Biophysical *Journal*
Article

Biophysical Society

Coevolutionary Landscape of Kinase Family Proteins: Sequence Probabilities and Functional Motifs

Allan Haldane,[1] William F. Flynn,[1,2] Peng He,[1] and Ronald M. Levy[1,*]
[1]Center for Biophysics and Computational Biology, Department of Chemistry, and Institute for Computational Molecular Science, Temple University, Philadelphia, Pennsylvania and [2]Department of Physics and Astronomy, Rutgers, The State University of New Jersey, Piscataway, New Jersey

ABSTRACT The protein kinase catalytic domain is one of the most abundant domains across all branches of life. Although kinases share a common core function of phosphoryl-transfer, they also have wide functional diversity and play varied roles in cell signaling networks, and for this reason are implicated in a number of human diseases. This functional diversity is primarily achieved through sequence variation, and uncovering the sequence-function relationships for the kinase family is a major challenge. In this study we use a statistical inference technique inspired by statistical physics, which builds a coevolutionary "Potts" Hamiltonian model of sequence variation in a protein family. We show how this model has sufficient power to predict the probability of specific subsequences in the highly diverged kinase family, which we verify by comparing the model's predictions with experimental observations in the Uniprot database. We show that the pairwise (residue-residue) interaction terms of the statistical model are necessary and sufficient to capture higher-than-pairwise mutation patterns of natural kinase sequences. We observe that previously identified functional sets of residues have much stronger correlated interaction scores than are typical.

INTRODUCTION

About 2% of the human genome belongs to the protein kinase family and over 10^5 different kinases have been sequenced from many species (1). Protein kinases' common catalytic role in protein phosphorylation is carried out by a conserved catalytic structural motif, but individual kinases are specialized to phosphorylate particular substrates and are bound by different regulatory partners as part of cell signaling networks. Kinases are implicated in many human diseases, and understanding how a particular kinase's sequence determines its individual function has clinical applications. The ability to predict the sequence-dependent effect of specific mutations is relevant for the treatment of kinase-related cancers (2), and understanding the differences in functionality between kinases can aid in selective drug design (3).

One approach to understanding the effects of particular kinase sequence variations has been by structural analysis, based on thousands of observed kinase crystal structures and comparison of their sequences. Patterns of structural variation and conservation within and between protein kinase subfamilies has led to the identification of various functional motifs such as the HRD and DFG motifs necessary for catalysis, networks of stabilizing interactions formed in the kinase active catalytic state known as the C-spine and R-spine, and the importance of the C and F helices in acting as rigid foundations on which the catalytic core rests (4–10). Two conformational states, the catalytically active "DFG-in" and the inactive "DFG-out" states have been discovered to be important in controlling kinase activation and regulation (11). An important goal of these studies is to understand the sequence-dependent ligand-binding properties of different kinases for therapeutic purposes; however, ligand binding affinities are still difficult to predict (12–15), and crystal structures only give a partial view of kinase function.

Another way to extract information about function from kinase sequence variation is to construct a statistical (Potts) model from a multiple sequence alignment (MSA) of sequences collected from many organisms. The idea of using sequence statistics to understand protein structure and function has been motivated and justified by the observation that strongly covarying positions in an MSA correspond well to contacts in structure, a fact used for protein contact

Submitted March 20, 2017, and accepted for publication October 17, 2017.
*Correspondence: ronlevy@temple.edu
Editor: Nathan Baker.
https://doi.org/10.1016/j.bpj.2017.10.028

prediction with significant success (16–21). Using concepts from statistical physics, this idea has evolved and led to the Potts model of protein sequence variation, which is able to capture the pairwise and higher-order mutational correlation patterns, although the model is inferred only from pairwise interaction terms. The Potts model has wider potential applications beyond protein family contact prediction, and can be used to predict sequence-specific properties (22–24). Statistical energies computed using the Potts model can be used to predict the relative probability of any sequence in the family, including sequences not seen in the data set, and can be used to predict the effect of mutations on the probability of a sequence (25–28). The probability is often interpreted as a fitness. The sequence-space landscapes predicted by the Potts model have been found to correlate to experimentally measured fitness landscapes and free energy landscapes (24,29–32). For example, in human immunodeficiency virus (HIV) sequences, Potts statistical energies correlate well with in vitro fitness measurements for tens of sequence variants with multiple mutations relative to the well-defined wild-type sequence (26,33), and Potts models inferred on one HIV sequence database predict sequence frequencies in an independent database (25,34). Similarly, the Potts probability is found to correlate well with measurements of the free energy of folding of proteins in a family (24,29,35–37). This connection between Potts probabilities and fitnesses suggests that the Potts model can be used to predict some features of the relationship between protein sequence and function.

The physical interpretation of the Potts model parameters and the capabilities and limitations of the Potts model are still being explored. Potts model predictions of the effect of mutations in particular sequences have often been limited to a relatively small number of mutations at a time, typically single and double mutants, or in systems with high sequence conservation (29,33). Other studies have shown that higher-than-pairwise variations are well described by Potts models in a number of biological systems; however, these tests were limited to systems with very small, explicitly enumerable state spaces (38–40). Modeling the sequence landscape of the highly diverged protein kinase family is a challenge because kinase sequences have an average of only 30% identity to each other, vary at many positions at once, and cover a vast span of sequence space. In this work, we focus on the model's ability to reconstruct kinase sequence-specific statistics, particularly subsequence probabilities, and illustrate how highly correlated patterns can be associated with functional sets of positions.

We use a previously described Monte Carlo inference method designed to obtain the Potts model parameters for diverse protein families such as the protein kinase family (22). We demonstrate the ability of the inferred model to describe a large sequence landscape by showing that it captures the observed higher-order marginals (subsequence probabilities) of the original MSA, which are not directly fitted. Using in silico tests, we show that when the MSA contains a few thousand effective sequences, the inferred statistical energies of the model are not sensitive to the size of the MSA. Through comparison to site-independent (uncorrelated) models of sequence variation, we show the that epistatic effects of correlations are essential to accurately predict higher-order marginals, i.e., subsequences that vary at many positions simultaneously. We show how well the statistical energies of the Potts model for the kinase family reflects the frequency of subsequences observed in the Uniprot database and in the much larger data set constructed in silico. We then use the subsequence statistics predicted by the Potts model to illustrate how highly correlated patterns can be associated with functional motifs, and to identify motifs within the kinase sequence with strong correlated signals. We illustrate how functional units of kinase family proteins are more conserved and exhibit strong epistatic effects.

Potts covariation analysis

Potts covariation analysis models the distribution $P(S)$ for the probability of observing a sequence in an MSA of a protein family, incorporating pairwise correlated effects to parametrize the model. $P(S)$ has been interpreted as a fitness, and sometimes as the probability of the protein's native fold in thermodynamic equilibrium (24,29,41–43). Because of the enormous size of sequence space (roughly estimated to be 10^{140} sequences for the kinase family in Supporting Material), this distribution cannot be directly measured from an MSA of only a few thousand sequences. An alternative is to solve for the maximum entropy distribution, subject to the constraints that the univariate and bivariate marginals $f^{ij}_{\alpha\beta}$ of sequences generated from the model (for residues α, β at positions i, j) match those of the MSA data set, which can be accurately measured. The maximum entropy distribution is found to be $P(S) \propto e^{-E(S)}$ for the Potts Hamiltonian $E(S) = \sum_i^L h^i_{S_i} + \sum_{i<j}^L J^{ij}_{S_iS_j}$, which contains pairwise "coupling" terms J and single-site "fields" h, which may be solved for by maximum likelihood inference. One could in principle build a Hamiltonian that includes higher-order terms by fitting triplet correlations in the data, but not only is there insufficient data to build such a model, it does not appear to be necessary as we discuss below.

Given a parametrized model, the Hamiltonian $E(S)$ defines a statistical energy landscape over sequence space, computable for any sequence such that lower values are more favorable, and the coupling parameters $J^{ij}_{S_iS_j}$ give information about the statistical interaction between two residues in a sequence. The $J^{ij}_{S_iS_j}$ have been related to folding or binding free energy contributions (43,44). From $P(S)$ we can estimate the probability of any sequence, and by similar computation we may also predict the probability of subsequences in particular (sub)sets of positions (not necessarily contiguous) of the MSA. The Potts model allows us to explore aspects of the statistics that we do not have the power to measure from the raw data because of sample size. For instance, given a data set of N sequences, it is not possible to directly measure the probability of a (sub) sequence that appears in nature with frequency of roughly $1/N$ or less. This sampling noise (or "shot noise") issue is particularly a problem for longer sets of positions and for the full-length sequences because the probability of individual subsequences decreases rapidly with increasing number of positions due to the increased size of the sequence space. The correlated nature of the model is also important. The collective effect of the pairwise terms J mean that the statistics of the Potts model can be significantly different from a site-independent or uncorrelated model that ignores correlated effects, particularly for longer sets of positions where more pairwise terms come into play. We will compare the Potts model to the maximum entropy independent model fitted to the univariate marginals of the data, which is exactly solvable and takes a "log odds" form where $h^i_\alpha = -\log f^i_\alpha$ and $J = 0$.

Inference of the Potts model parameters is nontrivial. The Potts landscape has primarily been used for the purpose of protein structure contact prediction, and the approximations and algorithms developed to solve for the parameters J have mostly been tailored for this application (19,36,45–49). For the purpose of understanding kinase sequence variation, the distribution $P(S)$ itself is more central, and more accurate inference techniques are necessary to model this distribution as illustrated in a recent benchmark (43). For this reason, we use a Monte Carlo inference technique that makes fewer approximations (22).

METHODS

In this study, we focus on the statistical properties of a Potts model for the kinase family. We use a Potts model and kinase sequence data set that we have previously prepared using methods of parameter inference, MSA preprocessing, alphabet reduction, interaction scoring, and Protein Data Bank (PDB) contact analysis described in (22). These methods are recapitulated in additional detail for this study below. In the current study, we additionally develop methods to analyze the subsequence statistics of this Potts model.

Potts model inference

We use Markov Chain Monte Carlo (MCMC) methods to perform the Potts parameter inference, a method developed in previous studies (33,36,50). Our implementation is based on the one described in reference (33). This method makes few analytic approximations such as the weak-coupling approximation used in mean-field methods (49), approximate likelihood functions (19), or truncated cluster entropies (51), at the expense of increased computation time. We compare our results to mean-field methods below. In the MCMC method, we generate sequences from the model according to the equilibrium distribution $P(S)$ by MCMC, given a trial set of couplings J, and update the parameters J based on the discrepancy between the model and data set bivariate marginals. Our graphic processing unit-based implementation decreases the computation time, and also allows efficient generation and analysis of the large simulated MSAs used in this study. A description of the MCMC algorithm is provided in the Supporting Material of (22). Convergence of the parameters is shown in Fig. S8.

MSA preprocessing

We obtain kinase sequences using HHblits (52) to search the Uniprot database starting from the Pfam kinase family seed (PF00069). We remove any sequences with gaps in the "HRD" or "DFG" triplets, sequences missing the aspartic acid required for Mg^{2+} binding, more than 10 gaps, more than 40 inserts, or with invalid/unknown amino acids, leaving 127,113 sequences of length 241. These sequences are phylogenetically related and sampled with experimental biases, and therefore do not represent independent samples from the distribution $P(S)$. We correct for this as described in (49) by downweighting similar sequences. We assign a weight $w = 1/n$ to each sequence, where n is the number of sequences in the alignment with $>$ 60% sequence identity to it. This cutoff was chosen based on analysis of the distribution of pairwise sequence similarities in the kinase data set (see Supporting Material). This leaves an effective number of sequences $N_{\text{eff}} = \sum w$ of 8149. We then trim the first 5 and last 61 positions from the alignment that contain variable secondary structures, leaving 175 positions.

Alphabet reduction

We reduce the alphabet size q from 21 residue types (20 amino acids plus gap) to 8 in a way that preserves the correlation structure of the MSA, unlike amino acid reduction schemes based on physiochemical properties (53,54). For each position (processed in random order) we merge the pair of letters that gives the best least-squares fit between the $\binom{L}{2}$ Mutual Information (MI) scores across all position pairs of the MSA in the eight-letter and 21-letter alphabets. MI is a measure of correlation strength between two MSA columns i,j, given by $\text{MI}^{ij} = \sum^q_{\alpha\beta} f^{ij}_{\alpha\beta} \log f^{ij}_{\alpha\beta}/f^i_\alpha f^j_\beta$ (55). This merging is repeated until all positions have been reduced to eight letters. In practice, this procedure often first merges the very low-frequency residue types at a position into a single "mutant" residue. After computing bivariate marginals from the weighted eight-letter sequence set, we add a small pseudocount of roughly $1/N$ as a finite size correction.

