Article

Protein Conformational States—A First Principles Bayesian Method [†]

David M. Rogers (ID)

National Center for Computational Sciences, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA; rogersdm@ornl.gov

check for updates

Abstract: Automated identification of protein conformational states from simulation of an ensemble of structures is a hard problem because it requires teaching a computer to recognize shapes. We adapt the naïve Bayes classifier from the machine learning community for use on atom-to-atom pairwise contacts. The result is an unsupervised learning algorithm that samples a 'distribution' over potential classification schemes. We apply the classifier to a series of test structures and one real protein, showing that it identifies the conformational transition with $> 95\%$ accuracy in most cases. A nontrivial feature of our adaptation is a new connection to information entropy that allows us to vary the level of structural detail without spoiling the categorization. This is confirmed by comparing results as the number of atoms and time-samples are varied over 1.5 orders of magnitude. Further, the method's derivation from Bayesian analysis on the set of inter-atomic contacts makes it easy to understand and extend to more complex cases.

Keywords: Bernoulli mixture; Bayesian clustering; unsupervised classification

1. Introduction

The conventional description of protein dynamics asserts that proteins posses intrinsic conformational states [1]. An enzyme may cycle between catalytic and open states [2]. An ion channel may open and close its central pore [3]. A chaperone protein assists transformation of large, hydrophobic proteins from initial, linear, to folded shapes [4]. X-ray and cryo electron-microscopy reveals conformations with small motions on the 1–5 Ångstrom level for those proteins that crystallize [5]. Neutron scattering and nuclear magnetic resonance structures of room temperature proteins show greater shape variability, but are usually able to classify structures into a few 'canonical' structures.

Advances in molecular modeling have made it possible to simulate the protein folding process, generating very large numbers of samples, free energy landscapes, and information on kinetics. Nevertheless, computational identification of distinct conformational states from molecular simulations has remained an active area of methodological research. Approaches fall roughly into two classes – linear and nonlinear. Linear methods work with the real $3N$-dimensional vector of atomic coordinates as a linear space, using pairwise alignments and Euclidean (RMSD) distances between structures. Nonlinear methods work with 'features' derived from nonlinear functions of the atomic coordinates – like pairwise distances between alpha-carbons. Well-known linear techniques include principle

component analysis and linkage clustering based on pair-RMSDs. Nonlinear techniques include methods based on internal coordinates (relative distances and angles between groups of atoms) and T-distributed stochastic neighbor embedding [6].

Both linear and nonlinear techniques can be applied to subsets or groups of atoms to classify site-specific or inter-domain structures, respectively. This can be done trivially by handing off different point-sets to the analysis. Thus, most methods can be adapted to run with reasonable compute time and yield classifications most relevant to a motion under investigation. Automatically detecting protein secondary and tertiary structures, however, remains a challenging problem which we cannot address here.

One of the principle methods developed for visualizing domain motions in proteins is DynDom [7, 8]. DynDom works from two input structures and determines relative domain rotations (using internal coordinates). This allows predicting transition motions from experimentally observed conformers, but not conformations from observed motions. On the other hand, many structure-to-structure similarity classification methods have emerged for this problem [9–11]. A recent review of such dimensionality reduction methods for protein conformational spaces noted that nonlinear methods are generally better than Cartesian or linear ones, but that the complexity of assumptions behind those models makes them difficult to work with and adapt [6,12].

This work presents a complete inference method derived from one single statistical hypothesis: that conformational states are defined by sets of contacting residues. Specifically, we hypothesize that the conformational state, k, uniquely determines which pairs of residues u,v, will be touching. Like a weighted coin flip, the contact probability is $\mu_{k,(u:v)}$ – independently from all the other contacting pairs. Each conformational state is thus characterized by a vector, μ_k, encoding the set of contacting pairs in state k.

The statistical model derived from this problem statement is termed a Bernoulli mixture model for binary feature classification [13]. The problem setup is similar to the Naive Bayes method [14]. However, because the categories are not known in advance, this is an unsupervised learning and classification problem.

Bernoulli mixture models have been applied extensively in the field of text subject analysis [15], optical character recognition [14], and image feature classification [13]. Essentially all of these applications have been successful at building extremely accurate classification models. The latter work also presents a thorough summary of sampling methods.

However, there remain difficulties sampling the distribution over categories, μ, especially when the number of categories and reference classifications are not known in advance. The well-known expectation-maximization algorithm (EM) [16] is available in principle, but is not a replacement for sampling. Theoretical work on the EM method [17] shows that redundant categories will result in many circumstances. In this work, we have introduced a prior that eliminates redundant categorizations.

