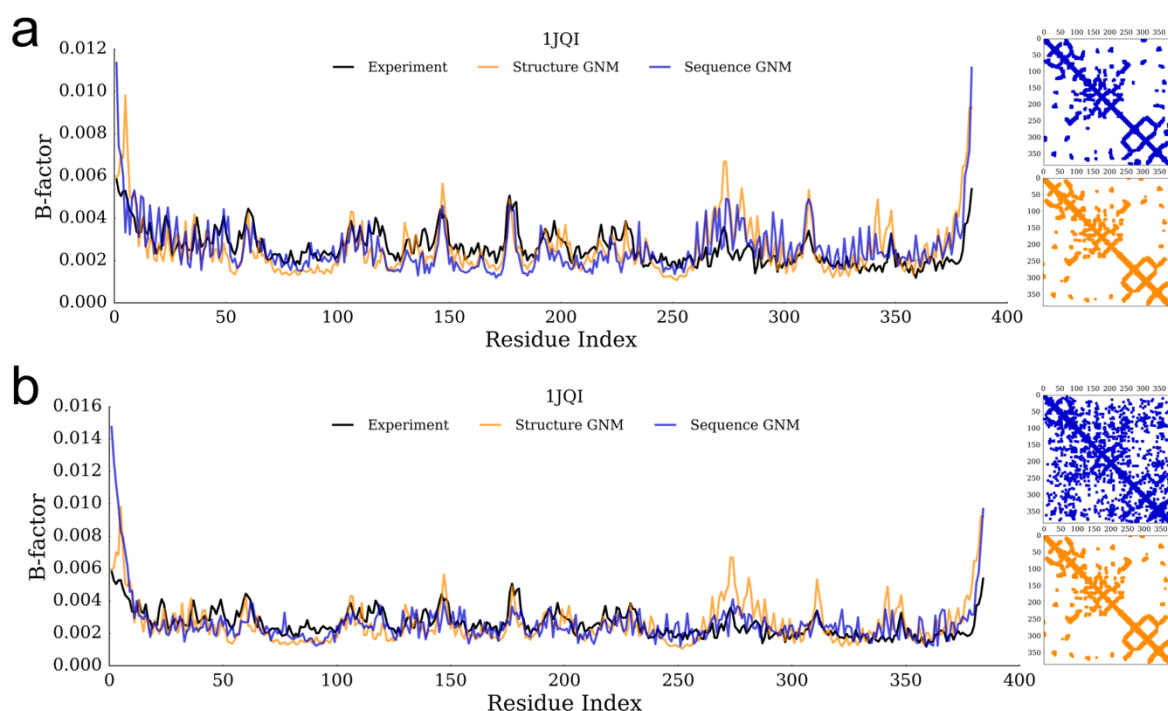


Fig 1. B-factors plots.

A plot of theoretical B-factors as calculated by our Seq-GNM (blue), the original GNM obtained from structure (orange), and observed experimental B-factors (black) for acyl-CoA dehydrogenase (PDB id: 1JQI) along with predicted contact maps by Seq-GNM (using a threshold score, shown as blue) and the contact map of the structure (using 10Å cut-off distance) (a) The Seq-GNM with values obtained from RaptorX produced a correlation of 0.56 with experiment, and 0.77 with the GNM obtained from structure. Moreover, the contact maps reveal the predicted contacts between the Seq-GNM and structural GNM approaches are remarkably similar. (b) The Seq-GNM that uses values obtained from EVcouplings produced a correlation of 0.60 with experiment, and 0.68 with the GNM obtained from structure. The B-factor obtained by applying GNM to the experimental structure yields a correlation of 0.57. The contact map captures the dominant contacts with noise coming from poorly predicted EVcouplings scores

148

149 The Seq-GNM produces a correlation with crystallographic B-factors of 0.60, which

150 is within the same range as those produced by the GNM from structure of 0.57. Moreover,

151 theoretical B-factor profiles obtained from both methods were able to identify the catalytic

152 sites on all of the proteins.

As a further test of the efficacy of the Seq-GNM, we superimposed the predicted B-factors onto the structures of three diverse proteins— 5'(3')-deoxyribonucleotidase (2JAO), acyl protein thioesterase (1FJ2), and NADH-cytochrome b(5) reductase (1UMK)— to visually contrast the predicted B-factors with that of experiment. Fig 2 shows each protein color-coded according to their B-factor profile on a spectrum of blue–white–red, where blue represents the lowest B-factors (less mobility) and red represents the highest B-factors (more mobility). The left panel shows the experimental B-factors for each protein, while the right panel shows the theoretical values predicted by the Seq-GNM. We investigated whether secondary structure was a factor in how the B-factors were distributed across the protein, and if certain secondary structure domains would exhibit less agreement with experiment. In this context, the proteins were selected so that they had a variety of secondary structure components—2JAO contains primarily alpha helices, 1UMK is mainly composed of beta-sheets, and 1F2J is a combination of alpha helices and beta-sheets. For 2JAO, the exterior helices that are flexible (red) in the observed structure are all reproduced in the predicted structure. The one highly rigid (blue) helix in the observed structure was more flexible in the predicted structure but was still in overall agreement. There is a surprising amount of similarity between the observed and predicted structure of 1F2J, considering that it contains both alpha-helix and beta-sheet elements. Similarly, 1UMK showed good agreement, except for some miniscule differences. This gives further evidence that the magnitudes of residue fluctuations predicted by the Seq-GNM model is representative of the crystallographic B-factor profiles for many proteins.

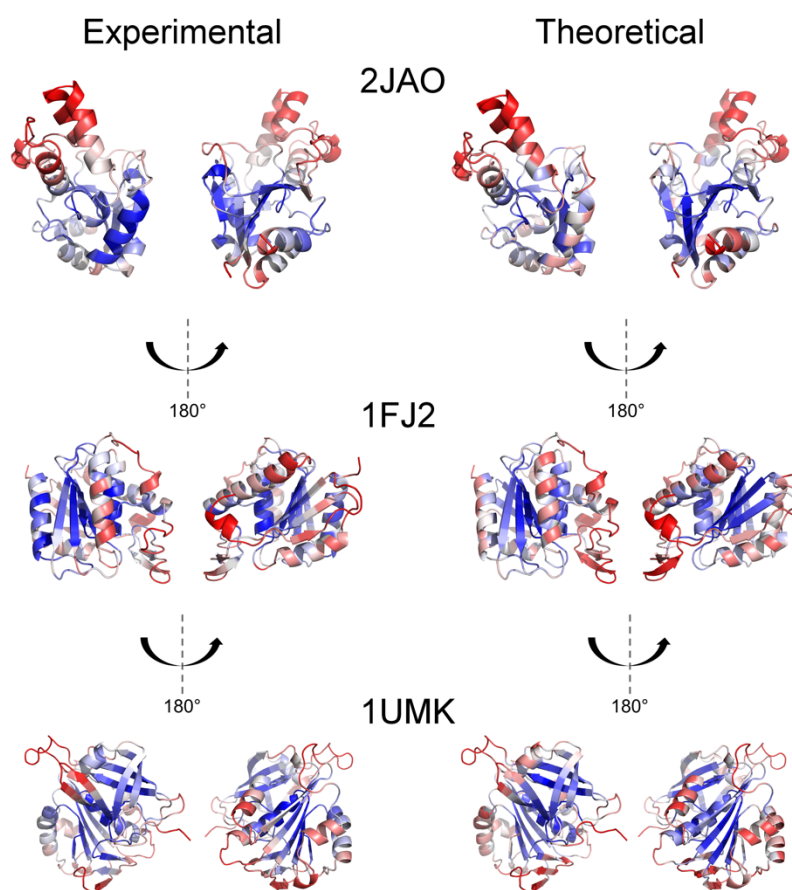


Fig 2. Color-coded ribbon diagrams using experimental and theoretical B-factors obtained by Seq-GNM.