Alphabet reduction has the benefit of eliminating many small marginals (rare residue types) from the system and thus decreases the computational cost of inference, which scales as q^2. For the kinase MSA, we find that reduction to eight letters is a suitable compromise between reducing the problem size and preserving the sequence correlations (Fig. S1 *B*), and captures almost all the sequence variation; kinase sequences in our data set have 27% average pairwise identity with 21 letters but still only 31% identity after reduction to eight (Fig. S1 *A*). Further justifying this choice, the mean effective number of amino acids at each position of our raw data set is 8.9, computed by exponentiating the site entropy (see Supporting Material). The Pearson correlation between the 21-letter and eight-letter MI scores is 0.97.

Interaction score: weighted Frobenius norm

A number of different methods have been suggested for obtaining a position pair interaction score from the Potts model parameters, including the "Direct information" (45), Frobenius norm (19), and average product corrected Frobenius norm (56). To control and reduce the contribution of marginals with high sampling error, we score interactions using a weighted Frobenius norm computed as $I^{ij} = \sqrt{\sum_{\alpha\beta}(w^{ij}_{\alpha\beta}J^{ij}_{\alpha\beta})^2}$ where $w^{ij}_{\alpha\beta} > 0$ are tunable weights. In the case where the weights $w^{ij}_{\alpha\beta} = 1$, this reproduces the unweighted Frobenius norm calculation. Both the Frobenius norm and weighted Frobenius norm depend on the choice of "gauge" of the model, referring to the fact that the Potts model described above with $\binom{L}{2}q^2$ couplings contains superfluous parameters, such that compensatory transformations of the $J^{ij}_{\alpha\beta}$ parameters can leave the distribution $P(S)$ unchanged. In fact, there are only $\binom{L}{2}(q-1)^2+L(q-1)$ independent parameters, fitted based on an equal number of independent marginals. These gauge transformations have been described in other publications (45,49,51). Typically, the Frobenius norm is computed in the "zero-mean" gauge, which minimizes the Frobenius norm and guarantees that uncorrelated positions have an interaction score of 0. For the weighted Frobenius norm, we instead transform the model to a gauge that satisfies the gauge constraint $\sum_\alpha w^{ij}_{\alpha\beta}J^{ij}_{\alpha\beta}=0$, which similarly minimizes the weighted norm. To downweight the influence of couplings corresponding to infrequently observed mutant pairs that have high sampling error, we heuristically choose $w^{ij}_{\alpha\beta}=f^{ij}_{\alpha\beta}$, which gives good correspondence between the interaction score and observed contacts in crystal structures (see Fig. S5; Supporting Material).

PDB contact frequency analysis

To measure contact frequencies in the kinase DFG-out and DFG-in conformational states, we obtain 2896 kinase structures from the PDB classified into the DFG-in and DFG-out state collected as described in a previous publication (22) and aligned them to our kinase MSA. A contact is defined as a nearest heavy-atom distance between two residues of less than 6 Å. See reference (22) for further details. When compiling statistics of the residue identities in the sequences of the PDB data set, the sequences are weighted to account for similarity at a 10% similarity threshold after applying the method described above for MSA preprocessing.

In silico sequence data set

We generate our main in silico data set by sampling from the kinase Potts Hamiltonian by MCMC. To roughly simulate the effect of the phylogenetic corrections, we take sequence samples after only a short interval of 175 MCMC steps, giving a nonindependent set of sequences. We then apply the phylogenetic filter at 40% identity, giving 9990 effective sequences. We infer a new in silico set of Potts model parameters using this in silico data set as input, which may differ from the original kinase model due to the effects of finite sampling, phylogeny, and other potential sources of error.

Estimating subsequence frequencies

To test the Potts model's ability to describe the probability of variations over many positions, we need to estimate the frequency of subsequences (higher-order marginals) predicted by the model. We use two methods to do this. For shorter sets of positions with $L \leq 10$, we generate a large in silico MSA of 4×10^6 sequences by Monte Carlo sampling of the kinase Potts Hamiltonian and simply count the subsequence frequencies. For longer sets of positions, this method is insufficient because the probability of generating a particular subsequence falls far below $1/10^6$. Instead, we use a reweighting procedure that allows us to compute relative subsequence frequencies from a generated in silico MSA even if the subsequence does not appear in it. The procedure is described next.

Dividing the MSA into a set of positions whose subsequence probabilities we wish to estimate and a remainder set of "background" positions, the equilibrium probability of a subsequence A is given by $f_A=\sum_b p^A_b$, where p^A_b is the Potts probability of a sequence with background b and subsequence A. Since $p^A_b=e^{-E^A_b}/Z$ and $p^B_b=e^{-E^B_b}/Z$ for subsequence B at the same positions, we can also write $f_A=\sum_b p^B_b e^{E^B_b-E^A_b}$. It follows that given a large enough equilibrium sample of sequences $\{S\}$, we can approximate the frequency of subsequence A as $f_A \propto \sum_S e^{E_S-E_{A\rightarrow S}}$, where E_S is the Potts energy of sequence S and $E_{A\rightarrow S}$ is the energy after substituting subsequence A, up to an unknown normalization constant. The ratio of subsequences frequencies, e.g., f_A/f_B, can then be unambiguously obtained as the unknown normalization factor cancels. This approximation becomes exact in the limit of large in silico MSAs, and should be valid as long as the distributions of sampled backgrounds for each subsequence, with the subsequence held fixed, would overlap significantly with each other. Using an in silico MSA of size 4×10^6, we confirm that this approximation is accurate, first for shorter subsequences tested for lengths 2–10 by comparing the frequency predicted by this method to the counted frequency in the raw MSA, and second for longer subsequences of length $L-8$ to L (i.e., those with short backgrounds), by comparing to the exact frequencies computed by enumerating the backgrounds b and summing Potts probabilities as $f_A \propto \sum_b e^{-E^A_b}$.

RESULTS AND DISCUSSION

We infer a model for a data set of $N=8149$ effective kinase sequences of length 175, and quantify the quality of fit through the sum of squared residuals (SSR) of the bivariate marginals. Due to the finite sample size, there is error in each measured bivariate marginal f around its true (unknown) value, and due to this error we estimate an expected SSR of 1.69 between the data set marginals and the (unknown) true marginals. This estimate is obtained by summing over the expected binomial variances of each bivariate marginal of $f(1-f)/N$ (approximating the observed bivariate marginals as independent), and we also confirm this by generating MSAs of size 8149 from the inferred model and comparing these MSA's SSR relative to the model's marginals. The SSR between the inferred Potts model's marginals and the observed MSA's marginals is close to 1.69, which suggests that the inferred model approximates the "true" bivariate marginals as well as finite sampling effects allow. In contrast, the SSR of 36.4 between the independent model and the data set is much larger. This shows that the independent model must have significant error in addition to finite sampling error, and demonstrates the importance of modeling correlated effects.

Probability distributions of kinase subsequences

Although the Potts model is fitted to the bivariate marginals of a data set of $N \sim 10^4$ sequences, it is able to capture higher-order marginals of the data set involving simultaneous variation at many positions. To test this, we would ideally directly compare predicted higher-order marginals (equivalent to subsequence probabilities) to the corresponding frequency observed in an MSA. However, the "shot noise" effect makes this impossible for long sequences, as the probability of seeing an individual kinase sequence of length 175 is always many orders of magnitude smaller than $1/N$ (the smallest observable frequency). We may nevertheless verify the Potts model predictions by examining shorter sets of positions whose MSA statistics can still be measured with reasonable accuracy given the sample size of the data set, but long enough that they encompass a large sequence space. To quantify model error for a set of positions, we measure the Pearson correlation r_{20} between the frequency of the top 20 subsequences most frequently observed at those positions in the kinase family MSA to the probability predicted by the Potts model. We estimate the Potts probability of a subsequence from a generated MSA of 4×10^6 sequences as described in the Methods. We use the top 20 subsequences for each set of positions in this comparison because the remaining rarer subsequences have high sampling error.

For sets of positions up to about length 10 for which there are sufficient statistics to test the model, the Potts model correctly predicts the observed frequencies and the independent model performs very poorly. In Fig. 1 *A* we illustrate subsequence frequencies for a specific set of seven positions associated with the DFG-in versus DFG-out conformational transition, described in more detail in another section. The r_{20} score for this set is very high (0.94), which means that the predicted probabilities of the subsequences (seventh-order marginals) agree very well with the corresponding frequencies observed in the data set MSA. In contrast, there is essentially no r_{20} correlation with the independent model.

We verify this more generally by choosing 1000 random sets of positions of length 2–10 from the 175 positions of the full sequence, and compute r_{20}, as shown in Fig. 1 *B*. We also compute the expected r_{20} due to finite sampling alone, by comparing subsequence frequencies in a synthetic MSA of size 8149 generated by the Potts model to those predicted by a second Potts model fitted to this synthetic MSA, estimated from a sample of 4×10^6 sequences, shown as a dashed line. The Potts model r_{20} correlation decreases for increasing subsequence length, but it closely follows the expected r_{20} due to finite sampling, which shows that this decrease reflects the increasing statistical error in the finite sample data set observed marginals rather than increasing error in the model. Furthermore, the r_{20} between the Potts model and the data set is entirely accounted for by the finite sample size of the data. If third- or higher-order terms affected the subsequence frequency distributions for the lengths we tested, on average, this could manifest as additional error in the model past that we observe due to the finite sample size of the reference data set. For instance, in Fig. 1 *B* one analogously sees how the lack of second-order terms in the independent model manifests as additional error of this model relative to the finite sample estimate. The absence of such additional error in the Potts model estimates suggests that higher-than-pairwise terms do not play a significant role here. This is a striking result. We note, however, that absence of evidence is not necessarily evidence of absence: It remains possible that a large number of weaker higher-order interaction terms have a small effect for subsequences with $L < 10$, but a greater effect for larger L.

Nevertheless, these observations support an interpretation that the collective effects of the pairwise terms of the Potts model are necessary and also sufficient to predict higher-order statistics (marginals) of the data set. The fact that the Potts model captures the higher-order marginals of the data set significantly beyond the pair marginals (up to

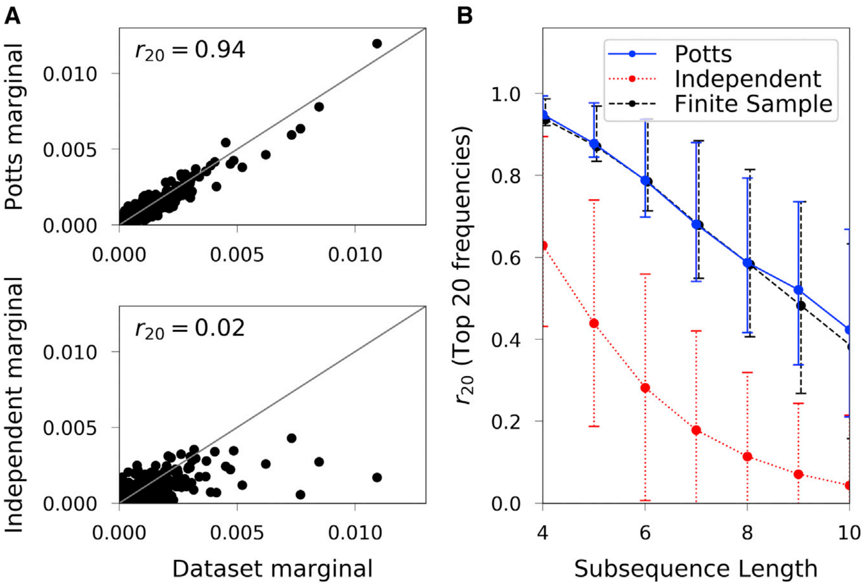


FIGURE 1 Subsequence frequency predictions. (*A*) Predicted subsequence frequencies for a set of seven positions known to be important for kinase activity, compared to the data set frequencies. The Potts distribution (*top*) models the observed distribution well, in contrast to the independent model (*bottom*). (*B*) Average correlation between observed and predicted frequencies for the top 20 subsequences for large samples of subsequences of varying length, for observed subsequence frequencies with the Potts model (*blue*), and with the independent model (*red*, *dotted*). Circles show the means, and error bars show the range of first to third quartile values (25–75% of sets of positions). The dashed line (*black*) is an estimate of the expected correlation due only to finite sampling, computed by comparing the subsequence frequencies of a finite synthetic data set MSA of size 8149 to the frequencies of a large MSA of 4×10^6 sequences generated from a Potts model fitted to the synthetic MSA of size 8149. Both the trend and range of the expected correlations due to the effects of the sample size (8149) are consistent with the correlation between the observed frequencies and those predicted by the Potts model. To see this figure in color, go online.