This work is structured as follows. Section 2 presents the underlying probability distribution of categories, then outlines a novel method for quickly sampling parameter space – achieving category inference. Full technical details are present in Appendices A.1-A.2. Section 3 describes test problems on which the method is demonstrated. Section 4 presents results demonstrating that the method creates structurally meaningful categories with $> 90\%$ accuracy. Although the potential application space is vast, this work focuses on proving method robustness using well-defined synthetic test problems. Each follows a time sequence mimicking domain motions in proteins – so that the classification accuracy can be judged by correctly assigning categories in time-order. For the practitioner interested in trying the method directly, full source code and scripts reproducing the test cases in this work are available under the GPLv3 license [18].

2. Theory

A naïve Bayes model (for bit-strings) assumes that structural input samples, $i = 1, \ldots, N$ are generated by first selecting a conformational state, $z_i \in \{1, \ldots, K\}$, with probability π_{z_i}, and then

independently deciding whether each point-to-point contact ($x_{ij}, j = 1, \ldots M$), is made with probability $\mu_{z_i j}$. If contact j is made in sample number i, then $x_{ij} = 1$. The bit-vector, x_i, is said to posess feature j. Otherwise $x_{ij} = 0$, and the feature is absent. The model parameters are thus $\theta = (K, \pi, \mu)$.

It leads to a sampling distribution,

$$P(xz|\theta) = \prod_i \pi_{z_i} \prod_{j=1}^{F} \mu_{z_i j}^{x_{ij}} (1 - \mu_{z_i j})^{1 - x_{ij}} \tag{1}$$

$$= \prod_{k=1}^{K} \pi_k^{N_k} \prod_{j=1}^{M} \mu_{kj}^{N_{kj}} (1 - \mu_{kj})^{N_k - N_{kj}}. \tag{2}$$

The second line above notes that, once the categorization, z, is known, the sampling distribution is easy to express in terms of feature counts in $\{i : z_i = k\}$ – the set of samples assigned to category k,

$$N_k = |\{i : z_i = k\}|, \qquad\qquad N_{kj} = \sum_{\{i : z_i = k\}} x_{ij}. \tag{3}$$

The first is the number of samples in set k, and the second is the number of times each contact is seen in that set.

According to Bayes' theorem [19], we can turn this around to predict two important things – the probability that sample i belongs to category k, (read z given x and θ),

$$P(z_i = k | x_i \theta) \propto \prod_{j=1}^{M} \mu_{kj}^{x_{ij}} (1 - \mu_{kj})^{1 - x_{ij}}, \tag{4}$$

and also the probability distribution over all possible parameters,

$$P(\theta z | x I) = C(x) P(xz|\theta) P(\theta|I), \tag{5}$$

where $C(x)$ is an x-dependent normalization constant. Sampling this distribution provides everything—the categorizations, z, the conformational states, π, μ, and even a predicted number of categories, K.

In Bayesian probability, a prior distribution has to be assumed by the researcher. The prior characterizes the parameter space, independently from any sampled data. Our prior distribution over parameters, introduced below, is $P(\theta|I)$. Since the parameters directly determine the sampling distribution, the prior does not affect it [$P(xz|\theta) = P(xz|\theta I)$]. Note that this work juggles between two different priors, I and U, because the inference problem is simpler using $P(\theta|U)$, but $P(\theta|I)$ eliminates redundant solutions.

We choose a prior probability,

$$P(\theta|I) \propto P(\theta|U) \prod_{k<l} (1 - B(\mu_k, \mu_l)), \tag{6}$$

$$P(\theta|U) = \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \prod_{k=1}^{K} \pi_k^{\alpha - 1} \tag{7}$$

that enforces uniqueness of the categories. Here $B(\mu_k, \mu_l) = B_{kl}$ is the Bhattacharyya similarity between distributions p and q,

$$B(p, q) = \prod_{j=1}^{M} \sqrt{p_j q_j} + \sqrt{(1 - p_j)(1 - q_j)}. \tag{8}$$

If two categories share the same distribution, then $p = q$, and $B(p, q) = 1$. This forces our estimator to return zero likelihood that $\mu_k = \mu_l$ for any $k \neq l$.

The second part of Equation (7), $P(\theta|U)$, is a conventional prior used for Bernoulli mixture inference in the literature. We use $\alpha = M + 1$ throughout this work. Appendix A.1 contains a detailed justification for this choice. Essentially, it forces the likelihood for dividing a category into two parts to be asymptotically insensitive to the number of features, M. A proof of this fact, as well as a useful connection to the information entropy of compression is also present in Appendix A.1.

2.1. Sampling Method

Our sampling method is traditional Markov Chain Monte Carlo using four types of moves: a recategorization move, where categories, z, are assigned according to $P(z|\theta x I)$, a reclassification move, where θ is sampled from $P(\theta|zxU)$ and accepted with probability $\prod_{k<l}(1 - B_{kl})$, and one split and one join rule. The function, $P(\theta|zxU)$, referred to here is just $P(\theta|zxI)$ without the Bhattacharyya distance terms, and with a different constant prefactor.