The observed crystallographic B-factors (left) and the predicted B-factors from the Seq-GNM superimposed on the structure. The three proteins selected—2JAO, 1F2J, and 1UMK—were high-resolution structures are better than 2.0 Å. The B-factors are color-coded according to their B-factor profile on a spectrum of blue–white–red where blue represents the lowest B-factors (less mobility) and red represents the highest B-factors (more mobility). The B-factor scores were converted to a percentile rank so that they could be compared across different proteins. Each protein was also rotated 180° so that both sides could be visualized and compared. Moreover, the proteins were selected so that they had a variety of secondary structure components—2JAO contains primarily alpha helices, 1UMK is mainly composed of beta-sheets, and 1F2J is a combination of alpha helices and beta-sheets.

174

175

In order to compare predicted B-factors with crystallographic B-factors, we extracted a subset of 39 structures that had a resolution better than 2.0Å to obtain more realistic crystallographic B-factors (unreliable B-factors are common for many PDB structures) [18,46]. The same cutoff of 2.0Å was used in an earlier study to compare GNM predicted B-factors with those determined by crystallography [47]. For all 39 structures, the Seq-GNM (using EC values from RaptorX) and structure GNM were used to estimate their B-factors, which were then compared with the observed B-factors by calculating the correlation for each protein. The mean correlation coefficient for the Seq-GNM was 0.53 while the mean correlation coefficient for the structure GNM was 0.58. The correlation of 0.58 for structural GNM of our smaller data set is consistent with the findings of Kundu et al. where 113 high-resolution structures (resolution <2.0 Å) were used and, the mean correlation coefficient with observed B-factors was 0.59 [47].

As shown in Fig 3a, boxplot distributions reveal that correlations are not significantly different between the sequence and structure GNM ($p = 0.055$ in a student t-test). The structure GNM appears to perform only slightly better than the Seq-GNM. Fig 3b shows the same distribution separated into 10 individual bins of size 0.1. The overall shapes of the two distributions are similar, except for the exaggerated relative lower second peak of the Seq-GNM at 0.4. It should also be noted that for these cases where Seq-GNM had low correlations, the EC threshold could be tuned to yield much higher correlations. If this were done on a case-by-case basis, the overall correlation distributions would be even more similar. Thus, the EC threshold may be used as a tuning parameter to enhance the correlation coefficient for purposes of model optimization.

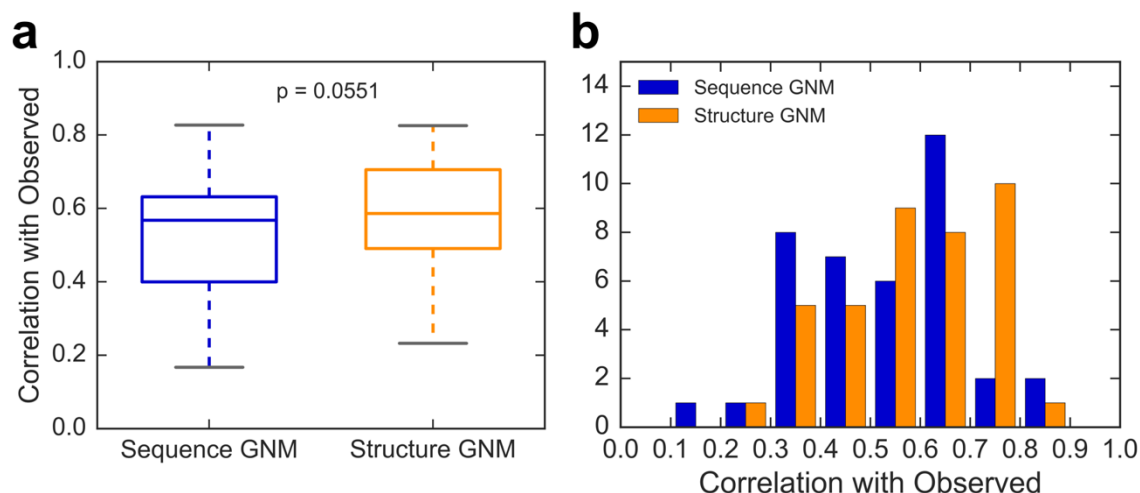


Fig 3. Comparison of B-factors obtained by GNM and Seq-GNM with experimental B-factors.

(a) Boxplot showing the correlation of predicted B-factors by the Seq-GNM with experimentally observed B-factors (blue) in comparison to that of the GNM obtained from structure (orange) for a subset of 39 structures with resolution better than 2.0 Å. (b) A distribution plot of the same correlations binned into 10 bins with sizes of 0.1. A student t-test revealed no significant difference between the two distributions ($p=0.055$) indicating that the Seq-GNM is producing competitive results compared to the original GNM from structure. The mean correlation of the Seq-GNM is 0.53 while that of the GNM from structure is 0.58.

199

200 Interestingly, for the cases where predicted B-factors by Seq-GNM yielded
 201 significantly better correlations with the experimental B-factors than those obtained by
 202 GNM from structures, we observed that biological units of these proteins are assigned as
 203 oligomeric forms. While predicted B-factors obtained using Seq-GNM does not retain this
 204 information, it successfully predicts the experimentally low B-factor values of interface
 205 positions as shown for protein 5'(3')-deoxyribonucleotidase (2JAO) and protein aldehyde
 206 Dehydrogenase 7A1 (2J6L) in Fig 4. It is indeed shown in earlier work of direct contact
 207 analysis that co-evolution can identify positions of protein interfaces and protein-protein

interaction partners and successfully reconstruct protein complexes and interaction network [23,30,48]. Thus, it is not surprising to see that it yields good correlations with the experimental B-factors. Conversely, predicted B-factors from structure can only improve when the oligomeric structure is used for the GNM analysis.

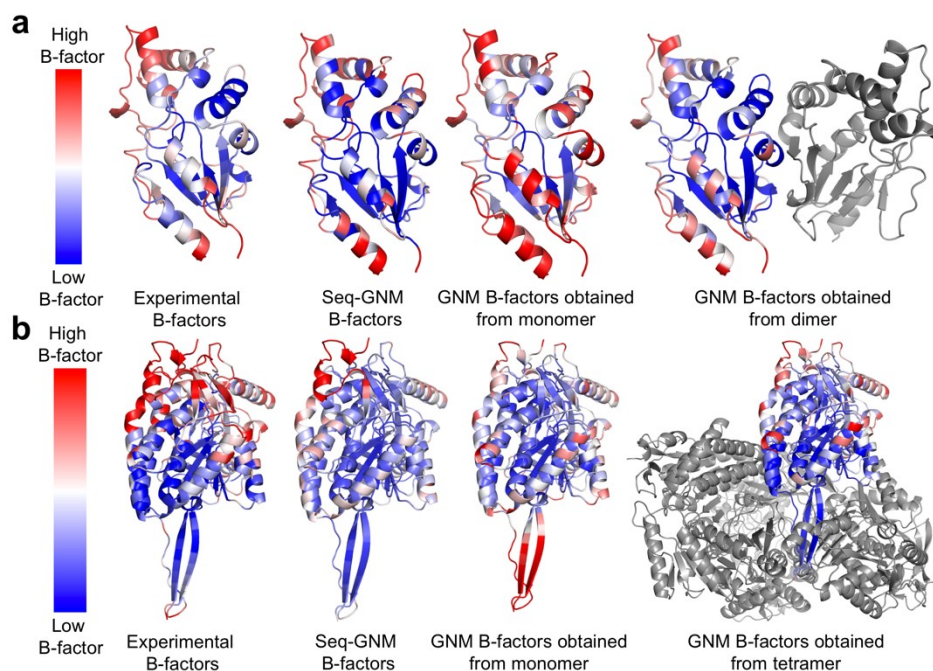


Fig 4. Comparison of B-factors obtained from Experiments, Seq-GNM, GNM from monomeric structure, and GNM from oligomeric structure.

