

Shape Analysis with Applications in Bioinformatics

Chris Fallaize

November 2019



The University of
Nottingham

UNITED KINGDOM • CHINA • MALAYSIA

1 Statistical Shape Analysis

- Representation and Alignment
- Unlabelled Shape Analysis
- Bayesian Alignment

2 Multiple Alignment

- Motivation — Molecular Alignment
- Hierarchical Template Model
- Application

3 Sequence-Structure Alignment

- Motivation — Protein Alignment
- Sequence-Structure Model
- Modelling Evolutionary Distance

Registration:landmark-based objects

- Identify L landmarks, with coordinates $x_j \in \mathbb{R}^d$, $j = 1, \dots, L$.
- Represent as $(L \times d)$ configuration of points, with rows x_j^T giving coordinates of landmark j .
- Seek optimal alignment with another $(L \times d)$ configuration Y .
- *Labelled* case, where x_j is known to match y_j , $j = 1, \dots, L$.
- Minimise objective function

$$f(A, c, \delta) = \sum_{j=1}^L \|x_j - cAy_j - \delta\|^2$$

for rotation matrix A , translation vector $\delta \in \mathbb{R}^d$ and scale factor $c > 0$.

- An important problem is the *unlabelled* case, where correspondence between landmarks on different configurations is unknown (e.g. molecular data).

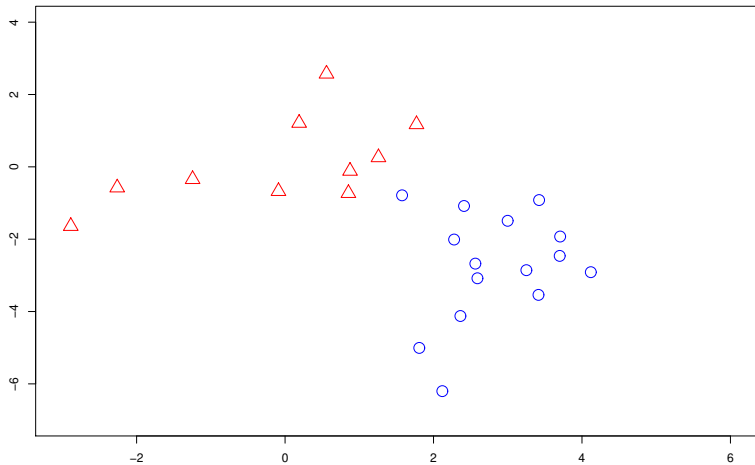
Unlabelled shape analysis

- Configurations X, Y of sizes m and n . In general $m \neq n$.
- What if correspondence between landmarks on different configurations unknown?
- Introduce matching matrix M such that

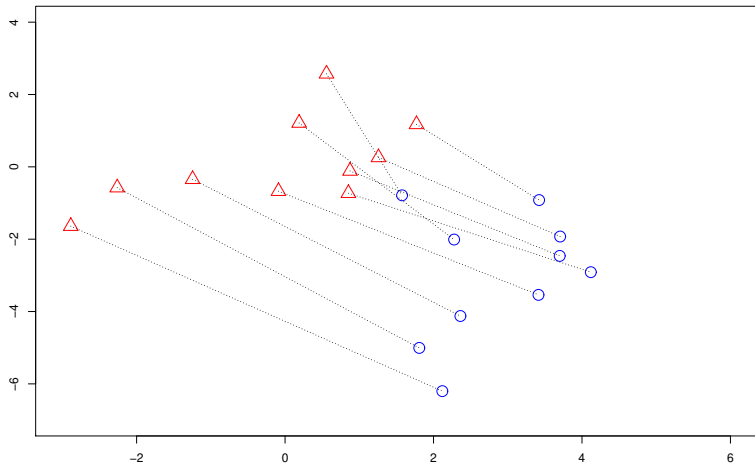
$$M_{jk} = \begin{cases} 1 & x_j \text{ corresponds to } y_k \\ 0 & \text{otherwise.} \end{cases}$$

- Require simultaneous inference for M and transformation parameters.
 - EM algorithm (Kent, Mardia and Taylor, 2010).
 - Bayesian-Procrustes model (Dryden, Hirst and Melville, 2007; Schmidler, 2007; Rodriguez and Schmidler, 2014).
 - Bayesian hierarchical model, rigid-body (Green and Mardia, 2006).
 - Full similarity shape (Mardia et.al., 2013).
- Example applications:
 - Molecular alignment.
 - Fingerprint matching (Forbes and Lauritzen, 2013).