10th order, see Fig. 4) which it fits directly supports its use in predicting properties of the sequence space landscape.

In silico tests: shot noise and importance of pairwise terms

To further demonstrate the ability of the model to describe large sequence spaces for longer sets of positions with $L > 10$, we perform in silico tests to show that the statistical energies of sequences are not strongly affected by finite sampling effects given a sequence sample size of $\sim 10^4$. For longer sequence lengths it is conceivable that the effects of sampling noise in the data or inference errors become more pronounced, as the number of pairwise terms J used in the computation of statistical energy grows quadratically in L. To test this, we generate an in silico data set MSA consisting of 9990 effective sequences generated from the original Potts model as described in the Methods, to which we fit a new in silico Potts model, and then compare the two models. The in silico data set represents a finite resampling process that scrambles small bivariate marginals that have large relative error, and serves to demonstrate that the inferred model is not sensitive to their precise values.

We first examine subsequence statistics of longer position sets. In sets longer than length 10, the subsequence frequencies become minute and cannot be measured even by generating simulated MSAs of up to $\sim 10^6$ sequences, but instead we are able to compute their relative frequencies using an algorithm described in the Methods. We compute these frequencies using both the original Potts model parameterized on the kinase family MSA and the in silico Potts model, and take the logarithm, giving an effective statistical energy of each subsequence for both models. We find that for subsequences from length 4–175, the two Potts models agree with an average correlation of 0.9 in statistical energy (Fig. S2). The independent model, in contrast, predicts statistical energies for short subsequences of length 4 with similar correlation, but as position set length increases its power drops dramatically, and for sequences of length 128 it has no predictive power ($r = 0.08$).

The importance of correlated effects is most pronounced for full kinase sequences varying over all 175 positions. In Fig. 2 *A* we compare the statistical energies of the 127,113 kinase family sequences of our unweighted data set computed using the in silico Potts model with those computed using the original Potts model (the "reference energy"), finding a Pearson correlation of $r = 0.92$. Most of the sequences in this plot are highly dissimilar from the 9990 effective in silico sequences used to parametrize the in silico model. On average, a sequence in the in silico data set has only 52% sequence identity to its most similar sequence in the unweighted kinase data set, and 31% similarity on average to the whole data set, demonstrating the Pott model's ability to model variation far in sequence space from the sequences it is parametrized with. In contrast, the independent model is unable to predict statistical energies, showing no correlation $(r = 0.05)$ between its predictions and the original statistical energy values (Fig. 2 *B*). Most dramatically, sequences predicted to be lowest probability (high statistical energy) in the independent model include some of the highest probability (low statistical energy) sequences predicted by the Potts model. These are sequences with multiple rare mutations that the independent model necessarily assigns a low probability, but which the Potts model predicts are very favorably coupled. These results strongly support the importance of the correlated terms and show that they become necessary for predicting statistics of full sequences with many mutations.

This test does not probe whether triplet and higher-order terms in the Hamiltonian are needed to predict full sequence probabilities because the in silico data set MSA is generated from a pairwise model. However, the lack of a need to parametrize higher-order terms in the Hamiltonian is justified by the results of the previous section for $L < 10$. The in silico tests do show that, given such a pairwise model, the statistical energy predictions for sequences with $L > 10$, for which sampling error is more significant, are robust given an MSA of thousands of sequences, and the independent model is grossly inadequate.

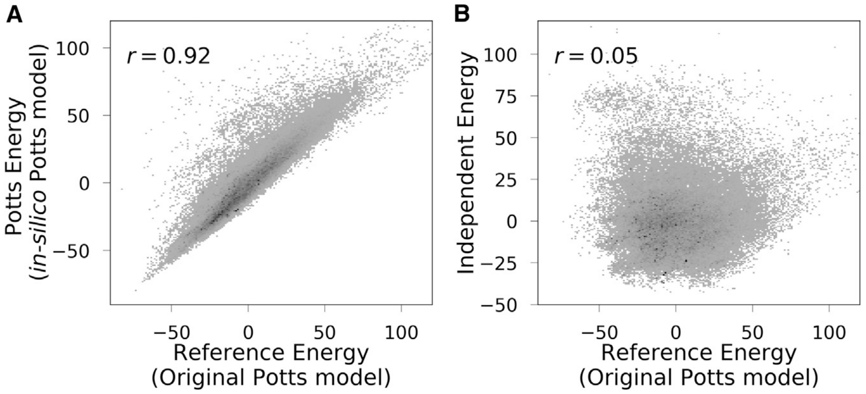


FIGURE 2 Statistical energies computed for kinase sequences taken from Uniprot. (*A*) Statistical energies computed using the original Potts model compared to those computed using a Potts model refitted to a finite size in silico sample of 9990 effective sequences generated from the first model, and (*B*) computed using the original Potts model compared to the those computed using an independent model fit to the in silico sequences. Lower energies are more favorable. The darkness of a plotted point reflects the log of the number of sequences at that point, and most sequences are concentrated near the center of the distribution.

Although we have focused on the kinase family, we expect these results generalize to other protein families. We have also analyzed the trypsin and photoactive yellow protein families using the same methods as for the kinase family, and obtain similar results (see Fig. S4). We also analyze the kinase in silico data set using the more approximate mean-field methods for parameter inference. We find that the correlation between the energies computed with this model and those computed with the original model is 0.7, compared with 0.92 found by MCMC (Fig. S7, compare to Fig. 2 *A*).

Identifying highly correlated sets of mutations and functional motifs

Statistical energies calculated from the Potts model can be used to investigate kinase function, and allow us to probe statistics not measurable from the data alone because the finite size of the MSA prevents direct measurement of the frequencies of subsequences or full sequences. As an example, we examine subsequence statistics of particular kinase position sets, and investigate how functional sets of positions (motifs) have strong correlated interactions contributing to their statistical energy, and can be identified because their marginals are more accurately predicted by the Potts model than by the independent model as measured by the r_{20} scores.

To use the r_{20} scores in this way, it is useful to understand that r_{20} scores for a particular position set, in either the independent model or the Potts model, can be lower (reflecting poorer model-data correspondence) due to two different effects. First, due to inaccuracy of the model itself (i.e., due to ignoring correlations), and second, due to sampling error (shot noise) in the data set used as the benchmark for the model predictions, due the finite size of the data MSA.

The degree of model inaccuracy depends on the nature of the correlated interactions within the set of positions. If the "true" Hamiltonian describing the MSA involves higher-order terms than those included in the model (e.g., third-order terms) this will lower the r_{20} score for the Potts model, particularly for sets of positions in which the higher-order interactions contribute significantly to the statistics. At least for $L < 10$, our results above suggest that these terms are not important. The independent model does not include second-order terms, so we expect it to perform more poorly for motifs that have functional constraints and therefore correlation is expected to be important. We expect highly correlated (potentially functional) sets of positions to have higher r_{20} score with the Potts model than with the independent model.

Data set sampling error, on the other hand, will often be smaller in functional motifs because they have greater conservation. The sampling error for a subsequence of frequency f can be modeled as the binomial distribution SD $\sqrt{f(1-f)/N}$ for MSA size N. For small f, the relative error in a statistical energy (obtained by dividing by f) is approximately $\sqrt{1/Nf}$, meaning that higher-frequency subsequences have lower relative statistical error. Highly conserved sets of positions, whose statistics are dominated by a small number of high-frequency subsequences, will therefore have lower sampling error as measured by r_{20}. We expect more highly conserved sets of positions to have higher r_{20} scores with both the Potts model and the independent model.

These observations suggest that we can identify strongly correlated motifs by comparing the r_{20} statistics using the Potts model with the corresponding results for the independent model. A high r_{20} score for the Potts model and a low score for the independent model is a sign that the set of positions is more conserved and more correlated than typical, suggesting that it may be an important functional motif.

Previously identified functional set of positions has high correlation

We first examine a motif of length 7 formed from a set positions previously identified in the literature to control kinase function by structure-based analysis (11), illustrated in Fig. 1 *A*. Its Potts $r_{20} = 0.94$ score is much higher than the typical score for sequences of the same length ($\overline{r}_{20} = 0.65$, see Fig. 1 *B*), yet the independent model's $r_{20} = 0.02$ is much lower than is typical ($\overline{r}_{20} = 0.18$). These are the positions 24, 42, 67, 112, 113, 115, and 127 in our alignment, which correspond to PDB residue indices K72, L95, M120, L167, K168, E170, and V182 for the protein kinase A PDB: 2CPK (57), as tabulated in Tables S1 and S2. These seven residues are highlighted in the kinase structure in Fig. 3 *A*. Residues 112, 113, 115, and 127 form a small subgroup anchoring the catalytic loop, and 24 (known as the β-3 Lysine), 42, and 67 (the gatekeeper residue) form a group on the opposite side of the DFG motif. This example motif demonstrates how sets of positions identified to be important structurally are also found by examining the Potts sequence statistics, and both conservation and correlation are important in the statistics of functional motifs.

The Potts model also gives us insights into the important interactions among these residues. The high Potts interaction scores (see Methods) between pairs of these residues suggest that position pairs 112–127 and 113–115 interact strongly, and that the gatekeeper (67) and position 42 on the α-C helix also have a moderate-to-strong interaction. Position 112 is an important residue known to anchor the N-terminal of the catalytic loop to the F-helix (11), whereas position 127 is in the β-8 loop at the N-terminal of the activation loop. The strong Potts interaction score between 112 and 127 suggests a, to our knowledge, new interpretation that the start of the activation loop is indirectly anchored to the F-helix through the intermediary residue 112, thus stabilizing the activation loop. Positions 113 and 115 are known to be involved in catalysis and substrate binding,

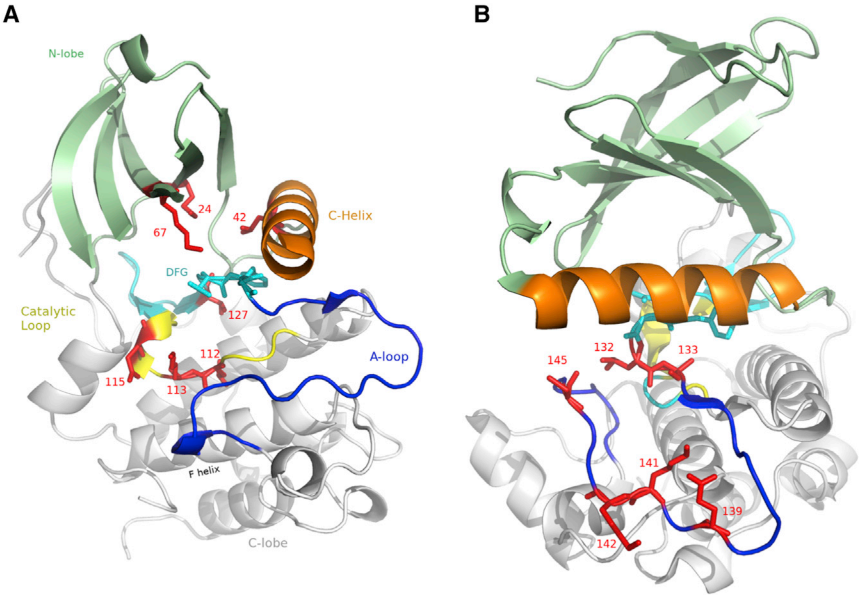


FIGURE 3 (*A*) Seven positions (*red*) identified as important for kinase function in previous literature based on structural analysis shown in crystal structure (PDB: 2CPK), which we identify to be a highly correlated motif. The C-lobe (*white*) and N-lobe (*light green*) are shown with the A-loop in blue, the DFG motif and β-7-8 loops in cyan, catalytic loop in yellow, and the α-C helix in orange. The seven positions are shown in red with their alignment index, (*B*) six positions in the activation loop identified to form a correlated motif (*red*), and other colors as in (*A*) (PDB: 2YAC, in the DFG-in state). Residue numbers correspond to positions in our alignment, and map to PDB residue indices as listed in Tables S1 and S2. To see this figure in color, go online.

respectively (11). The predicted interaction between the gatekeeper (67) and position 42, a residue in the α-C helix and part of what is called the "hydrophobic spine," supports previous results suggesting that the gatekeeper can stabilize this spine (58) and anchor the α-C helix, whose positioning is important for catalysis. The Potts model recapitulates previously identified interactions between important residues, but also suggests, to our knowledge, new interactions among them.