B-factors are shown on the respective structures for (a) 5'(3')-deoxyribonucleotidase (2JAO) and (b) Aldehyde Dehydrogenase 7A1 (2J6L). (a) The correlation of Seq-GNM to experimental B-factors is 0.83 while correlation of GNM B-factors obtained from monomer to experimental B-factors is 0.63. When dimer is used for GNM analysis the correlation of GNM B-factors obtained from monomer to experimental B-factors increase to 0.72. (b) The correlation of Seq-GNM to experimental B-factors is 0.61 while correlation of GNM B-factors obtained from monomer to experimental B-factors is 0.37. When tetramer is used for GNM analysis the correlation of GNM B-factors obtained from monomer to experimental B-factors increase to 0.76. The change in correlation for GNM between monomer and oligomer clearly shows the drawback for

dependence on the crystal structure of biounits. However, Seq-GNM captures the interface B-factors correctly.

Even when using high-resolution X-ray structures, there is still some uncertainty about the realistic nature of crystallographic B-factors. For this reason, we thought a more plausible way to determine the efficacy of the Seq-GNM was to compare it directly with the structure GNM. The structure GNM is a robust method to describe thermal fluctuations in a protein, and in many cases, it performs as good or better than the ANM or MD [47,49]. We systematically evaluated the performance of the Seq-GNM and structure GNM for the entire set of 139 structures and obtained the correlation coefficients for each protein (Fig 5).

The average correlation of B-factors between the Seq-GNM and structure GNM model is 0.63 when using EC contacts from RaptorX and 0.43 when using contacts from EVcouplings. As seen in Fig 5a, the distribution of correlation coefficients increases until 0.8, and then subsequently decreases. Interestingly, there are still an appreciable number of sequences yielding high correlations from 0.8 to 1.0. A distinguishing feature of the distribution is the pronounced peak in the bin from 0.7 to 0.8, indicating that significant fraction of our data set yields high correlations between 0.7 and 0.8. This is evidence that the Seq-GNM is efficiently capturing protein dynamics and supports the theory that ECs can be used as a substitute to 3D structure contacts in the GNM and still produce reliable dynamics profiles. The results of Seq-GNM based on contacts predicted by RaptorX usually yields B-factors that are closer to experimental B-factors as it uses structural

information in its neural networks leading to better EC values and correlations with structure [44].

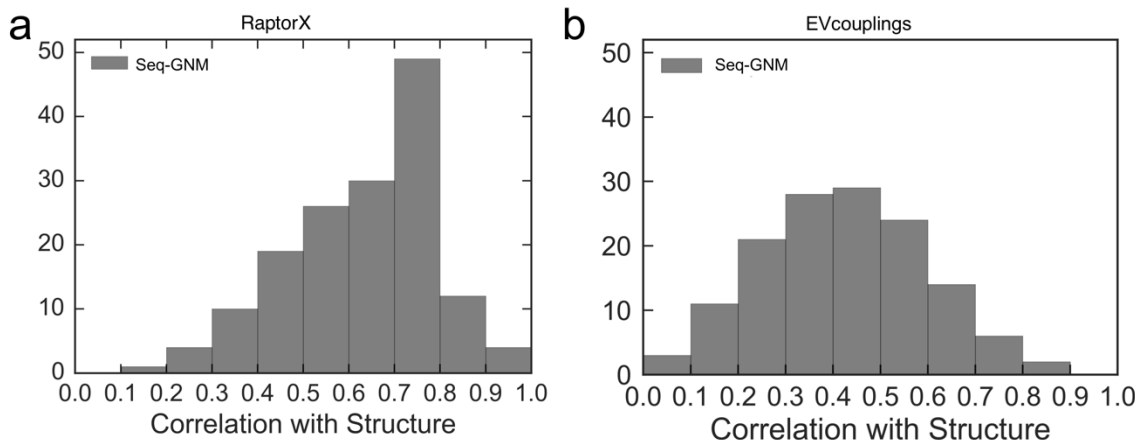


Fig 5. Distribution of correlation coefficients.

The distribution of correlation coefficients between B-factors from Seq-GNM and GNM from structure. (a) The average correlation coefficient is 0.63 with RaptorX EC values. (b) The average correlation coefficient is 0.43 by using EVcouplings EC values.

Assessing nSNV Phenotypes Using the Seq-GNM

Crystallographic B-factors have previously been used to assess the impact of nSNVs on protein function [18,50–54]. A study [51] found that mutations on lysozyme that impaired function exhibited lower than average temperature factors, suggesting that rigid sites on the protein are more susceptible to destabilizing nSNVs than flexible sites [55]. Another study revealed a relationship between crystallographic B-factors and the impact of nSNVs on protein function [56]. A commonly used tool to diagnose neutral and disease associated nSNVs, PolyPhen-2, uses evolutionary information, structural information,

and crystallographic B-factors in its prediction model [49]. These studies indicate that crystallographic B-factors can be used to predict the tolerance of a given residue to an nSNV (i.e., whether or not the occurrence of an nSNV would impact function).

We investigated whether B-factors predicted by the Seq-GNM were indicative of biological phenotype for nSNVs in the human population. A total of 738 nSNVs were mapped to the 139 enzymes, where 436 are disease-associated and 302 are neutral. S1 Table shows the number of disease and neutral nSNVs that occur on each protein. The Seq-GNM (using EC contacts from **RaptorX** and **EVcouplings**) was computed systematically for all 139 enzymes to obtain their dynamics profiles. The theoretical B-factors scores were converted into a percentile rank so that the values could be compared across different proteins.

We initially looked at two human enzymes, human lysozyme (PDB: 1C7P) and human cytochrome reductase (PDB: 1UMK). They were chosen because they were short proteins that each contain a disease and neutral nSNV. Human lysozyme is a glycoside hydrolase that functions in the immune system by causing damage to cell walls of bacteria. Human cytochrome b5 reductase is involved in many oxidation/reduction reactions including converting methemoglobin to hemoglobin [55].

Each structure is color-coded according to its theoretical B-factor profile on a spectrum of blue–white–red. Sites that exhibit high mobility (flexible) are red, and sites that have low mobility (rigid) are blue. Regions that are characterized by low mobility are usually important for maintaining stability and function, thus a mutation could act to destabilize the protein and impair its function. Fig 6a show the disease mutation I56T

occurring on a rigid site with a B-factor of 0.0075. The neutral mutation T70N has a B-factor of 0.96 indicating that it is a highly mobile site. Both I56T and T70N occur on loop regions. Although loops are generally more flexible, three alpha-helical domains encompass the loop containing I56T, which implies that it may be involved in interactions that contribute to stabilizing the functional conformation. Thus, the I56T mutation may disrupt these critical interactions and impair the enzymatic function. In the case of cytochrome reductase (Fig 6b), the disease mutation R57Q is also on a rigid site with a B-factor of 0.14. Instead of being located near the core, R57Q is highly exposed protruding outwardly from a beta-barrel. However, since beta-barrels often harbor functional residues, the R57Q mutation may disrupt certain interactions critical for modulating function. The neutral mutation T116S is located on a loop and has a B-factor of 0.96, indicating that it has a high mobility. In our earlier proteome wide study of over 100 human protein structures, we have shown that sites that are highly flexible (e.g., loop regions, or superficial sites) are typically more robust to mutations. Conversely, rigid sites are more susceptible to mutations that may disrupt function [18,19]. For these two cases, the B-factors produced by Seq-GNM successfully distinguished between the disease and neutral nSNVs, without using the 3D structures.

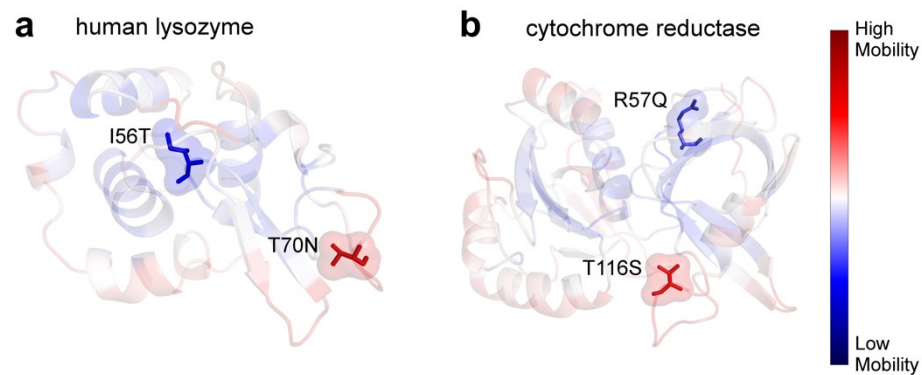


Fig 6. Comparison of theoretical B-factors on disease versus neutral mutant sites.

