BIOINFORMATICS

Vol. 00 no. 00
Pages 1–8

Probabilistic Modeling and Alignment of Protein Structure Families

Rui Wang[1] and Scott C. Schmidler[1,2*]

[1]Computational Biology and Bioinformatics Program, Institute for Genome Sciences and Policy of Duke University, USA
[2]Departments of Statistical Science and Computer Science, Duke University, USA

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation:

Multiple protein structure alignment is an important tool in bioinformatics. Although several algorithms exist for this purpose, recent publications highlight inconsistencies among alignments from different algorithms and, increasingly, the recognition that alignments by a single algorithm may be highly unstable under small fluctuations of the input protein structures arising from thermal fluctuations, measurement error, and intrinsic flexibility. Lack of robustness to such uncertainties is a barrier to routine use to obtain reliable and reproducible biological insights. Similar issues in biopolymer sequence alignment are routinely handled by probabilistic models and statistical alignment techniques. Extending these principles from 1D sequence to 3D structure requires new statistical models and computational algorithms.

Results:

We present a fully probabilistic approach to the problem of multiple structure alignment in a protein family. In contrast to existing methods based on optimization of heuristic score functions, our approach is based on an explicit statistical model with testable assumptions. The resulting algorithm produces a Bayesian posterior distribution over possible alignments which accounts for alignment uncertainty arising from evolutionary variability, experimental noise, and thermal fluctuation, as well as the unknown parameters of the alignment algorithm itself. We demonstrate the robustness of this approach on alignments identified previously in the literature as "difficult" for existing algorithms. We also show the potential for significant stabilization of tree reconstruction in structural phylogenetics. We conclude by applying the algorithm deduce an important region responsible for functional diversification in two structurally similar paralogous neuronal guidance proteins.

Availability:

Program will be available at: http://www.duke.edu/ scs/

Contact: schmidler@stat.duke.edu

1 INTRODUCTION

Uncertainty in biological sequence alignments has received considerable attention recently, particularly in regard to its effect on phylogeny reconstruction (40; 25). Alignment uncertainty arises from multiple sources: the stochasticity of the underlying evolutionary model, the limited information contained in the pair or set of sequences to be aligned, and the sensitivity to input parameters of alignment algorithms. In this paper we consider uncertainty arising in the context of protein *structure* alignment. Structural alignment of proteins is a key tool for understanding protein function, mechanism, and evolution (see (11; 17) for reviews), and are commonly used as a "gold-standard" for evaluating or calibrating sequence alignments (24). Uncertainties in pairwise structure alignments have recently been considered by (34; 35; 30) and again arise from multiple sources. First, most current algorithms formulate structure alignment as an optimization problem with respect to some similarity metric, and different metrics weight various structural properties differently (often with tunable weights), leading to considerable subjective or empirical bias (14; 18). In addition, 3D protein structures are intrinsically flexible and dynamic - variability exceeding 1Å is common, and conformational changes may vary over tens of angstroms (6) - but high resolution (X-ray) structures are static snapshots. Sub-angstrom structural variation can cause substantial inconsistencies and apparently incorrect alignments by existing methods (28). Such shortcomings have led to calls for new approaches to the multiple structure alignment problem (4). Here we present a probabilistic approach to multiple structure alignment which addresses these issues explicitly.

Probabilistic modeling is a natural framework for accounting for uncertainty arising from multiple sources. For biological sequence analysis, probabilistic modeling has yielded highly effective tools for global and local pairwise sequence alignment (42; 39), multiple sequence alignment (20; 3; 23; 22), secondary structure prediction of proteins (33; 36) and RNA (10; 32), and protein tertiary structure prediction (37; 21; 41). See (9) for an introduction emphasizing sequence alignment and RNA base pairing. In the case of multiple sequence alignment, probabilistic models such as hidden Markov models (HMMs) and multinomial mixture models (23; 22) also have a computational advantage over optimizing all pairwise distances, which is NP-hard (38). The HMM approach instead aligns each input sequence to a (albeit unknown) profile model, reducing the multiple alignment problem to one of estimating the common profile using standard statistical algorithms, then pairwise aligning each input sequence to this model. Thus formulating the problem in terms of an explicit probabilistic model yields practical algorithms.

*to whom correspondence should be addressed

Here we describe a probabilistic model-based approach to multiple *structural* alignment. As with multiple sequence alignment, we build on the machinery of HMMs. However, as the application of HMMs to 3D structures requires significant generalization and algorithmic development, such models have not been applied to structure alignment previously. (Alexandrov and Gerstein (1) use an HMM to represent the core residue profile in a multiple structure alignment obtained by other means, but this does not address the alignment problem itself.) Our approach directly accounts for multiple sources of uncertainty in the alignment process, using Bayesian statistical methods to identify multiple possible alignments and their relative probabilities. Formally, the alignment is obtained by marginalizing over all remaining uncertainty in the model. This approach handles unknown model parameters in a coherent statistical estimation framework, allowing alignment parameters such as gap penalties and thresholds to be adaptively estimated from the data. We also address an unresolved problem in sequence alignment HMMs - choosing the length of the model - via Bayesian model averaging. The resulting algorithm is significantly more robust; replaces heuristic optimization criteria with clear, testable statistical assumptions; and results in structural alignments that lead to significantly more stable, robust phylogenetic trees.

2 APPROACH

A probabilistic model for protein structure families

Our approach to multiple structure alignment constructs a probabilistic model of the underlying protein family. This generalizes the approach introduced by (34; 35; 30) for *pairwise* protein structure alignment, and the closely related approach for matching residues of two protein active sites developed by (16). Let X_j be an $n_j \times 3$ matrix containing the C_α coordinates of protein j, for $j = 1, \ldots, m$. Our stochastic model represents each input structure X_j as being generated from "mean" or model structure U, to which insertions are added or deletions made stochastically, random noise added to the coordinates, and then an arbitrary Euclidean transformation applied. In the special case that the noise is independent $\Sigma_\epsilon = \phi^{-1} I$, this can be written

$$X_j = Y_j R_j + \vec{1}^T \vec{v}_j \qquad\qquad Y_j \sim \text{HMM}\,(\Theta)$$

where R_j is a rotation (special orthogonal matrix), and v_j an arbtrary translation, applied to the coordinates Y_j of the jth protein. Here $\Theta = \{U, \vec{\phi}, \vec{\mu}_\text{I}, \phi_\text{I}, Q\}$ denotes the collection of profile HMM model parameters (described below), with $U = [\vec{\mu}_1^\text{M}, \ldots, \vec{\mu}_n^\text{M}]^T$ the matrix of mean structure coordinates, $\vec{\phi} = \{\phi_i^\text{M}\}_{i=1}^n$ the corresponding emission precisions, and $\vec{1}$ the column vector of ones.

