BAYESIAN PROTEIN STRUCTURE ALIGNMENT[1]

BY ABEL RODRIGUEZ AND SCOTT C. SCHMIDLER

University of California, Santa Cruz and Duke University

The analysis of the three-dimensional structure of proteins is an important topic in molecular biochemistry. Structure plays a critical role in defining the function of proteins and is more strongly conserved than amino acid sequence over evolutionary timescales. A key challenge is the identification and evaluation of structural similarity between proteins; such analysis can aid in understanding the role of newly discovered proteins and help elucidate evolutionary relationships between organisms. Computational biologists have developed many clever algorithmic techniques for comparing protein structures, however, all are based on heuristic optimization criteria, making statistical interpretation somewhat difficult. Here we present a fully probabilistic framework for pairwise structural alignment of proteins. Our approach has several advantages, including the ability to capture alignment uncertainty and to estimate key "gap" parameters which critically affect the quality of the alignment. We show that several existing alignment methods arise as *maximum a posteriori* estimates under specific choices of prior distributions and error models. Our probabilistic framework is also easily extended to incorporate additional information, which we demonstrate by including primary sequence information to generate simultaneous sequence–structure alignments that can resolve ambiguities obtained using structure alone. This combined model also provides a natural approach for the difficult task of estimating evolutionary distance based on structural alignments. The model is illustrated by comparison with well-established methods on several challenging protein alignment examples.

1. Introduction. Protein alignment is among the most powerful and widely used tools available for inferring homology and function of gene products, as well as determining evolutionary relationships between organisms. In particular, protein sequence alignment uses information about the identity of amino acids to establish regions of similarity, and has a long history of providing valuable insights. For example, the alignment of a putative human colon cancer gene with a yeast mismatch repair gene played a crucial rule in its identification and characterization [Bronner et al. (1994), Papadopoulos et al. (1994), Zhu, Liu and Lawrence (1998)].

Sequence alignment is most useful for shorter evolutionary distances, when amino acid composition has not drifted dramatically from a common ancestor.

However, when comparing proteins that are distantly related, sequence conservation may be too dilute to establish meaningful relationships. Because a protein's function is largely determined by its three-dimensional structure, and significant sequence mutation can occur while maintaining this structure, it is widely recognized that structural similarity is conserved over much longer evolutionary timescales than sequence similarity. In addition, sequence alignment cannot detect convergent evolution, when proteins with similar 3D structure and carrying out similar functions have evolved from unrelated genes.

Aligning 3D structures requires choosing which amino acids to match as in sequence alignment, but has the added complexity of handling coordinate frames arising from arbitrary rotation and translation. Early work in structural alignment [Rao and Rossmann (1973), Rossmann and Argos (1975, 1976)] developed techniques that iterate between a rigid body registration and an alignment step, and Satow et al. (1986) introduced the use of dynamic programming [applied to sequence alignment by Needleman and Wunsch (1970)] as an efficient way to construct the alignment given a registration. Similar methods have been adopted by many authors [Cohen (1997), Gerstein and Levitt (1998), Wu et al. (1998)]. Most work uses a penalized root mean squared deviation (RMSD) between corresponding backbone α-carbon (C_α) atoms to measure quality of the alignments, but several other measures have been proposed, including soap-bubble surface metrics [Falicov and Cohen (1996)], differential geometry [Kotlovyi, Nichols and Eyck (2003)], and heuristic rules like the SSAP method of Taylor and Orengo (1989).

An alternative to iterative methods is the use of distance geometry to avoid the registration problem, thus representing each protein by a pairwise distance matrix between all C_α atoms. The popular DALI [Holm and Sander (1993)] method is an example of this approach. Other techniques are specially tailored for the large-scale computational demands of rapid searching of large protein databases, sometimes employing highly redundant representations of the data; these include geometric hashing [Altschul et al. (1990), Fischer et al. (1994), Wallace, Laskowsi and Thornton (1996)], graph algorithms [Taylor (2002)] and clustering methods like VAST [Gibrat, Madej and Bryant (1996)]. Finally, some authors combine these ideas with additional heuristics to produce faster or more accurate algorithms, including CE [Shindyalov and Bourne (1998)] and PROSUP [Lackner et al. (2000)]. Detailed reviews on pairwise structural alignment methods can be found in Brown, Orengo and Taylor (1996), Eidhammer, Jonassen and Taylor (2000) and Lemmen and Lengauer (2000).

The profusion of methods shows the difficulties involved in performing structural alignments: in defining how to measure alignment quality and in computing "best" alignments efficiently. It has been well documented in the literature that different algorithms can produce alignments sharing very few amino acid pairings, and are sensitive to both the initial alignment and the specific choice of algorithm parameters [Gerstein and Levitt (1998), Godzik (1996), Zu-Kang and Sippl (1996)]. Additional complications arise when trying to determine the significance of the resulting alignments. Although substantial effort has been devoted to this

point and important progress made [Gerstein and Levitt (1998), Levitt and Gerstein (1998), Lipman and Pearson (1985), Mizuguchi and Go (1995)], the solutions remain based on heuristics and upper bounds that are difficult to interpret. Finally, all the methods described above approach the structural alignment as an optimization problem, finding a single best alignment. However, structural comparisons are subject to substantial uncertainties arising from evolutionary divergence, population variability, experimental measurement error and protein conformational variability, not to mention sensitivity to parameters of comparison metrics and optimization algorithms. To address these sources of variability, approaches based on explicit statistical modeling are desirable, and the results of structural comparisons require careful analysis to understand the impact of uncertainty.

In this paper, we develop a Bayesian statistical approach to pairwise protein structure alignment, combining techniques from statistical shape analysis [Dryden and Mardia (1998), Kendall et al. (1999), Small (1996)] and Bayesian sequence alignment [Liu and Lawrence (1999), Webb, Liu and Lawrence (2002), Zhu, Liu and Lawrence (1998)]. This represents one aspect of a general Bayesian framework developed here and elsewhere [Schmidler (2003, 2004)], and subsequently extended by Schmidler (2007a, 2007b), Wang and Schmidler (2008). Green and Mardia (2006) and Dryden, Hirst and Melville (2007) independently developed related approaches for hierarchical Bayesian alignment of protein active sites rather than whole proteins, and for small molecules, respectively. However, our approach differs in a number of important points: we introduce hierarchical priors on the space of alignments that are equivalent to the standard affine gap penalty of classical alignment approaches, but allow us to estimate the parameters controlling the complexity of the alignment. We also introduce an efficient computational approach that allows rapid computation and sampling, which both enables identification of alternative alignments and provides direct measures of alignment uncertainty. A significant advantage of our formulation is the unification of many existing alternative methods for structural alignment, which can be seen as special cases of MAP alignment under different specific choices of error models or alignment priors. This provides valuable insight into the relationships and properties of existing algorithms.

Another powerful advantage of a fully probabilistic framework is the ability to incorporate disparate sources of information in a natural and coherent fashion. Using our Bayesian structural alignment model as a platform, we also develop a fully probabilistic approach for *simultaneous sequence-and-structure* alignment, which combines information from both primary sequence *and* 3D structure. In the presence of unambiguity in geometric matching for highly-divergent proteins or low-resolution structural data, amino acid identities or preferred substitutions can significantly alleviate the remaining uncertainty. We demonstrate this approach on some difficult structural alignment problems from the literature. Finally, we show that our simultaneous alignment approach provides a natural method for estimating evolutionary distances directly from structure comparison, a notoriously difficult task.
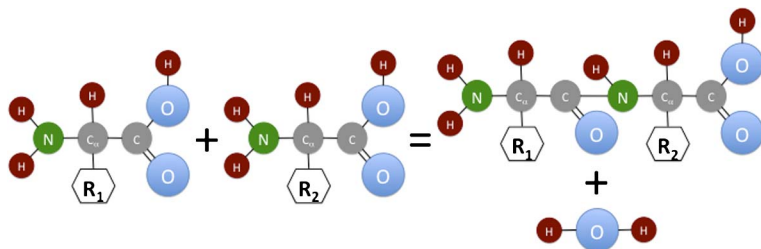
FIG. 1. *The chemical structure of proteins, showing the combination of two amino acids to form a peptide bond. Repeated applications of this process form a linear chain to make a protein. The identities of the R-groups determine the protein sequence and thus its properties, including 3D structure and biochemical function.*

2. Proteins and their structure. Proteins are the most diverse macromolecules in organisms, playing a wide range of roles: as enzymes, molecular receptors, antibodies, hormones, structural proteins, and molecular transporters. Proteins are linear polymers, molecular chains created by stringing together amino acids using peptide bonds to form a polypeptide. The constituent amino acids are themselves small molecules characterized by a central carbon atom (C_α) to which additional chemical groups are attached, including a carboxyl group (COOH), an amino group (NH_2) and an organic side chain (see Figure 1). There are 20 distinct naturally occurring types of side chains, ranging from very simple (G) to relatively complex (F), giving the 20 naturally-occurring amino acids their identities. During the process of peptide-bond formation, a water molecule is shed and, as a result, amino acids occurring within a protein chain are often referred to as "residues." Because residues are not symmetric, the chain is directional, with the beginning end having a free amino group known as the amino- (or N-) terminus and the end having a free carboxyl group known as the carboxy- (or C-) terminus. The sequences of amino acids making up proteins are encoded in DNA by the universal genetic code; Figure 2 shows a simple classification of these amino acids including some of their chemical properties. It is the combinatorics of combining these properties in different numbers and orderings that gives rise to the diversity of protein structures and functions.

The linear sequence of amino acid identities makes up the *primary* structure of a protein and, like DNA, can be encoded using strings of letters. Primary protein sequences can be aligned to identify evolutionarily related or otherwise similar regions, using algorithms for string comparison. This requires an amino acid distance metric, often summarized in the form of substitution matrices such as PAM [Dayhoff and Eck (1968)] or BLOSUM [Henikoff and Henikoff (1992)]. Sequence alignments can provide important insights into the function of proteins and the evolution of organisms.