A ribbon diagram for two human enzymes, human lysozyme (a) and cytochrome reductase (b) colored according to their predicted B-factors by the Seq-GNM. Red indicates high mobility sites, and blue indicates low mobility sites. Each protein contains two known nSNVs. I56T and R57Q are disease-associated, and they occur on low mobility (rigid) sites. Conversely, the neutral nSNVs T116S and T70N occur on high mobility sites.

These findings prompted us to analyze the proteome-wide set of 139 enzymes to determine if the B-factors were indicative of phenotype for all 436 disease and 302 neutral nSNVs. The raw B-factor values were converted into a percentile rank (%B-factor) and then binned into 5 bins of size 0.2. We computed the observed-to-expected ratio of B-factors, where the expected values were based on the B-factor distribution of all 51,618 sites across all 139 proteins, and the observed values were based on the B-factors of the 436 disease sites. The same process was done for the 302 neutral nSNVs. Under the null hypothesis that predicted B-factor of the disease associated nSNVs yields similar distribution of all the positions gathered from 139 enzyme sequences, the ratio of expected and observed sites harboring disease mutations for each %B-factor bin should be close to 1, which would imply that B-factor does not distinguish sites that are prone to

disease. This is the null hypothesis that disease sites are distributed uniformly between sites with low and high mobility. However, the null hypothesis was rejected for the 436 disease nSNVs ($p < 0.001$). Fig 7 shows the observed-to-expected ratio plot of disease and neutral nSNVs, which indicates that disease nSNVs are overabundant at low %B-factor sites (< 0.4) and under abundant at high %B-factor sites. Conversely, neutral nSNVs are overabundant at high %B-factor sites (> 0.6) and under abundant at low %B-factor sites. This evidence suggests that the occurrence of an nSNV on a site with a low B-factor is likely damaging based on the position irrelative of the substitution. This is in agreement with our previous proteome-wide study showing that substitutions at rigid sites are more often associated with diseases [18]. Conversely, an nSNV on a high B-factor site is usually benign. Low B-factors usually signify a residue that is crucial for modulating functional motions (e.g., a hinge). Thus, mutations on these sites can severely impact function. High B-factor sites are more flexible (e.g., loops) and more robust to mutations. Fig 7 suggest that it is possible to use the predicted B-factors to discriminate between disease and neutral nSNVs using co-evolution obtained from only multiple sequence alignment. Moreover, it can be used as an additional feature for *in silico* predictions [12].

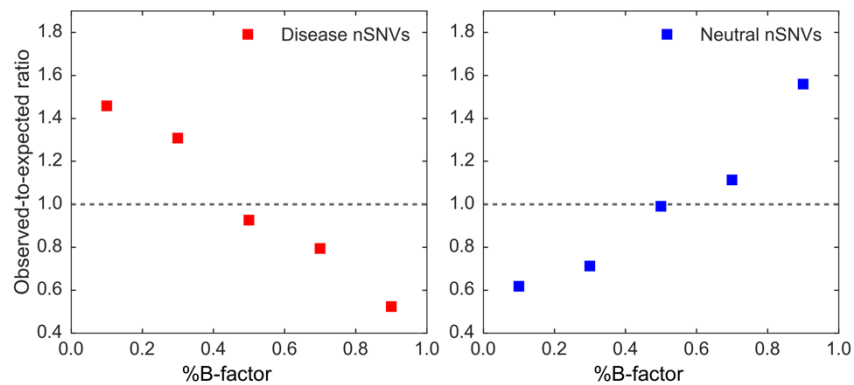


Fig 7. Observed to expected ratio plots for disease and neutral nSNVs.

The relationship of observed-to-expected numbers between 436 disease nSNVs (red) and 302 neutral nSNVs (blue) from 139 human enzymes. The %B-factor scores derived from the Seq-GNM are binned into 5 bins of size 0.2.

Predictive models were created using logistic regression as the classification algorithm, 80% of the data was used for training and 20% for testing for 10 randomized sets. Models were evaluated based on ROC curves and their respective area under curve (AUC), the best performance is labeled as AUC_max and average performance as AUC. Theoretical B-factors obtained by Seq-GNM, experimental B-factors, and evolutionary parameters were used as predictive variables for training and testing (Fig 8). Seq-GNM and experimental B-factors have similar performance (maximum AUC of best 0.76 and 0.75, respectively), with Seq-GNM overshadowing experimental B-factors on average (AUC of 0.69 and 0.60, respectively). The ~0.70 AUC of B-factors obtained from Seq-GNM is impressive, as it has been shown that majority of state-of-art methods also yields similar AUC in independent tests [5,13]. Moreover, incorporation of Seq-GNM as an additional feature with evolutionary parameters resulted in higher prediction performance. While the AUC scores obtained using the evolutionary features for classification gives

0.76, this is increased to 0.81 after including the B-factors of Seq-GNM (Fig 8c-d). This result also demonstrates the efficacy of Seq-GNM in disease prediction as a complementary metric to other metrics used as features in classifiers.

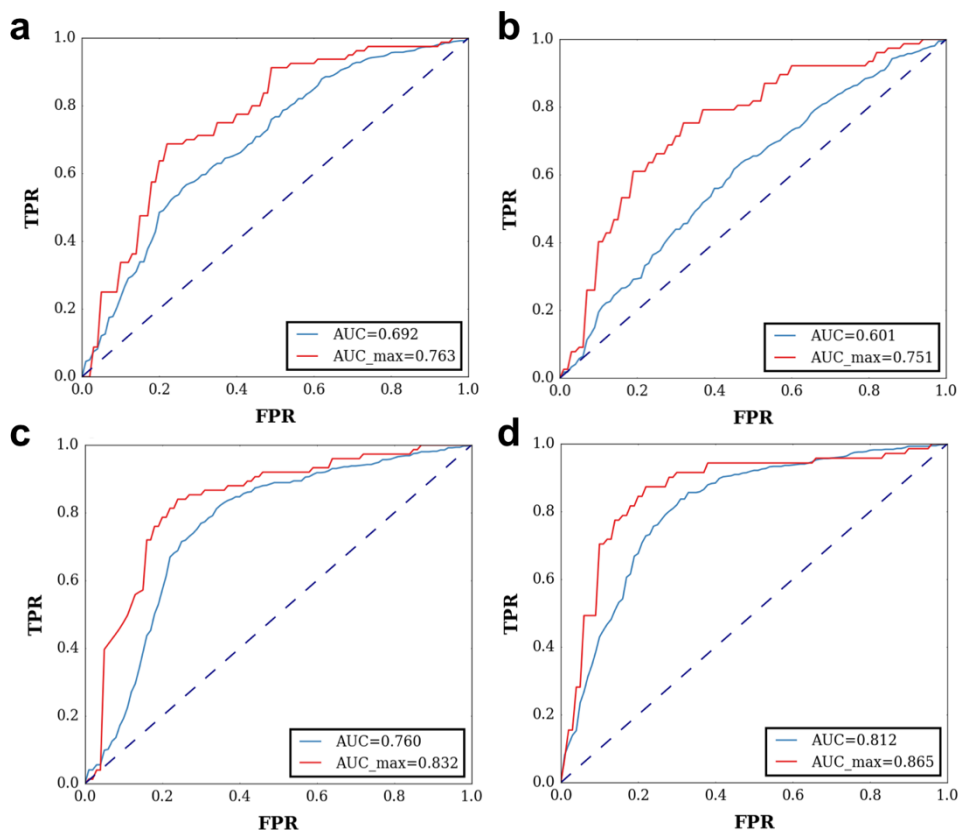


Fig 8. ROC curves for disease prediction performance comparing Seq-GNM, experimental B-factors and evolutionary parameters.