Correlated motif within the activation loop

We next investigate functionality of the activation loop. It is well known that the activation loop conformation consisting of ~23 residues is important in controlling kinase activation and signaling (22,56). Phosphorylation of residues in the activation loop causes kinase activation in vivo. The activation loop has different conformations in different functional states (e.g., active, src-like inactive, and DFG-out inactive) and the residues are intricately coupled (Fig. 3 *B*). In the active state, this loop becomes more structured and stabilizes the catalytic residues in preparation for catalysis, and forms more extensive contacts with other parts of the protein. An important catalytically inactive state is known as DFG-out, in which the activation loop becomes more flexible and frequently cannot be resolved in DFG-out crystal structures, and forms more intraloop contacts and fewer contacts with the rest of the protein. This conformation has clinical significance because certain inhibitors stabilize the DFG-out state, rendering the kinase inactive.

To investigate networks of interactions within the activation loop that are likely to contribute to kinase function, we searched for sets of positions within the loop with the largest differences in r_{20} between the Potts and independent models. These correspond to motifs that are both more conserved and more correlated than observed on average for subsequences of that length, leading us to a motif of six positions, with statistics shown in Fig. 4 and structure shown in Fig. 3 *B*. To understand the possible functional significance of these residues, we investigated whether they are related to the DFG-in and DFG-out conformational transition, by comparing the interaction scores for pairs of these residues to contact frequencies in the DFG-in and DFG-out conformations measured from a set of 4129 PDB structures, shown in Fig. 5. We find that out of the 15 possible pair interactions, six of these have high interaction scores

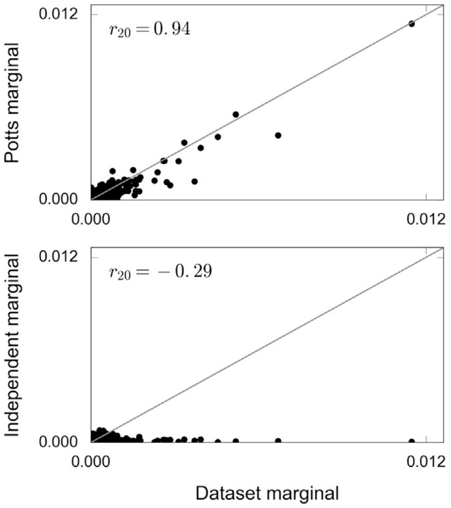


FIGURE 4 Observed and predicted marginals for a set of six positions in the activation loop. Top: Potts model predictions. Bottom: Independent model predictions. The slightly negative correlation coefficient means that the independent model predicts a low frequency for some of the most frequent subsequences observed in the data, an effect already noted in the discussion of Fig. 2 *B*.

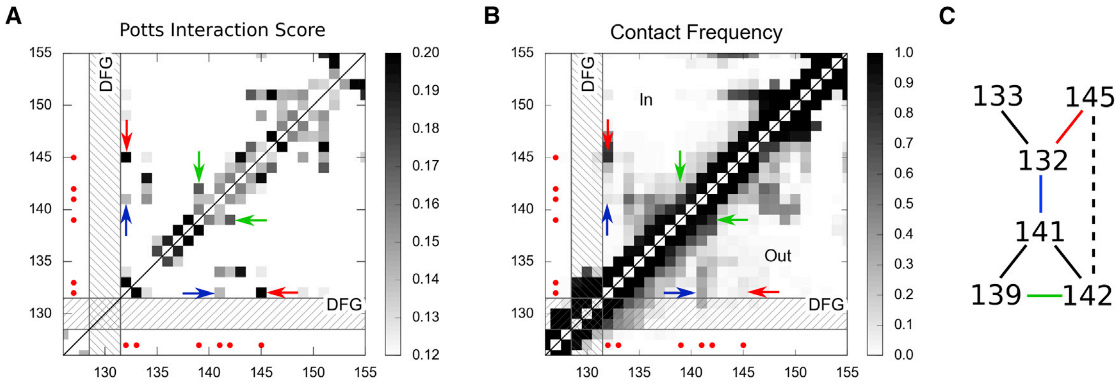


FIGURE 5 Interaction map and Contact map focusing on the activation loop region. (*A*) Potts Interaction Score map. The activation loop spans positions 132–151, and is preceded by the DFG motif (which is shown as a *hashed area*). Position pairs are shaded by their interaction score (see Methods). The six-residue motif identified as highly correlated is marked by red points, and three particular interactions are pointed out with colored arrows: the pair 132,145 (*red*) is an interaction in the DFG-in state only, and 132,141 (*blue*) and 139,142 (*green*) are interactions in the DFG-out state only. (*B*) Contact frequency map constructed from analysis of the PDB database. The upper triangle shows pair-contact frequency (6 Å closest heavy atom-atom cutoff) in DFG-in conformations, and lower triangle in DFG-out conformation. (*C*) Network interaction structure of the six-residue motif, showing a link for pairs with high interaction score, or a dotted line for intermediate to weak interaction score. To see this figure in color, go online.

above a cutoff that is used to distinguish contacts during contact prediction. They are connected together, forming a network illustrated in Fig. 5 *C*.

We further investigate the residue pair 132–145, as it has a very strong Potts interaction score and forms a contact in 81% of DFG-in structures and only 8% of DFG-out structures, yet to our knowledge has not been previously identified as functionally important. Position 132 is the DFG + 1 residue, and position 145 is a residue closer to the C-terminal end of the activation loop. An example crystal structure in which this pair is in contact in the DFG-in state (PDB: 2YAC) is shown in Fig. 3 *B*. To better understand why this interaction may be important, we examine kinase structure and sequence statistics for sequences observed in our sequence data set and in the DFG-in or DFG-out state in the PDB. Interactions between a leucine or phenylalanine at position 132 and cysteine at position 145 are present in ~20% of DFG-in structures, and none of the DFG-out structures. The “LC” residue combination also gives one of the most positive $C^{ij}_{\alpha\beta}$ correlations from among the 64 possibilities for this position pair in the kinase alignment, as well as one of the most positive contributions to the interaction score (see Fig. S6, *D* and *E*). The “FC” residue combination behaves similarly. Crystal structures involving these interactions show that the LC and FC residue combinations often form a hydrophobic interaction, and that the more polar cysteine is more solvent exposed and shields the L or F from solvent. Other residue combinations more prevalent in the DFG-in sequences similarly involve hydrophobic residues (see Fig. S6 *B*). In total, the 132–145 pair appears to form interactions that stabilize the DFG-in state, based on Potts model scores and crystal structure conformations.

This example illustrates first how functional motifs within a protein might be identified, and second how the Potts model can help suggest the biophysical basis for the functional role of the motif. In future work, we will develop more systematic methods of identifying functional groups of residues. Previous studies have shown how covariation-based techniques can give information about protein architecture and groups of coevolving residues, which have been termed “protein sectors” (59,60). Our present results suggest that the Potts model may be used in a similar way, in addition to accounting for the collective effects of many pairwise interactions at once.

CONCLUSIONS

The protein kinase catalytic domain is one of the most abundant domains across all branches of life. Although kinases share a common core function of phosphoryl-transfer, they also have wide functional diversity, which is primarily achieved through sequence variation. In this study, we use a statistical inference technique to build a maximum entropy coevolutionary Potts Hamiltonian model of sequence variation in the kinase protein family. Our results show that the kinase sequence statistics (higher-order marginals) calculated with a Potts model containing only two-body interactions in the Hamiltonian, and inferred using the MCMC algorithm as we have done, recapitulate the observed marginals for the kinase family up to the observable limit imposed by the shot noise effects inherent in the data because of the sample size. The higher-order marginals (beyond bivariate marginals) are not fitted.

We have shown that the pairwise terms of the Potts model are necessary, and also appear to be sufficient, to model the kinase sequence landscape, particularly for the purpose of modeling the higher-order marginals. The discrepancies we observe between the kinase family subsequence probabilities predicted by the Potts model, with only pairwise terms and the observed subsequence frequency counts in the MSA, can be accounted for by the finite size effect of the MSA. Other groups have explored how shot noise can affect the univariate and bivariate marginals, and individual coupling parameters in toy models (46,61), and we have previously studied the effects of finite sampling for Potts models fitted to HIV sequence data (40,62). Here, we examine how shot noise affects the prediction of subsequence probabilities and the statistical energies of full sequences, using real data from the kinase protein family.

Although the finite size of the kinase sequence database and MSA constructed from the database places a limit on the ability of the Potts model to recapitulate the statistics of the higher-order marginals actually observed in the sequence database, it has only a small effect on the statistical energies of the Potts model itself. To show this, we carried out an in silico test. In this test, we used our Potts model of the kinase family to construct an in silico MSA data set, of size ~8000 sequences. This in silico MSA has only 31% sequence similarity to the original MSA that we generated from the Uniprot database. We then parameterized a new Potts model from the set of in silico kinase sequences, and showed that the scoring of Uniprot sequences with the new Potts model was highly correlated with that of the original Potts model (see Fig. 2 *A*).

We propose that kinase family protein functional motifs may be identified as sets of positions where the sequence covariation is much more correlated than is typical for subsequences of that length, and which also exhibit larger than average sequence conservation. Those two criteria can be quantified by identifying sets of positions where the Potts statistical energies are much more favorable than the average Potts statistical energy of a marginal of that same length, whereas the statistical energy of the independent model is much less favorable than the average. We have shown how a set of previously identified functional residues have higher correlation and conservation than typical random sets of positions, and we have also identified a highly correlated and conserved set of positions in the activation loop, which is potentially important in controlling activation loop function. We hope that Potts models used in this and similar ways will help increase our understanding about the deep connections between protein sequence covariation on one hand, and protein structure and function on the other.

SUPPORTING MATERIAL

Supporting Materials and Methods, eight figures, and two tables are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(17)31449-9.

AUTHOR CONTRIBUTIONS

R.M.L and A.H. designed the research. A.H., W.F.F., and P.H. performed the research. R.M.L. and A.H. wrote the manuscript.

ACKNOWLEDGMENTS

This work has been supported by National Institutes of Health grants GM030580-36, P50-GM103368-05, and computer grant S10-OD020095-01.