This hierarchical model combines ideas from two distinct fields: probabilistic sequence analysis (9), and statistical shape analysis (8). The additive error ϵ models the combined effects of thermal fluctuation and conformational variability, evolutionary drift, and experimental measurement error. The HMM consists of Match (**Mat**), Insert (**Ins**), and Delete (**Del**) states, organized as in profile HMM sequence alignment (20; 9). However in our model the **Mat** and **Ins** states emit 3D coordinate vectors from multivariate Gaussian distributions $y_i \sim \mathcal{N}(\vec{\mu}_i^\text{M}, \Sigma_i^\text{M})$ with state-specific mean positions, rather than letters of a nucleotide or amino acid sequence from discrete distributions. To simplify, we assume **Ins** states share a common mean position $\vec{\mu}^\text{I}$ and that covariance matrices Σ^I and Σ_j^M are diagonal, i.e. $\Sigma_j = \phi_j^{-1} \text{I}_3$. Importantly however, we allow the precision parameters ϕ_j for each input structure $j \in \{1, \ldots, m\}$ to be distinct. As the ϕ_j's are estimated via Bayesian inference along with all other parameters (see below), this enables the algorithm to adaptively determine the precision of each input structure, allowing us to analyze structures of varying experimental resolution and/or having a wide range of evolutionary divergence times or rates. As demonstrated in Results, this provides significant benefits over existing algorithms that implicitly weight each input structure equally. It also helps determine the core conserved residues by evaluating fluctuations relative to the variance in each structure. In addition, it aids in computation by preventing kinetic trapping of the MCMC chain in regions having only a subset of proteins aligned. Lastly, the Markov transition matrix Q assigns probabilities to transitions between the three types of states. Since the transitions (**Ins** $\rightarrow$ **Del**) and (**Del** $\rightarrow$ **Ins**) yield the same alignment, we constrain Q(**Del** $\rightarrow$ **Ins**) $= 0$ as commonly done (39; 30). Note that the transition probabilities are not currently site-dependent, but could be made so in future versions of the model.

Bayesian multiple structure alignment and MCMC sampling

In our probabilistic framework, multiple structure alignment amounts to simultaneously estimating the parameters of the probabilistic model (including "mean" structure U) and the alignments of each input structure to the model. We do so via Bayesian inference. Let $A = \{A_j\}_{j=1}^m$ where each A_j denotes an adjacency matrix specifying the alignment of protein j to the model, and $(R, v) = \{R_j, v_j\}_{j=1}^m$ the corresponding rotations and translations. and $\Phi = (A, R, v)$. We compute the posterior distribution:

$$\pi(\Theta, \Phi | X_1, \ldots, X_m) \propto p_0(n)\pi_0(\Theta, \Phi | n)$$
$$\textstyle\prod_{j=0}^{m-1} f(X_j | R_j, \vec{v}_j, A_j, n, \Theta) p(A_j | \Theta, n) \pi_0(R_j, \vec{v}_j)$$

where $f()$ is the likelihood function obtained from the Gaussian emission distributions. Here $p(A_j \mid \Theta, n)$ is the prior on alignments/correspondences/matchings implied by the Markov chain indel process of the HMM. To compute (1) we construct a Markov chain Monte Carlo (MCMC) algorithm (13) using Gibbs sampling and Metropolis-Hastings steps to sample the alignments, translations, rotations and the models from their joint posterior distribution. Here we emphasize non-standard steps in the sampling which require specialized solutions, especially the sampling of rotations and changes to model dimension.

A key distinction between sequence and structure alignment is the need for invariance under Euclidean transformations. Although an HMM that emits 3D coordinates is easily defined, alignment of 3D coordinate sequences cannot be done by straightforward application of sequence HMM techniques because each matching implies a distinct (distribution of) rotation and translation which depends on the matching globally. This requires joint evaluation of the likelihood (emission probability) simultaneously rather than locally, and this global dependence destroys the conditional independence structure required for efficient forward/backward algorithm in HMMs; the recursive marginalization of the forward step cannot be achieved.

Some structural alignment algorithms simultaneously optimize over matchings and rotation or translation transformations by iteratively maximizing the alignment give the superposition, and then the superposition given the alignment (reviewed in (11)). This suggests an iterative *sampling* scheme for probabilistic inference, a type of MCMC algorithm known as a Gibbs sampler (13), whereby having defined a joint distribution $\pi(\Theta, \Phi \mid X)$ over alignments and superpositions, one iteratively samples from the *conditional* distributions:

$$\pi(\Theta \mid \Phi, X), \quad \pi(A \mid \Theta, R, v, X), \quad \text{and} \quad \pi(R, v \mid \Theta, A, X)$$

Given Θ and (R, v), the alignments A_j are conditionally independent and can be sampled using standard stochastic dynamic programming recursions well-known in the sequence alignment HMM literature (9).

Although natural, such alternation of alignments and superpositions (Gibbs sampling) does not necessarily yield an efficient sampling method. Alignment/matching and superposition are strongly correlated, and the convergence rate the Gibbs sampler is determined by strength of that correlation. Thus an important consideration is the development of efficient sampling moves which can modify the two simultaneously.

