HHS Public Access

Author manuscript

Curr Opin Struct Biol. Author manuscript; available in PMC 2018 March 27.

Published in final edited form as:

Curr Opin Struct Biol. 2017 April ; 43: 55–62. doi:10.1016/j.sbi.2016.11.004.

Potts Hamiltonian models of protein co-variation, free energy landscapes, and evolutionary fitness

Ronald M Levy[1], **Allan Haldane**[1], and **William F Flynn**[1,2]

[1]Center for Biophysics and Computational Biology, Department of Chemistry, and Institute for Computational Molecular Science, Temple University, Philadelphia, PA 19122, United States

[2]Department of Physics and Astronomy, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, United States

Abstract

Potts Hamiltonian models of protein sequence co-variation are statistical models constructed from the pair correlations observed in a multiple sequence alignment (MSA) of a protein family. These models are powerful because they capture higher order correlations induced by mutations evolving under constraints and help quantify the connections between protein sequence, structure, and function maintained through evolution. We review recent work with Potts models to predict protein structure and sequence-dependent conformational free energy landscapes, to survey protein fitness landscapes and to explore the effects of epistasis on fitness. We also comment on the numerical methods used to infer these models for each application.

Introduction

There is a long history of the use of coevolutionary information in the form of protein sequence variation to probe the relationship between protein structure, function, and fitness, to understand how allosteric interactions are transmitted through proteins, and to predict protein structure from sequence. Reviews of this history can be found in [1–4]. More recently, powerful inverse inference statistical approaches have been developed to study these relationships, which encode sequence variation extracted from multiple sequence alignments (MSAs) into spin-glass Hamiltonian models of the sequence space. The Hamiltonians take the form

$$H(S)=\sum_{i=1}^{L}h_i(s_i)+\sum_{i=1}^{L}\sum_{i<j}^{L}J_{ij}(s_i,s_j) \quad (1)$$

where the sequence S is a string of letters corresponding to the amino acid types $\{s\}$ at each of L positions, encoded in an alphabet of q letters, and $h_i(s_i)$ ('fields') and $J_{ij}(s_i, s_j)$ ('couplings') are single-site and pair-site parameters [5,6]. For some problems it is sufficient

Corresponding author: Levy, Ronald M (ronlevy@temple.edu).

to use a $q = 2$ binary alphabet {0, 1} at each position, where 0 corresponds to the wildtype or consensus amino acid type and 1 corresponds to a mutant; this Hamiltonian is referred to as an Ising model by analogy with the Hamiltonian which describes magnetic spin systems in condensed matter physics. To describe the 20 amino acid types (plus a gap character) at each position the generalized model is referred to as a Potts Hamiltonian model, describing spin glasses. As with spin glasses, this Hamiltonian can describe complex landscapes with multiple local minima. The model is based on the maximum entropy principle that seeks to construct the minimally biased sequence probability distribution that reproduces the one-site $\langle s_i \rangle$ and two site $\langle s_i s_j \rangle$ mutational probabilities (or marginal probabilities) of a protein MSA, giving a distribution $P(S) \propto_{exp} H(S)$. Its parameters can be determined by maximum likelihood inference given data, by methods reviewed below.

This area of research is a rapidly developing field of computational biology at the intersection with structural biology, biological physics, and branches of biology concerned with evolutionary protein fitness. The fast paced developments in sequencing, including the sequencing of large numbers of whole genomes for many species, and the development of deep sequencing techniques [7–9], has provided a rich source of data for the construction of correlated mutation models of protein structure, energetics, and evolution. The purpose of this article is to review the most recent developments using Ising/Potts models in computational biology with a particular focus on: firstly protein structure and the mapping of (free) energy landscapes; secondly modeling of fitness landscapes as proteins evolve under selective pressure, and finally a review of the most recent numerical methods for solving the inverse inference problem, and how well they perform in the context of the kinds of problems reviewed in the following sections.

Protein structure and free energy landscapes

Evolutionary sequence patterns contain information about protein structure, which can complement experimental data such as crystallographic and NMR structures. Before the introduction of Ising/Potts spin models of protein sequence variation, covariance matrices evaluated from MSAs were used in this way for protein structure modeling [10–12]. Covariance occurs because interacting residues are constrained by physical proximity effects and their mutations are therefore correlated. But correlations are induced by both direct and indirect effective interactions between residues that need not be close in space. For example, allosteric effects often involve long range interactions mediated by networks of interacting residues. One of the prime motivations for developing Potts models of protein sequence space has been to disentangle the direct from the indirect correlations as part of a procedure known as ‘direct coupling analysis’ (DCA), which improves on traditional covariance methods by providing a mapping between strongly directly interacting residues inferred from the model and contacts in 3D protein structure [13].

For contact prediction, the Potts model parameters are not used directly but an ‘interaction score’ summarizing the q^2 residue-dependent coupling parameters $J_{ij}(s_i, s_j)$ is computed for each position-pair i, j, and the top-ranking position-pairs are interpreted as predicted contacts for the family as a whole. These contacts can then be used as input in further computations to study protein structures and conformational landscapes, for example for

abinitio structure prediction of the native (folded) conformation using NMR (distance geometry) structure determination algorithms [14,15], or using the contacts to bias or constrain molecular dynamics simulations and Go-models using a homologous crystal structure as a starting point [16–18]. Progressively more detailed aspects of the conformational-energy landscapes of proteins have been explored with this strategy. The inferred contact constraints have been used to predict alternate (experimentally unobserved) intermediate conformations [19•], to sample conformational space along a conformational transition, predict conformational flexibility of parts of a protein and simulate the folding transition [20•], and to predict conformations of dimers, multimers, and repeat proteins [21–24]. Other studies investigate ways of combining the predicted contacts with additional structural data [25–27].