REFERENCES

1. Oruganty, K., and N. Kannan. 2013. Evolutionary variation and adaptation in a conserved protein kinase allosteric network: implications for inhibitor design. *Biochim. Biophys. Acta.* 1834:1322–1329.
2. Lahiry, P., A. Torkamani, …, R. A. Hegele. 2010. Kinase mutations in human disease: interpreting genotype-phenotype relationships. *Nat. Rev. Genet.* 11:60–74.
3. Zhang, J., P. L. Yang, and N. S. Gray. 2009. Targeting cancer with small molecule kinase inhibitors. *Nat. Rev. Cancer.* 9:28–39.
4. Taylor, S. S., and A. P. Kornev. 2011. Protein kinases: evolution of dynamic regulatory proteins. *Trends Biochem. Sci.* 36:65–77.
5. Kornev, A. P., N. M. Haste, …, L. F. T. Eyck. 2006. Surface comparison of active and inactive protein kinases identifies a conserved activation mechanism. *Proc. Natl. Acad. Sci. USA.* 103:17783–17788.
6. Kannan, N., and A. F. Neuwald. 2005. Did protein kinase regulatory mechanisms evolve through elaboration of a simple structural component? *J. Mol. Biol.* 351:956–972.
7. Endicott, J. A., M. E. Noble, and L. N. Johnson. 2012. The structural basis for control of eukaryotic protein kinases. *Annu. Rev. Biochem.* 81:587–613.
8. Leonard, C. J., L. Aravind, and E. V. Koonin. 1998. Novel families of putative protein kinases in bacteria and archaea: evolution of the "eukaryotic" protein kinase superfamily. *Genome Res.* 8:1038–1047.
9. Kannan, N., S. S. Taylor, …, G. Manning. 2007. Structural and functional diversity of the microbial kinome. *PLoS Biol.* 5:e17.
10. Hanks, S. K., and T. Hunter. 1995. Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. *FASEB J.* 9:576–596.
11. Kornev, A. P., S. S. Taylor, and L. F. Ten Eyck. 2008. A helix scaffold for the assembly of active protein kinases. *Proc. Natl. Acad. Sci. USA.* 105:14377–14382.
12. Gani, O. A., B. Thakkar, …, R. A. Engh. 2015. Assessing protein kinase target similarity: comparing sequence, structure, and cheminformatics approaches. *Biochim. Biophys. Acta.* 1854:1605–1616.
13. Vijayan, R. S. K., P. He, …, R. M. Levy. 2015. Conformational analysis of the DFG-out kinase motif and biochemical profiling of structurally validated type II inhibitors. *J. Med. Chem.* 58:466–479.
14. Lovera, S., M. Morando, …, F. L. Gervasio. 2015. Towards a molecular understanding of the link between imatinib resistance and kinase conformational dynamics. *PLoS Comput. Biol.* 11:e1004578.
15. Lin, Y.-L., Y. Meng, …, B. Roux. 2013. Explaining why gleevec is a specific and potent inhibitor of Abl kinase. *Proc. Natl. Acad. Sci. USA.* 110:1664–1669.
16. Shindyalov, I. N., N. A. Kolchanov, and C. Sander. 1994. Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Eng.* 7:349–358.
17. Socolich, M., S. W. Lockless, …, R. Ranganathan. 2005. Evolutionary information for specifying a protein fold. *Nature.* 437:512–518.
18. Sułkowska, J. I., F. Morcos, …, J. N. Onuchic. 2012. Genomics-aided structure prediction. *Proc. Natl. Acad. Sci. USA.* 109:10340–10345.
19. Ekeberg, M., C. Lövkvist, …, E. Aurell. 2013. Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 87:012707.
20. Marks, D. S., T. A. Hopf, and C. Sander. 2012. Protein structure prediction from sequence variation. *Nat. Biotechnol.* 30:1072–1080.
21. Marks, D. S., L. J. Colwell, …, C. Sander. 2011. Protein 3D structure computed from evolutionary sequence variation. *PLoS One.* 6:e28766.

22. Haldane, A., W. F. Flynn, …, R. M. Levy. 2016. Structural propensities of kinase family proteins from a Potts model of residue co-variation. *Protein Sci.* 25:1378–1384.

23. Cheng, R. R., M. Raghunathan, …, J. N. Onuchic. 2016. Constructing sequence-dependent protein models using coevolutionary information. *Protein Sci.* 25:111–122.

24. Morcos, F., N. P. Schafer, …, P. G. Wolynes. 2014. Coevolutionary information, protein folding landscapes, and the thermodynamics of natural selection. *Proc. Natl. Acad. Sci. USA.* 111:12408–12413.

25. Haq, O., M. Andrec, …, R. M. Levy. 2012. Correlated electrostatic mutations provide a reservoir of stability in HIV protease. *PLoS Comput. Biol.* 8:e1002675.

26. Mann, J. K., J. P. Barton, …, T. Ndung'u. 2014. The fitness landscape of HIV-1 gag: advanced modeling approaches and validation of model predictions by in vitro testing. *PLoS Comput. Biol.* 10:e1003776.

27. Shekhar, K., C. F. Ruberman, …, A. K. Chakraborty. 2013. Spin models inferred from patient data faithfully describe HIV fitness landscapes and enable rational vaccine design. *Phys. Rev. E.* 88:1539–3755.

28. Hopf, T. A., J. B. Ingraham, …, D. S. Marks. 2017. Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.* 35:128–135.

29. Figliuzzi, M., H. Jacquier, …, M. Weigt. 2016. Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase TEM-1. *Mol. Biol. Evol.* 33:268–280.

30. Dwyer, R. S., D. P. Ricci, …, N. S. Wingreen. 2013. Predicting functionally informative mutations in Escherichia coli BamA using evolutionary covariance analysis. *Genetics.* 195:443–455.

31. Cheng, R. R., F. Morcos, …, J. N. Onuchic. 2014. Toward rationally redesigning bacterial two-component signaling systems using coevolutionary information. *Proc. Natl. Acad. Sci. USA.* 111:E563–E571.

32. Cheng, R. R., O. Nordesjö, …, F. Morcos. 2016. Connecting the sequence-space of bacterial signaling proteins to phenotypes using coevolutionary landscapes. *Mol. Biol. Evol.* 33:3054–3064.

33. Ferguson, A. L., J. K. Mann, …, A. K. Chakraborty. 2013. Translating HIV sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design. *Immunity.* 38:606–617.

34. Barton, J. P., M. Kardar, and A. K. Chakraborty. 2015. Scaling laws describe memories of host-pathogen riposte in the HIV population. *Proc. Natl. Acad. Sci. USA.* 112:1965–1970.

35. Contini, A., and G. Tiana. 2015. A many-body term improves the accuracy of effective potentials based on protein coevolutionary data. *J. Chem. Phys.* 143:025103.

36. Lapedes, A., B. Giraud, and C. Jarzynski, 2002. Using sequence alignments to predict protein structure and stability with high accuracy. arXiv, arXiv:1207.2484v1, https://arxiv.org/abs/1207.2484.

37. Hopf, T. A., J. B. Ingraham, …, D. S. Marks, 2015. Quantification of the effect of mutations using a global probability model of natural sequence variation. arXiv, arXiv:1512.04612v1, https://arxiv.org/abs/1510.04612.

38. Schneidman, E., S. Still, …, W. Bialek. 2003. Network information and connected correlations. *Phys. Rev. Lett.* 91:238701.

39. Schneidman, E., M. J. Berry, 2nd, …, W. Bialek. 2006. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature.* 440:1007–1012.

40. Haq, O., R. M. Levy, …, M. Andrec. 2009. Pairwise and higher-order correlations among drug-resistance mutations in HIV-1 subtype B protease. *BMC Bioinformatics.* 10:S10.

41. van Nimwegen, E. 2016. Inferring contacting residues within and between proteins: what do the probabilities mean? *PLoS Comput. Biol.* 12:e1004726.

42. Aurell, E. 2016. The maximum entropy fallacy redux? *PLoS Comput. Biol.* 12:e1004777.

43. Jacquin, H., A. Gilson, …, R. Monasson. 2016. Benchmarking inverse statistical approaches for protein structure and design with exactly solvable models. *PLoS Comput. Biol.* 12:e1004889.

44. Coucke, A., G. Uguzzoni, …, M. Weigt. 2016. Direct coevolutionary couplings reflect biophysical residue interactions in proteins. *J. Chem. Phys.* 145:174102.

45. Weigt, M., R. A. White, …, T. Hwa. 2009. Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl. Acad. Sci. USA.* 106:67–72.

46. Cocco, S., and R. Monasson. 2011. Adaptive cluster expansion for inferring boltzmann machines with noisy data. *Phys. Rev. Lett.* 106:090601.

47. Jones, D. T., D. W. A. Buchan, …, M. Pontil. 2012. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics.* 28:184–190.

48. Balakrishnan, S., H. Kamisetty, …, C. J. Langmead. 2011. Learning generative models for protein fold families. *Proteins.* 79:1061–1078.

49. Morcos, F., A. Pagnani, …, M. Weigt. 2011. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. USA.* 108:E1293–E1301.

50. Mora, T., and W. Bialek. 2011. Are biological systems poised at criticality? *J. Stat. Phys.* 144:268–302.

51. Barton, J. P., E. De Leonardis, …, S. Cocco. 2016. ACE: adaptive cluster expansion for maximum entropy graphical model inference. *Bioinformatics.* 32:3089–3097.

52. Remmert, M., A. Biegert, …, J. Söding. 2011. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods.* 9:173–175.

53. Murphy, L. R., A. Wallqvist, and R. M. Levy. 2000. Simplified amino acid alphabets for protein fold recognition and implications for folding. *Protein Eng.* 13:149–152.

54. Solis, A. D. 2015. Amino acid alphabet reduction preserves fold information contained in contact interactions in proteins. *Proteins.* 83:2198–2216.

55. Wollenberg, K. R., and W. R. Atchley. 2000. Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap. *Proc. Natl. Acad. Sci. USA.* 97:3288–3291.

56. Sutto, L., S. Marsili, …, F. L. Gervasio. 2015. From residue coevolution to protein conformational ensembles and functional dynamics. *Proc. Natl. Acad. Sci. USA.* 112:13567–13572.

57. Knighton, D. R., J. H. Zheng, …, J. M. Sowadski. 1991. Crystal structure of the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase. *Science.* 253:407–414.

58. Azam, M., M. A. Seeliger, …, G. Q. Daley. 2008. Activation of tyrosine kinases by mutation of the gatekeeper threonine. *Nat. Struct. Mol. Biol.* 15:1109–1118.

59. Halabi, N., O. Rivoire, …, R. Ranganathan. 2009. Protein sectors: evolutionary units of three-dimensional structure. *Cell.* 138:774–786.

60. McLaughlin, R. N., Jr., F. J. Poelwijk, …, R. Ranganathan. 2012. The spatial architecture of protein function and adaptation. *Nature.* 491:138–142.

61. Cocco, S., and R. Monasson. 2012. Adaptive cluster expansion for the inverse ising problem: convergence, algorithm and tests. *J. Stat. Phys.* 147:252–314.

62. Flynn, W. F., A. Haldane, …, R. M. Levy. 2017. Inference of epistatic effects leading to entrenchment and drug resistance in HIV-1 protease. *Mol. Biol. Evol.* 34:1291–1306.

Biophysical Journal, Volume *114*

Supplemental Information

Coevolutionary Landscape of Kinase Family Proteins: Sequence Probabilities and Functional Motifs

Allan Haldane, William F. Flynn, Peng He, and Ronald M. Levy

Supplementary Information for: Co-Evolutionary landscape of Kinase Family Proteins: Correlated Mutations, sequence probabilities, and functional motifs

Allan Haldane, William F. Flynn, Peng He, and Ronald M. Levy

Details of the Potts model inference methods used in this study are published in the supplementary information of [1]. Source code for our software is available at https://github.com/ahaldane/IvoGPU. Here we describe additional analysis discussed in the main text.

I. PHYLOGENETIC WEIGHTING

As described in the main text, ideally the sequences in our dataset MSA would represent independent samples but in practice some sequences are related due to phylogeny and experimental biases, which we correct for by weighting each sequence using a sequence similarity cutoff. In many publications a sequence identity cutoff of 80% is used to detect sequences which are non-independent.

In this study we instead use a cutoff determined by examining the distribution of pairwise sequence identities between all pairs of sequences in the MSA (figure S1A). The fact that the upper tail of this distribution becomes negligible in size near about 60% sequence identity leads us to use this value as the cutoff for the phylogenetic weighting, as any sequences more similar that this are unlikely to occur independently in nature by chance.

II. KINASE EFFECTIVE ALPHABET SIZE

In the main text we justify our reduction of the amino alphabet from 21 letters to 8 letters on the basis that the Mutual Information (MI) values between residue pairs are preserved, as illustrated in figure S1B, and based on the small change in average sequence dissimilarity seen in figure S1A. The reduction to 8 letters can be further justified by a statistical estimate of the "effective" number of amino acids. This calculation also allows us to estimate the size of the kinase sequence space.

We estimate the effective number of amino acids at each position of the alignment as the exponential of the entropy per site, $q_i^{\text{eff}} = e^{-\sum_\alpha f_\alpha^i \log f_\alpha^i}$, since entropy can be interpreted as the log of the number of states. For the kinase dataset we find this gives an average of 8.9 effective amino acids per position on average. The size of the evolutionary accessible sequence space assuming site independence can be estimated as $\prod_i q_i^{\text{eff}}$, giving an evolutionarily accessible sequence space of roughly 10^{149} for kinase sequences. These results illustrate the large size of the kinase sequence space and the high degree of variation of kinase sequences.