For the split trial move, one of the categories, k, is split into two new categories. Every member of k is re-assigned into one of two new sets, labaled L or R. Join moves are the opposite of split moves. This re-categorization changes the set labels, z. Specifically, z_i goes from k to L or R for every i formerly at $z_i = k$.

Generating split or join trial moves was done by randomly choosing either one category to split or choosing two categories to join. For splits, member i of category k is moved to set L with probability $\eta^{x_{ij}}(1 - \eta)^{1 - x_{ij}}$. We used $\eta = 0.9$, but any $\eta \in (0.5, 1.0)$ should work in principle. If all elements end up in L or R, the partitioning is re-done. To concentrate splitting on productive cases, we did not attempt to split categories with $N_{kj} = 0$ or $N_{kj} = N_k$. Immediately after splitting or joining categories, a reclassification move (re-assigning θ) was performed. Category split moves were accepted using the Metropolis criterion, which is the smaller of $P_{\text{gen,join}}P_{\text{split}}/P_{\text{gen,split}}P_{\text{joined}}$ or one. Explicit formulas for the move generation probabilities ($P_{\text{gen,split}}$, etc.) are provided in Appendix A.2.

Figure 1 provides a graphical summary of this inference scheme. Each conformational sample is mapped to a bit-string, which is used as the basis for inferring μ. Inference proceeds by sampling potential parameters until a good explanation for the data is found. Trial moves that re-categorize and update μ look horizontally to find better category prototypes. Trial moves that split categories based on presence or absence of some features allow us to traverse category space vertically.
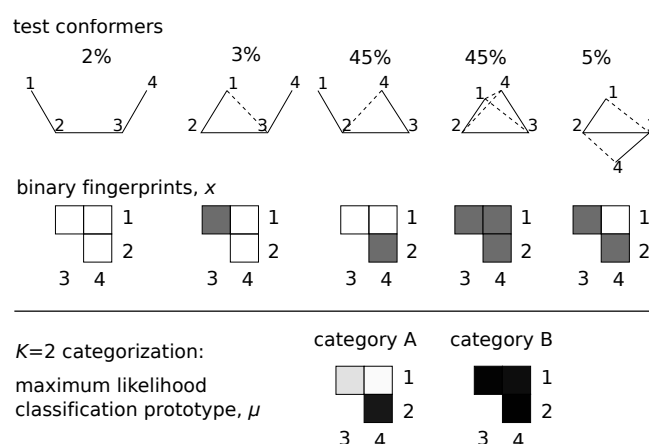

Figure 1. Classification applied to a 4-site structure. Five conformers (with frequencies shown) are input to the method. Bit-strings, x, for each conformer represent pairs of contacting residues (black squares correspond to dashed lines). Sampling the posterior distribution over (K, π, μ) provides a maximum-likelihood categorization at each K. We show both category prototypes, μ, determined by the algorithm for the $K = 2$ classification. Darker shading in μ represents contacts with higher probability of forming.

3. Test Systems

The ability of our sampling procedure to predict categories was tested on three geometrical systems: 'chomp' (a closing angle), 'heli' (a rotating line), and 'glob' (three rotating spheres). Each system was generated as a time-series of 1000 frames for approximately P total particles in 2 or 3-dimensional space. After generation, Gaussian random noise of width $\sigma = 0.1$ was added to every degree of freedom.

Figure 2 shows images of these three test trajectories. A complete description of the coordinate generation methods is present in Appendix A.3. All three systems were processed into binary feature data by calculating pairwise distances between all points. Pairs of points within 2 distance units were translated to 1 (representing contact). When forming the feature vectors, x, we removed features, j, for which every sample showed the same result (all contacting or all disconnected). It is important to note that this removal changes M seen by the algorithm – usually decreasing it well below $P(P-1)/2$.
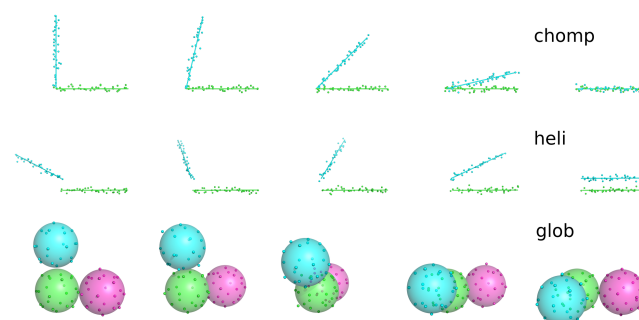

Figure 2. Sketch of the three synthetic motions used for testing. Lines, spheres, and colors are used to guide the eye, but the classification engine sees only the unlabeled points. The time axis proceeds left to right. For the 'glob' system, the sphere on the right moves behind, and then back to the right again while the top sphere moves slowly downward.

Critically, we repeated these classifications for a range of material points, P. This tested robustness of the unsupervised classification problem with respect to the amount of features available. Adding more points without changing the geometry of the problem should not change the number of categories detected.