Recent studies have begun to mine the residue-specific parameters of the Potts model to map out sequence-dependent conformational free energy landscapes. These applications implicitly relate the coupling parameters $J_{ij}(s_i, s_j)$ of the Potts model to pairwise physical interaction strengths or free energy contributions. In one approach the sequence-specific coupling strengths $J_{ij}(s_i, s_j)$ are used as interaction strengths (biases) in coarse grained molecular simulations [28]. By examining the residue-dependent coupling parameters for sequences in a common family, it is possible to detect which interaction pairs contribute to the stability of particular conformations (not necessarily the native conformation) and the sequence dependence of these conformational preferences (see Figure 1) [29•].

The correspondence between the Potts Hamiltonian and free energy landscapes has been made even more explicit in other studies, which compare $H(S)$ to the free energy of folding of a protein sequence in its native fold $\Delta G(S)$, motivated by the idea that the folding process is a major determinant of protein fitness and therefore prevalence of sequences in the MSA. Experimental evidence supports this relationship. A number of studies have shown that for single and double mutants to a sequence, the change in the sequence's statistical Potts score ΔH is linearly related to the experimental change in free energy of folding upon mutation $\Delta\Delta G$ [30••,31••,32]. The Potts energy $H(S)$ of entire sequences has also been found to be correlated to protein melting temperature [33,34]. The distribution of free energies of folding predicted by the Potts model can be interpreted using the energy landscape theory of protein folding [35,36], and it has been shown that the Potts model gives results consistent with energy landscapes predicted from physiochemical considerations [31••]. Other studies have sought to better understand the physical origin of individual Potts model parameters. The value of certain coupling parameters can be rationalized from charge–charge interactions modulated by distance [37], and using lattice models the relationship between the coupling terms $\{J\}$ and pairwise interaction strengths in the native folded state has been shown to be modulated by the effect of competing non-native (decoy) folds and interactions [38].

Epistasis and fitness

The relationship between conformational free energy and $H(S)$ is mediated by effects related to protein (and organismal) fitness, and fitness may be affected by other protein properties such as enzyme efficiency and the specificity for interactions with other proteins in signaling

networks [39]. This motivates another avenue of research using Potts models in which the Potts Hamiltonian is used to probe protein fitness.

The Potts Hamiltonian is a mapping of sequences to statistical scores in which sequences with lower Potts statistical energy are more probable, generating a landscape termed the 'prevalence landscape' [40,41,42••, 43••]. Recent studies have demonstrated that experimental measurements of fitness differences are empirically well correlated with the change in Potts statistical energy ΔH for sequences from large protein families [30••,34]. Potts models have also been used to interrogate the fitness landscapes of viral systems, most notably HIV proteins escaping immune pressure [40,41,42••,44] or HIV enzymes evolving under drug pressure [45•]. These studies have demonstrated linear relationships between Potts energy and experimental measurements of viral fitness, and these combined with the previously mentioned studies provide strong evidence that there exists a correspondence between the Potts prevalence landscape and the protein fitness landscape.

Interpreting the meaning of individual Potts model parameters in terms of evolutionary quantities is complicated by the nontrivial differences between the evolution of viruses and protein families to which the Potts model has been applied, such as the strength of selection, mutation rate, effective population size and population dynamics [46,47]. A systematic comparison of the implications of these features on model building is lacking. For proteins evolving nearly neutrally the suggestive relationship between the evolutionary fitness landscape and the prevalence landscape described by statistical physics (with several caveats) [48] reinforces the observed correlations between large experimental fitness datasets and Potts statistical energies in studies of protein families [30••,34]. In viral systems the Potts model parameters have been related to parameters in viral evolution models, such as Eigen's model of quasispecies, showing that there exist complex mappings from Potts prevalence landscape parameters to those of the fitness function [41,40]. Other work has incorporated the Potts Hamiltonian as a fitness function in population genetics simulations; a recent study has demonstrated that the relationship between *in vivo* fitness of HIV proteins targeted by the immune system and Potts equilibrium predictions of fitness can be improved by creating a dynamical model combining the Potts Hamiltonian with Wright-Fisher evolutionary simulations [43••]

A key aspect of the Potts model in this application is its ability to model epistasis, the intuitively understandable phenomenon that the effect of a mutation is modulated by the background sequence in which it is acquired, because of the collective effects of pairwise couplings. Experimental technologies like deep mutational scanning [49–52] have allowed for the full mapping of the genotype space located one or two mutations from the native sequence to the phenotype space, yielding insights into the local fitness landscape around the native sequence. However, the size of the sequence space increases exponentially as the number of sites included in high throughput mutagenesis increases, and thus only a tiny portion of the mutational landscape and the associated epistatic effects can be sampled. Organizing principles discovered through statistical modeling are needed to interpret the massive amounts of data becoming available (see Figure 2).