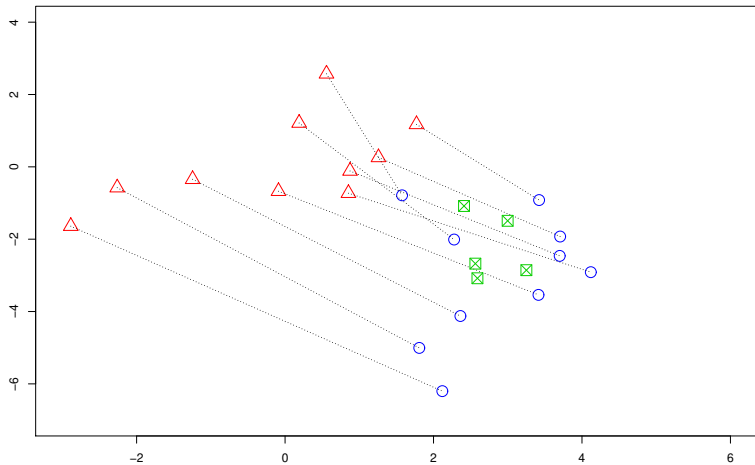
Simulated illustration



Simulated illustration



Simulated illustration



Green-Mardia model

- The observed points are

$$x_j \sim N(\mu_{\xi_j}, \sigma^2 I_d), \quad j = 1, \dots, m,$$

$$Ay_k + \delta \sim N(\mu_{\eta_k}, \sigma^2 I_d), \quad k = 1, \dots, n,$$

derived from hidden configuration μ .

- Transformation parameters are rotation matrix A and translation vector δ .
- Joint posterior: $p(M, A, \delta, \sigma, x, y) \propto p(A)p(\delta)p(\sigma)p(M, x, y)$, where

$$p(M, A, \delta, \sigma, x, y) \propto p(A)p(\delta)p(\sigma)\kappa^L\sigma^{-Ld} \\ \times \exp \left\{ -\frac{1}{4\sigma^2} \sum_{j,k:M_{jk}=1} \|(x_j - Ay_k - \delta)\|^2 \right\}.$$

- Given L matches, uniform prior on M .
- $p(M|L) \propto 1$, $p(M) \propto \kappa^L$.

Multiple Alignment

Multiple alignment

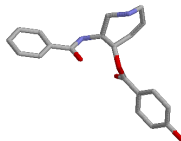
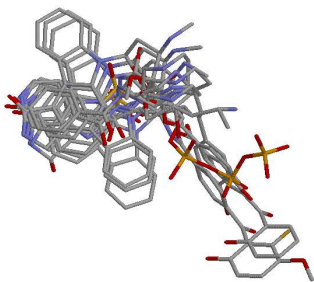
- Suppose we want to align $C \geq 3$ configurations.
- Ruffieux and Green (2009) and Dryden, Hirst and Melville (2007) describe models for simultaneous multiple alignment.
- Assumes one common underlying structure.
- We propose a multi-stage pairwise alignment algorithm.
 - Successively builds hierarchy of templates representing matched points.
 - Allows possibility of multiple subsets of configurations.
 - Similarities (and differences) with hierarchical clustering.
 - Motivation - multiple alignment of ligands in bioinformatics.

Pharmacophore models

- Pharmacophore models are a key ingredient in the discovery of new drugs.
- Drug activity controlled by interaction of active molecule, called a *ligand*, with a protein active site.
- A ligand must have certain chemical features in the correct spatial orientation to be recognised at the active site.
- A pharmacophore model simultaneously captures this chemical and geometric information. This model can be used to screen for other potentially-active molecules.

Motivation

- Since shape is key to activity of a molecule, we could use ideas from (unlabelled) statistical shape analysis.
- Chemical features common to a set of active ligands, in the same orientation, could be responsible for activity.
- Objective is to obtain estimate(s) of mean shape — for use as *templates* for pharmacophore models.
- Could be *clusters* with different mean shape (core features).
- Example - diverse set of protein kinase inhibitors.