For each run, five independent MCMC chains were started from an assignment of all points to a single category. Each chain ran for 1000 steps. Samples were collected every 10 steps – starting from step 500. The acceptance probabilities for category split/join Monte Carlo moves varied around 10-13%.

4. Results

We analyzed the results of MC in two different ways. First, the categories assigned were tested for grouping in time. Since the contact lists (on which the categorization is based) varied slowly over time, we expect categories to come in 'runs.' Second, we computed histograms over the number of categories, K. This is a strong test of the method's sensitivity to the number of material points, P.

As expected, we found a high degree of correlation between categories and time for every case. Similar time-points were grouped into similar categories. To quantify these results, we counted transitions between category indices in time-order. For a perfect categorization, the number of transitions should equal the number of categories minus one. We computed the categorization accuracy in two ways. For each system, the left columns of Tables 1 and 2 are 100 minus the percent of mis-categorized frames. For lossy categorization at time-boundaries, we expect oscillation between two values. We quantified this by forming a transition matrix between categories, and removing transitions along the 'most likely path'. Excluding this boundary oscillation we found nearly 100% accuracy for the categorizations. Those are shown in the right columns of Tables 1 and 2.

Table 1. Categorization accuracy for each system type and number of points, *P*. The left column for each system is the percent of category assignments identical to their previous time (not including $K - 1$ required). The right column for each is the percent of category assignments that do not switch between runs.

P	chomp		heli		glob	
18	89.2	98.0	89.9	98.9	91.1	96.7
42	94.7	98.9	93.7	99.0	93.5	97.3
78	96.3	99.3	95.3	99.4	95.5	98.0
114	96.8	99.4	96.5	99.5	96.4	98.4
150	97.2	99.4	96.7	99.4	97.0	98.5
222	97.8	99.5	97.0	99.7	98.2	99.1
258	98.4	99.7	97.7	99.5	98.6	99.4
330	98.4	99.6	97.8	99.7	98.7	99.4
438	98.5	99.7	98.1	99.5	98.8	99.4

Table 2. Categorization accuracy for each test system as the sample number, *N*, varies. The number of points remains fixed at $P = 222$. Column labels are as in Table 1.

N	chomp		heli		glob	
125	99.3	99.2	99.8	99.2	99.5	99.3
250	99.4	99.6	99.3	99.6	99.2	99.4
500	98.8	99.6	97.7	99.7	98.6	99.3
1000	97.7	99.6	96.9	99.6	98.2	99.2
2000	96.9	99.4	96.5	99.6	98.0	99.1

Integrating the posterior probability (Equation (2) times Equation (7)) over π leads to factors like $\Gamma(N + \alpha K)^{-1}$, which seem prohibitively costly as *N* increases. We therefore wanted to check that the number of categories does not decrease as features or samples are added. Figure 3 shows the sampled probability distributions over *K*, the number of categories, for increasing *P* (Figure 3a) and *N* (Figure 3b). Interestingly, for every system, about five conformational states were deduced at $N = 1000$. As *N* increased, however, more categories were deduced and the distribution spread to higher numbers. This is probably reasonable, since more values of the 'time' coordinate generated a more fine-grained motion.
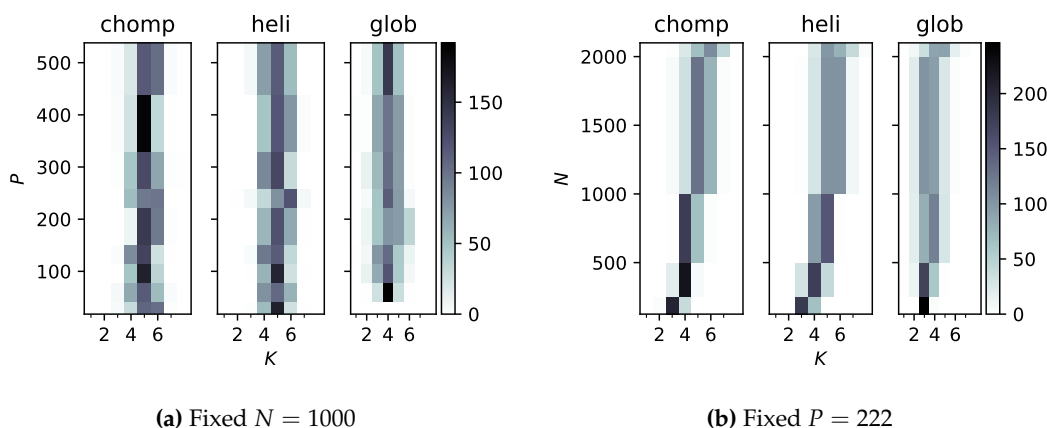

(a) Fixed $N = 1000$ **(b)** Fixed $P = 222$

Figure 3. Counts (out of 250 samples each) of the number of categories, *K*, as *P* or *N* are varied. The shape of the distribution function remains essentially constant as *P* varies, but tends to spread toward higher *K* with more samples.