A central premise of Potts spin glass models applied to biological systems is the assumption that modeling pair correlations is both necessary and sufficient to describe higher order correlation patterns beyond pairs [53]. For Potts models of protein sequence variation, this includes mutation patterns up to and including whole sequences, including those with many simultaneous mutations and those not observed in the multiple sequence alignment. This makes the Potts model a capable probe for exploring the regional mutational landscape around the native sequence, pushing the size of the surveyable sequence space to much larger landscapes than are currently accessible experimentally. A vivid illustration of the predictive power of spin models of protein sequence variation is the successful prediction of the occurrence of HIV protease sequences in a second curated database which were not present in the one used to parameterize the Ising model [37]. Another study has recently demonstrated that deleterious mutations in HIV-1 protease needed to escape drug pressure become advantageous in the context of specific sequence backgrounds that are observed when a sufficient number of mutations have accumulated [45•] and become more advantageous (or entrenched [54–56]) as additional mutations accumulate in the presence of these formerly deleterious mutations. Fit variants with multiple resistance mutations are less likely to revert to wildtype in infected individuals not undergoing antiretroviral therapy, and thus pose a significant risk to transmit resistance.

The use of (a properly parameterized) Potts model to survey the prevalence landscape far from the native sequence is justifiable given the model's ability to capture higher order correlations in the input dataset. While the model is trained on pair correlations, studies have demonstrated the model's ability to reproduce marginal probability distributions higher than 2nd order, typically showing the preservation of 3-body or 4-body correlation statistics [20•, 37,38,42••,43••,57]. These studies have also demonstrated that the distribution of observed hamming distances from a reference sequence can be accurately modeled by the Potts Hamiltonian, and studies demonstrating the accurate prediction of higher order marginal distributions are forthcoming [45•]. Further, these results indicate that models without pair couplings, often termed independent models because they model the effect on fitness of each site individually, are poor predictors of these experimental quantities, meaning the incorporation of pairwise epistatic terms are necessary to accurately model a mutation's impact on fitness [30••,34,37,45•]. Potts models provide an unprecedented glimpse into the fitness landscape around naturally observed protein sequences which is not yet achievable with current mutagenesis experiments.

Inverse inference methods

Inferring the maximum likelihood set of parameters of the Potts model $J_{ij}(s_i, s_j)$ given the MSA statistics $\langle s_i s_j \rangle$ is a significant computational challenge, and a variety of methods have been developed to solve it including message passing [13,37], mean field approximations [17,58], pseudolikelihood methods (PLM) [59,60], Monte Carlo methods [6,20•,29•,42••], and adaptive cluster expansion (ACE) [57]. These have each been implemented with particular applications in mind, generally either contact prediction or fitness mapping (prediction of statistical energies), potentially at the expense of other uses.

Many inference techniques were developed for protein contact prediction. The accuracy of the Potts model's estimate of probabilities $P(S)$ and $\langle s_i s_j \rangle$, and the value of the individual residue-dependent parameters $J_{ij}(s_i, s_j)$ are not directly relevant in this application, and relaxing the requirements on their accuracy allows stronger approximations to be made. Message passing, used in some of the earliest studies, assumes the interaction network is mostly 'loop-free', making the computation of $\langle s_i s_j \rangle$ given trial $\{J\}$ computationally efficient so that the maximum likelihood values $\{J\}$ may be solved for iteratively. This method has been superseded by mean-field methods, which using the weak-coupling limit relate the $J_{ij}(s_i, s_j)$ and the correlation coefficients $C_{ij}(s_i, s_j) = \langle s_i s_j \rangle - \langle s_i \rangle \langle s_j \rangle$ by a simple and efficient matrix inversion [61]. This method has proven particularly popular because of its speed. Pseudolikelihood methods, which approximate the likelihood function and maximize likelihood by gradient descent, have been shown to give more accurate contact prediction at moderately higher computational cost [62]. While these methods reliably distinguish the most strongly interacting position-pairs from the rest, they generally do not correctly model sequence statistics, including the univariate and bivariate statistics the model is based on [57]. A benchmarking study comparing statistics of an input MSA to MSAs generated from models inferred from the input by different methods shows that mean-field methods underestimate the amount of sequence diversity and generate overly fit sequences, while PLM tends to generate models giving less stable sequences than in the original dataset [38].

In evolutionary applications other methods are favored which more faithfully model the dataset sequence statistics, including Monte Carlo methods, which solve for $\langle s_i s_j \rangle$ given trial $\{J\}$ by Monte Carlo sampling of sequences from the model and solve for the optimal $\{J\}$ iteratively, as well as Cluster Expansion techniques which recursively solve sub-networks of the full problem, allowing uncorrelated sub-networks to be left out of the computation. These methods are much more computationally intensive than mean-field and pseudolikelihood methods, but can accurately model the bivariate (and higher) statistics of the dataset and the distribution of sequence probabilities of the MSA as described in the previous section, though the ACE method may generate models with sequences with slightly higher statistical energies than the original data-set [38]. Studies using Monte Carlo methods have also shown that low-dimensional (Principal Component) representations of the inferred Potts model landscape match those of the dataset [20•,41]. Monte Carlo methods have been used for the purpose of contact prediction, as well [20•,29•,33].

In addition to the Potts model inference itself, data preprocessing steps can have a significant influence on the inferred model. The Potts model is an equilibrium model assuming each sequence in the dataset is generated by an independent process according to the distribution $P(S)$, however in practice MSAs are constructed from datasets with experimental and phylogenetic structure [20•]. For MSAs constructed from large cross-species alignments a correction should be applied for over-counting of sequences or organisms which received more experimental focus or which are phylogenetically related. Current methods to account for this are rudimentary, involving simple sequence similarity cutoffs, and it has been suggested that a more rigorous approach is desirable [20•,63]. In HIV datasets, in contrast, corrections are applied to sequences originating from the same individual but otherwise such phylogenetic filtering is not performed, because evidence suggests HIV data is not distorted by phylogenetic effects [9,43••,64].