The diverse chemical properties of amino acids lead proteins to "fold" reproducibly into complicated, sequence-specific bundles. This three-dimensional struc-
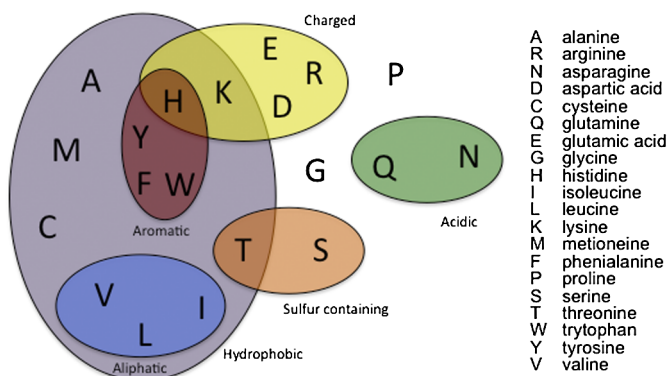
A	alanine
R	arginine
N	asparagine
D	aspartic acid
C	cysteine
Q	glutamine
E	glutamic acid
G	glycine
H	histidine
I	isoleucine
L	leucine
K	lysine
M	metioneine
F	phenialanine
P	proline
S	serine
T	threonine
W	trytophan
Y	tyrosine
V	valine

FIG. 2. *The twenty amino acids encoded by the universal genetic code, and their chemical properties.*

ture enables a protein to perform its functions (such as specific binding of a target). Within this fold are often smaller, recognizable structural "motifs" occurring across many proteins, known as secondary structures. These are regularly repeating local structural patterns, with the most famous being α-helices (successive backbone atoms following a right-handed helical path though space) and β-sheets (extended stretches of backbone connected laterally to form sheets). Because these secondary structure elements are local, many regions of different secondary structure can be present in the same protein molecule. In contrast, the *tertiary* structure of a protein refers the overall 3D shape, including the relative locations of secondary structures in space. In this paper we are concerned with the problem of alignment between this 3D structure of two proteins, as 3D structures tend to be much more highly conserved across evolution than the sequence itself. This 3D shape is well summarized by the positions of the C_α carbons, giving a path through space known as the backbone of the protein.

Protein structural data arises most frequently from the experimental methods of X-ray crystallography and NMR spectroscopy. Although these experimental techniques differ greatly, the end result of each is a set of 3D coordinates for the protein atoms in an arbitrary coordinate system. These coordinates, which are publicly available at the Protein Data Bank repository (http://www.pdb.org/) along with the primary sequence, are the data we use in developing our models.

3. Bayesian protein structure alignment. Let $X_{n \times 3}$ and $Y_{m \times 3}$ be coordinate matrices for two proteins, with rows x_i (y_i) containing coordinates of the C_α of the ith amino acid. An alignment between X and Y is a $n \times m$ match matrix $M = (m_{ij})$ such that $m_{ij} = 1$ if residues X_i and Y_j are matched, and 0 otherwise. Each position in X can be matched to at most one position in Y, so each row and column of M contains at most one nonzero entry, thus, M is the adjacency matrix for a matching (a subset of edges, no two sharing an endpoint) on a complete

bipartite graph between sets X and Y. In the sequel we denote by X_M and Y_M the $|M| = \sum_{ij} m_{ij}$ nonzero rows of $M'X$ and MY giving the coordinates of the matched positions, and by $X_{\bar{M}}$ and $Y_{\bar{M}}$ the rows of X and Y not included in X_M and Y_M giving coordinates of the unmatched position.

We adopt a Bayesian approach to structure alignment which defines a prior distribution on alignments $P(M)$ and, given a probability model for the coordinates matrices X and Y conditional on M, obtains the posterior distribution:

$$P(M|X, Y) = \frac{P(X, Y|M)P(M)}{\sum_M P(X, Y|M)P(M)},$$

where the marginal likelihood $P(X, Y) = \sum_M P(X, Y|M)P(M)$ involves a sum over all possible alignments. Although the number of matchings is exponential in n and m, inferences may be obtained by Monte Carlo sampling from the posterior $P(M|X, Y)$ to approximate posterior summaries such as the posterior mode,

$$(1) \qquad \hat{M} = \arg\max_M P(M|X, Y),$$

or the *marginal alignment matrix*, (p_{ij}), giving the marginal posterior probability $p_{ij} = \sum_M m_{ij} P(M|X, Y)$ of matching position X_i with Y_i summing over all possible alignments. MAP estimates have well-known drawbacks: they ignore the variability in the posterior, including that arising from uncertainty in hyperparameters or potential multimodality. However, they are simple to obtain and provide a convenient "representative" alignment in which each residue matches at most one other. The marginal alignment matrix is easily obtained by sampling alignments from the posterior distribution, but is somewhat harder to visualize. In Section 6 we use heatmaps for this purpose, but it is also possible to generate a point estimator by maximizing an appropriate utility function, as an alternative to the MAP alignment.

3.1. *Likelihood.* Given a matching matrix M, we factor the joint likelihood of the observed structures X and Y into (dependent) matched and (independent) unmatched positions:

$$(2) \qquad P(X, Y|M) = P(Y_M|X_M)P(Y_{\bar{M}})P(X).$$

This arises naturally, for example, by viewing the aligned positions as homologous (having a common evolutionary ancestor) and the unmatched positions as random insertions and deletions occurring independently in each protein after divergence. Note that while assuming $Y_M \perp Y_{\bar{M}}|M$ ignores the physical constraints of neighboring bonds, it simplifies the calculations in important ways described below.

We adopt a probabilistic model for matched regions of the proteins which assumes that deviations are independent and normally distributed,

$$(3) \qquad Y_M = X_M + \varepsilon, \qquad \varepsilon \sim \mathsf{N}(\mathbf{0}, \sigma^2 I),$$

and, to complete the model, we assume that the rows in $Y_{\bar{M}}$ follow common distribution f. (This normality assumption is discussed with possible relaxations in Section 3.4.)

However, both X and Y are observed only up to arbitrary coordinate frames. That is, we observe $[X]$ and $[Y]$, where $[X]$ denotes the *size-and-shape* of X, formally defined [Dryden and Mardia (1998), Kendall et al. (1999)] as an equivalence class of invariant matrices under the group of Euclidean transformations:

$$[X] = \{XR + \mu : R \in \mathsf{SO}(3), \mu \in \mathsf{R}^3\}.$$

Here $\mathsf{SO}(3)$ is the special orthogonal group of 3×3 rotation matrices. [Alignment using a somewhat more general class of nonrigid transformations is considered by Schmidler (2007b).] We therefore define the likelihood by

$$
\begin{aligned}
(4) \quad P(X, Y|M) = {}& \left(2\pi\sigma^2\right)^{-3|M|/2} \\
& \times \exp -\frac{1}{2\sigma^2} \left\| Y_M - \left(X_M \hat{R}_M + \mathbf{1}\hat{\mu}'_M\right) \right\|_F^2 P(X) \prod_{y_i \in Y_{\bar{M}}} f(y_i|\lambda),
\end{aligned}
$$

where $\|X\|_F = \mathrm{tr}(X'X)^{1/2}$ is the Frobenius norm, $f(\cdot; \lambda)$ is a one-parameter density for inserted/deleted positions, and $P(X)$ is a probability distribution describing the shape of the reference protein X. Here $(\hat{R}_M, \hat{\mu}_M)$ are the optimal least-squares rotation and translation placing X and Y on a common coordinate system, given by $\hat{R}_M = U_M V_M^T$ and $\hat{\mu}_M = \bar{Y}_M - \bar{X}_M \hat{R}_M$, where $V_M, U_M \in \mathsf{SO}(3)$ are obtained from the singular value decomposition $Y_M^T C_M X_M = U_M D_M V_M^T$ for centering matrix $C_M = I - \frac{1}{|M|} 11^T$.

The appearance of $(\hat{\mu}_M, \hat{R}_M)$ in likelihood (4) may be interpreted in two different ways. First, (4) may be viewed as a *profile* likelihood for M, maximizing over nuisance parameters (μ, R) corresponding to the unknown translation and rotation conditional on M, under the model

$$Y_M = (X_M + \varepsilon)R + \mu, \qquad \varepsilon \sim \mathsf{N}(\mathbf{0}, \sigma^2 I).$$

A fully Bayesian approach would instead assign prior distributions to these nuisance parameters and integrate them out. Green and Mardia (2006), Dryden, Hirst and Melville (2007) and Wang and Schmidler (2008) adopt this integration approach, and Schmidler (2007a, 2007b) considers both (maximization and integration) approaches for handling the unknown registration parameters. However, in our experience the uncertainty on (μ, R) given M is minimal for most structural alignments, with the posterior heavily concentrated about the mode, making the two approaches perform nearly identically. Kenobi and Dryden (2012) report similar findings and discuss this issue in detail.

We may also interpret (4) as a sampling density defined directly on (a local tangent space approximation to) the underlying shape space of the configurations,

replacing $3|M|$ in the normalizing constant with $3|M| - 6$, the dimension of the shape manifold. The exponent $\|Y_M - (X_M \hat{R}_M + \mathbf{1}\hat{\mu}'_M)\|_F^2 = d_P^2(X, Y)$ is known as the (squared) *partial Procrustes distance*, and serves as the Riemannian metric on this (size-and) shape space [Dryden and Mardia (1998), Kendall et al. (1999)]. This metric effectively *defines* a one-to-one correspondence between matchings M and Euclidean transformations (μ, R), enabling inference to be performed directly in the space of matchings. Under this interpretation, our approach is fully Bayesian, but the likelihood is approximated by evaluating the density in the tangent space.

In what follows, we take $f(\cdot|\lambda) = \lambda = 1/|\Omega|$ uniform over a bounded region Ω; then λ can be interpreted as a lower bound for the gap penalty as discussed in Section 3.4. Note that the factorization (2) implies that the marginal distribution $P(X)$ cancels in the posterior distribution, and may be left unspecified so long as it is assumed to be functionally independent of parameters M and σ^2. This is similar to a proportional hazards model where the baseline risk is left unspecified to obtain a semi-parametric survival model. In addition, the isotropic error model ensures the model is symmetric in X and Y if we take

$$P(X) = \prod_{x_i \in X} f(x_i|\lambda).$$

3.2. *Prior on the alignment matrix.* Prior distributions on matchings $P(M)$ may be specified in a variety of ways; here we adopt a *gap-penalty* formulation familiar in the sequence and structure alignment literature, where unmatched stretches of amino acids are penalized by the affine function:

$$u(M; g, h) = gs(M) + h \sum_{i=1}^{s(M)} l_i(M)$$

with gap-opening penalty g and gap-extension penalty h, where $s(M)$ is the number of gaps in alignment M and $l_i(M)$ is the length of the ith gap. Exponentiating and normalizing this function provides a prior on M [Liu and Lawrence (1999)], essentially a Markov chain parametrized as a "Boltzmann chain" Gibbs random field [Saul and Jordan (1995), Schmidler, Lucas and Oas (2007)]:

(5) $$P(M|g, h) = Z(g, h)e^{-u(M;g,h)}$$

with normalizing constant Z. This prior encourages grouping of matches together along the protein backbone. It allows for explicit control over the number of gaps, compared to, for example, the prior of Green and Mardia (2006) which controls only the expected total length.