III. ANALYSIS OF TRYPSIN AND PHOTOACTIVE YELLOW PROTEIN

To test whether our results may apply to systems besides kinase, we also analyze the Trypsin and Photoactive Yellow Protein (PYP) families, using the MSAs for these families generated in [3]. The mean effective number of amino acids in these families is 8.8 and 10.4 respectively. We reduce both families to 8 letters as described in the main text, and based on the upper tail of the distributions shown in figure S3 we choose sequence identity cutoffs of 60% and 40% respectively for phylogenetic weighting, giving an effective number of sequences of 4806 and 5720. We fit Potts models to these datasets and generate *in silico* MSAs from these models of size 4806 and 5720 sequences, respectively. In figure S4C we compare subsequence frequencies of the dataset to those of the model and to those of an independent model, again showing that the Potts model outperforms the independent model and that its performance is close to that expected due to finite-sampling effects alone.

We then fit *in silico* Potts models to the *in silico* datasets, and compare the statistical energies of the original and *in silico* Potts models and the independent model (figure S4A and B). In the case of PYP, the independent model energies still have a small correlation with the reference energies, though much lower than for the energies of the *in silico* Potts model. The number of effective sequences in these families is smaller than for kinase, and we correspondingly observe a decrease in the correlation of the *in silico* energies to the reference energies. Overall, the behavior of these two families is similar to that of the kinase family.

[1] A. Haldane, W. F. Flynn, P. He, R. Vijayan, and R. M. Levy, Protein Science **25**, 1378 (2016).

[2] S.-Y. Rhee, M. J. Gonzales, R. Kantor, B. J. Betts, J. Ravela, and R. W. Shafer, Nucleic Acids Research **31**, 298 (2003).

[3] T. A. Hopf, J. B. Ingraham, F. J. Poelwijk, C. P. I. Scharfe, M. Springer, C. Sander, and D. S. Marks, Nat Biotech **35**, 128 (2017).

[4] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt, Proceedings of the National Academy of Sciences **108**, E1293 (2011).

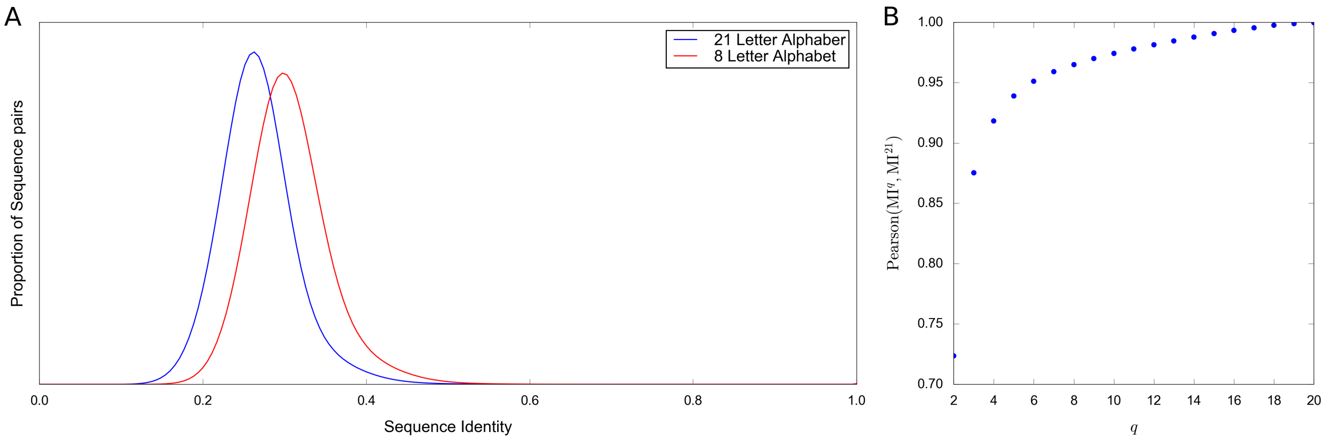


FIG. S1. (A) Distribution of sequence identity scores (normalized inverse Hamming distance) between all pairs of sequences in the kinase dataset, computed for the original sequences using a 21 letter alphabet of 20 residues plus gap with phylogenetic weighting, and for the reduced 8 letter alphabet. The mean sequence identity is 27% for 21 letters and 31% for 8 letters. (B) Pearson correlation between the $\binom{L}{2}$ MI values for reduction from 21 letters to alphabet size q, for varying q. This figure was previously published in the supplementary information of [1].

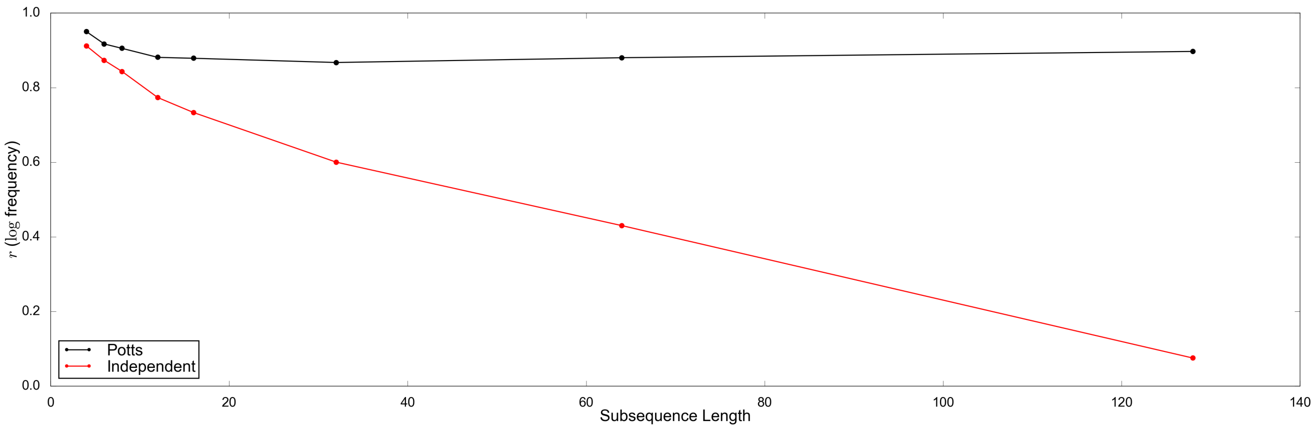


FIG. S2. Using the *in silico* model we demonstrate the accuracy of the Potts model for longer subsequences with $L > 10$. Here we show the correlation between the "true" log frequencies of subsequences and the log frequencies predicted by the independent model (red) and the *in silico* Potts model (black), computed using the approximation described in the Methods of the main text. The correlation decreases quickly for longer subsequence lengths for the independent model, but stays roughly constant for the Potts model.

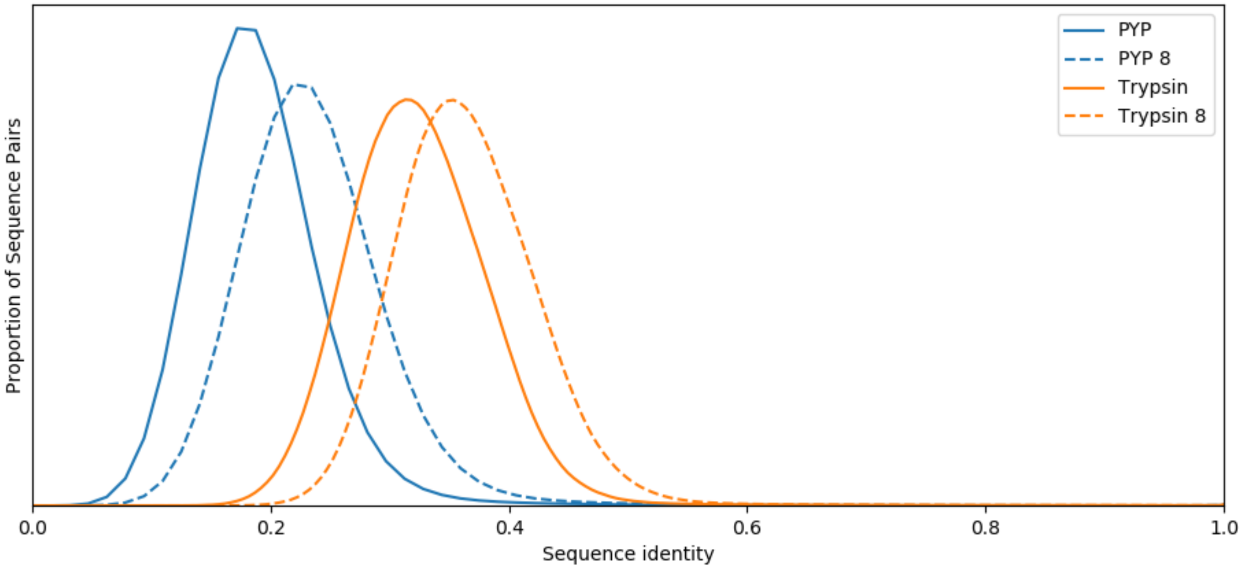


FIG. S3. Distribution of sequence identity scores (normalized inverse Hamming distance) between all pairs of sequences in the PYP and trypsin datasets, computed for the original sequences using a 21 letter alphabet and for the reduced 8 letter alphabet.

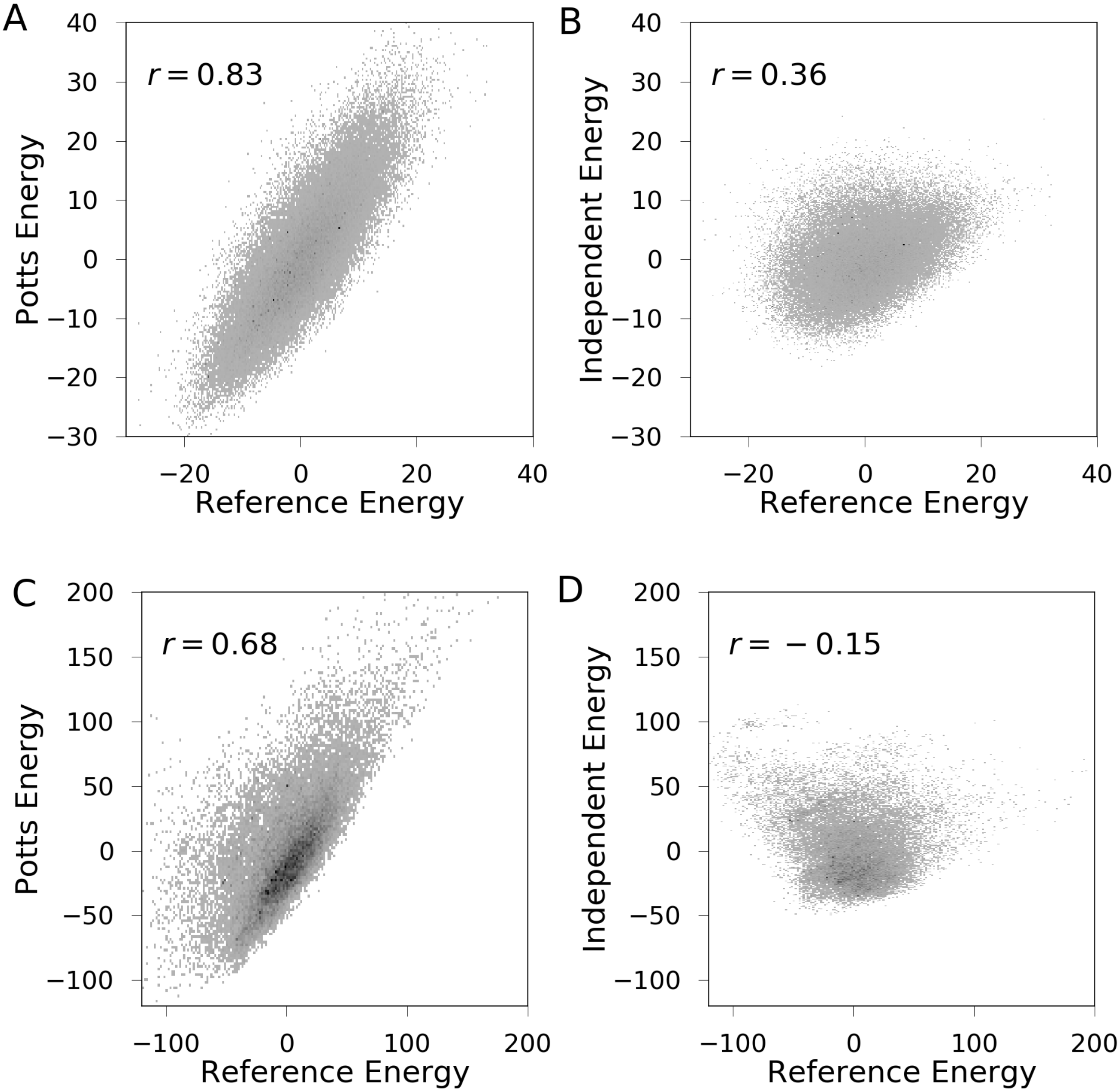


FIG. S4. Analysis of Potts models fit to two different protein families. (A,B,C) Photoactive Yellow Protein, and (D, E, F) Trypsin. (A and D) Statistical energies computed using the original Potts model compared to those computed using a Potts model refit to a finite size sample of sequences from the first model, as in figure 2 in the main text. (B, E) Comparison of statistical energies computed using the original Potts model compared to the those computed using the independent model. (C, F) Average correlation between observed and predicted frequencies for the top 20 subsequences for large samples of subsequences of varying length, as described in figure 1 in the main text.