Extensions to the Potts Hamiltonian have also been proposed, such as inclusion of higher order terms [32,65,66], inclusion of population-genetics terms [67] to account for phylogenetic structure. Other studies examine how to combine data from multiple data sources or inference methods [68,69].

Concluding remarks

Potts statistical models of protein sequence variation are being used to map the conformational energy and fitness landscapes of proteins with impressive results, but the full potential of this modeling is just beginning to be tapped. Through constraints imposed by evolution so as to maintain the fitness of the organism, pair and higher order correlations are induced in the sequence patterns that are captured by the models, even though the Potts Hamiltonian only includes pair interaction energy terms. Potts modeling can be used to interrogate epistatic effects whereby the probability of observing a particular mutation in a protein sequence depends strongly on the background. There are many opportunities and challenges to pursue; they include establishing closer synergies between protein conformational free energy landscapes and Potts model evolutionary landscapes, clarifying the relationship between Potts model parameters and physical free energy terms, and incorporating population genetics concepts into preprocessing steps which are performed on MSAs, as well as the interpretation of the model itself.

Acknowledgments

This work was supported by the National Institutes of Health P50 GM103368 and R01 GM30580.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

• of special interest

•• of outstanding interest

1. de Juan D, Pazos F, Valencia A. Emerging methods in protein co-evolution. Nat Rev Genet. 2013; 14:249–261. http://dx.doi.org/10.1038/nrg3414. [PubMed: 23458856]
2. Stein RR, Marks DS, Sander C. Inferring pairwise interactions from biological data using maximum-entropy probability models. PLoS Comput Biol. 2015; 11:e1004182. http://dx.doi.org/10.1371/journal.pcbi.1004182. [PubMed: 26225866]
3. Serohijos AW, Shakhnovich EI. Merging molecular mechanism and evolution: theory and computation at the interface of biophysics and evolutionary population genetics. Curr Opin Struct Biol. 2014; 26:84–91. http://dx.doi.org/10.1016/j.sbi.2014.05.005. [PubMed: 24952216]
4. Neuwald AF. Gleaning structural and functional information from correlations in protein multiple sequence alignments. Curr Opin Struct Biol. 2016; 38:1–8. http://dx.doi.org/10.1016/j.sbi.2016.04.006. [PubMed: 27179293]
5. Mzard M, Mora T. Constraint satisfaction problems and neural networks: a statistical physics perspective. J Physiol Paris. 2009; 103:107–113. http://dx.doi.org/10.1016/j.jphysparis.2009.05.013. [PubMed: 19616623]
6. Mora T, Bialek W. Are biological systems poised at criticality? J Stat Phys. 2011; 144:268–302. http://dx.doi.org/10.1007/s10955-011-0229-4.

7. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008; 5:621–628. http://dx.doi.org/10.1038/nmeth.1226. [PubMed: 18516045]
8. Jabara CB, Jones CD, Roach J, Anderson JA, Swanstrom R. Accurate sampling and deep sequencing of the HIV-1 protease gene using a primer ID. Proc Natl Acad Sci U S A. 2011; 108:20166–20171. http://dx.doi.org/10.1073/pnas.1110064108. [PubMed: 22135472]
9. Flynn WF, Chang MW, Tan Z, Oliveira G, Yuan J, Okulicz JF, Torbett BE, Levy RM. Deep sequencing of protease inhibitor resistant HIV patient isolates reveals patterns of correlated mutations in Gag and protease. PLOS Comput Biol. 2015; 11:e1004249. http://dx.doi.org/10.1371/journal.pcbi.1004249. [PubMed: 25894830]
10. Socolich M, Lockless SW, Russ WP, Lee H, Gardner KH, Ranganathan R. Evolutionary information for specifying a protein fold. Nature. 2005; 437:512–518. http://dx.doi.org/10.1038/nature03991. [PubMed: 16177782]
11. Russ WP, Lowery DM, Mishra P, Yaffe MB, Ranganathan R. Natural-like function in artificial WW domains. Nature. 2005; 437:579–583. http://dx.doi.org/10.1038/nature03990. [PubMed: 16177795]
12. Liu Z, Chen J, Thirumalai D. On the accuracy of inferring energetic coupling between distant sites in protein families from evolutionary imprints: illustrations using lattice model. Proteins Struct Funct Bioinform. 2009; 77:823–831. http://dx.doi.org/10.1002/prot.22498.
13. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T. Identification of direct residue contacts in protein–protein interaction by message passing. Proc Natl Acad Sci U S A. 2009; 106:67–72. http://dx.doi.org/10.1073/pnas.0805923106. [PubMed: 19116270]
14. Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, Sander C. Protein 3D structure computed from evolutionary sequence variation. PLoS ONE. 2011; 6:e28766EP. http://dx.doi.org/10.1371/journal.pone.0028766. [PubMed: 22163331]
15. Marks DS, Hopf TA, Sander C. Protein structure prediction from sequence variation. Nat Biotechnol. 2012; 30:1072–1080. http://dx.doi.org/10.1038/nbt.2419. [PubMed: 23138306]
16. Ovchinnikov S, Kamisetty H, Baker D, Roux B. Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. eLife. 2014; 3:e02030. http://dx.doi.org/10.7554/eLife.02030. [PubMed: 24842992]
17. Sukowska JI, Morcos F, Weigt M, Hwa T, Onuchic JN. Genomics-aided structure prediction. Proc Natl Acad Sci U S A. 2012; 109:10340–10345. http://dx.doi.org/10.1073/pnas.1207864109. [PubMed: 22691493]
18. Ovchinnikov S, Kinch L, Park H, Liao Y, Pei J, Kim DE, Kamisetty H, Grishin NV, Baker D, Shan Y. Large-scale determination of previously unsolved protein structures using evolutionary information. eLife. 2015; 4:e09248. http://dx.doi.org/10.7554/eLife.09248. [PubMed: 26335199]
19•. Morcos F, Jana B, Hwa T, Onuchic JN. Coevolutionary signals across protein lineages help capture multiple protein conformations. Proc Natl Acad Sci U S A. 2013; 110:20533–20538. http://dx.doi.org/10.1073/pnas.1315625110. Uses Potts model constraints in structure-based models to explore conformational landscapes of three proteins, revealing multi-welled landscapes which capture previously-proposed intermediates in addition to known functional states. [PubMed: 24297889]
20•. Sutto L, Marsili S, Valencia A, Gervasio FL. From residue coevolution to protein conformational ensembles and functional dynamics. Proc Natl Acad Sci U S A. 2015; 112:13567–13572. http://dx.doi.org/10.1073/pnas.1508584112. This study uses Monte Carlo methods to infer a Potts model. Using the predicted contacts as constraints in a coarse-grained simulation, the conformational landscape between two end-states is mapped, and conformational dynamics of subsets of the protein are examined. [PubMed: 26487681]
21. dos Santos RN, Morcos F, Jana B, Andricopulo AD, Onuchic JN. Dimeric interactions and complex formation using direct coevolutionary couplings. Sci Rep. 2015; 5:13652. http://dx.doi.org/10.1038/srep13652. [PubMed: 26338201]
22. Malinverni D, Marsili S, Barducci A, Rios PDL. Large-scale conformational transitions and dimerization are encoded in the amino-acid sequences of Hsp70 chaperones. PLoS Comput Biol. 2015; 11:e1004262. http://dx.doi.org/10.1371/journal.pcbi.1004262. [PubMed: 26046683]