Hierarchical template model

- Start with a set of C configurations $\{x_i\}$, $i = 1 \dots C$.
- Suppose there is a mean configuration of n_0 points, μ say, of points common to all configurations.
- Define matching array M_{ijk} , where

$$M_{ijk} = \begin{cases} 1 & \text{if } x_{ij} \text{ corresponds to } \mu_k \\ 0 & \text{otherwise,} \end{cases}$$

and x_{ij} is point j on configuration i .

Hierarchical template model

- Conditional on μ , for i, j, k such that $M_{ijk} = 1$,

$$A_i x_{ij} + \delta_i | \mu_k \sim N(\mu_k, \sigma^2 I_d).$$

- A_i is a rotation matrix and δ_i is a translation vector aligning configuration x_i and μ .
- We can then proceed with a pairwise alignment, using a pairwise method which provides estimates of the transformation parameters and M .
- A point estimate of μ is then given by

$$\hat{\mu}_k = \left(\sum_{i=1}^C \sum_{j=1}^{n_i} \hat{M}_{ijk} (\hat{A}_i x_{ij} + \hat{\delta}_i) \right) / \sum_{i=1}^C \sum_{j=1}^{n_i} \hat{M}_{ijk}$$

Implementation

- Initial set of configurations $\{x_1, \dots, x_C\}$.
- Consider all possible pairwise alignments.
- Evaluate the alignment for each pair (i, i') using a score based on geometric mean, \mathcal{G} , which on the log scale is

$$n_0^{-1} \sum_{j,k: \hat{M}_{jk}=1} \log \hat{p}_{jk},$$

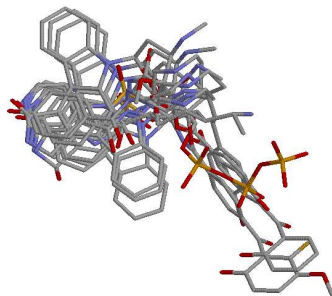
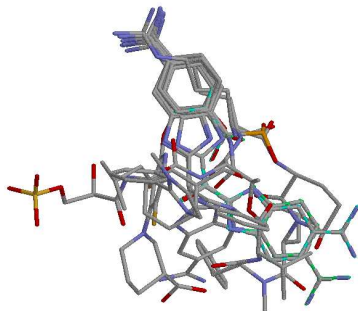
where n_0 is the number of matched points and \hat{p}_{jk} is the estimated probability that points x_{ij} and $x_{i'k}$ are matched.

- Each alignment gives an estimate $\hat{\mu}_{(i,i')}$.
- Take the pair with the best score, $(1, 2)$ say, and use these to obtain an estimate $\hat{\mu}_{(1,2)} = T_{12}$, say.
- Add T_{12} to the set of configurations, removing the two configurations merged to produce it.

Implementation

- New set is $\{x_3, \dots, x_C, T_{12}\}$.
- Now evaluate all pairwise alignments in the new set.
- Proceed successively, taking the best pairwise alignment at each stage to produce a new estimate $\hat{\mu}$, removing the two elements which are merged to produce it.
- Stop when no further pairwise alignments exceed some chosen threshold, \mathcal{G}_{min} say.
- May output one cluster of configurations, or multiple clusters each providing a different estimate of the mean shape.

- 2 datasets derived from SitesBase (Gold and Jackson, 2006) of ligands binding at structurally-related protein active sites.



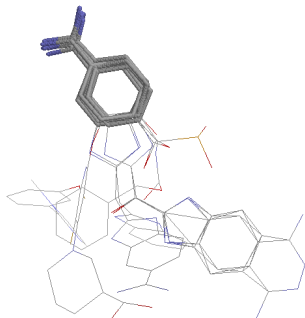
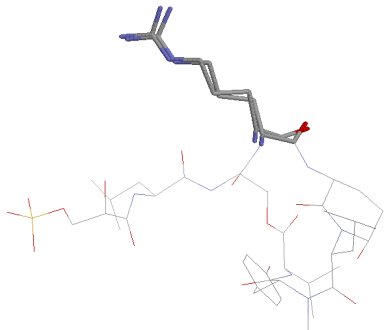
Ligands	\mathcal{G}
8 10	0.84
1 3 4 5 6 7 9 11	0.49