4.1. Adenylate kinase open/closed transition

Finally, we tested the classification method against a protein with a well-characterized conformational transition. Adenylate kinase (ADK) converts ATP, ADP, and AMP by closing around substrate molecules [20]. The transition from closed to open was simulated in Reference [5] using steered molecular dynamics on a reaction coordinate interpolating between the electron density maps of PDB IDs 1AKE (Reference [20], closed) and 4AKE (Reference [2], open). The simulation data we used did not contain ligands, but did contain water and ions. Our analysis only made use of the alpha carbon (C_α) positions.

Features were calculated for each of $N = 3900$ equally spaced frames during steered dynamics by testing whether C_α to C_α distances were less than 5 Å. These structures contain $P = 214$ C_α-s. Sampling was carried out as described in Section 3, but 8 independent MC chains were sampled for 1250 steps (instead of 5 for 1000). The acceptance probability of split/join moves was 17%. During sampling, we saved the parameters, θ, possessing maximum likelihood values at each K.

Our implementation (Reference [18]) is parallelized so that each thread carries out an independent Monte Carlo chain. The run-time for the ADK example was less than 30 minutes. Larger runs on proteins up to 400 residues using 75,000 sampled protein conformations have been carried out using this method in under an hour. The implementation also contains a series of vizualization tools to highlight the protein regions responsible for detected conformation-to-conformation differences.

We then extracted conformational states with the highest probability for landing in each category as representative points for that category. Since the reference open and closed PDB conformations formed extreme points, our representative structures approached them more nearly as K increased. Figure 4 shows that the two end-point conformations ended up very close to the open and closed states from the PDB.
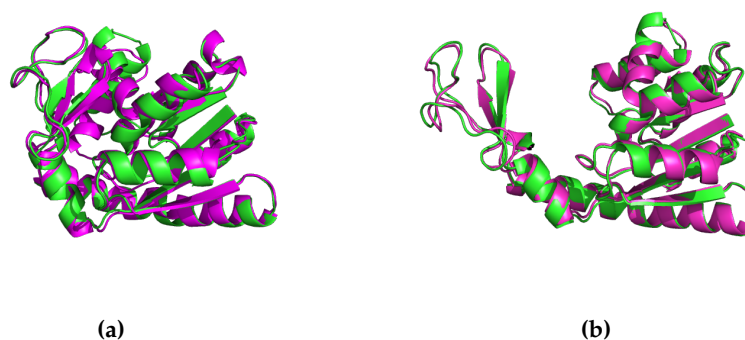

(a)   **(b)**

Figure 4. Adenylate kinase (ADK) closed (a) and open (b) configurations. PDB structures (magenta) are compared to the most similar molecular dynamics classification results (green) found among the five representative structures for the $K = 5$ classification.

5. Conclusions

The method developed here is ideally suited for the unsupervised structural classification problem. It has been derived from a first-principles Bayesian analysis of the set of atoms which interact within a structure. This work solved the central problem of defining an appropriate prior distribution over parameter space and implementing an efficient sampling method.

Tests on sample conformational transitions identified more categories than naïvely expected because it generated milestones along the motion's time-coordinate. However, all categorizations were shown to have excellent accuracy as judged by picking out the correct time-sequence. Finally, a test on Adenylate kinase verified that these conclusions easily generalize to protein motions.

The central results of this work showed that the method behaves well under large variations in the number of features and samples. These observations validate our choice for the prior distribution, since changes in $P(\theta|I)$ will have large effects on the distribution over number of categories, $P(K|xI)$. They also show robustness of the MC sampling method itself, since relatively high acceptance rates were achieved.

Many future applications and the development of this method are possible. Changes in the pair contacts between states could be analyzed more thoroughly, as in the DynDom method [7]. Probabilities of assignment to each conformational state can be used as reaction coordinates. We are presently applying the method to classify chemical compound space using binary MACCS fingerprints [21], and to characterize conformational space of SARS-CoV-2 proteins simulated using replica exchange molecular dynamics [22].

Now that the principle has been demonstrated, more informative classification schemes can be devised. Adding information like hydrogen-bonding, salt bridge formation, and secondary structure annotation will allow the Bayesian framework able to recognize categories more like a biochemist would. The method can also be focused on active domains and binding sites by adding more points and shorter distance cutoffs for key residues. The insensitivity to M shown in this work provides a high degree of confidence that any amount of additional data will improve the overall categorization without spoiling the classification already achieved using coarser-level data.

Appendix A.

Appendix A.1. Behavior of the Posterior PDF with Increasing M

We argue here that the implicit uniform prior for all μ_{kj} as well as the choice $\alpha = M + 1$ is most appropriate for this problem. This is the conclusion we arrive at by requiring that the likelihood for dividing a category into two parts should be insensitive to M.