23. Feinauer C, Szurmant H, Weigt M, Pagnani A. Inter-protein sequence co-evolution predicts known physical interactions in bacterial ribosomes and the Trp operon. PLOS ONE. 2016; 11:e0149166. http://dx.doi.org/10.1371/journal.pone.0149166. [PubMed: 26882169]

24. Espada R, Parra RG, Mora T, Walczak AM, Ferreiro DU. Capturing coevolutionary signals inrepeat proteins. BMC Bioinform. 2015; 16:1–10.

25. Jeong C-S, Kim D. Structure-based Markov random field model for representing evolutionary constraints on functional sites. BMC Bioinform. 2016; 17:1–11. http://dx.doi.org/10.1186/s12859-016-0948-2.

26. Tang Y, Huang YJ, Hopf TA, Sander C, Marks DS, Montelione GT. Protein structure determination by combining sparse NMR data with evolutionary couplings. Nat Methods. 2015; 12:751–754. http://dx.doi.org/10.1038/nmeth.3455. [PubMed: 26121406]

27. Hayat S, Sander C, Marks DS, Elofsson A. All-atom 3D structure prediction of transmembrane-barrel proteins from sequences. Proc Natl Acad Sci U S A. 2015; 112:5413–5418. http://dx.doi.org/10.1073/pnas.1419956112. [PubMed: 25858953]

28. Cheng RR, Raghunathan M, Noel JK, Onuchic JN. Constructing sequence-dependent protein models using coevolutionary information. Protein Sci. 2015; 25:111–122. http://dx.doi.org/10.1002/pro.2758. [PubMed: 26223372]

29•. Haldane A, Flynn WF, He P, Vijayan R, Levy RM. Structural propensities of kinase family proteins from a Potts model of residue co-variation. Protein Sci. 2016; 25:1378–1384. http://dx.doi.org/10.1002/pro.2954. This study demonstrates how sequence-dependent conformational preferences can be predicted using the Potts model, for many sequences in the studied family. [PubMed: 27241634]

30••. Figliuzzi M, Jacquier H, Schug A, Tenaillon O, Weigt M. Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase TEM-1. Mol Biol Evol. 2015; 33:268–280. http://dx.doi.org/10.1093/molbev/msv211. A comparison of several thousand experimental fitness measurements of beta lactamase TEM-1 with statistical energies from a Potts model parameterized via mean-field DCA, an independent model, and other state-of-the-art structure-based and sequence-based methodologies used to predict effects of mutations. Results show that the Potts model and independent model outperform other methods, especially structure based methods, at predicting fitness of individual sequences, and that residues up to 20 away in conformation space contribute to the fitness upon mutation. [PubMed: 26446903]

31••. Morcos F, Schafer NP, Cheng RR, Onuchic JN, Wolynes PG. Coevolutionary information, protein folding landscapes, and the thermodynamics of natural selection. Proc Natl Acad Sci U S A. 2014; 111:12408–12413. http://dx.doi.org/10.1073/pnas.1413575111. A study interpreting the Potts model in light of energy landscape theory, showing how the Potts Hamiltonian may be related to the free energy of folding, and how parameters in energy landscape theory such as the glass temperature can be deduced from a Potts analysis. [PubMed: 25114242]

32. Contini A, Tiana G. A many-body term improves the accuracy of effective potentials based on protein coevolutionary data. J Chem Phys. 2015; 143:025103. http://dx.doi.org/10.1063/1.4926665. [PubMed: 26178131]

33. Lapedes A, Giraud B, Jarzynski C. Using sequence alignments to predict protein structure and stability with high accuracy. 2002 arXiv:1207.2484.