ROC curves are plotted using 10 randomly selected training and testing data sets using 80%, and 20% of the data, respectively. (a) ROC curve of Seq-GNM. (b) ROC curve of experimental B-factors. (c) ROC curve of evolutionary parameters, where primate, mammal, and vertebrate fitch rates using Fitch Algorithm [57]; and Entropy2 are used as features for training. (d) ROC curve of evolutionary parameters used in (c) with the addition of Seq-GNM.

We also compared the performance of Seq-GNM with common *in silico* prediction tools like Polymorphism Phenotyping v2 (PolyPhen-2), and Sorting Intolerant from Tolerant (SIFT) [6,58]. The accuracy, sensitivity, and selectivity of disease predictions for nSNVs with experimental B-factors, B-factors from SIFT, PolyPhen-2, evolutionary parameters, and Seq-GNM are tabulated in Table 1. The accuracy of Seq-GNM using both EC values from EVcouplings and RaptorX is ~0.70. This accuracy is similar to using experimental B-factors for prediction (0.69) and also very close to prediction with evolutionary parameters (0.75), suggesting that Seq-GNM allows us to incorporate protein dynamics in nSNV predictions when the 3D experimental structures are not available. Moreover, accuracy of Seq-GNM approach is greater than SIFT (0.65) and PolyPhen-2 (0.64). Interestingly, Seq-GNM obtained by EVcouplings and RaptorX yields similar accuracies indicating that evolutionary couplings without the inclusion of structure could be utilized to predict B-factors to include as a feature to *in silico* prediction tools. Seq-GNM sensitivity (~0.90) surpasses other methods (0.80 for SIFT, 0.63 for PolyPhen-2, and 0.85 for evolutionary parameters), but it has a shortcoming in selectivity (~0.36) as other methods reach higher (~0.59). Conversely, training Seq-GNM combined with evolutionary parameters enhances the selectivity (0.66) to its highest value compared to others. Seq-GNM with evolutionary parameters predicted disease related nSNVs with accuracy 0.78 and sensitivity of 0.84, reaching beyond predictions of other metrics solely. These results suggest the incorporation of Seq-GNM with other prediction metrics can augment accuracy, sensitivity, and selectivity of prediction.

Table 1. The disease prediction data showing the accuracy, sensitivity, and the selectivity of Seq-GNM compared with experimental B-factors, SIFT, PolyPhen-2 and evolutionary parameters.

Classifier Feature and Methods	Accuracy	Sensitivity	Selectivity
Experimental B-factors	0.69	0.88	0.35
SIFT	0.65	0.70	0.56
PolyPhen-2	0.64	0.63	0.64
Seq-GNM B-factors (EVcouplings)	0.70	0.89	0.37
Seq-GNM B-factors (RaptorX)	0.71	0.91	0.35
Evolutionary parameters	0.75	0.85	0.58
Seq-GNM (EVcouplings) with evolutionary parameters	0.78	0.84	0.66

Prediction accuracy of Seq-GNM is further tested using 323 nSNVs (187 disease-associated, 136 neutral) of 22 proteins where their 3D experimental structures are not available (S2 Table). We used the trained classifier model of Seq-GNM B-factors for this test. While the B-factors obtained solely from Seq-GNM are used, it reached an accuracy, sensitivity, and selectivity of 0.82, 0.82, 0.83, respectively. This result further suggests that Seq-GNM allows us to incorporate protein dynamics as additional feature in *in silico* prediction tools without a known 3D structure.

Discussion

While we and others [5,19,59–63] have shown that the integration of conformational dynamics into genomic analysis will help next generation of approaches to predict the impact of novel missense mutations on the human proteome, the inherent limitations in availability of 3D structures compared to the vast number of sequences must be

addressed. This begs the question: how can protein dynamics be used in genome-wide analysis to predict functional impacts of nSNVs? There is, therefore, a need to be able to obtain protein dynamics by leveraging only sequence information, without *a priori* knowledge of a 3D structure. For this reason, we have developed this novel method to estimate the dynamics profile of a protein by using only a sequence as input. The method uses the coevolution of amino acids through multiple sequence (which tend to be spatially close in the 3D tertiary structure) and a simple Gaussian network model (GNM) to obtain dynamics. The original GNM based on the 3D structure is well-known for its ability to describe residue dynamics profiles due to thermal motions in proteins (i.e., B-factors). We showed that our sequence-based GNM model is able to adequately reproduce the mean-square fluctuations (B-factors) calculated by the original GNM, particularly outperforms for the cases where biological functional state is oligomeric. Our estimates of B-factors for a proteome-wide set of proteins exhibited good correlation with the structure GNM. Moreover, our estimated B-factors were in reasonable agreement with crystallographic B-factors for many cases. To address the issue of how protein dynamics can determine the impact of nSNVs across the genome where there are no known 3D structures, we tested the ability of our predicted dynamics from the Seq-GNM to assess nSNV phenotypes. A plot of the observed-to-expected ratio of the predicted B-factors revealed distributions of disease and neutral nSNVs that are similar to those in a previous protein dynamics analysis work [18]. The predicted B-factors using the Seq-GNM was able to discriminate between disease and neutral nSNVs with an accuracy of 0.70 and incorporating the Seq-GNM predicted B-factors with evolutionary parameters increased

overall accuracy to 0.78. This analysis demonstrates that the Seq-GNM makes it possible to obtain estimates of dynamics without using a 3D structure, which will allow for the integration of conformational dynamics into large-scale analysis of genomic variants.

Methods

Dataset

A curated set of 139 structures was selected for several reasons. First, they have high query coverage (>80%) and sequence identity (>80%) as found from a BLAST search, and the structures had already been modeled using the Modeller software package [64] to account for any missing residues. Second, genetic variants were previously mapped onto these structures, such that the positions containing known nSNVs were already determined, enabling us to easily compare our results using sequence coevolution with the genetic variation data. A total of 738 genetic variants were obtained from the HumVar database [58], which was comprised of 436 disease and 302 neutral nSNVs. Finally, the structures were either monomers or the single-chain unit of a multimer with <600 residues, allowing for tractable calculations of residue coevolution using the RaptorX web server [42,44], and EVfold (EVcouplings) [21]. A table summarizing the dataset is presented in S1 Table.

Obtaining Coevolved Residues

The amino acid sequence from each of the 139 structures was used as input for the evolutionary coupling (EC) analysis. The choice of taking the amino acid sequence from the structure was done so that the predicted EC contacts could be compared directly to the experimentally observed structure contacts as verification that the model was producing realistic contact maps. Moreover, the theoretical B-factors predicted by our sequence-based model could be directly compared to the experimental B-factors for each protein. If the structure was unknown, however, sequence databases (e.g. UniProt, PFAM, etc.) could be used. The PDB sequences were given to the RaptorX web server [42,43], which computed the relative probability of each residue pair i, j of being in 3D contact based on their coevolution strength. The sequences were also used to generate MSAs using phmmer [65]. Using MSAs, DI values are calculated by EVcouplings. In order to ensure consistency between different proteins of varying lengths, we converted the raw scores into percentile ranks. We then used a threshold value, taking only the top scoring evolutionary couplings (i.e., the strongest couplings are more likely to be in spatial contact). An optimized threshold value was systematically evaluated and is discussed in the Methods.

Sequence-based GNM Model (Seq-GNM)

The Gaussian network model (GNM) is an isotropic approach based on the contact topology of a crystal protein structure to obtain the equilibrium fluctuations of residues due to thermal motion. It uses a specified cutoff distance to define interacting pairs that

are connected by springs with a single-parameter harmonic potential. In this structure-based GNM, the interacting residue pairs within the cutoff range are represented as contacts in the Kirchhoff (connectivity matrix).

