Statistical alignment of multiple protein structures under a dynamics-based model of structural evolution

J. L. Herman*, R. Lyngsø and J. Hein

Department of Statistics, University of Oxford

1 Introduction

Successful models for the evolution of protein sequences have been in existence since the 1960s, and are widely used in a large range of applications, ranging from inference of phylogeny to identification of binding site motifs. The importance of sequence arises largely from its determination of structure, and this structure may be conserved despite major changes in sequence. Indeed, for some families the sequence identity may fall as low as 10% while still preserving the overall fold to a remarkable degree, such that structure comparison may enable the detection of distant homology that cannot be recognised from sequence alone. However, despite decades of research seeking to bridge the gap between sequence and structure, current methods for studying structural evolution are much less well developed, and the origins of the empirical relationships between sequence and structural divergence first characterised by Chothia and Lesk (1986) remain to a large extent a mystery.

Central to the comparative study of protein structures is the requirement of a measure of similarity in structure space. Most widely used distance metrics treat protein structures as static entities, despite the fact that studies carried out in biological conditions suggest that most proteins exhibit flexibility of some kind. Indeed, many protein functions are mediated through changes in conformation of the structure, and there is increasing evidence to suggest that these conformational changes can in fact be regarded as redistributions of the populations of thermally sampled states (Kern and Zuiderweg, 2003).

Following observations that the space sampled by a set of homologous structures is strikingly similar to the subspace spanned by the low frequency dynamics of the individual proteins (e.g. Leo-Macias *et al.*, 2005), Echave introduced a linear response model—termed the *linearly forced elastic network model*, or LFENM—to investigate the expected pattern of structural divergence upon mutation. The model stipulates that mutational perturbations are dissipated along energetically favourable directions in a similar fashion to the effects of ligand binding events (Echave and Fernández, 2010). This is an example of the more general *fluctuation-dissipation* theorem relating relaxation from non-equilibrium states to thermally accessible fluctuations at equilibrium.

Implicit in the extension of this theorem to evolutionary perturbations is the assumption that the folding pathway is not significantly altered by mutation events, since a change of folding pathway could lead the structure from one native state to another without the two being interconvertible in the folded state. This is often assumed in the analysis of mutagenesis experiments such as double mutant cycles, and is reasonable provided that the structures are similar.

Under a model of dynamically mediated structural evolution such as the LFENM, assuming at this stage a star phylogeny between the organisms, we can model a set of homologous structures $\mathcal{X} = \left\{X^{(1)}, \ldots, X^{(K)}\right\}$ as if each member of the set were sampled from a common distribution, such that the likelihood of an alignment can be written as a product of the form

$$p(\mathcal{X}|\mathcal{A}, \theta) = \prod_k f(A^{(k)}X^{(k)}|\theta) \tag{1}$$

where $f(\,\cdot\,)$ describes the common equilibrium distribution of the structures represented by the $L^{(k)} \times 3$ matrices $X^{(k)}$, and $A^{(k)} \in \mathcal{A}$ is an $L \times L^{(k)}$ binary matrix that maps each residue (row) in the kth configuration to the corresponding index in the L-dimensional joint model, with $L = \sum_k L^{(k)}$. We also impose the restriction that no more than one residue from each structure can map to a given index (i.e. each row of $A^{(k)}$ contains at most one non-zero entry). The set of maps $\mathcal{A}$ defines an *alignment*. Ultimately one might wish to use a physically realistic atomistic potential as the function f, but given the approximations involved in the model, such detail would almost certainly be inappropriate, and we may choose instead to examine the utility of more coarse-grain energy functions.

2 Structural alignment

2.1 Rossmann method

First introduced by Rao and Rossmann (1973), and developed further during the 1970s, this approach remains at the core of many commonly used methods of structural alignment, consisting of iterative alternating least squares superposition and alignment based upon the coordinate RMSD (which is closely related to the Procrustes distance):

$$\text{cRMSD}(k,l) = ||A^{(k)}X^{(k)} - A^{(l)}X^{(l)}\hat{R} + \hat{T}||_F \;/\; \sqrt{\text{tr}(A^{(k)T}A^{(l)})} \tag{2}$$

where $\hat{T}$ is a translation to bring the centres of mass of the two proteins into superposition, and $\hat{R}$ is the 3D rotation matrix that minimises the cRMSD. The quantity $\text{tr}(A^{(k)T}A^{(l)})$ in the normalising factor corresponds to the number of aligned residues. Since the cRMSD^2 is linearly additive over the matched residues in the alignment, then conditional on $\hat{T}$ and $\hat{R}$, and given a gap penalty parameter, λ, the optimal alignment can be found by dynamic programming using algorithms such as Needleman-Wunsch, minimising a score of the form cRMSD$^2 + \lambda(L - \text{tr}(A^{(k)T}A^{(l)}))$. The basic Rossmann algorithm for pairwise alignment proceeds by iterating between superposition and alignment steps until convergence is achieved.