Rotations A random-walk Metropolis proposal was constructed on the space $SO(3)$ of 3D rotations using a unit quaternion parametrization:

$$\vec{q} = [q_0, \vec{q}] = [\cos(\theta/2), \vec{v} \sin\theta/2]$$

where θ is an angle around unit vector $\vec{v} \in R_3$. New rotations are proposed by independently sampling a vector $\vec{v}'$ uniformly on the unit sphere $\mathbb{S}_2$, and a small angle of rotation around that vector $\theta' \sim \text{Gamma}(1, 40)$ to form a rotation $q' = [\cos(\theta'/2), \vec{v}' \sin\theta'/2]$. The proposed rotation is then obtained by composition of $\vec{q}'$ with the current rotation $\vec{q}$ via quaternion multiplication: $\vec{q}^* = \vec{q}' \cdot \vec{q}$. Note that this proposal yields a symmetric (geometric) random walk, as the inverse rotation $(\theta, \mathbf{v})^{-1} = (\theta, -\mathbf{v})$. This approach performs much better than a random walk on $\mathbf{q}$ itself; Supplementary Figure 3 shows that the Riemannian metric $d(R_1, R_2) = d(I, R_1^{-1} R_2) = \theta(\vec{q}_1 \cdot \vec{q}_2) = 2\arccos(\vec{q}_1 \cdot \vec{q}_2)$ is much more strongly correlated with log-likelihood than is Euclidean distance between quaternions.

Because a change in rotation affects all atoms and can dramatically increase the RMSD, the rotations are sampled jointly with alignments. Conditional on the proposed rotation R', a new alignment A'_j is drawn by dynamic programming, and the pair (R'_j, A'_j) are accepted or rejected together. This overcomes the strong dependency between R_j and A_j that is problematic for a Gibbs sampler updating $R_j \mid A_j$ and $A_j \mid R_j$. Since this proposal is symmetric, the joint acceptance probability is given by

$$\alpha\left((R_j, A_j), (R'_j, A'_j)\right) = \min(1, \frac{f(X_j | R'_j, \vec{v}_j, \Theta, n)}{f(X_j | R_j, \vec{v}_j, \Theta, n)})$$

Sampling of translations is achieved in an analogous manner, proposing $\vec{v}'_j \sim N(0, \sigma_v^2 I_3)$ (in practice $\sigma^2 = .1$ works well), then proposing a new alignment $A'_j \mid \vec{v}'_j$ and accepting or rejecting jointly.

The above moves involve only local perturbations to the rotation/translation. For strongly multimodal posteriors, we have previously developed a "library sampling" technique (30) which allows jumps between significantly different rotation/translation pairs. We did not find this necessary here, perhaps due to the additional mixing achieved by the transdimensional moves below. However for strongly multimodal examples (such as matching a single domain to a homo-dimer), this may still be necessary.

Transdimensional moves Because the number of "core" positions in a protein family is unknown *a priori*, the number of states n in the HMM cannot be fixed in advanced, and is subject to inference. The dimensions of the parameter vector Θ and alignments $\{A_j\}_{j=1}^m$ depend on n, and we use a reversible-jump step (15) to allow n to vary. We update $(n, \Theta, \{A_j\})$ jointly by first sampling a new (n', Θ'), and then new alignments $\{A'_j\}$ conditional on (n', Θ'). n is proposed to increase or decrease by 1 with equal probability:

- $n \to n+1$: Insert a new position of three states (**Mat**,**Del**,**Ins**)
- $n \to n-1$: Delete an existing position of three states

The proposed location $i \in (1, \ldots, n+1)$ for an inserted position is randomly sampled with probability proportional to $\lambda^{\min(m, k_i^I)}$ where k_i^I is the total number of C_α's in all proteins emitted from the ith **Ins** state and $\lambda > 1$ is a constant set at 1.2 to achieve a reasonable acceptance rate. This tends to propose new positions into the model in locations where there are many insertions. The mean of the proposed **Mat** state is sampled as follows:

- When the **Ins** state of the ith position has no emissions we set $\vec{\mu}^{\text{M}}_{i,\text{new}} = (\vec{\mu}^{\text{M}}_{i-1} + \vec{\mu}^{\text{M}}_i)/2 + \vec{z}$ for $\vec{z} \sim \mathcal{N}(\mathbf{0}, \sigma_\epsilon^2 I_3)$
- Otherwise we set $\vec{\mu}^{\text{M}}_{i,\text{new}} = \hat{\omega} + \vec{z}$ for $\vec{z} \sim \mathcal{N}(\mathbf{0}, \hat{\sigma}^2 I_3)$, where $\hat{\omega}$ is the sample mean and $\hat{\sigma}$ the sample s.d. of the coordinates emitted from the ith **Ins** state

In practice setting $\sigma_\epsilon = 6$ achieves reasonable acceptance rates. New alignments $\{A'_j\}$ are then sampled from their conditional distributions $P(A_j | \Theta', R_j, \vec{v}_j, X_j)$ as above. Deletions are proposed by sampling $i \in (1, \ldots n)$ with weight $\lambda^{-\min(m, k_i^M)}$ where k_i^M is the number emissions from the ith **Mat** state, and then sampling the A_j's from their conditional distributions. The Metropolis-Hastings acceptance ratio for these transdimensional moves is: $\alpha^{\text{Ins}} = \min(1, \gamma)$ for

$$\gamma = \frac{\pi(\Theta', A', R, v, \mid X) p(u'_1) \prod_{j=0}^{m-1} p(A_j \mid \Theta, R_j, \vec{v}_j, X_j)}{\pi(\Theta, A, R, v \mid X) p(u_1) p(u_2 | u_1) \prod_{j=0}^{m-1} p(A'_j | \Theta, R_j, \vec{v}_j, X_j)}$$

since $\left| \frac{\partial \Theta'}{\partial(\Theta, u_2)} \right| = 1$. Conversely, when deleting a layer the acceptance ratio is $\alpha^{\text{Del}} = \min(1, \gamma^{-1})$.

Prior distributions

Prior distributions for model parameters are taken as follows. Markov transition probability vectors Q(Mat $\to \cdot$) and Q(Ins $\to \cdot$) are given Dir(α, α, α) priors with $\alpha = 1$, and Q(Del $\to \cdot$) is given Dir$(\alpha, 0, \alpha)$ to enforce the alignment uniqueness constraint discussed previously. Prior distributions for means of the **Mat** and **Ins** state emission distributions were constructed from quartiles (q_{a1}, q_{a2}, q_{a3}) of the input structure C_α atom coordinates along each axis $a \in \{x, y, z\}$. Prior distributions for μ_{ia}^M's are independently

normal with mean q_{a2} (median) and variance $1.5(q_{a3} - q_{a1})$ (interquartile range). Priors for the precisions of these states' emission distributions are taken to be $Ga(.1, .01)$. The profile model length n is given uniform prior distribution over the range $[\lceil 0.5 n_{\min} \rceil, \lceil 1.5 n_{\max} \rceil]$, where $n_{\min}$ and $n_{\max}$ are the shortest and longest input protein lengths, respectively. Uniform prior distributions over rotation matrices (Haar measure) and translation vectors (improper) are assigned independently for each input structure, so $p(R_j, \vec{v}_j) \propto 1$ for $R_j \in SO(3)$.