34. Hopf TA, Ingraham JB, Poelwijk FJ, Springer M, Sander C, Marks DS. Quantification of the effect of mutations using a global probability model of natural sequence variation. 2015 arXiv:1510.04612.

35. Onuchic JN, Luthey-Schulten Z, Wolynes PG. Theory of protein folding: the energy landscape perspective. Annu Rev Phys Chem. 1997; 48:545–600. http://dx.doi.org/10.1146/annurev.physchem.48.1.545. [PubMed: 9348663]

36. Bryngelson JD, Wolynes PG. Spin glasses and the statistical mechanics of protein folding. Proc Natl Acad Sci U S A. 1987; 84:7524–7528. [PubMed: 3478708]

37•. Haq O, Andrec M, Morozov AV, Levy RM. Correlated electrostatic mutations provide a reservoir of stability in HIV protease. PLoS Comput Biol. 2012; 8:e1002675EP. http://dx.doi.org/10.1371/journal.pcbi.1002675. Parameterizes a Potts model on the charged residues of HIV-1 protease using message-passing and demonstrates that the model can predict sequence probabilities in a

separate database. This is one of the first studies to show that the Potts model captures higher order statistics up to those of full sequences. [PubMed: 22969420]

38. Jacquin H, Gilson A, Shakhnovich E, Cocco S, Monasson R. Benchmarking inverse statistical approaches for protein structure and design with exactly solvable models. PLoS Comput Biol. 2016; 12:e1004889. http://dx.doi.org/10.1371/journal.pcbi.1004889. [PubMed: 27177270]
39. Boucher JI, Bolon DNA, Tawfik DS. Quantifying and understanding the fitness effects of protein mutations: laboratory versus nature. Protein Sci. 2016; 25:1219–1226. http://dx.doi.org/10.1002/pro.2928. [PubMed: 27010590]
40. Butler TC, Barton JP, Kardar M, Chakraborty AK. Identification of drug resistance mutations in HIV from constraints on natural evolution. Phys Rev E. 2016; 93:022412. http://dx.doi.org/10.1103/PhysRevE.93.022412. [PubMed: 26986367]
41. Shekhar K, Ruberman CF, Ferguson AL, Barton JP, Kardar M, Chakraborty AK. Spin models inferred from patient data faithfully describe HIV fitness landscapes and enable rational vaccine design. Phys Rev E. 2013; 88:1539–3755. http://dx.doi.org/10.1103/PhysRevE.88.062705.
42••. Ferguson AL, Mann JK, Omarjee S, Ndung'u T, Walker BD, Chakraborty AK. Translating HIV sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design. Immunity. 2013; 38:606–617. http://dx.doi.org/10.1016/j.immuni.2012.11.022. Using Ising models of the 4 principal structural proteins of HIV-1, this manuscript shows that experimental measurements of replicative fitness is proportional to the Ising statistical energy for these proteins. Further, clinically observed sequence fragments that lead to immune escape are computed to be highly fit in the model. They leverage the results to design a Gag sequence fragment which their model suggests will allow the host immune system to recognize the most vulnerable Gag epitopes. [PubMed: 23521886]
43••. Barton JP, Goonetilleke N, Butler TC, Walker BD, McMichael AJ, Chakraborty AK. Relative rate and location of intra-host HIV evolution to evade cellular immunity are predictable. Nat Commun. 2016; 7 http://dx.doi.org/10.1038/ncomms11660. A concise study which shows escape from immune pressure in HIV is dependent on the sequence background; identical HLA epitopes targeted by the immune system in some patients are primed for escape by compensatory mutations in the background sequence. This study also demonstrates that equilibrium predictions of in vivo protein fitness using Potts model can be improved by incorporating the Potts Hamiltonian into dynamic Wright-Fisher evolutionary simulations.
44. Mann JK, Barton JP, Ferguson AL, Omarjee S, Walker BD, Chakraborty A, Ndung'u T. The fitness landscape of HIV-1 Gag: advanced modeling approaches and validation of model predictions by in vitro testing. PLoS Comput Biol. 2014; 10:e1003776EP. http://dx.doi.org/10.1371/journal.pcbi.1003776. [PubMed: 25102049]
45•. Flynn, WF., Haldane, A., Torbett, BE., Levy, RM. Inference of epistatic effects leading to entrenchment and drug resistance in HIV-1 protease. 2016. bioRxivhttp://dx.doi.org/10.1101/063750. A forthcoming study using Potts models to demonstrate the context-dependent effects promoting resistance mutations in drug-experienced HIV-1 protease, emphasizing that deleterious primary resistance mutations become stabilizing for the sequence as background mutations accumulate. Experimental and computational validation of the Potts model is also shown
46. Grenfell BT, Pybus OG, Gog JR, Wood JLN, Daly JM, Mumford JA, Holmes EC. Unifying the epidemiological and evolutionary dynamics of pathogens. Science. 2004; 303:327–332. http://dx.doi.org/10.1126/science.1090727. [PubMed: 14726583]
47. Nowak, MA. Evolutionary dynamics: exploring the equations of life. Harvard University Press; 2006.
48. Sella G, Hirsh AE. The application of statistical physics to evolutionary biology. Proc Natl Acad Sci U S A. 2005; 102:9541–9546. http://dx.doi.org/10.1073/pnas.0501865102. [PubMed: 15980155]
49. Fowler DM, Fields S. Deep mutational scanning: a new style of protein science. Nat Methods. 2014; 11:801–807. http://dx.doi.org/10.1038/nmeth.3027. [PubMed: 25075907]
50. Olson CA, Wu NC, Sun R. A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. Curr Biol. 2014; 24:2643–2651. http://dx.doi.org/10.1016/j.cub.2014.09.072. [PubMed: 25455030]