A variant of the Rossmann method can be seen to be equivalent to an EM scheme (Kent *et al.*, 2010), and this method (or variations thereof) is implemented in a number of well-known structural alignment programs, such as STRUCTAL, LSQMAN, SSM, and MAMMOTH. The superposition step is also known as *Procrustes registration*, and can be generalised to multiple structures, as implemented in MUSTANG and MAMMOTH Mult.

2.2 Extensions to Bayesian framework

One of the problems with many of the aforementioned methods is that the scoring schemes are not based on probabilistic models, producing a point estimate with little information of the uncertainty surrounding this estimate. A common approach is to fit the scores to an extreme value distribution, such that empirical p-values can be calculated, giving a rough measure of significance, although there are a number of issues associated with such approaches. Another shortcoming of most existing methods is that the choice of parameters such as gap opening and gap extension penalties may have a large effect on the output, and is usually left to the user, who may have no good intuition as to how to specify these parameters.

Over the last twenty years, probabilistic evolutionary models have been successfully developed for protein sequences to address such problems. More recently, several methods (e.g. Green and Mardia, 2006; Rodriguez and Schmidler, *submitted*) have been developed for addressing the structural alignment problem within a statistical framework. These methods essentially use the Rossmann method, albeit within a Bayesian probabilistic model. Since the likelihood in such models involves maximisation or integration over rotations, it is invariant to

rotations of the data. In the case of Green and Mardia (2006), the integration is carried out through an MCMC procedure.

However, the above methods generally assume that the location of each residue is distributed with independent, isotropic variance. In the unweighted Procrustes schemes mentioned in the previous section, a high variance at a particular region of the protein will therefore be spread out over the whole structure by the least-squares rotation, giving a misleading picture of any covariance structure within the two conformations (Walker, 2000). Similarly, in the Bayesian procedure, the posterior induced on rotations is affected equally by the deviation at each residue.

Since the covariance structure is critically important to any model of dynamically constrained evolution, it is essential that we allow for interresidue correlations in any alignment scheme. Various methods have been proposed to account for heteroskedasticity and correlated dynamics when superposing protein structures, for example using weighted Procrustes, either with a theoretical estimate of the covariance matrix as the weights, or estimating the covariance and superposition jointly from the data (Theobald and Wuttke, 2008).

However, it can be shown (Kent and Mardia, 1997) that only if the covariance matrix is equal to the identity matrix (a highly unrealistic scenario for protein molecules) will superposition-based methods yield a consistent estimator for the mean structure; such methods will, therefore, in general lead to biased estimators of the covariance. Indeed, any method that relies on superposition will encounter significant problems in situations where it is impossible to simultaneously line up all the corresponding regions, for example in the case of an open and closed conformation of a protein such as calmodulin that undergoes a major hinge motion. This problem can be addressed by splitting the protein up into domains or structural units and superposing these separately, inferring breakpoints between domains either manually (e.g. Lesk and Chothia, 1980) or through automated methods (Mechelke and Habeck, 2010; Schmidler, *submitted*). However, such methods inevitably lose information about the *inter*domain correlation structure, and are highly sensitive to the choice of what constitutes a superposable unit.

2.3 Distance-based scoring functions

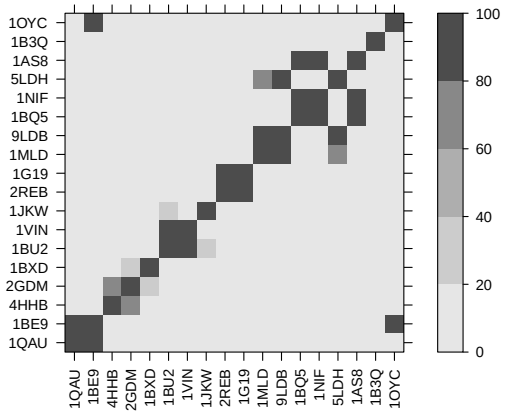
While distance-based methods have the advantage that there is no need for superposition, there is the complication that distance-based scoring functions are generally non-local, and therefore not linearly additive over the residues. This makes optimisation of such scores difficult, with all existing implementations using a heuristic at some point during the procedure. Although the class of distance-based methods includes some of the most successful structural alignment algorithms, including programmes such as SSAP (Taylor and Orengo, 1989), DALI (Holm and Sander, 1993), and CE (Shindyalov and Bourne, 1998), the heuristics involved make the significance of resulting alignments difficult to determine, and convergence properties are often far from obvious.

2.4 Reflection size-and-shape

As discussed by Goodall and Mardia (1993), and more recently by Dryden *et al.* (2008), the size-and-shape (up to reflection) of a configuration can be represented by the centred Gram matrix $G(X)$, or a projection onto the column space thereof. Goodall and Mardia (1993) focus on the case in which X is distributed about a mean configuration M, with diagonal covariance, allowing the rotations to be integrated out analytically, yielding a hypergeometric distribution for G. However, this marginal distribution is difficult to work with directly, due to the presence of the hypergeometric function. Nevertheless, as argued by Prompers and Brüschweiler (2002), for any three-dimensional object free to rotate in space, the population mean of the coordinate matrix X (without superposition) will in fact be zero, such that we need only consider the central case, which is simply a Wishart distribution on G. It should be noted that, in the case

of proteins, since amino acids are found in only one enantiomer, and helices are therefore all right-handed, reflection invariance is unlikely to lead to erroneous ascription of homology.