3 METHODS

Let $C_j = (X_j - \vec{v_j}) R_j^{-1} = [\vec{c}_{j1}, \ldots, \vec{c}_{jn_j}]$ denote the coordinate matrix after inverting the Euclidean transformation. The likelihood is then

$$f(X_j | R_j, \vec{v}_j, A_j, n, \Theta) \propto \exp(-\frac{1}{2}\mathbf{s}^2)(\phi^{\mathrm{I}})^{\frac{1}{2}\eta_I}(\phi_j^{\mathrm{M}})^{\frac{1}{2}\eta_{jM}}$$

where

$$\eta_{jM} = \Sigma_l \Sigma_k \delta_{jkl}^{\mathrm{M}}$$

$$\eta_I = \Sigma_k \delta_{jk}^{\mathrm{I}}$$

$$s^2 = \Sigma_k \delta_{jk}^{\mathrm{I}} \phi^{\mathrm{I}} \|\vec{c}_{jk} - \vec{\mu}^{\mathrm{I}}\|^2 + \Sigma_k \Sigma_l \delta_{jkl}^{\mathrm{M}} \phi_j^{\mathrm{M}} \|\vec{c}_{jk} - \vec{\mu}_l^{\mathrm{M}}\|^2$$

Here $\delta_{jkl}^{\mathrm{M}}$ equals 1 if c_{jk} is emitted from the lth **Mat** state in alignment A_j, and zero otherwise, and δ_{jk}^{I} equals 1 if c_{jk} is emitted from any **Ins** state.

The full conditionals obtained from (1) for elements of Θ are given by:

$$\mu_l^{\mathrm{M}} | \cdot \sim \mathcal{N}\left((\frac{2}{3} q_{a2}(q_{a3} - q_{a1}) I^3 + \Sigma_j \phi_j^{\mathrm{M}} I^3 \Sigma_k \vec{c}_{jk} \delta_{jkl}^{\mathrm{M}}) / T_{\mathrm{M}_l}, T_{\mathrm{M}_l} \right)$$

$$\mu^{\mathrm{I}} | \cdot \sim \mathcal{N}\left((\frac{2}{3} q_{a2}(q_{a3} - q_{a1}) I^3 + \phi^{\mathrm{I}} I^3 \Sigma_j \Sigma_k \delta_{jk}^{\mathrm{I}} \vec{c}_{jk}) / T_{\mathrm{I}}, T_{\mathrm{I}} \right)$$

$$\phi_j^{\mathrm{M}} | \cdot \sim \Gamma\left(a' + \frac{1}{2} \Sigma_l \Sigma_k \delta_{jkl}^{\mathrm{M}}, b' + \frac{1}{2} \Sigma_l \Sigma_k \delta_{jkl}^{\mathrm{M}} (\vec{c}_{jk} - \vec{\mu}_l^{\mathrm{M}})^2 \right)$$

$$\phi^{\mathrm{I}} | \cdot \sim \Gamma\left(a' + \frac{1}{2} \Sigma_j \Sigma_k \delta_{jk}^{\mathrm{I}}, b' + \frac{1}{2} \Sigma_j \Sigma_k \delta_{jk}^{\mathrm{I}} (\vec{c}_{jk} - \vec{\mu}^{\mathrm{I}})^2 \right)$$

for match and insertion position means and precisions respectively, where

$$T_{\mathrm{M}_l} = \frac{2}{3}(q_{a3} - q_{a1}) I^3 + \Sigma_j \phi_j^{\mathrm{M}} I^3 \Sigma_k \delta_{jkl}^{\mathrm{M}}$$

$$T_{\mathrm{I}} = \frac{2}{3}(q_{a3} - q_{a1}) I^3 + \phi_{\mathrm{I}} I^3 \Sigma_j \Sigma_k \delta_{jk}^{\mathrm{I}}$$

Parameters of the transition matrix follow conditionally independent Dirichlet distributions:

$$\mathrm{Q}(\mathrm{Mat} \to \mathrm{Mat}, \mathrm{Del}, \mathrm{Ins}) \sim \mathrm{Dir}(\alpha + n_{\mathrm{MM}}, \alpha + n_{\mathrm{MD}}, \alpha + n_{\mathrm{MI}})$$

$$\mathrm{Q}(\mathrm{Del} \to \mathrm{Mat}, \mathrm{Del}) \sim \mathrm{Dir}(\alpha + n_{\mathrm{DM}}, 1 + n_{\mathrm{DD}})$$

$$\mathrm{Q}(\mathrm{Ins} \to \mathrm{Mat}, \mathrm{Del}, \mathrm{Ins}) \sim \mathrm{Dir}(\alpha + n_{\mathrm{IM}}, \alpha + n_{\mathrm{ID}}, \alpha + n_{\mathrm{II}})$$

and the conditional distributions of alignments are given by

$$A_j | \cdot \sim f(X_j | R_j, \vec{v}_j, A_j, \Theta, n) \mathrm{p}(A_j | \Theta, n)$$

Here n_{AB} is the number of transitions from states of type A to states of type B. Parameters are sampled directly from these conditional distributions; A_j's are sampled by dynamic programming. HMM size and rotations R_j are updated as described in above sections. Translation vectors $\vec{v}_j$ are updated similarly to rotations, except $\vec{v}'_j \sim \mathrm{Norm}(\vec{v}_j, \epsilon_0 I_3)$, where ϵ_0 is small (examples use $\epsilon_0 = 0.1 - 1$). The acceptance ratio is calculated as aforementioned.

For all examples mentioned in this paper, unless otherwise described we run at least three independent MCMC chains, each using a different input structure to initialize the profile HMM model and pre-aligning other input structures to the model with the pairwise structural alignment program FAST (43). We monitor convergence using the Gelman-Rubin diagnostic (7).