Under the affine-gap-penalty prior, sampling may be done efficiently using stochastic recursions analogous to those of standard sequence alignment algorithms [Liu and Lawrence (1999)], along with additional Monte Carlo steps, as described below. Note that this prior requires the alignment to preserve the sequential order along the polypeptide backbone, requiring topological equivalence

of the two proteins. More general priors applicable for comparing proteins of potentially different topologies (convergent evolution) are easily accommodated with the introduction of additional Monte Carlo steps, but will be described elsewhere. Although the prior allows for simultaneous gaps on both proteins, for identifiability purposes we do not allow gaps in X to be followed by gaps in Y [see Webb, Liu and Lawrence (2002) for details].

3.3. *Hyperpriors.* In standard sequence and structure alignment algorithms, the gap parameters g and h are assigned fixed values. However, they have a critical effect on the resulting alignment, with large opening gap penalties g tending to produce alignments with few gaps and vice versa. In the context of sequence alignment, Liu and Lawrence (1999) treat (g, h) as nuisance parameters and assign hyperpriors, integrating them out to obtain a marginal posterior distribution over alignments. We similarly assign g and h Gamma hyperpriors,

$$(6) \qquad g \sim \mathsf{Ga}(a_g, b_g), \qquad h \sim \mathsf{Ga}(a_h, b_h),$$

with hyperparameters (a_g, b_g, a_h, b_h) chosen to be diffuse (but proper). An alternative is to utilize manually-obtained reference alignments [e.g., BAliBASE, see Thompson, Plewniak and Poch (1999) and Thompson et al. (2005)] to obtain informative priors for g and h. The model is completed by specifying inverse-gamma prior $\sigma^2 \sim \mathsf{IGa}(a_\sigma, b_\sigma)$ on variance parameter σ^2.

3.4. *Many existing structure alignment algorithms are special cases.* Rather than summarize the posterior $P(M|X, Y)$ by Monte Carlo sampling, we may instead obtain a single MAP alignment (1) by maximizing the (log-) posterior. Conditioning on parameters $\theta = (\sigma^2, g, h, \lambda)$, we obtain

$$\log\big(P(M|X, Y, \theta)\big)$$

$$(7) \qquad = -\frac{3}{2}|M|\log(2\pi\sigma) - \frac{1}{2\sigma^2}d_P^2(X_M, Y_M) + (n + m - |M|)\log(\lambda)$$

$$\qquad + \log\big(Z(g, h)\big) - u(M; g, h)$$

and noting that $\sum_{i=1}^{s(M)} l_i(M) = (n + m) - 2|M|$, this is equivalent to minimizing

$$d_P^2(X_M, Y_M) + u\big(M; g^*, h^*\big) + C\big(\sigma^2, \lambda, g, h\big),$$

where $g^* = \sigma^2(g + \frac{3}{2}\log(2\pi\sigma) + \log\lambda)$ and $h^* = \sigma^2 h$, and $C(\sigma^2, \lambda, g, h)$ is independent of M. Therefore, the MAP estimate for M with $(g, h, \lambda, \sigma^2)$ fixed corresponds to a global alignment obtained via dynamic programming [Needleman and Wunsch (1970)], using RMSD under optimal least-squares rotation/translation as the dissimilarity metric, and with gap opening and extension penalties given by g^* and h^*. Since $g \geq 0$, the term $(\frac{3}{2}\log(2\pi\sigma) + \log\lambda)$ serves as a lower bound on g^*, the "effective" gap extension penalty.

Note that the relative posterior probability of two alignments which differ by an unmatched pair (x_i, y_j) is greater than one if and only if

$$\left|y_j - (x_i R + \mu')\right| < g^*(1 - \xi_{ij}) + h^* \xi_{ij},$$

where ξ_{ij} is an indicator taking value 1 if removing the pair (x_i, y_j) creates a new gap in the alignment and 0 otherwise. Thus, the model favors inclusion of pairs below a dynamically estimated threshold given by g^* and h^*. Since these threshold parameters are estimated from the data (in contrast to standard optimization-based alignment algorithms where they are fixed a priori), our approach automatically controls for the error rates associated with multiple comparisons.

It is also worth noting that the assumption of normally distributed errors in (3) may be replaced with an alternative error model, altering the d_P term in (4) and the corresponding posterior distribution. In particular, robust error models with heavy tails may be considered (e.g., Student-t or double exponential distributions) to account for possible outliers. In this way, our probabilistic formulation provides statistical insight into various existing optimization-based algorithms.

For example, Gerstein and Levitt (1998) define the similarity between residues x_i and y_j by

$$S_{ij} = \frac{c}{1 + (d_{ij}/d_0)^2},$$

where d_{ij} denotes the distance between i and j under the current optimal registration and c and d_0 are arbitrarily chosen constants. Then dynamic programming is employed to obtain the alignment M maximizing the similarity between proteins, defined by $\sum_{(i,j) \in M} S_{ij}$. This is equivalent to obtaining the MAP estimate under our Bayesian model when the distribution of the error ε is given by $f(x) \propto \exp\{-M(1 + (x/d_0^2))^{-1}\}$, which is an exponentiated Cauchy density. [This is indeed a proper density as $\int_{-\infty}^{\infty} \exp\{-\frac{M}{1+(x/d_0)^2}\} dx < \infty$.] Thus, our unified probabilistic framework allows us to interpret such heuristics in terms of their underlying assumptions about the data generating process.

4. Computational algorithms. Combining (4), (5) and (6), we obtain the posterior distribution,

$$P(M, g, h, \sigma^2 | X, Y) \propto P(X, Y | M, \sigma^2) P(M | g, h) P(\sigma^2) P(g) P(h).$$

This posterior can be explored using a Markov chain Monte Carlo algorithm that iterates between sampling from the conditional distributions, $P(M | g, h, \sigma^2, X, Y)$, $P(g, h | M, X, Y)$ and $P(\sigma^2 | M, X, Y)$. The full-conditional posterior for the variance σ^2 is obtained by standard conjugate updating:

$$\sigma^2 | M, X, Y \sim \mathsf{IGa}\left(a_\sigma + \tfrac{3}{2} |M|, b_\sigma + \tfrac{1}{2} d_P^2(X_M, Y_M)\right).$$

The gap penalty parameters (g, h) are updated jointly by a two-dimensional geometric random walk proposal with Metropolis–Hastings acceptance probability

$$(8) \qquad 1 \wedge \frac{Z(g', h')e^{-u(M; g', h')}}{Z(g, h)e^{-u(M; g, h)}} \frac{g'h'}{gh} \left(\frac{g'}{g}\right)^{a_g - 1} \left(\frac{h'}{h}\right)^{a_h - 1} e^{-(b_g(g' - g) + b_h(h' - h))}.$$

In the previous expression (g', h') correspond to the proposed values and (g, h) correspond to the current values of the gap parameters. The required normalizing constants $Z(g, h)$ can be calculated efficiently via the recursions provided in Appendix.

As shown by Schmidler (2003), if we *condition* on registration parameters (R, μ), the alignment matrix M may be sampled from its full conditional distribution using a forward–backward algorithm similar to that of sequence alignment [Liu and Lawrence (1999), Zhu, Liu and Lawrence (1998)] and described in Appendix. Wang and Schmidler (2008) use this approach for structural alignment. However, here we have instead defined the likelihood (4) directly on shape space using maximal values $(\hat{R}_M, \hat{\mu}_M)$ associated with each distinct M, so this is no longer the case. But we may still use this efficient block Gibbs step to generate efficient Metropolis–Hastings proposals $P(M \to M')$ with distribution $q(M'; R_M, \mu_M)$, where (R_M, μ_M) is the registration associated with the *current* state M, and

$$q(M'; R_M, \mu_M)$$
$$\propto P(M'|g, h)(2\pi\sigma^2)^{-3|M'|/2}e^{-1/(2\sigma^2)\|Y_{M'} - \hat{X}_{M', M}\|_F^2} \prod_{y_i \in Y_{\bar{M}'}} f(y_i|\lambda),$$

where $\hat{X}_{M', M} = X_{M'}\hat{R}_M + \mathbf{1}\hat{\mu}'_M$. This q can be sampled efficiently using the recursions of Appendix. The proposed M' is then accepted according to the Metropolis–Hastings criteria

$$1 \wedge \frac{P(X, Y|M', \sigma^2)P(M'|g, h)q(M; R_{M'}, \mu_{M'})}{P(X, Y|M, \sigma^2)P(M|g, h)q(M'; R_M, \mu_M)},$$

with the required normalizing constants of q obtained from the sampling recursions.

These dynamic programming proposals are highly efficient for local sampling and sufficient for closely matched proteins. However, when multiple alternative alignments with distinct rotation/translations exist, mixing between them will be slow. We therefore add an additional Metropolized independence step where global moves are proposed without conditioning on the values of $(\hat{\mu}_M, \hat{R}_M)$ associated with the current alignment. To construct the independence proposal distribution, we first generate a library of viable registrations using the following procedure:

1. Compute the least-squares registration for each pair of consecutive 6-residue subsequences on protein X to each such subsequence on Y.

2. If the subsequence RMSD is less than threshold δ, include the corresponding registration in the library.

This library is computed once upon initialization of the algorithm and stored for use throughout the simulation, generating an efficient data-set-specific proposal distribution that deals effectively with potential multimodality in the posterior. A proposal is made from this distribution by drawing a registration (R', μ') uniformly at random from the library and proposing a new alignment M' from $q(M'; R', \mu')$ using the forward–backward algorithm. It is then accepted according to the Metropolis–Hastings criteria

$$1 \wedge \frac{P(X, Y|M', \sigma^2)P(M'|g, h)q(M; R', \mu')}{P(X, Y|M, \sigma^2)P(M|g, h)q(M'; R', \mu')},$$

leaving the posterior distribution invariant.

5. Bayesian synthesis of sequence and structure information. Another advantage of the Bayesian probabilistic framework given above is the ability to seamlessly incorporate additional information when available. For example, our approach leads to a natural algorithm for performing alignments based on primary sequence and tertiary structure simultaneously. This approach allows alignments which synthesize two types of information: geometric conservation of the protein architecture, and physico-chemical properties and evolutionary information provided by sidechain identities. As an important consequence, our approach enables the estimation of evolutionary distances from *structure comparison*, which has previously been quite difficult [Chothia and Lesk (1986), Grishin (1997), Johnson, Sutcliffe and Blundell (1990), Koehl and Levitt (2002), Levitt and Gerstein (1998), Wood and Pearson (1999)]. Being able to estimate evolutionary distances from structural information has important implications because structure is much more strongly conserved than sequence, enabling comparisons across much longer evolutionary timescales.