3 Results



We have recently developed a model of this type for multiple structures taken from a joint distribution of the form in equation (1), with f taken to be a zero-mean Gaussian with unknown covariance. Integrating out this covariance, we obtain a marginal posterior for the alignment. If we consider the alignment as specifying a particular model, then we can regard the ratios between posteriors as Bayes factors, and the alignment problem can be thought of as model selection.

Simulating from this model using MCMC, we can generate samples of alignments for any set of structures. Using a recently developed method for maximum posterior decoding (MPD) in the space of alignment columns (Herman *et al.*, *submitted*), it is then possible to generate a single representative alignment that minimises one of a family of loss functions similar to that proposed by Green and Mardia (2006).

As an example here we show results for a set of eighteen structures, comprising seven known homologous pairs or triplets and one outlier, taken from a variety of structural classes, determined by either NMR or crystallography. In the case of NMR we work with an ensemble of models, which provides additional information regarding the uncertainty in the structures. Aligning all against all and computing Bayes factors for the MPD alignment with respect to the unaligned model, we obtain a measure of homology between each of the structures (shown in the figure above). Aside from one false positive (1OYC) and one pair of false negatives (1BXD, 1B3Q), the method successfully clusters the data into the sets of known homologues.

References

Chothia, C. and Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *The EMBO Journal*, **5**(4):823–826.

Dryden, I. L., Kume, A., Le, H. and Wood, A. T. A. (2008). A multi-dimensional scaling approach to shape analysis. *Biometrika*, **95**(4):779–798.

Echave, J. and Fernández, F. M., (2010). A perturbative view of protein structural variation. *Proteins: Structure, Function, and Bioinformatics*, **78**(1):173–180.

Goodall, C. R. and Mardia, K. V. (1993). Multivariate Aspects of Shape Theory. *Annals of Statistics*, **21**(2):848–866.

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

Herman, J. L., Novák, Á., Lyngsø, R., Miklós, I. and Hein, J. (*submitted*). Efficient posterior decoding for statistical multiple sequence alignments.

Holm, L. and Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *Journal of Molecular Biology*, **233**(1):123–138.

Kent, J. T. and Mardia, K. V. (1997). Consistency of Procrustes estimators. *Journal of the Royal Statistical Society, Series B*, **59**(1):281–290.

Kent, J. T., Mardia, K. V. and Taylor, C. C. (2010). An EM interpretation of the Softassign algorithm for alignment problems. In A. Gusnanto, K. V. Mardia, C. J. Fallaize & J. Voss (eds), *High-throughput Sequencing, Proteins and Statistics*, pp.29–32. Leeds.

Kern, D. and Zuiderweg, E. R., (2003). The role of dynamics in allosteric regulation. *Current Opinion in Structural Biology*, **13**, 748–757.

Leo-Macias, A., Lopez-Romero, P., Lupyan, D., Zerbino, D. and Ortiz, A. R. (2005). An analysis of core deformations in protein superfamilies. *Biophys. J.*, **88**(2):1291–1299.

Lesk, A. M. and Chothia, C. (1980), How different amino acid sequences determine similar protein structures: The structure and evolutionary dynamics of the globins. *Journal of Molecular Biology*, **136**(3):225–230.

Mechelke, M. and Habeck, M. (2010). Robust probabilistic superposition and comparison of protein structures. *BMC Bioinformatics*, **11**:363.

Prompers, J. J. and Brüschweiler, R. (2002). Dynamic and structural analysis of isotropically distributed molecular ensembles. *Proteins: Structure, Function, and Bioinformatics*, **46**(2):177–189.

Rao, S. T. and Rossmann, M. G. (1973). Comparison of super-secondary structures in proteins. *Journal of Molecular Biology*, **76**(2):241–250.

Rodriguez, A. and Schmidler S. C. (*submitted*). Bayesian protein structure alignment http://www.isds.duke.edu/~scs/Papers/BayesStructAlignAAS.pdf

Schmidler, S. C. (*submitted*). Bayesian flexible shape matching with applications to structural bioinformatics. http://www.isds.duke.edu/~scs/Papers/BayesFlexShapeJASA.pdf

Shindyalov, I. N. and Bourne, P. E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.*, **11**(9):739–747.

Taylor, W. R. and Orengo, C. A. (1989). Protein structure alignment. *Journal of Molecular Biology*, **208**(1):1–22.

Theobald, D. L. and Wuttke, D. S. (2008). Accurate structural correlations from maximum likelihood superpositions. *PLoS Computational Biology*, **4**(2):e43

Walker, J. A. (2000). Ability of geometric morphometric methods to estimate a known covariance matrix. *Systematic Biology*, **49**(4), pp. 686–696.