Fig. 1. Examples of protein structures that are previously difficult to align due to structural variability. (a) Comparison of the conserved residues in protein d1k1ga_ identified by our algorithm (HMM), MUSTANG, MATT, POSA and Manual alignment (AL00054790) from SISYPHUS database. X-axis: the protein sequences; Y-axis: the probability of a residue being conserved in all input proteins. (b) The protein structure d1k1ga_ (red) superposed by d1j5ka_ based on SISYPHUS manual alignment. Region I and II corresponding to the starting residues of "IRGKGS" and "GEDEPLH" in (a), respectively. (c) Visualization of the aligned GroES proteins from *E. coli* and *M. tuberculosis*, adopted from the Figure 4c in (28).(d) Our alignment of the six GroES proteins. Residues in lower case: insertions. Red: identified insertions/deletions (including the mobile loop) *M. tuberculosis*. Boxes: Regions poorly aligned by existing algorithms (28).

4 RESULTS

We demonstrate our algorithm on three problems. The first involves alignment of two protein structure families identified in the literature as difficult to align. We then consider applications to alignment of a large globin family, which illustrates certain advantageous features of our approach. Finally, we conclude with an analysis of axon growth receptors which demonstrates the ability to obtain new biological insights.

Alignment of difficult cases

We first test our algorithm on two example sets of protein structures that are difficult to automatically align due to structural variability. The first is a set of eight KH-domain type I structures taken from the SISYPHUS database (alignment: AL00054790), previously identified as a difficult case for multiple alignment algorithms (2). Figures 1a and 1b highlight two regions in one of the structures (scop id: d1k1ga_) with great uncertainties about their structural conservation based on alignments from MUSTANG, MATT and POSA (recently evaluated as the most accurate alignment algorithms currently available (4)). Also shown is the manual alignment from SISYPHUS.

Alignments from these algorithms give a binary assignment to each position, either conserved or not conserved, in some cases jumping back and forth in an evolutionarily implausible manner. In contrast, our algorithm computes a smooth posterior probability of inclusion, reflecting the uncertainty from multiple possible good alignments. This avoids the instability of arbitrary cut-offs and indicates to the user where the alignment is ambiguous. We see that these two regions are assigned intermediate values which vary smoothly along the sequence; for example our method assigns 40% probability match to 'G' where Mustang includes and Matt/POSA exclude. Moreover, those positions with no uncertainty (probability of conservation essentially equal to one) provide the only alignment that is identical to the manual alignment in both regions.

In the second example, we aligned six GroES proteins, five from *E. coli* and one from *M. tuberculosis*. The five *E. coli* GroES proteins have highly conserved structures (overall RMSD within 0.5Å) including their mobile loops, but GroEs from M. tuberculosis differs significantly in the mobile loop (Figure 1c). Pirovano *et al* (28) show that this structural variation of *M. tuberculosis* causes

Table 1. Hemoglobin subunits aligned in our analysis

species name	PDB id	α	β	Å	type
Aldabra Giant Tortoise	1wmu	deoxy	deoxy	1.65	D
Bar-headed goose	1c40	aquo-met	aquo-met	2.3	A
Bar-headed goose	1hv4	deoxy	deoxy	2.8	A
Bluefin tuna	1v4u	cmo	cmo	2	A
Bluefin tuna	1v4x	deoxy	deoxy	1.6	A
Bovine	1fsx	cmo	cmo	2.1	A
Bovine	1hda	deoxy	deoxy	2.2	A
Chicken	1hbr	deoxy	deoxy	2.3	D
Dusky rockcod	1la6	cmo	deoxy	2	A
Emerald rockcod	1hbh	deoxy	deoxy	2.2	A
Emerald rockcod	2h8d	deoxy	deoxy	1.78	A
Horse	1g0b	cmo	cmo	1.9	A
Horse	2dhb	deoxy	deoxy	2.8	A
Human	2dn3	cmo	cmo	1.25	A
Human	1ird	cmo	cmo	1.25	A
Human	2dn2	deoxy	deoxy	1.25	A
Human	1a3n	deoxy	deoxy	1.8	A
Rainbow trout	1ouu	cmo	cmo	2.5	I
Rainbow trout	1out	deoxy	deoxy	2.3	I
Red Stingray	1cg8	cmo	cmo	1.9	A
Red Stingray	1cg5	deoxy	deoxy	1.6	A
Spot	1spg	cmo	cmo	1.95	A
Hound shark	1gcw	cmo	cmo	2	A
Hound shark	1gcv	deoxy	deoxy	2	A
Yellow perch	1xq5	met	met	1.9	A

Fig. 2. Examples of analyzing Hgb α subunits. (a) the trees of Hemoglobin α subunits built from sequence alignment by ClustalX, and three structural alignments by SSM, Mustang and our algorithm, respectively. (b) the tree of Hemoglobin α subunits based on multiple structural alignment by our algorithm, after replacing the human CMO-bound hemoglobin α subunit with a lower resolution structure. In (a) and (b), fonts in red: liganded; black: unliganded; ellipses in blue: bony fishes; green: cartilagineous fishes; yellow: reptiles and birds; brown: mammals. '*' Hgb D; '+' Hgb I; otherwise Hgb A. (c) a snapshot of the actual alignment of the structures.

all of several popular algorithms considered (DALI, CE and even flexible alignment algorithms, MATT and FATCAT) to fail to align the 5 *E. coli* structures in this region, as well as in two other regions (boxes in Fig. 1d)) containing an insertion and a deletion, respectively. In sharp contrast, our alignment is shown in Fig. 1d. We use upper and lower-case letters to denote residues from **Mat** or **Ins** states for a sampled alignment, and color red positions with marginal posterior probability 95% of being insertions. It is seen that our algorithm successfully aligns all five *E. coli* structures perfectly, identifying the mobile loop as one large insertion in *M. tuberculosis*, and also identifying the single insertion and single deletion.