Table: Subsets found in dataset 1

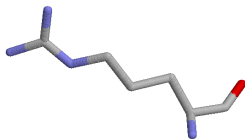
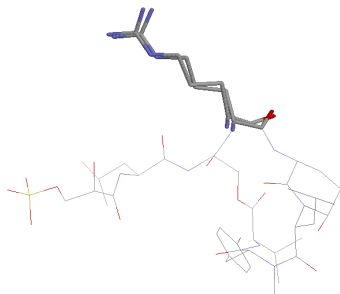
Ligands	\mathcal{G}
2 3 6	0.99
9 10 11	0.96
4 5 7	0.81

Table: Subsets found in dataset 2

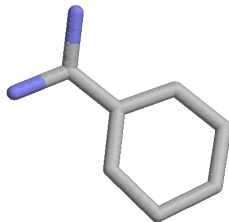
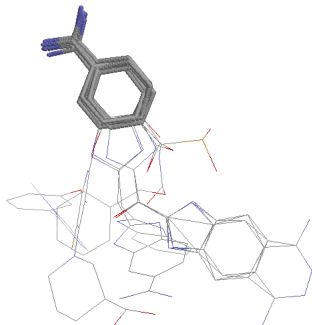
Results-dataset 1



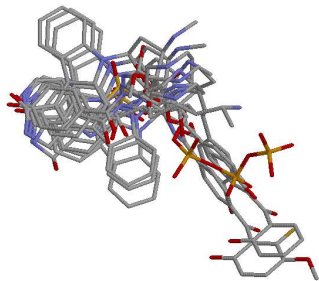
Results-dataset 1



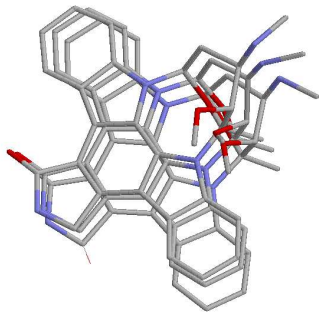
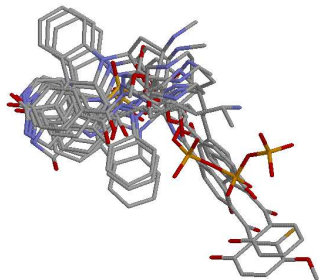
Results-dataset 1



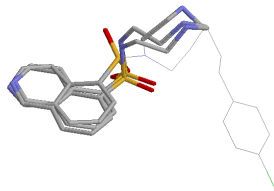
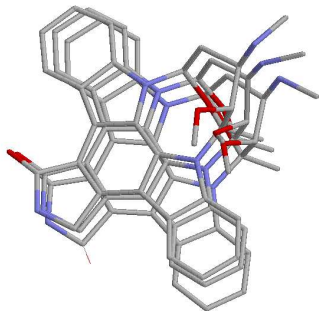
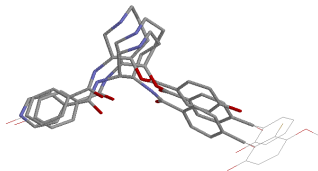
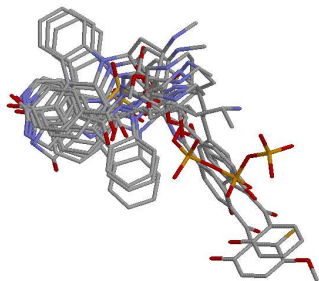
Results-dataset 2



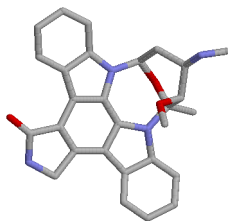
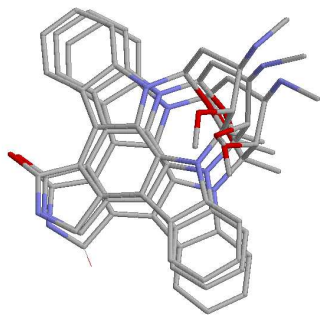
Results-dataset 2