The model given by (4) for structural observations is easily extended to account simultaneously for both sequence *and* structure information by assuming the structure and sequence to be conditionally independent given the alignment M, that is, $P(A^x, A^y, X, Y|M, \theta) = P(A^x, A^y|M, \theta)P(X, Y|M, \theta)$. We take the conditional likelihood of the sequences given the alignment to be

$$(9) \qquad P(A^x, A^y|M, \Theta) = \prod_{(i,j) \in M} \Theta(A_i^x, A_j^y) \prod_{i \notin M} \Theta(A_i^x, \cdot) \prod_{j \notin M} \Theta(\cdot, A_j^y),$$

where A_i^y is the ith amino acid in protein x, $\Theta(a, b)$ gives the probability of residues a and b being matched on related sequences, and $\Theta(a, \cdot) = \Theta(\cdot, a)$ gives the marginal probability for residue a. Equation (9) is the standard likelihood form of sequence alignment [Bishop and Thompson (1986)], and these joint and

marginal distributions form the bases of standard sequence alignment substitution matrices such as PAM and BLOSUM [Dayhoff and Eck (1968), Henikoff and Henikoff (1992)], where the distributions are estimated from alignments of closely related proteins. For example, the PAM-k substitution can be written as $\Psi_k = (\Psi_k(a, b))$, where

$$\Psi_k(a, b) = 10 \log_{10}\left(\frac{\Theta_k(a, b)}{\Theta(a, \cdot)\Theta(\cdot, b)}\right),$$

and k represents the expected percentage of amino acid replacements, most often between 30 and 250, with larger numbers used for sequences further away in the evolutionary scale. Sequence alignment may then be formulated as a maximum-likelihood or Bayesian inference problem [Bishop and Thompson (1986), Durbin et al. (1998), Liu and Lawrence (1999), Zhu, Liu and Lawrence (1998)], where the introduction of Θ_k amounts to the introduction of a number of new fixed hyperparameters. Inference on k can also be carried out by placing a prior distribution over members of the PAM family of matrices [Zhu, Liu and Lawrence (1998)]. The posterior distribution on k then provides an estimate of evolutionary distance between the two proteins.

Multiplication of equations (4) and (9) directly yields a joint likelihood for inferring M by combining both sequence and structure information. However, as we noted in Section 2, structure is generally much more strongly conserved than sequence, thus, we would like to weight the contribution of structure information in determining the alignment more heavily than that of sequence. In this way the sequences will serve primarily to provide supplementary information in regions of the alignment where structural information leaves uncertainty; as we will see, it also permits the estimation of evolutionary distance from the largely structure-based alignment.

To control the relative weighting of sequence and structure information, we introduce a concentration (or inverse temperature) parameter η, resulting in the modified sequence likelihood

$$\Pr\left(A^x, A^y | M, \Theta, \eta\right)$$

$$= \frac{\prod_{(i,j)\in M} \Theta(A_i^x, A_j^y)^\eta \prod_{i\notin M} \Theta(A_i^x, \cdot)^\eta \prod_{j\notin M} \Theta(\cdot, A_j^y)^\eta}{\sum_{A^{x*}, A^{y*}} \prod_{(i,j)\in M} \Theta(A_i^{x*}, A_j^{y*})^\eta \prod_{i\notin M} \Theta(A_i^x, \cdot)^\eta \prod_{j\notin M} \Theta(\cdot, A_j^y)^\eta}$$

$$= \prod_{(i,j)\in M} \frac{\Theta(A_i^x, A_j^y)^\eta}{\sum_{A_r, A_s} \Theta(A_r, A_s)^\eta} \prod_{i\notin M} \frac{\Theta(A_i^x, \cdot)^\eta}{\sum_{A_r} \Theta(A_r, \cdot)^\eta} \prod_{j\notin M} \frac{\Theta(\cdot, A_j^y,)^\eta}{\sum_{A_s} \Theta(\cdot, A_r)^\eta}.$$

Setting $\eta = 1$ corresponds to simple multiplication of the sequence and structure likelihoods (4) and (9), while as $\eta \to 0$, $\Pr(A^x, A^y | M, \Theta, \eta)$ approaches a uniform distribution for every Θ and $\eta = 0$ reduces to the structure-only model (9). Thus, η^{-1} plays the role of a dispersion parameter for the discrete observations A. We

can consider estimating η directly under the restriction $\hat{\eta} < 1$, in which case $\hat{\eta}$ can be interpreted as a measure of agreement between sequence and structure, which shrinks to down-weight the contribution of sequence information if sequence and structural information are in conflict.

6. Examples. We apply our Bayesian structural alignment algorithm to a number of illustrative examples. Hyperparameter values used are given in Table 1: the prior distribution for σ^2 has mean 1.5 Å and variance 1.0 Å, in line with the results for analogous proteins in Chothia and Lesk (1986), and following Gerstein and Levitt (1998), the prior mean for h is about 40 times larger than the prior mean for g. Results were mostly unaffected by changes in the prior mean for σ^2 between 0.5 Å and 4.0 Å, or by changes in the prior mean of g and h of around 50%. All inferences described are based on 100,000 samples obtained after a burn-in period of 20,000 iterations, with convergence verified by visual inspection of the trace plots and using the Gelman–Rubin convergence test [Gelman and Rubin (1992)]. Monitored quantities include the length of the alignment, the rotation angles corresponding to rotation matrix $\hat{R}_M$, the translation vector $\hat{\mu}_M$ and the two gap penalty parameters (g, h). We report MAP alignments unless otherwise noted.

We first analyze 16 pairs of proteins from Ortiz, Strauss and Olmea (2002). This list includes pairs of very different lengths and proteins from various structural classes, including α proteins containing primarily α-helical secondary structure, β proteins containing primarily β-sheets and $\alpha + \beta$ proteins containing significant fractions of both. Table 2 summarizes the results obtained using three different values for λ ranging from a relatively low (7.6) to the relatively high (9.6), and compares the Bayesian alignments against those obtained using the popular CE algorithm [Shindyalov and Bourne (1998)]. In most cases, the differences between Bayesian and CE alignments are important; in more than half the examples less than 20% of the matched residues coincide. These differences are mostly due to the way CE handles gaps: to reduce the computational complexity, CE assumes that gaps cannot be introduced simultaneously in both proteins. Similar restrictions can be easily introduced in our model [by setting $q_{i,j}(2, 3) = 0$ in Appendix], and when this is done the results for both methods tend to agree. Generally speaking, the added flexibility means that the quality of the Bayesian alignments is superior to CE: it tends to produce alignments containing more matched residues but with a lower RMSD.

TABLE 1
Hyperparameter values used in the examples

a_σ	b_σ	a_h	b_h	a_g	b_g
2.25	1.5	2	1/2	2	20

TABLE 2

Bayesian structural alignments of the 16 *pairs of proteins from Ortiz, Strauss and Olmea* (2002), *compared against the popular CE algorithm* [*Shindyalov and Bourne* (1998)]. |M| *is the length of the alignment,* N_{CE} *denotes the number of matches in common with CE, and RMSD is expressed in Å. The values of g and h reported correspond to posterior means*

			CE		$\lambda = 7.6$					$\lambda = 8.6$					$\lambda = 9.6$				
X–Y	*n*	*m*	|*M*|	RMSD	|*M*|	RMSD	N_{CE}	*h*	*g*	|*M*|	RMSD	N_{CE}	*h*	*g*	|*M*|	RMSD	N_{CE}	*h*	*g*
1ABA–1DSB	87	188	56	4.5	24	2.2	0	9.64	0.01	57	3.7	0	6.26	0.13	76	4.7	14	6.10	0.42
1ABA–1TRS	87	105	70	2.7	65	3.0	37	5.68	0.14	72	3.4	38	5.68	0.22	75	3.6	38	5.70	0.24
1ACX–1COB	108	151	92	4.0	66	2.1	49	6.89	0.06	86	3.8	57	6.60	0.15	93	4.1	57	6.38	0.21
1ACX–1RBE	108	104	56	7.3	25	2.5	6	9.02	0.01	31	2.8	0	7.89	0.02	50	4.2	15	7.45	0.03
1MJC–5TSS	69	194	61	2.7	52	2.3	25	7.89	0.03	60	3.0	29	7.24	0.36	66	3.9	15	7.36	0.44
1PGB–5TSS	56	194	48	2.9	39	2.3	19	6.52	0.56	55	3.3	40	6.60	0.87	55	3.1	34	6.65	0.94
1PLC–1ACX	102	108	80	3.3	71	3.4	23	5.92	0.10	84	4.0	23	5.80	0.20	89	4.6	23	6.30	0.22
1PTS–1MUP	119	157	80	4.1	76	3.0	0	6.80	0.06	83	3.1	0	6.60	0.09	88	3.5	0	6.77	0.12
1TNF–1BMV	152	185	115	4.1	70	2.7	3	7.86	0.02	109	4.2	40	7.10	0.08	113	4.3	35	6.94	0.11
1UBQ–1FRD	76	98	64	4.4	62	3.0	28	5.41	0.15	62	2.9	23	5.10	0.25	65	3.1	32	5.36	0.29
1UBQ–4FXC	76	98	64	4.0	46	2.3	22	5.46	0.13	61	2.9	34	5.20	0.24	66	3.4	42	5.43	0.29
2GB1–1UBQ	56	76	48	3.1	44	2.1	0	5.38	0.18	51	3.4	6	5.27	0.46	51	3.3	6	5.66	0.59
2GB1–4FXC	56	98	48	3.6	35	3.5	0	9.06	0.06	53	3.9	7	7.56	0.42	55	4.1	0	7.55	0.62
2RSL–3CHY	119	128	80	4.1	43	2.6	22	7.99	0.02	76	3.8	31	6.76	0.08	81	4.0	33	6.67	0.11
2TMV–256B	154	106	84	3.5	65	2.3	0	7.39	0.06	79	2.9	0	7.13	0.10	89	4.0	69	6.87	0.19
3CHY–1RCF	128	169	116	3.9	80	3.0	38	6.89	0.06	122	4.5	87	5.99	0.54	126	4.7	70	6.05	0.76

Some pairs of proteins seem to be somewhat sensitive to the choice of λ (e.g., 1ACX–1COB), while the alignment of other pairs seems to be remarkably robust (e.g., 1UBQ–FRD). In general, larger values of λ (which imply larger penalties for both opening and extension) tend to generate longer alignments. The sensitivity of the model to λ is not surprising; indeed, setting $\lambda = 0$ immediately implies that the optimal alignment is empty for any pair of proteins. The model is robust to small changes in λ (around 5% or so), which can be absorbed by g and h. However, when λ is increased by a large amount, we set a new baseline threshold that pairs of residues need to satisfy in order to be included in the alignment, favoring the alignment of sections that were previously not close enough to be aligned. When structural similarity varies dramatically along the alignment (as is the case for some pairs in our list), this "threshold effect" can produce important changes in the resulting alignment. In general, we can think of λ as controlling the tightness of the alignment. In our experience, a conservative value such as $\lambda = 7.6$ works well in most applications, and we use this value in further illustrations.