Hemoglobin isoform evolution

For the second test, we aligned 25 structures of vertebrate hemoglobin α subunit (Table 1) from the SCOP database (26). The phylogenetic tree constructed from the sequences of these proteins via NJ by ClustalX (Fig. 2a) reveals three major monophylies: mammals/reptile/birds, cartilaginous fishes (sharks/rays), and bony fishes, which agrees with the NCBI Taxonomy (http://www.ncbi.nlm.nih.gov/Taxonomy/) species tree and the highly conserved function in vertebrates. As emphasized above, our algorithm produces not just a single multiple structure alignment, but a posterior distribution over all alignments, allowing it to account for uncertainty. Here we also take advantage of this to construct evolutionary trees which account for alignment uncertainty, as follows. For each alignment sampled from the posterior, we calculated pairwise RMSD of matched residues for every pair of input structures, and used them construct a Neighbor-Joining (NJ) tree (31). We collapsed the resulting sample of trees into equivalence classes of zero nodal distance (5). This calculation gives the Bayesian marginal posterior distribution over (equivalence classes of) tree, integrating over all possible alignments according to their respective posterior probabilities.

In comparison to the sequence-based phylogenetic tree produced by ClustalX, our algorithm identifies two NJ trees which account for $> 98\%$ of the posterior probability (78% (Figure 2a) and 20%, respectively). They are 95.8% similar in topology (27), differing only in the placement of chicken and tortoise hemoglobins: the MAP (78%) tree places these with the two goose proteins, while the less probable (20%) tree places them with the shark proteins.

We also obtained NJ trees using alignments generated by MUSTANG and another deterministic alignment algorithm SSM (19) (Figure 2a). Of all three structure-based trees, only ours identifies the same three major monophylies as the sequence tree in the presence of allosteric conformational changes. The SSM tree fails to identify the boundary between fishes and non-fishes and proposes a mixed monophyly consisting of cartilaginous fishes, reptiles/birds and mammals. The MUSTANG tree has the same topology as ours for unliganded proteins; however, it clusters the liganded subunits of horse and cow with the goose subunits. The alignments obtained by the three methods are of similar length and quality, as measured by mean pairwise RMSD (number residues aligned): 1.02 (136) for SSM, 1.08 (140.6) for MUSTANG, and 0.918 ± 0.004 (136.8 ± 0.2) for our algorithm (posterior means). Thus the improvement comes not from finding a better alignment, but in averaging over alignment uncertainty to construct a more stable and accurate tree.

To demonstrate utility of our algorithm for constructing trees robust to input noise, we repeated the analysis with one of the human CMO-bound structures (2dn3, 1.25 Å) replaced by a lower resolution predecessor in the PDB (2hco, 2.7 Å). This change effected a 4.7% topological change in trees obtained from Mustang alignments (Fig 2b), placing 2hco in a separate branch from another human cmo-Hgb 1ird, and also grouping the D isoforms (1wmu,1hbr) with cartilagineous fish. In contrast, our algorithm retains the sistership of the two human cmo-Hgbs (1ird and 2hco)with posterior probability $> 80\%$, although their estimated divergence (0.4) is five times as large as that of 1ird-2dn3 (0.08). The estimated ϕ for the 2hco structure is 0.043 ± 0.005, similar with that of 2dn3: 0.044 ± 0.004. This is explained by Supplementary Figure 1 which shows that the low- and high-resolution structures differ significantly in only a few positions, which are highly probable to be insertions under our model. The two most probable tree topologies (and the only ones with posterior probability $> 20\%$) remain unchanged, simply altering their relative probabilities to 26% and 39%, respectively.

Fig. 3. Structural analysis of FN3 domains. (a) sequence-based tree built from ClustalX. Structure-based trees built from Mustang (b) and our algorithm (c), using shorter structures. Structure-based tree built from Mustang (d) and our algorithm (e), using longer structures. (f) analysis of structural variations using shorter structures. Squared in red: unique structural changes in 3rd FN3 domain of human neogenin. Squared in blue: putative netrin-1 binding regions for both neogenin and DCC. Underlined residues: predicted β strands (D, F or G).

Reliable structure-based phylogenetic trees make possible comparisons with sequence-based trees for additional insight about functional evolution. In the first monophyly (birds and mammals) of the sequence tree, the D isoform (chicken and tortoise) appears to have diverged first, followed by the divergence between avian and mammalian A isoforms. In the structure-based tree however, while the geese A and D isoforms form distinct monophylies, the mammal vs bird/reptile split is more prominent. This suggests that both A and D isoforms in birds and reptiles have adopted similar conformations distinctive from mammalian A isoforms, possibly due to the similar hypoxic environments where these species live. We also built structure-based trees for Hgb β subunits (Supplementary Figure 2); although both α and β trees contain three major monophylies, liganded subunits form monophylies in the β tree form compared with paraphylies in the α tree. This indicates two distinctive β subunit conformations in different allosteric states.

Neuronal axon growth guidance receptors

Finally, we applied our algorithm to analyze two neuronal guidance receptors DCC and neogenin, vertebrate homologues the invertebrate genes UNC-40 in *C. elegans* and frazzled in Drosophila. These membrane proteins are expressed at the growing tip of neuronal axons during nervous system development, and their extracellular domains interact with external cues via to 'sense' a path for the growing axon to reach its target. DCC and neogenin are paralogous and have similarly organized extracellular domains, with four immunoglobulin(Ig)-like loops followed by six fibronectin type-III (FN3) repeats, with the FN3 domains known to interact with guidance cues (12). Both DCC and neogenin bind the guidance cue netrin-1, at the FG loop of the 5th FN3 domain in DCC and putatively the 4th and/or 5th FN3 domain in neogenin (12). However only neogenin binds another cue RGMa, also mediated via its FN3 domains but independent of netrin-1 (29). This is a typical example of sub/neofunctionaliztion, with two copies having undergone a modest functional divergence at some domain after gene duplication. In many cases, such divergence cannot be detected by comparative sequence analysis: as can be seen from their sequence-based ClustalX phylogeny (Figure 3a), all six pairs of FN3 domains exhibit a high homology to each other.