First, multiply $P(xz|\theta)$ and $P(\theta|U)$ from the main text to provide $P(xz\theta|U)$,

$$\frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \prod_{k=1}^{K} \left(\frac{\Gamma(2\beta)}{\Gamma(\beta)^2} \right)^M \pi_k^{N_k+\alpha-1} \prod_{j=1}^{M} \mu_{kj}^{N_{kj}+\beta-1} (1 - \mu_{kj})^{N_k-N_{kj}+\beta-1}. \tag{A1}$$

This is generalized a bit from the main text so that a Beta prior for μ (with parameter β) is explicitly shown.

Observe that (without the uniqueness constraint) the posterior distribution over categories is proportional to,

$$P(xz|U) \propto f(N,K) \prod_k \Gamma(N_k + \alpha) \prod_j \frac{\Gamma(N_{kj} + \beta)\Gamma(N_k - N_{kj} + \beta)}{\Gamma(N_k + 2\beta)}, \tag{A2}$$

with $f(N,K) = \Gamma(\alpha K)/\Gamma(\alpha)^K \Gamma(N + \alpha K)$ (for $\beta = 1$). The probability per category on the far-right can be likened to an information entropy,

$$S_{\text{inf}}(kj) \equiv \ln \binom{N_k}{N_{kj}}, \tag{A3}$$

but is offset because of the limit,

$$\lim_{n \to \infty} \frac{\Gamma(n + a)}{\Gamma(n) n^a} = 1. \tag{A4}$$

Inserting this limit on the right, we find

$$\frac{\Gamma(N_{kj} + \beta) \Gamma(N_k - N_{kj} + \beta)}{\Gamma(N_k + 2\beta)} \approx e^{-S_{\text{inf}}(kj)} \frac{N_{kj}^{\beta-1}(N_k - N_{kj})^{\beta-1}}{N_k^{2\beta-1}}. \tag{A5}$$

This provides the motivation for choosing $\beta = 1$. This choice for β makes the categorization closer to an information entropy, depending only on the ratio N_{kj}/N_k.

The categorization probability is now asymptotically like an information entropy, except for the extra factor of $1/(N_k + 1)$. This factor is present for every feature, and creates an undesired dependence on M. We can cancel that factor by appealing to Equation (A4). This time using $n = N_k + 1$ and $a = \alpha - 1$ to approximate $\Gamma(N_k + \alpha)$,

$$P(xz|U) \approx f(N,K) \prod_k \frac{N_k!(N_k + 1)^{\alpha-1}}{(N_k + 1)^M} \prod_j e^{-S_{\text{inf}}(kj)}. \tag{A6}$$

Equation (A6) provides the motivation for choosing $\alpha = M + 1$. With this choice, the $(N_k + 1)^M$ term cancels. It leaves behind a relative information entropy for the N_k. Essentially,

$$P(xz|U) \approx f(N,K) N! \exp \left(S_{\text{inf}}(\{N_k\}) - \sum_{kj} S_{\text{inf}}(kj) \right). \tag{A7}$$

At this point, different choices for $f(N,K)$ lead to differences in the probability distribution over K, the number of categories. Rather than directly influencing these, we have chosen not to create additional assumptions about $f(N,K)$ beyond the above choices for α and β.

Appendix A.2. Sampling Method

Since π can be trivially integrated out of the posterior distribution (Equation (A1)), we re-generate π on every split/join move, and consider only the integrated probability $P(z\mu|xI)$ as our sampling target. Each time a split/join move is generated, a re-categorization of samples in category k ($\{x_k\}$) into $\{x_L\}$ and $\{x_R\}$. is immediately followed by generation of μ_L and μ_R (or μ_k for a join) from their respective Beta distributions (with samples N_{Lj} and $N_L - N_{Lk}$, etc.).

We set up the acceptance probability for these split / join moves to satisfy the detailed balance condition,

$$\frac{P_{\text{acc}}(z'\mu'|z\mu x)}{P_{\text{acc}}(z\mu|z'\mu'x)} = \frac{P_{\text{gen}}(z\mu|z'\mu'x)}{P_{\text{gen}}(z'\mu'|z\mu x)} \frac{P(z'\mu'|xI)}{P(z\mu|xI)}. \tag{A8}$$

Without loss of generality, assume z' contains $K + 1$ categories and z contains K. We already have explicit formulas for $P(z\theta|xI)$ (modulo $P(x|I)$), so the only thing needed to calculate P_{acc} is the move generation probabilities.

Given category k and feature j was chosen to define the split (see main text for a description of this process), the recategorization probability is,

$$P_{\text{gen}}(z'|zkj) = \frac{1}{2}\left(\begin{matrix} \eta^{N_{Lj}+N_R-N_{Rj}}(1-\eta)^{N_{Rj}+N_L-N_{Lj}} \\ +(1-\eta)^{N_{Lj}+N_R-N_{Rj}}\eta^{N_{Rj}+N_L-N_{Lj}} \end{matrix}\right)/(1-p_L-p_R). \tag{A9}$$

The factor of $\frac{1}{2}$ is necessary because the labels L and R are interchangeable. p_L (p_R) is the expression in the numerator when all z' move from k to L (R).