In the proposed sequence-based GNM (Seq-GNM) approach we will instead use coevolving residue pairs (evolutionary couplings) as contacts in the Kirchhoff. In this way, the 3D structure is no longer a prerequisite to form a GNM. To construct the Kirchhoff, a threshold is defined where any evolutionary coupling scores above that threshold are sufficiently coupled such that they are spatially close in 3D structure. If a given evolutionary coupling pair meets the threshold criteria, it is assigned a value in the Kirchhoff for non-bonded contacts of -1 multiplied by its evolutionary coupling score (i.e., $-1 \times EC_{\text{score}}$). This will permit that the strength of each connection will attenuate proportionally to the evolutionary coupling strength. The Kirchhoff can be decomposed into the individual contributions from the bonded contacts representing the chain connectivity (Rouse chain) and that from the non-bonded contacts [56]. In the Seq-GNM the contribution of non-bonded contacts to the Kirchhoff is constructed according to

$$\Gamma_{ij}^{nb} = \begin{cases} -1 \times EC_{\text{score}}, & i \neq j \text{ evolutionary coupling} \\ 0, & i \neq j \text{ no coupling} \\ -\sum_{i, i \neq j} \Gamma_{ij}, & i = j \end{cases} \quad (1)$$

For the local chain connectivity (Rouse chain), we don't take into account evolutionary couplings, and matrix was constructed such that every residue pair $i, i \pm 1$ to $i, i \pm 3$ is in contact as

$$\Gamma_{ij}^{cc} = \begin{cases} -1, & i \neq j \text{ and } \sum_{i|k=1,2,3}^L i, i \pm k \\ 0, & i \neq j \text{ else} \\ -\sum_{i,l \neq j} \Gamma_{il}, & i = j \end{cases} \quad (2)$$

Then the overall Kirchhoff is the combination of the two contributions $\Gamma_{ij} = \Gamma_{ij}^{cc} + \Gamma_{ij}^{nb}$. The vibrational dynamics due to thermal fluctuations can then be evaluated in the same way as the original GNM by inverting the Kirchhoff matrix. The magnitude of mean-square fluctuations is then written in terms of the inverse Kirchhoff as

$$\langle (\Delta \mathbf{R}_i)^2 \rangle \cong [\Gamma^{-1}]_{ii} \quad (3)$$

This is proportional to the Debye-Waller temperature factors or B-factors, which describe the attenuation of X-ray scattering due to the thermal motions of atoms ($B_i = 8\pi^2 \langle (\Delta \mathbf{R}_i)^2 \rangle / 3$). Here there is no single-parameter force constant as in the GNM obtained from structure [52], and the pair-wise interactions are simply the strength of the evolutionary couplings as given by their ranked scores. The theoretical predictions of our Seq-GNM can be compared to the predictions of the original GNM obtained from structure as well as observed crystallographic B-factors. A general workflow of our method is presented as a flow diagram in Fig 9.

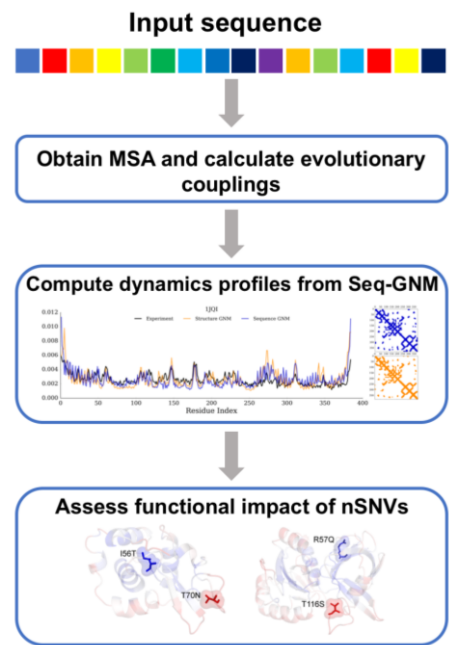


Fig 9. Flowchart of Seq-GNM method for nSNV predictions.

A workflow of our method to use predicted evolutionary couplings to determine protein dynamics and assess the functional impact of nSNVs. The initial input is an amino acid sequence, which is used to obtain MSA. Using MSA evolutionary coupling pairs are predicted through RaptorX and EVcouplings. The high scored evolutionary coupling pairs are assigned as contacts in our Seq-GNM to compute the dynamics profile of each protein. The dynamic profiles obtained from Seq-GNM can give insight into the functional impact of nSNVs. This was done for a curated set of 139 structures.

468

469

470 **Optimizing Threshold Value for EC Scores**

471

472

473

474

To ensure consistency when analyzing different proteins with varying lengths, we converted the raw scores of evolutionary couplings (EC) into a percentile rank. We computed the Seq-GNM for all 139 structures using a constant threshold percentile rank EC value to assign contacts and measured the correlation between the B-factors

475 predicted by our Seq-GNM to the GNM obtained from structure. We used only the top
476 percentile EC scores predicted by RaptorX and EVcouplings as predicted contacts,
477 because only certain fraction of high EC scores are true native contacts in 3D structure,
478 largely due to noisy artifacts in the MSA such as the transitivity of correlations and
479 phylogeny. To determine the optimal threshold value, we tested a range of threshold
480 values from 0.92 to 0.99. A threshold value ≤ 0.92 yields superfluous contacts leading to
481 a noisy contact map, and thus, a lower overall correlation (Fig 10). Conversely, a
482 threshold value ≥ 0.99 gives a deficient number of contacts, which yields an excessively
483 sparse contact map and a lower overall correlation. As Fig 10 shows, a threshold value
484 of 0.98 produced the best overall correlation with the GNM from structure and, thus, was
485 taken to be the optimal threshold value used in the analysis.

486

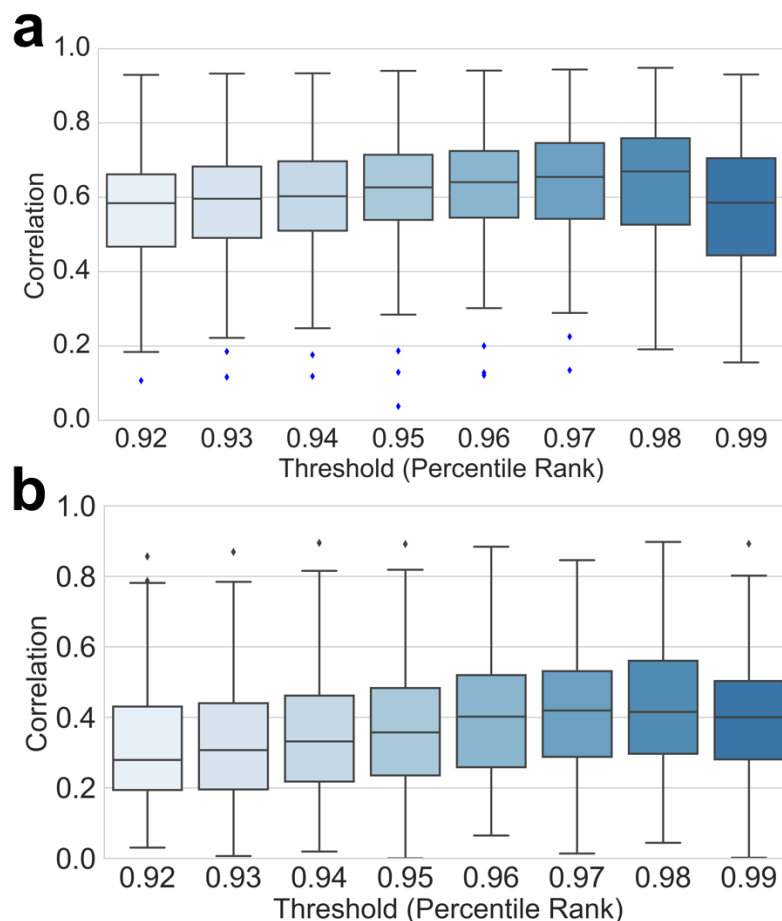


Fig 10. Comparison of theoretical B-factors.

Boxplots comparing the correlations of predicted B-factors by our Seq-GNM for (a) RaptorX and (b) EVcouplings with that of the structural GNM for all 139 structures using a constant threshold for EC contacts. The GNM analysis was conducted 8 times using a constant threshold (between 0.92 and 0.99) each time. The best average correlations were produced when the constant threshold value of 0.98 was used.