To gain insight into their differing binding behaviors, we generated the MAP structure-based NJ tree and compared with the MUSTANG results (Figure 3b,c). For computational speed, the N- and C-terminal residues outside the β-sheet structure are removed. Our algorithm identified 85.8 pairwise-aligned residues with a mean RMSD of 1.46Å. In all sampled posterior trees, we found the 1st and 4th-6th FN3 domains highly conserved (96% for 1st and 100% for others). However unlike in the sequence tree, the 2nd and 3rd FN3 domains in DCC form a monophylic paralogous pair with posterior probability 56.8%, and similar to the 4th FN3 domains of both DCC and neogenin (prob. 90.4%). And this four-member clade is further outgrouped by the 2nd FN3 domain of neogenin (prob. 84.8%), while the 3rd FN3 domain of neogenin closer to the 6th FN3 domains (prob. 90.4%). In contrast, the MUSTANG alignment yielded 94.3 pairwise-aligned residues with mean RMSD of 2.92, twice as high. The corresponding NJ tree identifies homologous pairs of 2nd, 4th, 5th and 6th FN3 domains and also suggests homology between neogenin 3rd FN3 and DCC 1st FN3 domains. This would require not only the 1st and 3rd FN3 domains to both have diverged, but also a second convergent evolutionary event.

To test the reliability of these trees we increased the N- and C-termini of all input structures by eight more residues and repeated the above analysis (Figure 3d,e). This should not significantly change the core structure; however, we see a significant topological change in the MUSTANG tree, with all homologous domains paired as in the sequence tree. In contrast, our MAP tree remains nearly unchanged with only a small positional change of the 3rd FN3 domain of neogenin around the basal position of the tree, suggesting the general topology is reliable and robust. Note that the added residues significantly increase the number of iterations needed for convergence (about 3×) according to our previously defined strict convergence criteria; although looking only at the MAP tree topologies convergence occurs in about the same time as before.

Figure 3f examines the alignment in detail. Here we color in red positions whose posterior mean squared deviation (standardized by the protein specific ϕ) from the model has a p-value ¡0.05 according to a χ^2_3 null distribution. These are residues identified as part of the structural core exhibiting significant structural fluctuations. Relevant to the netrin-1 binding ability of both proteins, our structural alignment identifies notable structural changes (in the boxes of Figure 3f) in the FG loops of domain 5 of both neogenin and DCC, but little change in other domains. In addition, two discontinuous but nearby regions are pinpointed as putatively responsible: the K(N/G)RR region and sites near the β-strand G. These findings are remarkably consistent with the results of previous experiments (12) that suggest the "KNRR" residues in the FG loop of the 5th domains contribute to, but are not wholly responsible for, binding netrin-1. We suggest that sites near the β-strand G are plausible candidates for the unidentified netrin-1 binding sites in this loop. Our alignment also suggests a unique structural feature (in the box of Figure 3f) of the 3rd FN3 domain of neogenin: its D β-strand is shifted by 4-5 residues towards the N-terminus by a single insertion relative to all other structures. This makes it a highly probable candidate binding region for the unique binding affinity of neogenin to RGMa that is independent of netrin-1 binding.

5 CONCLUSION

The algorithm described gives a fully probabilistic approach to multiple protein structure alignment and an explicit statistical model of variability in protein families. The Bayesian approach avoids sensitivity to alignment parameters by treating the associated statistical estimation problem, effectively providing adaptive learning of parameters. As described in (30) for the case of pairwise alignment, this approach also generalizes many existing algorithms for structural alignment, which can be seen to be MAP-type

alignments under various choices of prior and noise distributions. In addition, our results indicate that averaging over alignment uncertainty makes inference of phylogenetic trees from structural data significantly more robust.

Our algorithm may be extended to treat the phylogeny topology as an additional parameter and sample it according to its conditional posterior distribution. This would allow recent developments on simultaneous sequence alignment/phylogeny reconstruction (40; 25) to incorporate structural data, which is conserved over much longer evolutionary time scales than sequence. However, the current implementation of the algorithm takes considerable computing power: approximately 3 seconds per iteration on $m + 1$ parallel nodes (3.0G Hz, ¡512M RAM) to align m structures of length 100, and can be slower for lengths > 300 or with a high percentage ($> 30\%$) of flexible regions.

ACKNOWLEDGEMENT

Funding: This work was partially supported by NSF grant DMS-0605141 (SCS) and NIH grant 1R01GM090201-01 (SCS).

REFERENCES

[1]V. Alexandrov and M. Gerstein. Using 3D hidden Markov models that explicitly represent spatial coordinates to model and compare protein structures. *BMC Bioinformatics*, 5, 2004.

[2]A. Andreeva, A. Prlic, T. J. P. Hubbard, and A. G. Murzin. SISYPHUS - structural alignments for proteins with non-trivial relationships. *Nucleic Acids Research*, 35:D253–D259, 2007.

[3]P. Baldi, Y. Chauvin, T. Hunkapiller, and M. A. McClure. Hidden Markov-models of biological primary sequence information. *Proceedings of the National Academy of Sciences U.S.A*, 91(3):1059–1063, 1994.

[4]C. Berbalk, C. S. Schwaiger, and P. Lackner. Accuracy analysis of multiple structure alignments. *Protein Science*, 18(10):2027–2035, 2009.

[5]John Bluis and Dong-Guk Shin. Nodal distance algorithm: Calculating a phylogenetic tree comparison metric. In *Proc. 3rd IEEE Symposium on Bioinformatics and Bioengineering*, pp. 87–94, 2003.

[6]P. V. Burra, Y. Zhang, A. Godzik, and B. Stec. Global distribution of conformational states derived from redundant models in the PDB points to non-uniqueness of the protein structure. *Proceedings of the National Academy of Sciences U.S.A*, 106(26):10505–10510, 2009.

[7]John M. Castelloe and Dale L. Zimmerman. Convergence assessment for reversible jump MCMC samplers. Technical Report 313, University of Iowa, Dept. of Statistics and Actuarial Science., 2002.

[8]Mardia K.V. Dryden I.L., editor. *Statistical Shape Analysis*. Wiley, 1998.

[9]Richard Durbin, Sean Eddy, Anders Krogh, and G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998.

[10]Sean R. Eddy and Richard Durbin. RNA sequence analysis using covariance models. *Nucleic Acids Research*, 22(11):2079–2088, 1994.

[11]I. Eidhammer, I. Jonassen, and W. R. Taylor. Structure comparison and structure patterns. *Journal of Computational Biology*, 7(5):685–716, 2000.