51. Wu NC, Olson CA, Sun R. High-throughput identification of protein mutant stability computed from a double mutant fitness landscape. Protein Sci. 2016; 25:530–539. http://dx.doi.org/10.1002/pro.2840. [PubMed: 26540565]
52. Bank C, Hietpas RT, Jensen JD, Bolon DNA. A systematic survey of an intragenic epistatic landscape. Mol Biol Evol. 2015; 32:229–238. http://dx.doi.org/10.1093/molbev/msu301. [PubMed: 25371431]
53. Schneidman E, Berry MJ, Segev R, Bialek W. Weak pairwise correlations imply strongly correlated network states in a neural population. Nature. 2006; 440:1007–1012. http://dx.doi.org/10.1038/nature04701. [PubMed: 16625187]
54. Pollock DD, Thiltgen G, Goldstein RA. Amino acid coevolution induces an evolutionary Stokes shift. Proc Natl Acad Sci U S A. 2012; 109:E1352–E1359. http://dx.doi.org/10.1073/pnas.1120084109. [PubMed: 22547823]
55. Gong LI, Suchard MA, Bloom JD, Pascual M. Stability-mediated epistasis constrains the evolution of an influenza protein. eLife. 2013; 2:e00631. http://dx.doi.org/10.7554/eLife.00631. [PubMed: 23682315]
56. Shah P, McCandlish DM, Plotkin JB. Contingency and entrenchment in protein evolution under purifying selection. Proc Natl Acad Sci U S A. 2015; 112:E3226–E3235. http://dx.doi.org/10.1073/pnas.1412933112. [PubMed: 26056312]
57. Barton, JP., Leonardis, ED., Coucke, A., Cocco, S. ACE: adaptive cluster expansion for maximum entropy graphical model inference. Bioinformatics. 2016. http://dx.doi.org/10.1093/bioinformatics/btw328
58. Jones DT, Buchan DWA, Cozzetto D, Pontil M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics. 2012; 28:184–190. http://dx.doi.org/10.1093/bioinformatics/btr638. [PubMed: 22101153]
59. Lövkvist C, Lan Y, Weigt M, Aurell E, Ekeberg M. Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. Phys Rev E. 2013; 87:012707. http://dx.doi.org/10.1103/PhysRevE.87.012707.
60. Kamisetty, H., Ovchinnikov, S., Baker, D. Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. Proc Natl Acad Sci U S A. 2013. http://dx.doi.org/10.1073/pnas.1314045110
61. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc Natl Acad Sci U S A. 2011; 108:E1293–E1301. http://dx.doi.org/10.1073/pnas.1111471108. [PubMed: 22106262]
62. Ekeberg M, Hartonen T, Aurell E. Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences. J Comput Phys. 2014; 276:341–356.
63. Avila-Herrera A, Pollard KS. Coevolutionary analyses require phylogenetically deep alignments and better null models to accurately detect inter-protein contacts within and between species. BMC Bioinform. 2015; 16:1–18. http://dx.doi.org/10.1186/s12859-015-0677-y.
64. Gupta A, Adami C. Strong selection significantly increases epistatic interactions in the long-term evolution of a protein. PLoS Genet. 2016; 12:e1005960. http://dx.doi.org/10.1371/journal.pgen.1005960. [PubMed: 27028897]
65. Feinauer C, Skwark MJ, Pagnani A, Aurell E. Improving contact prediction along three dimensions. PLoS Comput Biol. 2014; 10:e1003847EP. http://dx.doi.org/10.1371/journal.pcbi.1003847. [PubMed: 25299132]
66. Haq O, Levy RM, Morozov AV, Andrec M. Pairwise and higher-order correlations among drug-resistance mutations in HIV-1 subtype B protease. BMC Bioinform. 2009; 14:1–14. http://dx.doi.org/10.1186/1471-2105-10-S8-S10.
67. Obermayer B, Levine E. Inverse Ising inference with correlated samples. New J Phys. 2014; 16:123017. http://dx.doi.org/10.1088/1367-2630/16/12/123017.
68. Ma J, Wang S, Wang Z, Xu J. Protein contact prediction by integrating joint evolutionary coupling analysis and supervised learning. Bioinformatics. 2015; 31:3506–3513. [PubMed: 26275894]

69. Jones DT, Singh T, Kosciolek T, Tetchner S. MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. Bioinformatics. 2015; 31:999–1006. http://dx.doi.org/10.1093/bioinformatics/btu791. [PubMed: 25431331]