Because any j could have been chosen, the overall probability of ending up with categories L and R is,

$$P_{\text{gen}}(z'|z) = \frac{\xi(K)}{\sum_{nl}Q_{nl}}\sum_j Q_{kj}P_{\text{gen}}(z'|zkj). \tag{A10}$$

To concentrate splitting on productive cases, we choose Q_{kj} to be one for all k,j unless $N_{kj}=0$ or $N_{kj}=N_k$. In those cases Q_{kj} is zero so that we do not split using a non-informative feature.

Generating μ_L and μ_R adds the factor,

$$P_{\text{gen}}(z'\mu'|zx) = P_{\text{gen}}(z'|z)P(\{x_L\}|\mu_L)(N_L+1)^M P(\{x_R\}|\mu_L)(N_R+1)^M e^{\sum_j S_{\text{inf}}(Lj)+S_{\text{inf}}(Rj)}. \tag{A11}$$

Here, we have made the intuitive definition,

$$P(\{x_k\}|\mu_k) \equiv \prod_j \mu_{kj}^{N_{kj}}(1-\mu_{kj})^{N_k-N_{kj}}. \tag{A12}$$

Most of this term will cancel $P(z\mu|xI)$ in the final expression for the acceptance probability.

The recombination rule is attempted with frequency $1-\xi(K+1)$. A set u is chosen at random with probability $1/(K+1)$ and recombined with set $v\neq u$ chosen with probability $2/K(K+1)$. The resulting generation probability is,

$$P_{\text{gen}}(z|z'x) = \frac{(1-\xi(K+1))2}{K(K+1)}. \tag{A13}$$

Generating μ_k adds the factor,

$$P_{\text{gen}}(z\mu|z'x) = P_{\text{gen}}(z|z'x)P(\{x_k\}|\mu_k)(N_k+1)^M e^{\sum_j S_{\text{inf}}(kj)}. \tag{A14}$$

As a final aid to computing the acceptance probability ratio (Equation (A8)), we provide a simplified expression for the likelihood ratio,

$$\begin{aligned}\frac{P(z'\mu'|xI)}{P(z\mu|xI)} = & \frac{(1-B_{LR})\prod_{k'\neq L,R}[(1-B_{Lk'})(1-B_{Rk'})]}{\prod_{k'\neq k}(1-B_{kk'})}\frac{\Gamma(\alpha(K+1))}{\Gamma(\alpha K)\Gamma(\alpha)}\frac{\Gamma(N+\alpha K)}{\Gamma(N+\alpha(K+1))} \\ & \frac{\Gamma(N_L+\alpha)\Gamma(N_R+\alpha)}{\Gamma(N_k+\alpha)}\frac{P(\{x_L\}|\mu_L)P(\{x_R\}|\mu_R)}{P(\{x_k\}|\mu_k)}\end{aligned} \tag{A15}$$

Appendix A.3. Coordinate Trajectory Generation for Synthetic Test Systems

The 'chomp' system was 2D, with two line segments, both of length 4 units. The bottom segment ($P/2$ uniformly spaced particles) stretched along the x-axis and remained stationary. The top segment ($P/2-1$ uniformly spaced particles, since the origin is excluded), moved at angle $\theta(t) = \frac{\pi}{4}\cos(t)$, $t\in[0,\pi]$. This motion mimics a harmonic oscillator—which spends a majority of its time at the two limits of its motion.

The 'heli' system was 3D. It also had two line segments with $P/2$ equally spaced particles each. The bottom segment remained as before, while the top segment was replaced by $P/2$ particles displaced

from the bottom one by 1 distance unit along z. The top segment was rotated in the xy plane with $\theta(t) = t, t \in [0, \pi]$.

The 'glob' system was 3D, consisting of 3 spheres, each represented by $P/3$ points at radius 2. The point locations were taken from Lebedev quadrature rules, which possess octahedral symmetry. Sphere 1 remained fixed at the origin. Spheres 2 and 3 started in locations displaced along z or x, respectively, by 2.1 distance units. As t traversed $[0, \pi]$, sphere 2 was rotated about the origin around the x-axis by angle $\theta_2(t) = \frac{\pi}{4}(1 - \cos(t))$, while sphere 3 was rotated about the origin around the z-axis by angle $\theta_3(t) = \frac{\pi}{4}(1 - \cos(2t))$. This motion creates something like three conformational states. Since sphere 3 moves twice as fast as sphere 2, sphere 2 can be at $+z$ or $-y$ while sphere 3 is at $+x$. A third conformation is at the center where sphere 3 is at $+y$.