Table 2 also shows the posterior median of the gap penalties for each alignment. Opening penalties range between 5 and 9, while extension penalties range from 0.01 to nearly 0.9, reflecting the differing levels of similarity across different pairs.

Next, we consider in detail the alignment of two α proteins from the globin family, 5MBN and 2HBG. Figure 3 presents both the marginal alignment matrix (which provides information on the uncertainty associated with the alignment) and the MAP alignment, comparing it against that obtained from CE. The most striking feature about this example is that different alignment methods tend to disagree in regions where the uncertainty in the Bayesian alignment is high (the regions surrounding the gap between residues 47 and 62 of 5MBN, the extremes of the helix between residues 81 and 98 of 5MBN, and at the very end of the alignment). This highlights the importance of using a probabilistic alignment framework, rather than relying on a single optimum. Figure 4 shows the prior and posterior distributions for both gap penalty parameters in this example, which demonstrate that the model does adaptively estimate parameters relevant to the data at hand.

Finally, we explore the alignment of the $\alpha + \beta$ proteins 1CEW_I and 1OUN_A. Lackner et al. (2000) describe two alternative alignments for these proteins having a comparable number of equivalent residues (70 vs. 68) and RMSD (2.4 Å both), which arise by shifts in the alignment of the secondary structures. Figure 5 shows the marginal alignment probabilities for all pairs of residues. Unlike the previous example, uncertainty levels in this alignment are very high, particularly in the α-helix region between residues 10 and 20. The two alternative alignments for this region correspond to the two alignments described in Lackner et al. (2000). However, the alignment of the rest of the proteins corresponds to the 70 residue alignment discussed by those authors. This example shows how the global sampling of the full posterior enables the model to automatically weight the relative importance of closely related alternative alignments, and how the estimation of gap penalties can further improve this.
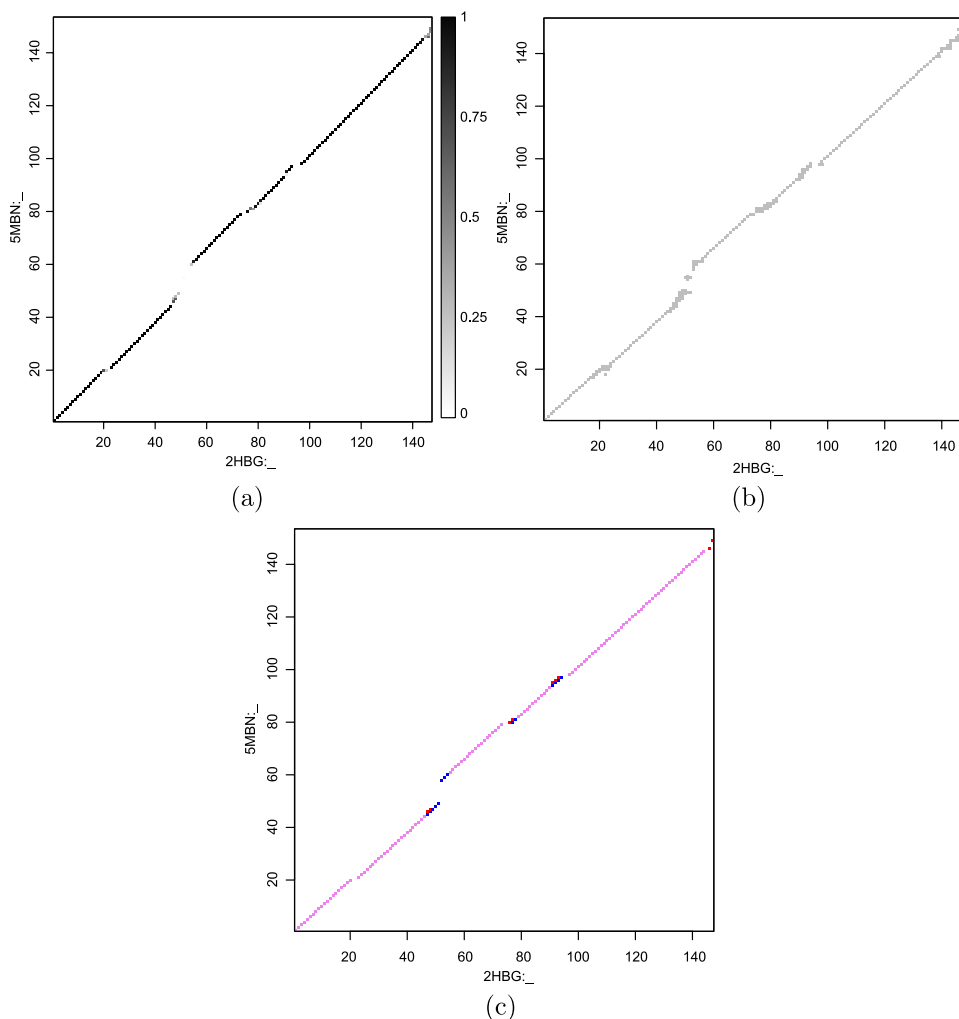
FIG. 3. *Bayesian structural alignment of* 5*MBN and* 2*HBG.* (a) *Marginal alignment probability matrix for all pairs of residues, showing uncertainty associated with the alignment.* (b) *Plot of all sampled alignments.* (c) *Comparison of the MAP alignment* (*red*) *with the CE alignment* (*blue*); *common regions are shown in purple.*

6.1. *Combined sequence–structure alignment.* To illustrate the performance of our simultaneous sequence-and-structure alignment approach, we consider two pairs of proteins that have been previously analyzed in the literature. For convenience we consider a discrete set of discount factors ranging from 0 to 1 in increments of 0.1, along with 21 PAM matrices ranging from PAM100 to PAM300. "Noninformative" uniform prior distributions are used for both discount factors and PAM matrices. All results are based on 130,000 iterations of the Gibbs sampler, after a burn-in period of 30,000 iterations.
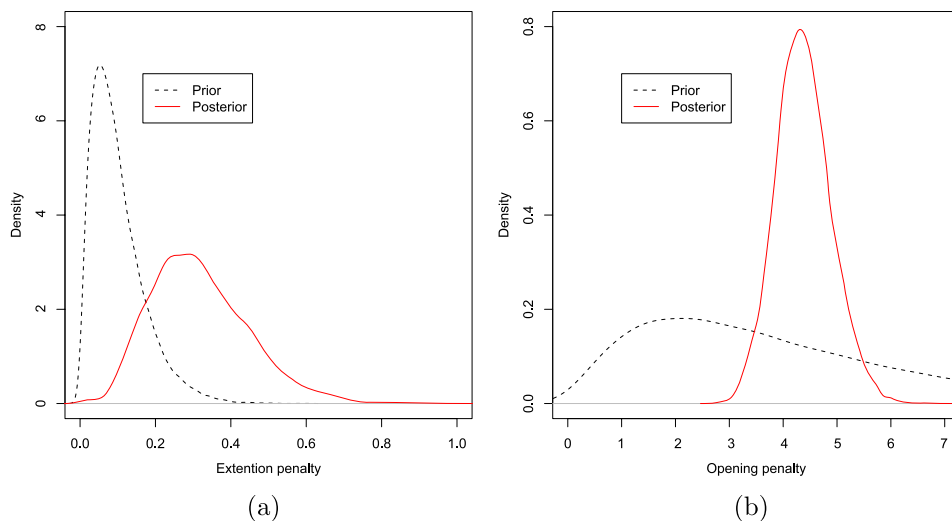
FIG. 4.   *Prior and posterior distributions for gap penalty parameters obtained for the Bayesian alignment of globins* 5*MBN and* 2*HBG. The Bayesian approach allows the algorithm to adaptively determine the appropriate gap parameters rather than treating them as fixed.*

In the first example we analyze two kinases studied by Bayesian sequence alignment in Zhu, Liu and Lawrence (1998); a guanylate kinase from yeast (1GKY) and
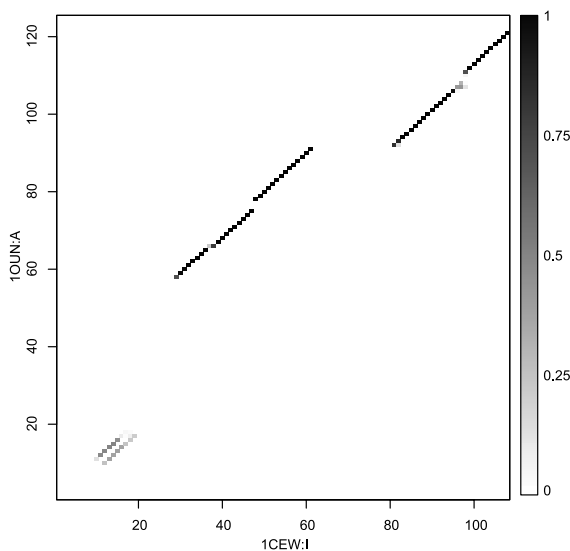