487

488

489

References

1. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. *Nucleic Acids Res.* 2000;28: 235–242.
2. Li MJ, Wang P, Liu X, Lim EL, Wang Z, Yeager M, et al. GWASdb: a database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res.* 2012;40: D1047–D1054. doi:10.1093/nar/gkr1182
3. Xie L, Ge X, Tan H, Xie L, Zhang Y, Hart T, et al. Towards Structural Systems Pharmacology to Study Complex Diseases and Personalized Medicine. *PLOS Computational Biology.* 2014;10: e1003554. doi:10.1371/journal.pcbi.1003554
4. Kumar S, Sanderford M, Gray VE, Ye J, Liu L. Evolutionary Diagnosis Method for Variants in Personal Exomes. *Nat Methods.* 2012;9. doi:10.1038/nmeth.2147
5. Kumar A, Butler BM, Kumar S, Ozkan SB. Integration of structural dynamics and molecular evolution via protein interaction networks: a new era in genomic medicine. *Current opinion in structural biology.* 2015;35: 135–142.
6. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 2003;31: 3812–3814.
7. Pers TH, Timshel P, Hirschhorn JN. SNPsnap: a Web-based tool for identification and annotation of matched SNPs. *Bioinformatics.* 2015;31: 418–420. doi:10.1093/bioinformatics/btu655
8. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the Functional Effect of Amino Acid Substitutions and Indels. *PLOS ONE.* 2012;7: e46688. doi:10.1371/journal.pone.0046688
9. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.* 2011;39: e118–e118. doi:10.1093/nar/gkr407
10. Wilkins A, Erdin S, Lua R, Lichtarge O. Evolutionary Trace for Prediction and Redesign of Protein Functional Sites. *Methods Mol Biol.* 2012;819: 29–42. doi:10.1007/978-1-61779-465-0_3
11. González-Pérez A, López-Bigas N. Improving the Assessment of the Outcome of Nonsynonymous SNVs with a Consensus Deleteriousness Score, Condel. *Am J Hum Genet.* 2011;88: 440–449. doi:10.1016/j.ajhg.2011.03.004

- 521 12. Kumar S, Dudley JT, Filipski A, Liu L. Phylomedicine: an evolutionary telescope to
522 explore and diagnose the universe of disease mutations. *Trends in Genetics*. 2011;27: 377–
523 386. doi:10.1016/j.tig.2011.06.004
- 524 13. Katsonis P, Koire A, Wilson SJ, Hsu T-K, Lua RC, Wilkins AD, et al. Single nucleotide
525 variations: Biological impact and theoretical interpretation. *Protein Science*. 2014;23:
526 1650–1666. doi:10.1002/pro.2552
- 527 14. Mahmood K, Jung C, Philip G, Georgeson P, Chung J, Pope BJ, et al. Variant effect
528 prediction tools assessed using independent, functional assay-based datasets: implications
529 for discovery and diagnostics. *Hum Genomics*. 2017;11. doi:10.1186/s40246-017-0104-8
- 530 15. Gray VE, Kukurba KR, Kumar S. Performance of computational tools in evaluating the
531 functional impact of laboratory-induced amino acid mutations. *Bioinformatics*. 2012;28:
532 2093–2096. doi:10.1093/bioinformatics/bts336
- 533 16. Miller M, Bromberg Y, Swint-Kruse L. Computational predictors fail to identify amino acid
534 substitution effects at rheostat positions. *Scientific Reports*. 2017;7: 41329.
535 doi:10.1038/srep41329
- 536 17. Kumar A, Glembo TJ, Ozkan SB. The Role of Conformational Dynamics and Allostery in
537 the Disease Development of Human Ferritin. *Biophysical Journal*. 2015;109: 1273–1281.
538 doi:10.1016/j.bpj.2015.06.060
- 539 18. Nevin Gerek Z, Kumar S, Banu Ozkan S. Structural dynamics flexibility informs function
540 and evolution at a proteome scale. *Evolutionary applications*. 2013;6: 423–433.
- 541 19. Butler BM, Gerek ZN, Kumar S, Ozkan SB. Conformational dynamics of nonsynonymous
542 variants at protein interfaces reveals disease association. *Proteins: Structure, Function, and*
543 *Bioinformatics*. 2015;83: 428–435.
- 544 20. Ponzoni L, Bahar I. Structural dynamics is a determinant of the functional significance of
545 missense variants. *PNAS*. 2018; 201715896. doi:10.1073/pnas.1715896115
- 546 21. Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, et al. Protein 3D
547 Structure Computed from Evolutionary Sequence Variation. *PLOS ONE*. 2011;6: e28766.
548 doi:10.1371/journal.pone.0028766
- 549 22. Hopf TA, Schärfe CPI, Rodrigues JPGLM, Green AG, Kohlbacher O, Sander C, et al.
550 Sequence co-evolution gives 3D contacts and structures of protein complexes. *eLife*. 3.
551 doi:10.7554/eLife.03430
- 552 23. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, et al. Direct-coupling
553 analysis of residue coevolution captures native contacts across many protein families.
554 *PNAS*. 2011;108: E1293–E1301. doi:10.1073/pnas.1111471108

- 555 24. Marks DS, Hopf TA, Sander C. Protein structure prediction from sequence variation.
556 Nature Biotechnology. 2012;30: 1072–1080. doi:10.1038/nbt.2419
- 557 25. Hopf TA, Colwell LJ, Sheridan R, Rost B, Sander C, Marks DS. Three-Dimensional
558 Structures of Membrane Proteins from Genomic Sequencing. Cell. 2012;149: 1607–1621.
559 doi:10.1016/j.cell.2012.04.012
- 560 26. Taylor WR, Sadowski MI. Structural Constraints on the Covariance Matrix Derived from
561 Multiple Aligned Protein Sequences. PLOS ONE. 2011;6: e28265.
562 doi:10.1371/journal.pone.0028265
- 563 27. Sułkowska JI, Morcos F, Weigt M, Hwa T, Onuchic JN. Genomics-aided structure
564 prediction. PNAS. 2012;109: 10340–10345. doi:10.1073/pnas.1207864109
- 565 28. Kim DE, DiMaio F, Wang RY-R, Song Y, Baker D. One contact for every twelve residues
566 allows robust and accurate topology-level protein structure modeling. Proteins: Structure,
567 Function, and Bioinformatics. 2014;82: 208–218. doi:10.1002/prot.24374
- 568 29. Smock RG, Rivoire O, Russ WP, Swain JF, Leibler S, Ranganathan R, et al. An
569 interdomain sector mediating allostery in Hsp70 molecular chaperones. Molecular Systems
570 Biology. 2010;6: 414. doi:10.1038/msb.2010.65
- 571 30. Juan D de, Pazos F, Valencia A. Emerging methods in protein co-evolution. Nature
572 Reviews Genetics. 2013;14: 249–261. doi:10.1038/nrg3414
- 573 31. Schug A, Weigt M, Onuchic JN, Hwa T, Szurmant H. High-resolution protein complexes
574 from integrating genomic information with molecular simulation. PNAS. 2009;106: 22124–
575 22129. doi:10.1073/pnas.0912100106
- 576 32. Jana B, Morcos F, Onuchic JN. From structure to function: the convergence of structure
577 based models and co-evolutionary information. Phys Chem Chem Phys. 2014;16: 6496–
578 6507. doi:10.1039/C3CP55275F
- 579 33. Morcos F, Jana B, Hwa T, Onuchic JN. Coevolutionary signals across protein lineages help
580 capture multiple protein conformations. PNAS. 2013;110: 20533–20538.
581 doi:10.1073/pnas.1315625110
- 582 34. Sutto L, Marsili S, Valencia A, Gervasio FL. From residue coevolution to protein
583 conformational ensembles and functional dynamics. PNAS. 2015;112: 13567–13572.
584 doi:10.1073/pnas.1508584112
- 585 35. Kass I, Horovitz A. Mapping pathways of allosteric communication in GroEL by analysis
586 of correlated mutations. Proteins. 2002;48: 611–617. doi:10.1002/prot.10180