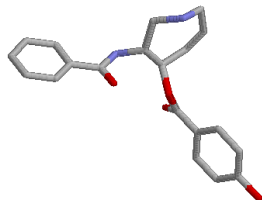
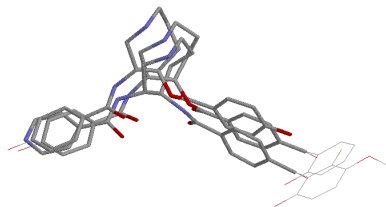
Results-dataset 2



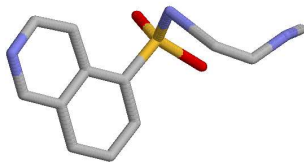
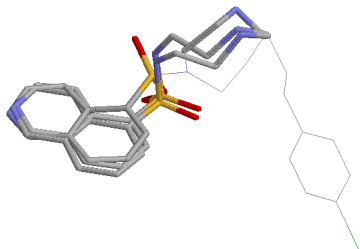
Results-dataset 2



Results-dataset 2



Results-dataset 2



Multiple Alignment: Summary

- Problems/challenges:
 - Not fully model-based (propagation of uncertainty in the clustering).
 - No “backtracking” — better solutions missed if “wrong path” taken early.
 - Non-overlapping clusters.
- Desiderata/questions:
 - Better representation of data?
 - Overlapping clusters.
 - Computational efficiency.
 - Mapping of mean shapes to realistic molecules/pharmacophore models.

Sequence-Structure Alignment

Proteins

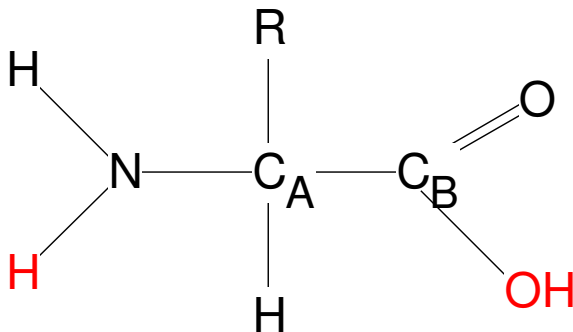
- Proteins are chains of amino acids, of which there are 20 types.
- Primary structure — a sequence of letters, one for each amino acid.

G T G K S T L L K K L F A

A section of a protein (Protein Data Bank ID: 1GKY).

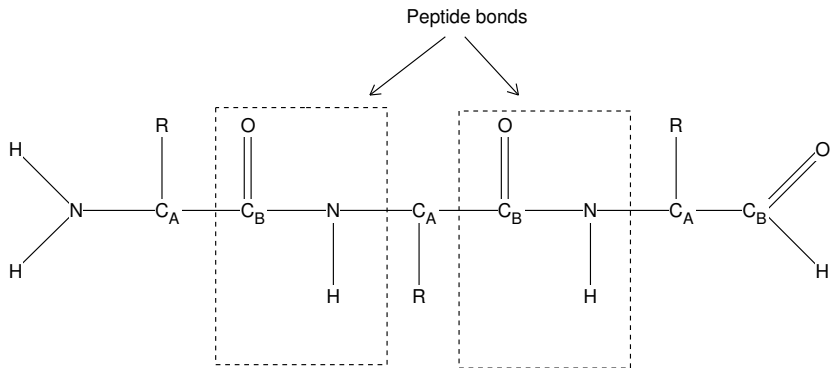
- Folds into a 3d structure, determined by the properties of the amino acids in the chain.
- Secondary structure elements are *beta strands* and *alpha helices*.
- The structure of a protein is much more conserved over long evolutionary periods than its sequence.

- Amino acids all have the following form, and the residue R determines which of the 20 types an amino acid is.

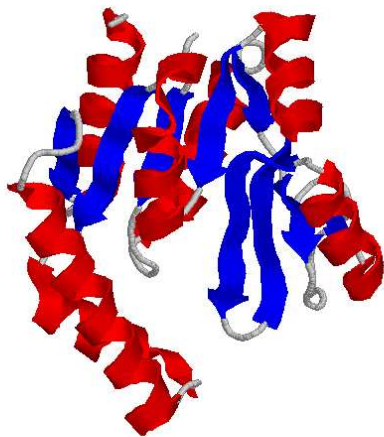


Peptide bonds

- Adjacent amino acids form peptide bonds to produce the protein chain.