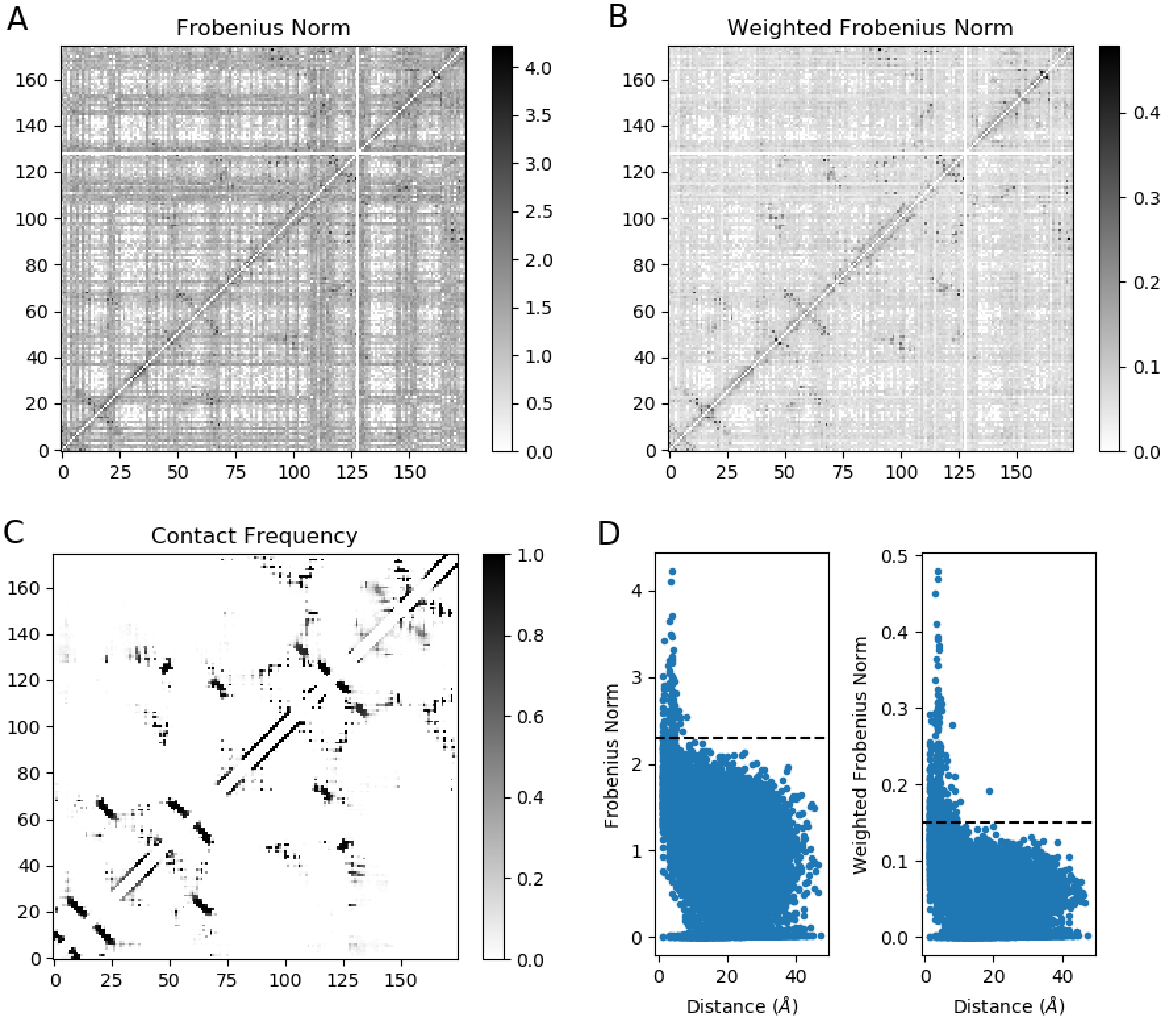
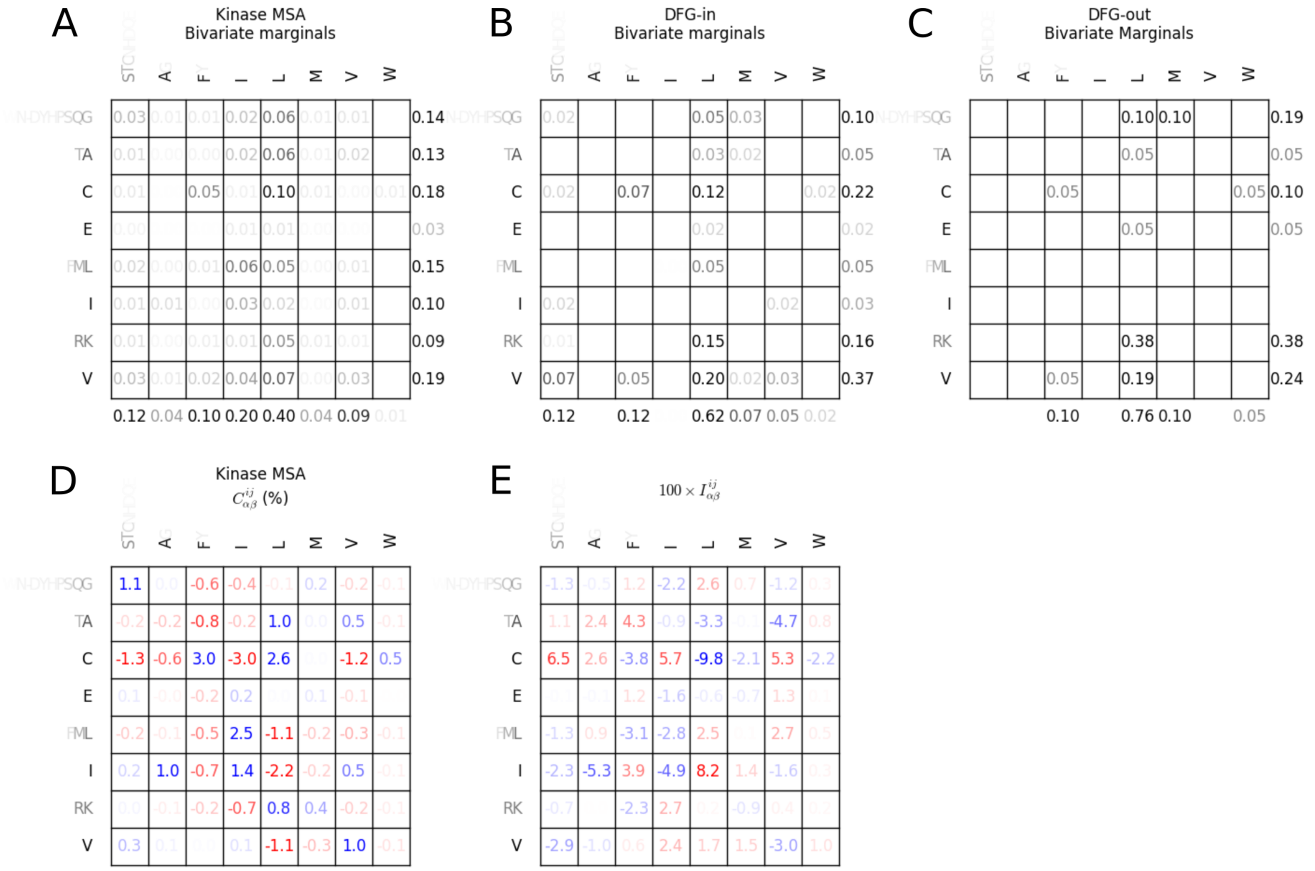


FIG. S5. Comparison of interaction scores with crystal structure contacts. (A) Interaction scores computed using the (unweighted) Frobenius Norm, shown as a pairwise interaction score map between all pairs of the 175 residues. The shading ranges from 0 to the maximum score. (B) Interaction scores computed using the Weighted Frobenius Norm, shown as in panel A. (C) Contact Frequency in the kinase PDB dataset (see main text), with contacts between residues 3 or less apart along the sequence removed. (D) Distances for each residue-pair, averaged over all PDB structures, compared to the interaction scores. The dotted lines represent a rough estimate of the noise threshold for interaction score significance, showing how the weighted norm better distinguishes contacts.



A Kinase MSA Bivariate marginals

	STCNHDQE	AG	FY	I	L	M	V	W	
WN-DYHPSQG	0.03	0.01	0.01	0.02	0.06	0.01	0.01		0.14
TA	0.01	0.00	0.00	0.02	0.06	0.01	0.02		0.13
C	0.01	0.00	0.05	0.01	0.10	0.01	0.00	0.01	0.18
E	0.00	0.00	0.00	0.01	0.01	0.00	0.00		0.03
FML	0.02	0.00	0.01	0.06	0.05	0.00	0.01		0.15
I	0.01	0.01	0.00	0.03	0.02	0.00	0.01		0.10
RK	0.01	0.00	0.01	0.01	0.05	0.01	0.01		0.09
V	0.03	0.01	0.02	0.04	0.07	0.00	0.03		0.19
	0.12	0.04	0.10	0.20	0.40	0.04	0.09	0.01	

B DFG-in Bivariate marginals

	STCNHDQE	AG	FY	I	L	M	V	W	
WN-DYHPSQG	0.02				0.05	0.03			0.10
TA					0.03	0.02			0.05
C	0.02		0.07		0.12			0.02	0.22
E					0.02				0.02
FML				[illegible]	0.05				0.05
I	0.02						0.02		0.03
RK	0.01				0.15				0.16
V	0.07		0.05		0.20	0.02	0.03		0.37
	0.12		0.12	[illegible]	0.62	0.07	0.05	0.02	

C DFG-out Bivariate Marginals

	STCNHDQE	AG	FY	I	L	M	V	W	
WN-DYHPSQG					0.10	0.10			0.19
TA					0.05				0.05
C			0.05					0.05	0.10
E					0.05				0.05
FML									
I									
RK					0.38				0.38
V			0.05		0.19				0.24
			0.10		0.76	0.10		0.05	

D Kinase MSA $C^{ij}_{\alpha\beta}$ (%)

	STCNHDQE	AG	FY	I	L	M	V	W
WN-DYHPSQG	1.1	0.0	-0.6	-0.4	-0.1	0.2	-0.2	-0.1
TA	-0.2	-0.2	-0.8	-0.2	1.0	0.0	0.5	-0.1
C	-1.3	-0.6	3.0	-3.0	2.6	0.0	-1.2	0.5
E	0.1	-0.0	-0.2	0.2	0.0	0.1	-0.1	0.0
FML	-0.2	-0.1	-0.5	2.5	-1.1	-0.2	-0.3	-0.1
I	0.2	1.0	-0.7	1.4	-2.2	-0.2	0.5	-0.1
RK	0.0	-0.1	-0.2	-0.7	0.8	0.4	-0.2	-0.1
V	0.3	0.1	0.0	0.1	-1.1	-0.3	1.0	-0.1

E $100 \times I^{ij}_{\alpha\beta}$

	STCNHDQE	AG	FY	I	L	M	V	W
WN-DYHPSQG	-1.3	-0.5	1.2	-2.2	2.6	0.7	-1.2	0.3
TA	1.1	2.4	4.3	-0.9	-3.3	-0.1	-4.7	0.8
C	6.5	2.6	-3.8	5.7	-9.8	-2.1	5.3	-2.2
E	-0.1	-0.1	1.2	-1.6	-0.6	-0.7	1.3	0.1
FML	-1.3	0.9	-3.1	-2.8	2.5	0.1	2.7	0.5
I	-2.3	-5.3	3.9	-4.9	8.2	1.4	-1.6	0.3
RK	-0.7	[illegible]	-2.3	2.7	0.2	-0.9	0.4	0.2
V	-2.9	-1.0	0.6	2.4	1.7	1.5	-3.0	1.0

FIG. S6. Statistics and parameters describing the position pair 132-145 in the kinase model. Each subplot is an 8x8 grid representing the entries in the bivariate marginal and coupling parameter matrices at this position pair, for each combination of the 8 letters in the reduced alphabet. Position 132 is the top (left-right) dimension, and 145 the left (up-down) dimension. For each column or row, the letter in the 8-letter alphabet is mapped back to the possibilities in the 21 letter alphabet with shading proportional to that letter's frequency. For example the 7th letter at position 145 appears as an R or K in the 21 letter alphabet with about equal frequency. (A) Bivariate marginals computed from the kinase MSA. Univariate marginals are shown in the margins. (B) Bivariate marginals for sequences in the PDB dataset, found to be in the DFG-in state by our PDB analysis. Note that when computing the marginals, a phylogenetic weighting of 0.1 was applied (see methods in main text). (C) Bivariate marginals for PDB sequences in the DFG-out conformation. (D) Correlation coefficients $C^{ij}_{\alpha\beta} = f^{ij}_{\alpha\beta} - f^{i}_{\alpha} f^{j}_{\beta}$ computed from marginals in the kinase MSA. (E) Interaction score elements from the inferred Potts model. $I^{ij}_{\alpha\beta} = w^{ij} J^{ij}_{\alpha\beta}$ computed in the weighted gauge, as described under "interaction score" in the main text, such that the weighted FFFFFFFFF norm is given by $I^{ij} = \sqrt{\sum_{\alpha\beta} (I^{ij}_{\alpha\beta})^2}$.