- 587 36. Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, et al. Protein 3D
588 Structure Computed from Evolutionary Sequence Variation. PLOS ONE. 2011;6: e28766.
589 doi:10.1371/journal.pone.0028766
- 590 37. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, et al. Direct-coupling
591 analysis of residue coevolution captures native contacts across many protein families.
592 PNAS. 2011;108: E1293–E1301. doi:10.1073/pnas.1111471108
- 593 38. Jones DT, Buchan DWA, Cozzetto D, Pontil M. PSICOV: precise structural contact
594 prediction using sparse inverse covariance estimation on large multiple sequence
595 alignments. Bioinformatics. 2012;28: 184–190. doi:10.1093/bioinformatics/btr638
- 596 39. Burger L, Nimwegen E van. Disentangling Direct from Indirect Co-Evolution of Residues
597 in Protein Alignments. PLOS Computational Biology. 2010;6: e1000633.
598 doi:10.1371/journal.pcbi.1000633
- 599 40. Park H, Ovchinnikov S, Kim DE, DiMaio F, Baker D. Protein homology model refinement
600 by large-scale energy optimization. PNAS. 2018; 201719115.
601 doi:10.1073/pnas.1719115115
- 602 41. Dunn SD, Wahl LM, Gloor GB. Mutual information without the influence of phylogeny or
603 entropy dramatically improves residue contact prediction. Bioinformatics. 2008;24: 333–
604 340. doi:10.1093/bioinformatics/btm604
- 605 42. Ma J, Wang S, Wang Z, Xu J. Protein contact prediction by integrating joint evolutionary
606 coupling analysis and supervised learning. Bioinformatics. 2015;31: 3506–3513.
607 doi:10.1093/bioinformatics/btv472
- 608 43. Wang S, Li W, Zhang R, Liu S, Xu J. CoinFold: a web server for protein contact prediction
609 and contact-assisted protein folding. Nucleic Acids Res. 2016;44: W361–W366.
610 doi:10.1093/nar/gkw307
- 611 44. Wang S, Sun S, Li Z, Zhang R, Xu J. Accurate De Novo Prediction of Protein Contact Map
612 by Ultra-Deep Learning Model. PLOS Computational Biology. 2017;13: e1005324.
613 doi:10.1371/journal.pcbi.1005324
- 614 45. Morcos F, Schafer NP, Cheng RR, Onuchic JN, Wolynes PG. Coevolutionary information,
615 protein folding landscapes, and the thermodynamics of natural selection. PNAS. 2014;111:
616 12408–12413. doi:10.1073/pnas.1413575111
- 617 46. Carugo O. How large B-factors can be in protein crystal structures. BMC Bioinformatics.
618 2018;19. doi:10.1186/s12859-018-2083-8

- 619 47. Kundu S, Melton JS, Sorensen DC, Phillips GN. Dynamics of Proteins in Crystals:
620 Comparison of Experiment with Simple Models. *Biophysical Journal*. 2002;83: 723–732.
621 doi:10.1016/S0006-3495(02)75203-X
- 622 48. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T. Identification of direct residue contacts
623 in protein–protein interaction by message passing. *PNAS*. 2009;106: 67–72.
624 doi:10.1073/pnas.0805923106
- 625 49. Doruker P, Atilgan AR, Bahar I. Dynamics of proteins predicted by molecular dynamics
626 simulations and analytical approaches: Application to α -amylase inhibitor. *Proteins:*
627 *Structure, Function, and Bioinformatics*. 2000;40: 512–524. doi:10.1002/1097-
628 0134(20000815)40:3<512::AID-PROT180>3.0.CO;2-M
- 629 50. Chasman D, Adams RM. Predicting the functional consequences of non-synonymous single
630 nucleotide polymorphisms: structure-based assessment of amino acid variation11Edited by
631 F. Cohen. *Journal of Molecular Biology*. 2001;307: 683–706. doi:10.1006/jmbi.2001.4510
- 632 51. Alber T, Sun DP, Nye JA, Muchmore DC, Matthews BW. Temperature-sensitive mutations
633 of bacteriophage T4 lysozyme occur at sites with low mobility and low solvent accessibility
634 in the folded protein. *Biochemistry*. 1987;26: 3754–3758. doi:10.1021/bi00387a002
- 635 52. Elahian F, Sepehrizadeh Z, Moghimi B, Mirzaei SA. Human cytochrome b5 reductase:
636 structure, function, and potential applications. *Critical Reviews in Biotechnology*. 2014;34:
637 134–143. doi:10.3109/07388551.2012.732031
- 638 53. Griffin JWD, Bradshaw PC. In silico prediction of novel residues involved in amyloid
639 primary nucleation of human I56T and D67H lysozyme. *BMC Structural Biology*. 2018;18:
640 9. doi:10.1186/s12900-018-0088-1
- 641 54. Gerek ZN, Ozkan SB. Change in Allosteric Network Affects Binding Affinities of PDZ
642 Domains: Analysis through Perturbation Response Scanning. *PLOS Computational*
643 *Biology*. 2011;7: e1002154. doi:10.1371/journal.pcbi.1002154
- 644 55. Tirion MM. Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic
645 Analysis. *Phys Rev Lett*. 1996;77: 1905–1908. doi:10.1103/PhysRevLett.77.1905
- 646 56. Bahar I, Atilgan AR, Erman B. Direct evaluation of thermal fluctuations in proteins using a
647 single-parameter harmonic potential. *Folding and Design*. 1997;2: 173–181.
648 doi:10.1016/S1359-0278(97)00024-2
- 649 57. Fitch WM. Toward Defining the Course of Evolution: Minimum Change for a Specific
650 Tree Topology. *Systematic Zoology*. 1971;20: 406–416. doi:10.2307/2412116

58. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. *Nature Methods*. 2010;7: 248–249. doi:10.1038/nmeth0410-248
59. Nussinov R, Tsai C-J. Allostery in disease and in drug discovery. *Cell*. 2013;153: 293–305.
60. Tsai C-J, Del Sol A, Nussinov R. Allostery: absence of a change in shape does not imply that allostery is not at play. *Journal of molecular biology*. 2008;378: 1–11.
61. Tsai C-J, Del Sol A, Nussinov R. Protein allostery, signal transmission and dynamics: a classification scheme of allosteric mechanisms. *Molecular Biosystems*. 2009;5: 207–216.
62. Swint-Kruse L, Matthews KS, Smith PE, Pettitt BM. Comparison of simulated and experimentally determined dynamics for a variant of the LacI DNA-binding domain, NLaC-P. *Biophysical journal*. 1998;74: 413–421.
63. Tungtur S, Skinner H, Zhan H, Swint-Kruse L, Beckett D. In vivo tests of thermodynamic models of transcription repressor function. *Biophysical chemistry*. 2011;159: 142–151.
64. Eswar Narayanan, Webb Ben, Marti-Renom Marc A., Madhusudhan M.S., Eramian David, Shen Min-yi, et al. Comparative Protein Structure Modeling Using Modeller. *Current Protocols in Bioinformatics*. 2014;15: 5.6.1-5.6.30. doi:10.1002/0471250953.bi0506s15
65. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res*. 2011;39: W29–W37. doi:10.1093/nar/gkr367

Supporting information

S1 Table. List of 139 proteins used in this study.

Proteins used to compute theoretical B-factors based on Seq-GNM and GNM obtained from structure. B-factor correlation values between GNM obtained from structure to experimental (Str GNM-Expt), Seq-GNM to experimental (Seq GNM-Expt), and Seq-GNM to GNM obtained from structure (Seq GNM-Str GNM) are given for two methods (RaptorX, EVfold (EVcouplings)).

678

679 **S2 Table. List of proteins without a known structure with disease related nSNVs**

680 **used for prediction.**

681 Seq-GNM predictions by using a set of 323 nSNVs without a known 3D structure from

682 22 proteins. The protein identifiers, residue numbers, amino acid substitutions Seq-

683 GNM scores, Seq-GNM predictions, and ground truth are provided.

684