[12]B. V. Geisbrecht, K. A. Dowd, R. W. Barfield, P. A. Longo, and D. J. Leahy. Netrin binds discrete subdomains of DCC and UNC5 and mediates interactions between DCC and heparin. *Journal of Biological Chemistry*, 278(35):32561–32568, 2003.

[13]Spiegelhalter D.J. Gilks W.R., Richardson S., editor. *Markov Chain Monte Carlo in Practice*. Chapman and Hall, 1996.

[14]A. Godzik. The structural alignment between two proteins: Is there a unique answer? *Protein Science*, 5(7):1325–1338, 1996.

[15]P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

[16]Peter J. Green and Kanti V. Mardia. Bayesian alignment using hierarchical models, with applications in protein bioinformatics. *Biometrika*, 93(2):235–254, 2006.

[17]H. Hasegawa and L. Holm. Advances and pitfalls of protein structural alignment. *Current Opinion in Structural Biology*, 19(3):341–348, 2009. Hasegawa, Hitomi Holm, Liisa.

[18]R. Kolodny, D. Petrey, and B. Honig. Protein structure comparison: implications for the nature of 'fold space', and structure and function prediction. *Current Opinion in Structural Biology*, 16(3):393–398, 2006.

[19]E. Krissinel and K. Henrick. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallographica Section D-Biological Crystallography*, 60:2256–2268, 2004.

[20]A. Krogh, M. Brown, I. S. Mian, K. Sjolander, and D. Haussler. Hidden Markov-models in computational biology - applications to protein modeling. *Journal of Molecular Biology*, 235(5):1501–1531, 1994.

[21]Richard Lathrop, Robert Rogers, Temple Smith, and James White. A Bayes-optimal sequence-structure theory that unifies protein sequence-structure recognition and alignment. *Bulletin of Mathematical Biology*, 60(6):1039–1071, 1998.

[22]CE Lawrence, SF Altschul, MS Boguski, JS Liu, AF Neuwald, and JC Wootton. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, 262(5131):208–214, 1993.

[23]Charles E. Lawrence and Andrew A. Reilly. An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins: Structure, Function, and Bioinformatics*, 7(1):41–51, 1990.

[24]Michael Levitt and Mark Gerstein. A unified statistical framework for sequence comparison and structure comparison. *Proceedings of the National Academy of Sciences, U.S.A*, 95(11):5913–5920, 1998.

[25]G. Lunter, A. Rocco, N. Mimouni, A. Heger, A. Caldeira, and J. Hein. Uncertainty in homology inferences: Assessing and improving genomic sequence alignment. *Genome Research*, 18(2):298–309, 2008.

[26]Alexey G. Murzin, Steven E. Brenner, Tim Hubbard, and Cyrus Chothia. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247(4):536–540, 1995.

[27]T. M. W. Nye, P. Lio, and W. R. Gilks. A novel algorithm and web-based tool for comparing two alternative phylogenetic trees. *Bioinformatics*, 22(1):117–119, 2006.

[28]W. Pirovano, K. A. Feenstra, and J. Heringa. The meaning of alignment: lessons from structural diversity. *BMC Bioinformatics*, 9, 2008.

[29]S. Rajagopalan, L. Deitinghoff, D. Davis, S. Conrad, T. Skutella, A. Chedotal, B. K. Mueller, and S. M. Strittmatter. Neogenin mediates the action of repulsive guidance molecule. *Nature Cell Biology*, 6(8):756–762, 2004.

[30]A. Rodriguez and Scott C. Schmidler. Bayesian protein structure alignment. revised for *Annal of Applied Statistics*, 2011.

[31]N. Saitou and M. Nei. The neighbor-joining method - a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425, 1987.

[32]Yasubumi Sakakibara, Michael Brown, Richard Hughey, I. Saira Mian, Kimmen Sj?lander, Rebecca C. Underwood, and David Haussler. Stochastic context-free grammers for tRNA modeling. *Nucleic Acids Research*, 22(23):5112–5120, 1994.

[33]S. C. Schmidler, J. S. Liu, and D. L. Brutlag. Bayesian segmentation of protein secondary structure. *Journal of Computational Biology*, 7(1-2):233–248, 2000.

[34]Scott C. Schmidler. Fast Bayesian shape matching using geometric algorithms (with discussion). In José M. Bernardo, Susie Bayarri, Jim O. Berger, A. P. Dawid, David Heckerman, Adrian F. M. Smith, and Mike West, editors, *Bayesian Statistics 8*, pp. 471–490, Oxford, 2006. Oxford University Press.

[35]Scott C. Schmidler. Bayesian flexible shape matching with applications to structural bioinformatics. submitted to *Journal of the American Statistical Association*, 2011.

[36]Scott C. Schmidler, Jun S. Liu, and Douglas L. Brutlag. Stochastic segment interaction models for biological sequence analysis. revised for*Journal of the American Statistical Association*, 2004.

[37]Kim T. Simons, Charles Kooperberg, Enoch Huang, and David Baker. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *Journal of Molecular Biology*, 268(1):209–225, 1997.

[38]L. Wang and T. Jiang. On the complexity of multiple sequence alignment. *Journal of Computational Biology*, 1(4):12, 1994.

[39]B. J. M. Webb, J. S. Liu, and C. E. Lawrence. BALSA: Bayesian algorithm for local sequence alignment. *Nucleic Acids Research*, 30(5):1268–1277, 2002.

[40]K. M. Wong, M. A. Suchard, and J. P. Huelsenbeck. Alignment uncertainty and genomic analysis. *Science*, 319(5862):473–476, 2008.

[41]Boomsma Wouter, Kanti V. Mardia, Charles C. Taylor, Jesper Ferkinghoff-Borg, Anders Krogh, and Thomas Hamelryck. A generative, probabilistic model of local protein structure. *Proceedings of the National Academy of Sciences, U.S.A*, 105(26):8932–37, 2008.

[42]J. Zhu, J. S. Liu, and C. E. Lawrence. Bayesian adaptive sequence alignment algorithms. *Bioinformatics*, 14(1):25–39, 1998.

[43]J. H. Zhu and Z. P. Weng. FAST: A novel protein structure alignment algorithm. *Proteins-Structure Function and Bioinformatics*, 58(3):618–627, 2005.