FIG. 5.   *Marginal alignment matrix for the Bayesian structural alignment of* 1*OUN:A and* 1*CEW:I. The posterior uncertainty in the alignment can be seen at the N-terminus, where two possible alignments of the α-helix at positions* 10–20 *have comparable posterior probabilities.*
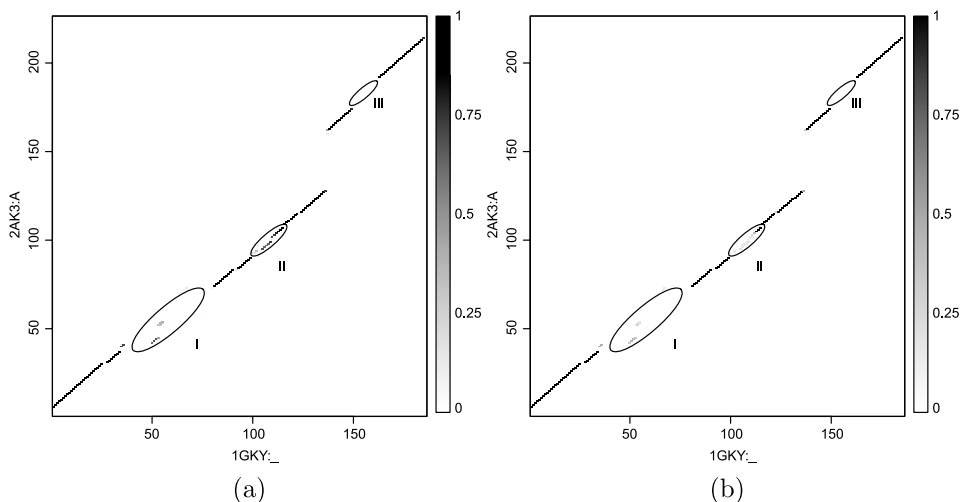
FIG. 6.   *Marginal probabilities over aligned pairs for* 1*GKY and* 2*AK*3*_A.* (a) *Shows alignments based only on structure*, *while* (b) *presents alignments that also incorporate sequence information. Although there is some structural similarity in regions* I *and* II, *sequence similarity in these areas is low* (*see Table* 3).

an adenylane kinase from the beef heart mitochondrial matrix (2AK3_A), which are VAST structural neighbors [Gibrat, Madej and Bryant (1996)]. A structural alignment for these proteins using the combinatorial extension (CE) algorithm [Shindyalov and Bourne (1998)] shows very little sequence similarity (under 13% identity). Figure 6 compares our structural and simultaneous sequence–structure alignments for these two proteins by showing the marginal probability of aligning any pair of residues integrated over all other parameters in the model (including PAM matrices and discount factors). Both algorithms tend to agree on which regions should be aligned. For example, both avoid aligning the section of the α-helix located between residues 150–162 in 1GKY and residues 175–191 in 2AK3_A (marked III in Figure 6). The axes for these helices are not parallel, producing a big divergence at the C terminus.

TABLE 3
Sequence alignment of corresponding to the best structural alignment between region II *of* 2*AK*3*_A and* 1*GKY, with residues* 93–100 *of* 2*AK*3*_A matched with residues* 103–111 *of* 1*GKY. Numbers correspond to the PAM* 250 (*log-odds*) *scores for each matched residue pair and clearly show that despite the shape similarity, there is little evidence of common ancestry in this region of the protein*

2AK3_A	R	T	L	P	Q	A	E	A
1GKY	G	V	K	S	V	K	A	I
	−3	0	−3	1	−2	−1	0	−1

In spite of the similarities, some differences are evident among both models. For example, a section of the alignment starting at residue 108 of 1GKY (marked II in Figure 6) is excluded when the sequence information is included in the analysis. Both proteins present a short helix in this region, and they can be structurally aligned reasonably well. However, there are important incompatibilities in the two sequences for these helices, which suggests that this section is not functionally important. Table 3 presents the sequence correspondence associated with the structural alignment of this section, along with the scores for each site. Note that the structural alignment implies no conserved residues in the area and the substitution of various basic and acidic amino acids by either hydrophobic or hydrophilic residues. Indeed, of the eight substitutions, only one happens between members of a common chemical group. This is a local discrepancy between sequence and structure that is not seen in other regions of the proteins, and suggests that the region should be dropped from the alignment. Similarly, a couple of short regions in the remote site for mono and triphosphate binding located between residues 35–80 for 1GKY and 38–73 in 2AK3_A (marked I in Figure 6), that show a moderate probability of being aligned under the structural model, are down-weighted (but not completely removed) when the sequence information is included. This region, which was probably functionally important in an ancestor, has degraded since both proteins diverged and does not seem functionally active in these proteins. These two minimal changes in the alignment lowers the RMSD from 3.5 Å to a median of 1.95 Å [with a 90% high posterior density interval of (1.84, 2.17)].

Figure 7 shows the marginal posterior probability distribution over PAM matrices that arises from our joint sequence–structure model, contrasting it with the
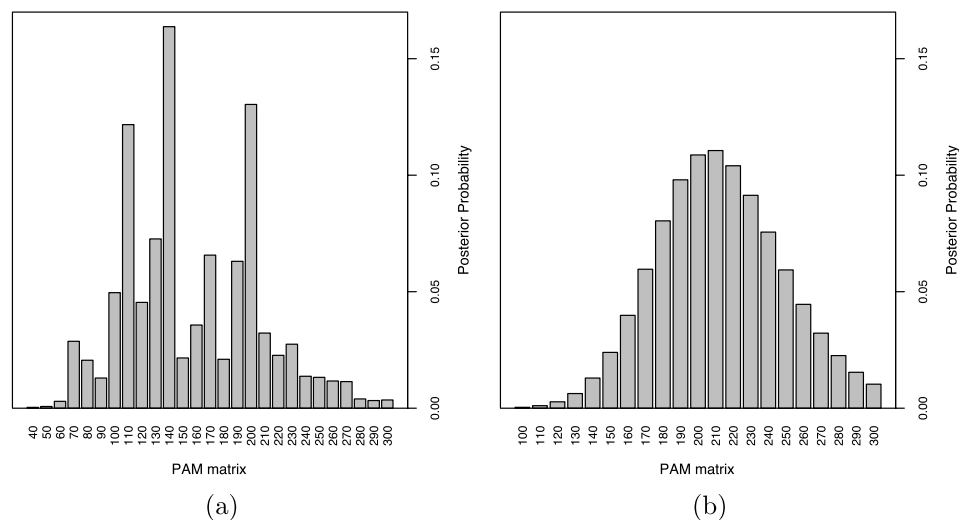

FIG. 7. *Posterior probabilities of PAM distances based on sequence information alone* (a) *and based on the Bayesian sequence–structure alignment.*

results in Zhu, Liu and Lawrence (1998). Whereas the sequence-based analysis in the original paper led to a multimodal posterior with modes at PAMs 110, 140 and 200, our posterior is smooth and unimodal, with its mode located between PAM200 and PAM210. This demonstrates the strong additional information obtained by aligning based on structure and sequence simultaneously: virtually all sequence alignments which are compatible with structural alignment indicate the larger evolutionary distance (posterior mean 212, median 206). The marginal mode for the temperature is 1 (posterior probability 0.57), indicating that there is very little need to discount sequence with respect to structure information.

Our second example focuses on comparing the single-chain fused Monellin from the Serendipity berry (1MOL_A) and the chicken egg white Cystatin (1CEW_I) analyzed previously in Lackner et al. (2000) and Kotlovyi, Nichols and Eyck (2003). Figure 8 shows the Bayesian alignments obtained with and without inclusion of sequence information. Again, the two alignments are quite similar as expected, but the sequence information leads to small refinements in the structural alignment. For example, two alternative alignments of the initial strand are supported by the structure-only alignment, with the one where 1MOL_A is shifted toward the C terminus being slightly preferred (this is also the one preferred by CE). However, incorporation of sequence information reverses this to prefer the N-terminus shifted alignment (approximate posterior probabilities of 0.85 vs 0.15), and examination of the sequences strongly supports this choice. Table 4 shows the sequence alignment under both alternatives, with amino acids colored by a simple classification according to physico-chemical properties (Table 5) to demonstrate the improved similarity on top of amino acid identity. The
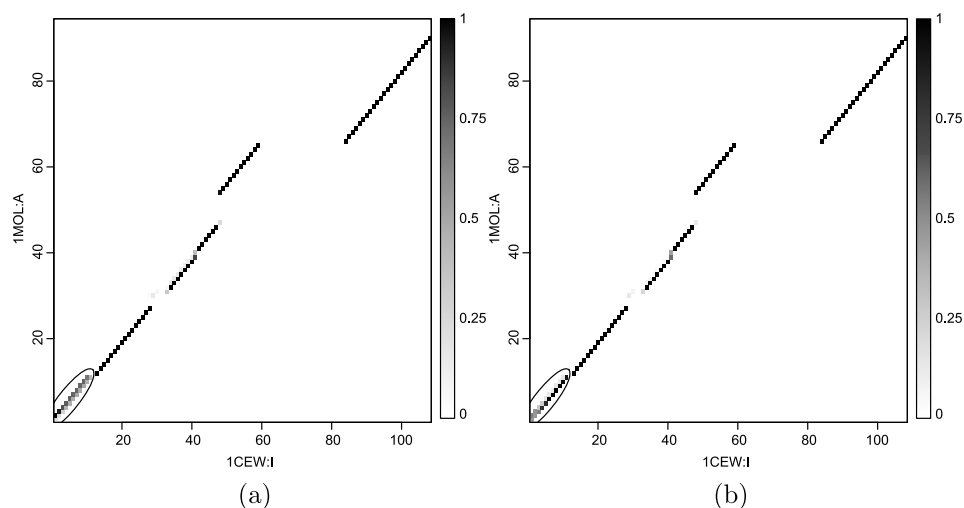

FIG. 8.   *Marginal probabilities over aligned pairs for* 1*MOL_A and* 1*CEW_I.* (a) *Shows alignments based only on structure, while* (b) *presents alignments that also incorporate sequence information. Circles show the strand region discussed in Table 4.*

TABLE 4

Sequence alignment of the first strand of 1*MOL_A and* 1*CEW_I induced by the two alternative models.* (a) *Mode using structural information only and* (b) *mode under the Bayesian simultaneous sequence–structure alignment. Colors refer to the classification in Table* 5; *note the improvement in matching of chemical classes*

(a)	1MOL_A	E	W	E	I	I	D	I	G	P
	1CEW_I	G	A	P	V	P	V	D	E	N

(b)	1MOL_A	G	E	W	E	I	I	D	I	G
	1CEW_I	G	A	P	V	P	V	D	E	N

sequence–structure alignment yields six matches in amino acid type, including an additional two identities and a hydrophilic match on top of the three hydrophobics achieved by the structural alignment. The corresponding price paid in structural distance [mean RMSD of 1.91 Å versus 1.89 Å, with both 90% h.p.d. regions being (1.81 Å, 2.05 Å)] is insignificant. This example clearly shows that incorporation of sequence information can refine structural alignments in areas where the structure alignment is ambiguous.