References

1. G, J.; Zhou, H.-X. Protein allostery and conformational dynamics. *Chem. Rev.* **2016**, *116*, 6503–6515.
2. Müller, C.W.; Schlauderer, G.J.; Reinstein, J.; Schulz, G.E. Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding. *Structure* **1996**, *4*, 147–156.
3. Amaral, C.; Carnevale, V.; Klein, M.L.; Treptow, W. Exploring conformational states of the bacterial voltage-gated sodium channel NavAb via molecular dynamics simulations. *Proc. Nat. Acad. Sci. USA* **2012**, *109*, 21336–21341.
4. Kim, Y.E.; Hipp, M.S.; Bracher, A.; Hayer-Hartl, M.; Hartl, F.U. Molecular chaperone functions in protein folding and proteostasis. *Annu. Rev. Biochem.* **2013**, *82*, 323–355.
5. Vant, J.W.; Sarkar, D.; Fiorin, G.; Skeel, R.; Vermaas, J.V.; Singharoy, A. Data-guided multi-map variables for ensemble refinement of molecular movies. 2020. submitted.
6. Spiwok, V.; Kříž, P. Time-lagged t-distributed stochastic neighbor embedding (t-SNE) of molecular simulation trajectories. *Frontiers Molecular Biosci.* **2020**, *7*, 132.
7. Lee, R.A.; Razaz, M.; Hayward, S. The DynDom database of protein domain motions. *Bioinformatics* **2003**, *19*, 1290–1291.
8. Girdlestone, C.; Hayward, S. The DynDom3D webserver for the analysis of domain movements in multimeric proteins. *J. Comput. Biol.* **2016**, *highlightvolume*, 21–26.
9. Stamati, H.; Clementi, C.; Kavraki, L.E. Application of nonlinear dimensionality reduction to characterize the conformational landscape of small peptides. *Protein. Struct., Funct. Bioinf.* **2010**, *78*, 223–235.
10. Ramanathan, A.; Savol, A.J.; Langmead, C.J.; Agarwal, P.K.; Chennubhotla, C.S. Discovering conformational sub-states relevant to protein function. *PLoS ONE* **2011**, *6*, e15827.
11. Ferguson, A.L.; Panagiotopoulos, A.Z.; Kevrekidis, I.G.; Debenedetti, P.G. Nonlinear dimensionality reduction in molecular simulation: The diffusion map approach. *Chem. Phys. Lett.* **2011**, *509*, 1–11.
12. Duan, M.; Fan, J.; Li, M.; Han, L.; Huo, S. Evaluation of dimensionality-reduction methods from peptide folding–unfolding simulations. *J. Chem. Theory. Comput.* **2013**, *9*, 2490–2497.
13. Li, C.; Wang, B.; Pavlu, V.; Aslam, J. Conditional Bernoulli mixtures for multi-label classification. In *Proceedings of the 33rd International Conference on Machine Learning*, 19–24 June 2016, New York, NY, USA; volume 48, 2016.
14. Alabau, V.; Andrés, J.; Casacuberta, F.; Civera, J.; Adrià Giménez, J.-H.; Juan, A.; Sanchis, A.; Vidal, E. The naive Bayes model, generalisations and applications. Available online: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.513.977&rep=rep1&type=pdf (accessed on 31 Ocotber 2020).
15. Novovičová, J.; Antonín Malík, A. Text document classification based on mixture models. *Kybernetika* **2004**, *40*, 293–304.
16. Kaji, D.; Watanabe, K.; Watanabe, S. Phase transition of variational bayes learning in Bernoulli mixture. *Australian J. Intell. Inf. Proc. Syst.* **2010**, *35*, 35–40.
17. Yamazaki, K.; Kaji, D. Comparing two Bayes methods based on the free energy functions in Bernoulli mixtures. *Neur. Network.* **2013**, *44*, 36–43.
18. Rogers.D.M. Available online: frobnitzem/classifier: Classifier Version 1.0 (Version v1.0). Zenodo. (accessed on 23 October 2020).
19. Jaynes, E.T. *Probability Theory, The Logic of Science*; Cambridge University Press: Cambridge, UK, 2003.
20. Müller, C.W.; Schulz, G.E. Structure of the complex between adenylate kinase from escherichia coli and the inhibitor ap5a refined at 1.9 å resolution. *J. Mol. Biol.* **1992**, *224*, 159–177.

21. Durant, J.L.; Leland, B.A.; Henry, D.R.; Nourse, J.G. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273–1280.

22. Acharya, A.; Agarwal,R.; Baker, M.; Baudry, J.; Bhowmik, D.; Boehm, S.; Byler, K.; Coates, L.; Chen, S.Y.; Cooper, C.J.; et al. Supercomputer-based ensemble docking drug discovery pipeline with application to Covid-19. Available online: https://chemrxiv.org/articles/preprint/Supercomputer-Based_Ensemble_Docking_Drug_Discovery_Pipeline_with_Application_to_Covid-19/12725465 (accessed on 31 October 2020).

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.