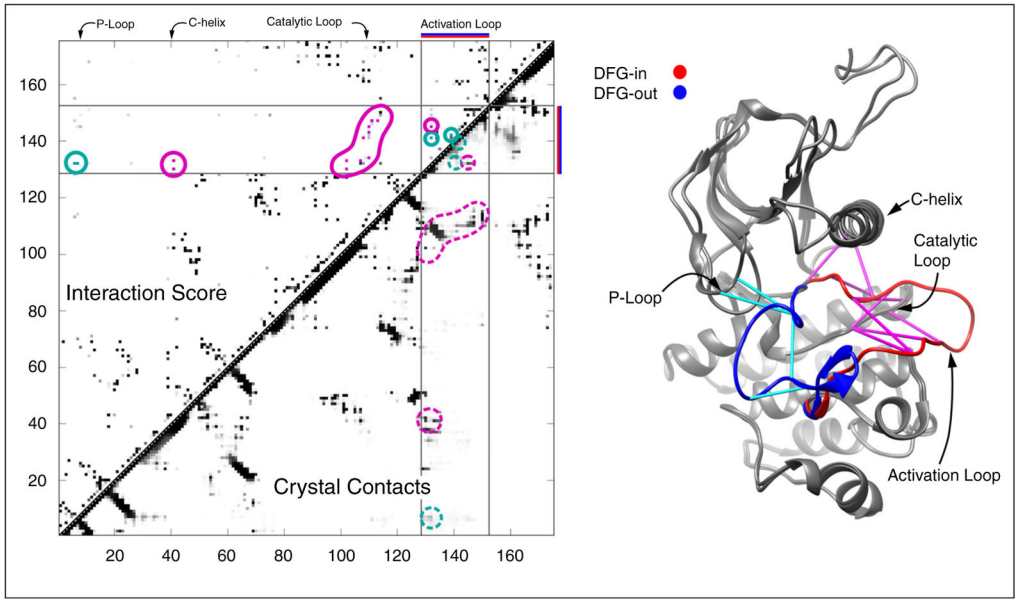

Figure 1.

The Potts model can be used to predict sequence-dependent conformational landscapes. **(Left)** Coevolutionary (Potts) interaction score map (upper triangle), and crystal structure contact frequency map (lower triangle, 6 Å cutoff) for the kinase family, showing high correspondence [29•]. Interactions predicted to be relevant to a conformational transition in kinases between a 'DFG-out' (cyan) and a 'DFG-in' (magenta) conformation are highlighted. **(Right)** Crystal structure of the two conformations (pdb-id 1IEP and 2GQG) illustrating the change in the activation loop (colored red/blue), showing C-α to C-α contacts relevant to the transition predicted by the Potts model (cyan/magenta, as in the contact map). The Potts model gives a sequence-dependent score $J_{ij}(s_i, s_j)$ for each interaction.

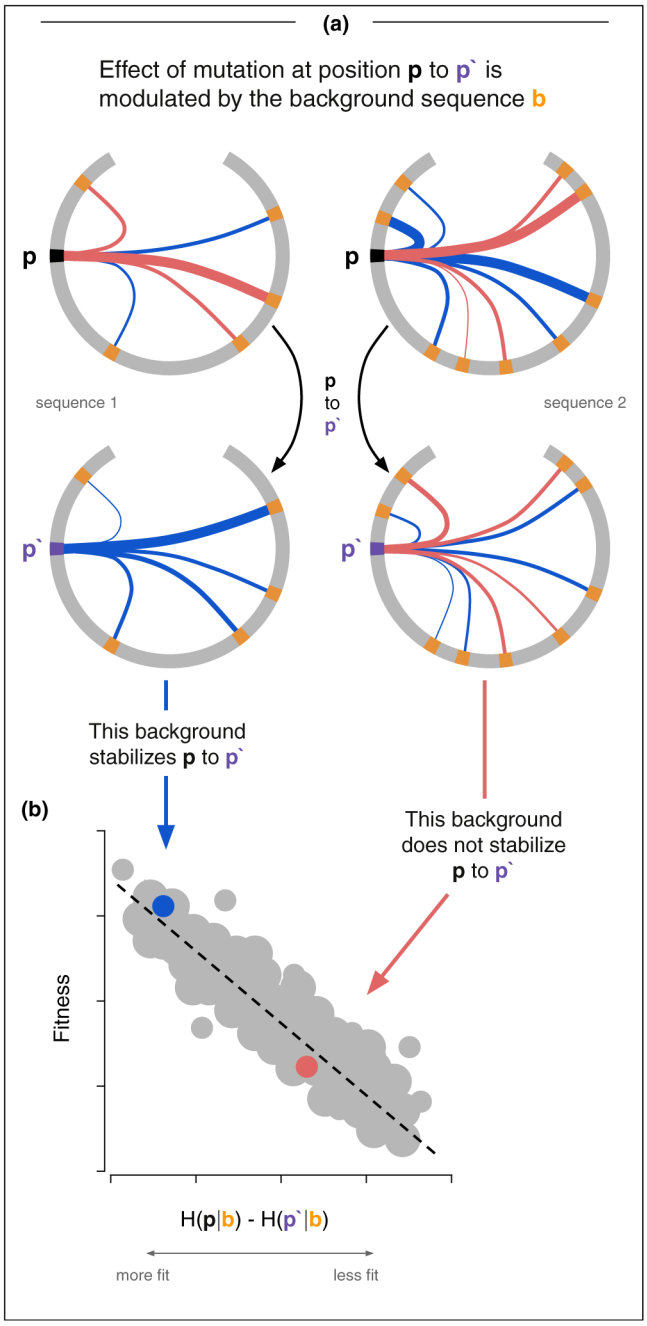

Figure 2.

The effect of a mutation depends on the background in which it's acquired. Shown in **(a)** are two sequences (gray annulus) with a focal residue p (black), and different background mutations (orange). Stabilizing (blue, $J < 0$) and destabilizing (red, $J > 0$) couplings between the focal residue and the background mutations are shown. The effect of mutation at the focal residue (p to p', black to purple) is dependent on the couplings between the mutation and the background sequence. Certain backgrounds are more accommodating than others of

this particular mutation, which leads to higher observed fitness **(b)**. Potts statistical energy H is well correlated with experimentally observed fitness measurements.