Figure 9 shows the joint posterior distribution over PAM matrices and discount factors for this example. Relative to the previous example, there is more uncertainty in both the evolutionary distance and the discount factor. The diagonal pattern in the plot suggests an obvious dependence between these two parameters. This is to be expected, as both η and evolutionary distance increase the entropy of the joint amino acid distribution. Nevertheless, the results point toward a relatively large divergence time (recall one is a plant protein and the other is an animal protein), with the mode of the distance at 210.

To avoid confounding of PAM and tempering parameters, one parameter may be chosen in advance and fixed. For example, the substitution matrix may be chosen to reflect prior information about evolutionary distance and inference performed only on the *discount factor* or vice-versa. When 1MOL_A and 1CEW_I are aligned using PAM250 as the fixed substitution matrix, the resulting distribution for *discount factor* is very similar: the mode is located at $\eta = 0.6$ with a posterior probabil-

TABLE 5

Simple amino acid classification based on chemical properties

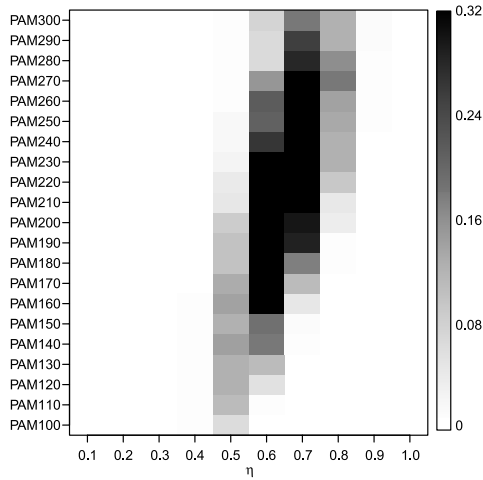| Group	Type	Amino acids
1	Non-polar, hydrophobic	A, V, L, I, P, M, F, W
2	Polar, hydrophilic	G, S, T, C, N, Q, Y
3	Acidic	D, E
4	Basic	K, R, H

FIG. 9.   *Heat map representation of the joint posterior distribution over discount factors and PAM matrices for* 1*MOL_A and* 1*CEW_I.*

ity of 0.32, and most of the remaining mass concentrates in $\eta = 0.5$ and $\eta = 0.7$, both with posterior probability of 0.24. Differences in the actual alignments are not obvious from the marginal distribution plot (not shown). However, a more detailed look at the values shows that fixing the PAM matrix further decreases the probability of the CE-like alignment below 10%.

The examples discussed in this section show that simultaneous estimation of PAM distance and discount factor may be difficult. Since larger evolutionary distances increase sequence divergence/decrease sequence conservation, both low discount values and high PAM distances imply more tolerance to substitutions. One way to measure substitution tolerance is via the Shannon entropy of the joint distributions (Figure 10). Although increasing η and k both increase this entropy, they do so in slightly different ways. We observe that PAM100 with a temperature of 0.8 has roughly equivalent entropy to untempered PAM200. However, PAM100/0.8 assign a much larger probability of match than does PAM200/1.0, as seen by the darker diagonal. In addition, temperature increases treat all combinations of amino acids in the same way, and thus low probability regions tend to disappear quickly. This is not so for increases in the evolutionary distances. In the limit of k, the PAM joint distribution will converge to the product of independent marginal distributions given by the stationary distribution of the underlying Markov chain (estimated as overall population frequencies). In contrast, as the discount factor approaches 0, the joint distribution and thus the marginal distributions converge to uniform. Figure 10(a) also shows that differences between PAM matrices grow weaker as the discount factor decreases. In the extreme case when $\eta = 0$, all matrices are equivalent.
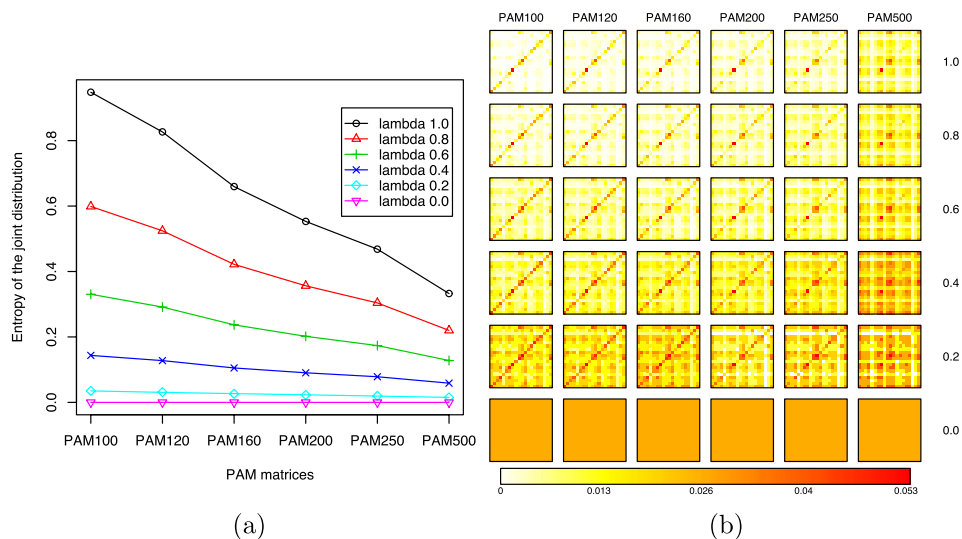
FIG. 10. (a) *Entropies of the joint distribution induced by different evolutionary distances and tempering parameters*. (b) *Heat map plots of the joint distributions. Amino acids are ordered alphabetically, starting with Alanine in the lower left.*

Finally, it is important to mention that we have not found alternative methodologies in the literature capable of this type of information synthesis, against which to compare our results. One of the few methods available is an extension of the combinatorial extension (CE) method Shindyalov and Bourne (1998), accessible via http://cl.sdsc.edu/ce.html. However, in this implementation there is little control on the choice of substitution matrices and, for the examples we have studied, the sequences seems to have little practical influence in the final results.

7. Conclusions. We have presented a unifying probabilistic framework for protein structure alignment based on Bayesian hierarchical modeling. Computationally efficient MCMC algorithms for sampling the posterior distribution enable us to directly account for uncertainty over alignments, including identification of alternative alignments and evaluation of their relative importance. Our model provides insights into the relations between and assumptions of standard optimization-based alignment techniques, along with a unifying framework that facilitates comparisons between them. It also naturally incorporates additional information, such as the inclusion of sequence information in structural alignments. As a byproduct of the latter, we obtained a model which can estimate evolutionary distance directly from structural alignment, an otherwise difficult task. The examples shown clearly highlight how these advantages of our model aid in identification of functionally relevant regions and in resolving ambiguities in alignments. By introducing a discount parameter, we are able to control the influence of the sequence information on the final alignment, an important characteristic missing in previous attempts

to combine sequence and structure. As noted, PAM distance and discount factor are correlated, and inference on evolutionary distance will therefore be more reliable if additional information is used to determine the discount factor; this is an area for additional study. Finally, we feel that sequence–structure alignments provide the most insight when used in conjunction with structure-only alignments as done in the examples. Comparisons between the two appear to provide more direct information on conservation than do comparisons between structure-only and sequence-only alignments.

APPENDIX: DYNAMIC PROGRAMMING FORWARD–BACKWARD SAMPLING

As shown by Schmidler (2003), if we *condition* on registration parameters (R, μ), the alignment matrix M may be sampled from its full conditional distribution using a forward–backward algorithm similar to that of sequence alignment [Liu and Lawrence (1999), Zhu, Liu and Lawrence (1998)]. Let $v_{i,j}(k)$ be the probability of the alignment of the ith prefix of X and the jth prefix of Y ending in type k, with $k = 1$ meaning that both final residues are aligned, $k = 2$ inserts a gap in X and $k = 3$ inserts a gap in Y. Then

$$v_{i,j}(1) = \sum_{k=1}^{3} q_{i,j}(k, 1) v_{i-1, j-1}(k), \qquad v_{i,j}(2) = \sum_{k=1}^{3} q_{i,j}(k, 2) v_{i-1, j}(k),$$

$$v_{i,j}(3) = \sum_{k=1}^{3} q_{i,j}(k, 3) v_{i, j-1}(k)$$

and letting $d_{ij}^2 = \| y_j - (x_i R + \mathbf{1}\mu') \|^2$, the transition weights are given by

$$q_{i,j}(l, k) = \begin{cases} \dfrac{\lambda}{(2\pi\sigma^2)^{3/2}} \exp\left\{ -\dfrac{1}{2\sigma^2} d_{ij}^2 \right\}, & k = 1, \\ \exp\{g + h\}, & (l, k) = (1, 2) \text{ or } (1, 3), \\ \exp\{g\}, & (l, k) = (2, 2) \text{ or } (3, 3) \text{ or } (2, 3), \\ 0, & (l, k) = (3, 2). \end{cases}$$

In order to ensure identifiability of the alignments, we do not allow a gap in Y to follow a gap in X, hence, $q_{3,2} = 0$. The initialization of these recursions are

$$v_{1,1}(1) = \frac{\lambda}{(2\pi\sigma^2)^{3/2}} \exp\left\{ -\frac{1}{2\sigma^2} d_{11}^2 \right\},$$

$$v_{i,1}(1) = \frac{\lambda}{(2\pi\sigma^2)^{3/2}} \exp\left\{ -\frac{1}{2\sigma^2} d_{1i}^2 + (i - 1)g + h \right\},$$

$$v_{1,j}(1) = \frac{\lambda}{(2\pi\sigma^2)^{3/2}} \exp\left\{ -\frac{1}{2\sigma^2} d_{j1}^2 + (j - 1)g + h \right\},$$

$$v_{1,j}(2) = \exp\{(j + 1)g + h\} \quad \text{and} \quad v_{i,1}(3) = 0.$$

Note that $v_{n,m}$ contains the sum over all alignments and, given (g, h), the same algorithm with $q_{i,j}(l, 1) = 1$ can be used to efficiently compute the normalizing constant $Z(g, h)$ in the gap-penalty prior (5), as required for the acceptance probability (8). Once $q_{i,j}(k)$ is available for all (i, j), the alignment is sampled backward, starting with

$$u_{n,m}(k) = \frac{v_{n,m}(k)}{\sum_{l=1}^{3} v_{n,m}(l)}$$

and then conditionally adding a matched pair or a gap on one of the proteins with probabilities:

$$u_{i,j}(k, l) = \frac{q_{i-1,j-1}(l, k)v_{i-1,j-1}(k)}{\sum_{k=1}^{3} q_{i-1,j-1}(l, k)v_{i-1,j-1}(k)}.$$

REFERENCES

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. and LIPMAN, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215** 403–410.