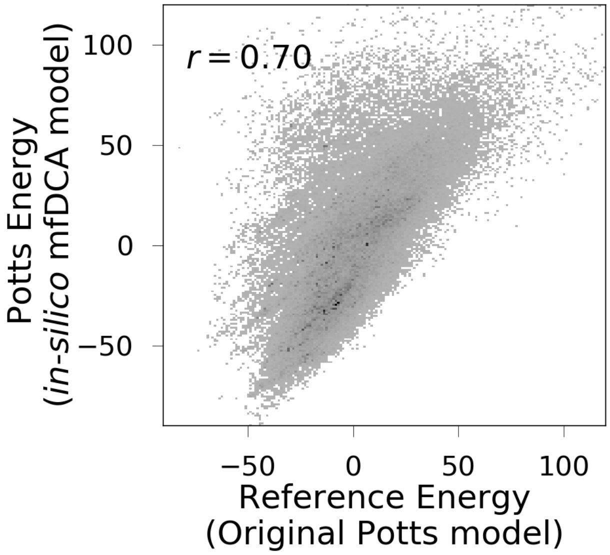


FIG. S7. Reconstruction of Potts energies as in figure 2A in the main text, but using the mfDCA inference method. We fit a Potts model to the *in silico* dataset using the mfDCA method. For this purpose, we use the same phylogenetic weighting as in the *in silico* test using the MCMC inference, but in contrast we add a very large pseudocount as prescribed in [4], using $\lambda = M_{\rm eff}$ as defined in that publication corresponding to a pseudocount equal in size the the dataset itself. We then compute Potts energies of the uniprot MSA using this Potts model, and compare to the energies computed using the original Potts model used to generate the *in silico* dataset.

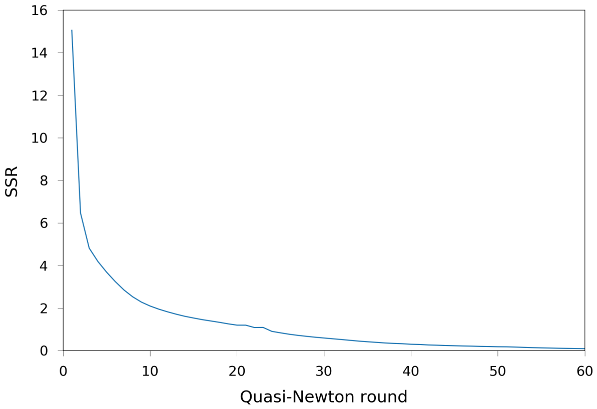


FIG. S8. Convergence of the MCMC algorithm for the Kinase dataset. This shows the SSR (sum of squared residuals) between the observed kinase MSA bivariate marginals, and those of the inferred Potts model decrease for increasing iterations of the quasi-Newton algorithm, described in [1]. Each iteration represents a round of MCMC sequence generation, followed by quasi-Newton optimization, for a total of 60 iterations.

Index	2CPK Index	Residue	Motif
1	48	T	P-loop
2	49	L	P-loop
3	50	G	P-loop
4	51	T	P-loop
5	52	G	P-loop
6	53	S	P-loop
7	54	F	P-loop
8	55	G	P-loop
9	56	R	P-loop
10	57	V	
11	58	M	
12	59	L	
13	60	V	
14	61	K	
15	62	H	
16	63	K	
17	64	E	
18	65	S	
19	66	G	
20	68	H	
21	69	Y	
22	70	A	
23	71	M	
24	72	K	β-3 Lysine
25	73	I	
26	74	L	
27	75	D	
28	76	K	α-C Helix
29	77	Q	α-C Helix
30	78	K	α-C Helix
31	79	V	α-C Helix
32	80	V	α-C Helix
33	81	K	α-C Helix
34	82	L	α-C Helix
35	83	K	α-C Helix
36	84	Q	α-C Helix
37	90	N	α-C Helix
38	91	E	α-C Helix
39	92	K	α-C Helix
40	93	R	α-C Helix
41	94	I	α-C Helix
42	95	L	
43	96	Q	
44	97	A	
45	98	V	
46	99	N	
47	100	F	
48	101	P	
49	102	F	
50	103	L	
51	104	V	
52	105	K	
53	106	L	
54	107	E	
55	108	F	
56	109	S	
57	110	F	
58	111	K	
59	112	D	
60	113	N	
61	114	S	
62	115	N	
63	116	L	
64	117	Y	
65	118	M	
66	119	V	
67	120	M	Gatekeeper
68	121	E	Hinge
69	122	Y	Hinge
70	123	V	Hinge
71	124	A	Hinge
72	126	G	Hinge
73	127	E	Hinge
74	128	M	Hinge
75	129	F	
76	130	S	
77	131	H	
78	132	L	
79	133	R	
80	134	R	
81	136	G	
82	137	R	
83	138	F	
84	-	S	
85	140	E	
86	141	P	
87	142	H	
88	143	A	
89	144	R	
90	145	F	
91	146	Y	
92	147	A	
93	148	A	
94	149	Q	
95	150	I	
96	151	V	
97	152	L	
98	153	T	
99	154	F	
100	155	E	
101	156	Y	
102	157	L	
103	158	H	
104	159	S	
105	160	L	
106	161	D	
107	162	L	
108	163	I	
109	164	Y	H
110	165	R	R
111	166	D	D
112	167	L	
113	168	K	
114	169	P	
115	170	E	
116	171	N	
117	172	L	
118	173	L	
119	174	I	
120	175	D	
121	176	Q	
122	177	Q	
123	178	G	
124	179	Y	
125	180	I	
126	181	Q	
127	182	V	
128	183	T	
129	184	D	D
130	185	F	F
131	186	G	G
132	187	F	Act. Loop
133	188	A	Act. Loop
134	189	K	Act. Loop
135	190	R	Act. Loop
136	191	V	Act. Loop
137	192	K	Act. Loop
138	-	-	Act. Loop
139	193	G	Act. Loop
140	194	R	Act. Loop
141	195	T	Act. Loop
142	196	W	Act. Loop
143	197	T	Act. Loop
144	198	L	Act. Loop
145	199	C	Act. Loop
146	200	G	Act. Loop
147	201	T	Act. Loop
148	202	P	Act. Loop
149	203	E	Act. Loop
150	204	Y	Act. Loop
151	205	L	Act. Loop
152	206	A	Act. Loop
153	207	P	
154	208	E	
155	209	I	
156	210	I	
157	211	L	
158	212	S	
159	213	K	
160	214	G	
161	215	Y	
162	216	N	
163	217	K	
164	218	A	
165	219	V	
166	220	D	
167	221	W	
168	222	W	
169	223	A	
170	224	L	
171	225	G	
172	226	V	
173	227	L	
174	228	I	
175	229	Y	

TABLE S1. Mapping from MSA alignment position to PDB residue index for PDB 2CPK. This table was provided in the supplementary information of [1].

Index	2YAC Index	Residue	Motif
1	58	F	P-loop
2	59	L	P-loop
3	60	G	P-loop
4	61	K	P-loop
5	62	G	P-loop
6	63	G	P-loop
7	64	F	P-loop
8	65	A	P-loop
9	66	K	P-loop
10	67	C	
11	68	F	
12	69	E	
13	70	I	
14	71	S	
15	72	D	
16	73	A	
17	74	D	
18	75	T	
19	76	K	
20	77	E	
21	78	V	
22	80	A	
23	81	G	
24	82	K	β-3 Lysine
25	83	I	
26	84	V	
27	85	P	
28	86	K	α-C Helix
29	87	S	α-C Helix
30	93	H	α-C Helix
31	94	Q	α-C Helix
32	95	R	α-C Helix
33	96	E	α-C Helix
34	97	K	α-C Helix
35	98	M	α-C Helix
36	99	S	α-C Helix
37	100	M	α-C Helix
38	101	E	α-C Helix
39	102	I	α-C Helix
40	103	S	α-C Helix
41	104	I	α-C Helix
42	105	H	
43	106	R	
44	107	S	
45	108	L	
46	109	A	
47	110	H	
48	111	Q	
49	112	H	
50	113	V	
51	114	V	
52	115	G	
53	116	F	
54	117	H	
55	118	G	
56	119	F	
57	120	F	
58	121	E	
59	122	D	
60	123	N	
61	124	D	
62	125	F	
63	126	V	
64	127	F	
65	128	V	
66	129	V	
67	130	L	Gatekeeper
68	131	E	Hinge
69	132	L	Hinge
70	133	C	Hinge
71	134	R	Hinge
72	135	R	Hinge
73	136	R	Hinge
74	139	L	Hinge
75	140	E	
76	141	L	
77	142	H	
78	143	K	
79	144	R	
80	145	R	
81	146	K	
82	147	A	
83	148	L	
84	149	T	
85	150	E	
86	151	P	
87	152	E	
88	153	A	
89	154	R	
90	155	Y	
91	156	Y	
92	157	L	
93	158	R	
94	159	Q	
95	160	I	
96	161	V	
97	162	L	
98	163	G	
99	164	C	
100	165	Q	
101	166	Y	
102	167	L	
103	168	H	
104	169	R	
105	170	N	
106	171	R	
107	172	V	
108	173	I	
109	174	H	H
110	175	R	R
111	176	D	D
112	177	L	
113	178	K	
114	179	L	
115	180	G	
116	181	N	
117	182	L	
118	183	F	
119	184	L	
120	185	N	
121	186	E	
122	187	D	
123	188	L	
124	189	E	
125	190	V	
126	191	K	
127	192	I	
128	193	G	
129	194	D	D
130	195	F	F
131	196	G	G
132	197	L	Act. Loop
133	198	A	Act. Loop
134	199	T	Act. Loop
135	200	K	Act. Loop
136	201	V	Act. Loop
137	202	E	Act. Loop
138	205	G	Act. Loop
139	206	E	Act. Loop
140	207	R	Act. Loop
141	208	K	Act. Loop
142	209	K	Act. Loop
143	210	T	Act. Loop
144	211	L	Act. Loop
145	212	C	Act. Loop
146	213	G	Act. Loop
147	214	T	Act. Loop
148	215	P	Act. Loop
149	216	N	Act. Loop
150	217	Y	Act. Loop
151	218	I	Act. Loop
152	219	A	Act. Loop
153	220	P	
154	221	E	
155	222	V	
156	223	L	
157	224	S	
158	225	K	
159	226	K	
160	227	G	
161	228	H	
162	229	S	
163	230	F	
164	231	E	
165	232	V	
166	233	D	
167	234	V	
168	235	W	
169	236	S	
170	237	I	
171	238	G	
172	239	C	
173	240	I	
174	241	M	
175	242	Y	

TABLE S2. Mapping from MSA alignment position to PDB residue index for PDB 2YAK.