BISHOP, M. J. and THOMPSON, E. A. (1986). Maximum likelihood alignment of DNA sequences. *J. Mol. Biol.* **190** 159–165.

BRONNER, C. E. et al. (1994). Mutation in the DNA mismatch repair gene homologue hMLH1 is associated with hereditary non-polyposis colon cancer. *Nature* **368** 258–261.

BROWN, N., ORENGO, C. and TAYLOR, W. (1996). A protein structure comparison methodology. *Compational Chemistry* **20** 359–380.

CHOTHIA, C. and LESK, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5** 823–826.

COHEN, G. (1997). ALIGN: A program to superimpose protein coordinates, accounting for insertions and deletions. *Acta Crystallographica* **30** 1160–1161.

DAYHOFF, M. and ECK, R., eds. (1968). *Atlas of Protein Sequence and Structure* **3**. National Biomedical Research Fundation, Silver Spring, MD.

DRYDEN, I. L., HIRST, J. D. and MELVILLE, J. L. (2007). Statistical analysis of unlabeled point sets: Comparing molecules in chemoinformatics. *Biometrics* **63** 237–251, 315. MR2345594

DRYDEN, I. L. and MARDIA, K. V. (1998). *Statistical Shape Analysis*. Wiley, Chichester. MR1646114

DURBIN, R., EDDY, S., KROGH, A. and MITCHISON, G. (1998). *Biological Sequence Analysis*: *Probabilistic Models of Proteins and Nucleic Acids*. Cambridge Univ. Press, Cambridge.

EIDHAMMER, I., JONASSEN, I. and TAYLOR, W. R. (2000). Structure comparison and structure patterns. *J. Comput. Biol.* **7** 685–716.

FALICOV, A. and COHEN, F. E. (1996). A surface of minimum area metric for the structural comparison of proteins. *J. Mol. Biol.* **258** 871–892.

FISCHER, D., WOLFSON, H., LIN, S. and NUSSINOV, R. (1994). Three-dimensional, sequence order-independent structural comparison of a serine protease against the crystallographic database reveals active site similaties: Potential implications to evolution and to protein folding. *Protein Sci.* **3** 768–778.

GELMAN, A. and RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7** 457–472.

GERSTEIN, M. and LEVITT, M. (1998). Comprehensive assessment of automatic structural alignment against a manual standard, the scop classification of proteins. *Protein Sci.* **7** 445–456.

GIBRAT, J.-F., MADEJ, T. and BRYANT, S. (1996). Surprising similarities in structure comparison. *Current Opinion in Structual Biology* **6** 377–385.

GODZIK, A. (1996). The structural alignment between two proteins: Is there a unique answer? *Protein Eng.* **5** 1325–1338.

GREEN, P. J. and MARDIA, K. V. (2006). Bayesian alignment using hierarchical models, with applications in protein bioinformatics. *Biometrika* **93** 235–254. MR2278080

GRISHIN, N. V. (1997). Estimation of evolutionary distances from protein spatial structures. *J. Mol. Evol.* **45** 359–369.

HENIKOFF, S. and HENIKOFF, J. (1992). Amino acid substituion matrices from protein blocks. *Procdings of National Academy of Science* **89** 10915–10919.

HOLM, L. and SANDER, C. (1993). Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233** 123–138.

JOHNSON, M. S., SUTCLIFFE, M. J. and BLUNDELL, T. L. (1990). Molecular anatomy: Phylectic relationships derived from three-dimensional structures of proteins. *J. Mol. Evol.* **30** 43–59.

KENDALL, D. G., BARDEN, D., CARNE, T. K. and LE, H. (1999). *Shape and Shape Theory*. Wiley, Chichester. MR1891212

KENOBI, K. and DRYDEN, I. L. (2012). Bayesian matching of unlabeled point sets using Procrustes and configuration models. *Bayesian Anal.* **7** 547–565. MR2981627

KOEHL, P. and LEVITT, M. (2002). Sequence variations within protein families are linearly related to structural variations. *J. Mol. Biol.* **323** 551–562.

KOTLOVYI, V., NICHOLS, W. L. and EYCK, L. F. T. (2003). Protein structural alignment for detection of maximally conserved regions. *Biophys. Chem.* **105** 595–608.

LACKNER, P., KOPPENSTEINER, W. A., SIPPL, M. J. and DOMINGUES, F. S. (2000). ProSup: A refined tool for protein structure alignment. *Protein Eng.* **13** 745–752.

LEMMEN, C. and LENGAUER, T. (2000). Computational methods for the structural alignment of molecules. *J. Comput.-Aided Mol. Des.* **14** 215–232.

LEVITT, M. and GERSTEIN, M. (1998). A unified statistical framework for sequence comparison and structure comparison. *Proceedings of the National Academy of Sciencies USA* **95** 5913–5920.

LIPMAN, D. J. and PEARSON, W. R. (1985). Rapid and sensitive protein similarity searches. *Science* **227** 1435–1441.

LIU, J. S. and LAWRENCE, C. E. (1999). Bayesian inference on biopolymer models. *Bioinformatics* **15** 38–52.

MIZUGUCHI, K. and GO, N. (1995). Seeking significance in three-dimensional protein structure comparisons. *Curr. Opin. Struck. Biol.* **5** 377–382.

NEEDLEMAN, S. B. and WUNSCH, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48** 443–453.

ORTIZ, A. R., STRAUSS, C. E. M. and OLMEA, O. (2002). Mammoth (Matching molecular models obtained from theory): An automated method for model comparison. *Protein Sci.* **11** 2606–2621.

PAPADOPOULOS, N., NICOLAIDES, N. C., WEI, Y. F., RUBEN, S. M., CARTER, K. C., ROSEN, C. A., HASELTINE, W. A., FLEISCHMANN, R. D., FRASER, C. M., ADAMS, M. D. et al. (1994). Mutation of a mutL homolog in hereditary colon cancer. *Science* **263** 1625–1629.

RAO, S. and ROSSMANN, M. (1973). Comparison of super-secondary structures in proteins. *J. Mol. Biol.* **105** 241–256.

ROSSMANN, M. and ARGOS, P. (1975). A comparison of heme binding pocket in globins and cytochrombe b5*. *J. Biol. Chem.* **250** 7523–7532.

ROSSMANN, M. and ARGOS, P. (1976). Exploring structural homology in proteins. *J. Mol. Biol.* **105** 75–96.

SATOW, Y., COHEN, G., PADLAN, E. and AVIES, D. (1986). Phosphocholine binding immunoglobulin fab mcpc603: An x-ray diffraction study at 2.7 a. *J. Mol. Biol.* **190** 593–604.

SAUL, L. K. and JORDAN, M. I. (1995). Boltzmann chains and hidden Markov models. In *Advances in Neural Information Processing Systems* (*NIPS*) 7 (G. Tesauro, D. S. Touretzky and T. Leen, eds.). MIT Press, Cambridge, MA.

SCHMIDLER, S. C. (2003) Statistical shape analysis of protein structure families. In *Presentation at the LASR workshop on Stochastic Geometry, Biological Structure and Images*, Leeds, UK.

SCHMIDLER, S. C. (2004). Bayesian shape matching and protein structure alignment. In *Presentation at the* 6*th World Congress of the Bernoulli Society and the* 67*th Annual Meeting of the Institute of Mathematical Statistics*, Barcelona, Spain.

SCHMIDLER, S. C. (2007a). Fast Bayesian shape matching using geometric algorithms. In *Bayesian Statistics* 8 (M. Bernardo, J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. Smith and M. West, eds.) 471–490. Oxford Univ. Press, Oxford. MR2433204

SCHMIDLER, S. C. (2007b). Bayesian flexible shape matching with applications to structural bioinformatics. Technical report, Dept. Statistical Sciences, Duke Univ., Durham, NC.

SCHMIDLER, S. C., LUCAS, J. E. and OAS, T. G. (2007). Statistical estimation of statistical mechanical models: Helix-coil theory and peptide helicity prediction. *J. Comput. Biol.* **14** 1287–1310. MR2373277

SHINDYALOV, I. N. and BOURNE, P. E. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* **11** 739–747.

SMALL, C. G. (1996). *The Statistical Theory of Shape*. Springer, New York. MR1418639

TAYLOR, W. R. (2002). Protein structure comparison using bipartite graph matching and its application to protein structure classification. *Mol. Cell. Proteomics* **1** 334–339.

TAYLOR, W. R. and ORENGO, C. A. (1989). Protein structure alignment. *J. Mol. Biol.* **208** 1–22.

THOMPSON, J. D., PLEWNIAK, F. and POCH, O. (1999). BAliBASE: A benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics* **15** 87–88.

THOMPSON, J. D., KOEHL, P., RIPP, R. and POCH, O. (2005). BAliBASE 3.0: Latest developments of the multiple sequence alignment benchmark. *Proteins* **61** 127–136.

WALLACE, A., LASKOWSI, R. and THORNTON, J. (1996). Tess: A geometric hashing algorithm for deriving 3d coordinate templates for searching structural databases: Applications to enzyme active sites. *Protein Sci.* **6** 2308–2323.

WANG, R. and SCHMIDLER, S. C. (2008). Bayesian multiple protein structure alignment and analysis of protein families. In preparation.

WEBB, B.-J. M., LIU, J. S. and LAWRENCE, C. E. (2002). BALSA: Bayesian algorithm for local sequence alignment. *Nucleic Acids Res.* **30** 1268–1277.

WOOD, T. C. and PEARSON, W. R. (1999). Evolution of pretein sequences and structures. *J. Mol. Biol.* **291** 977–995.

WU, T. D., SCHMIDLER, S. C., HASTIE, T. and BRUTLAG, D. L. (1998). Regression analysis of multiple protein structures. *J. Comput. Biol.* **5** 597–607.

ZHU, J., LIU, J. S. and LAWRENCE, C. E. (1998). Bayesian adaptive sequence alignment algorithms. *Bioinformatics* **14** 25–39.

ZU-KANG, F. and SIPPL, M. (1996). Optimum superimposition of protein structures: Ambiguities and implications. *Fold. Des.* **1** 123–132.

DEPARTMENT OF APPLIED MATHEMATICS
AND STATISTICS
UNIVERSITY OF CALIFORNIA, SANTA CRUZ
1156 HIGH STREET
MAILSTOP SOE2
SANTA CRUZ, CALIFORNIA 95064
USA
E-MAIL: abel@soe.ucsc.edu

DEPARTMENT OF STATISTICAL SCIENCE
DUKE UNIVERSITY
BOX 90251
DURHAM, NORTH CAROLINA 27708
USA