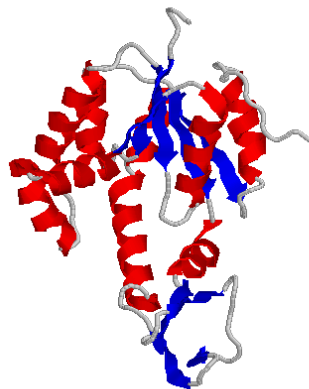
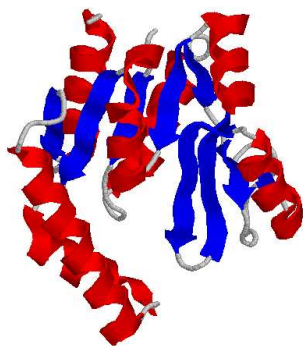


Protein structure



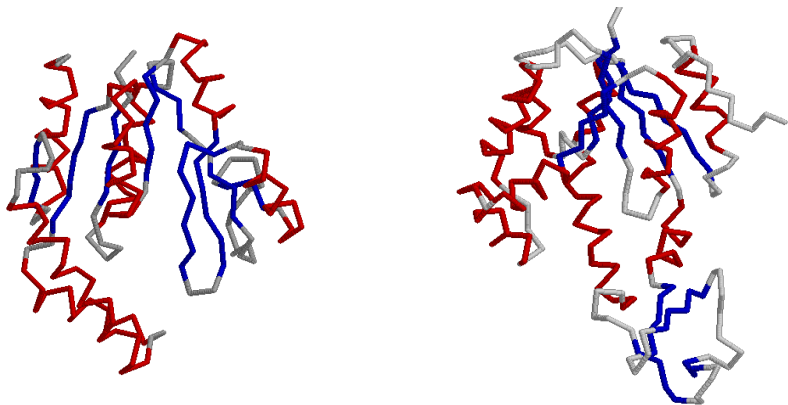
Protein alignment

- The goal is to align the structures, to assess structural similarity. More informative than assessing via sequences alone.
- We combine sequence information (sequence ordering and amino acid types) and structural information.



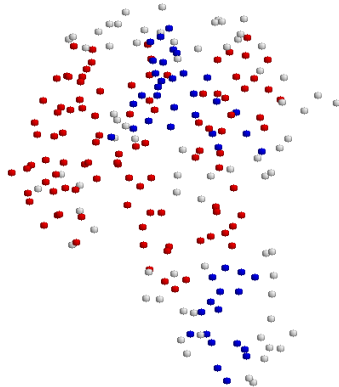
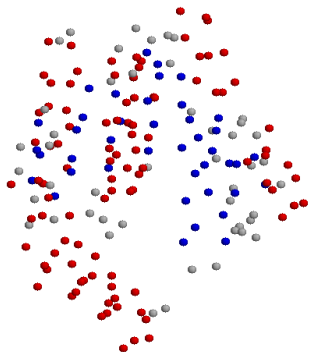
Protein alignment

- The goal is to align the structures, to assess structural similarity. More informative than assessing via sequences alone.
- We combine sequence information (sequence ordering and amino acid types) and structural information.



Protein alignment

- We represent each amino acid by the location of its C_α atom.



Sequence alignment

- Alignment problem: to identify the correspondence between amino acids on two proteins.
- Gaps enable the alignment of compatible amino acid types, allowing for insertions, deletions and substitutions in the protein sequences.

S^x		H	E	A	G	A	W	G	H	E	E
S^y		P	-	-	-	A	W	H	E	A	E

(a)

S^x		H	E	A	G	A	W	G	H	E	-	E
S^y		-	P	-	-	A	W	-	H	E	A	E

(b)

- Gaps can be in one sequence (a), or in both sequences (b).
- Sequence orders are preserved.

Scoring sequence alignments

- Scoring systems, such as PAM matrices, are used to score matches between each pair of amino acid types.
- Penalty functions penalise the number and length of gaps in an alignment.
- For a gap of length r , a commonly-used penalty function is

$$f(r) = -g - (r - 1)h,$$

where g is a gap opening penalty and h is an extension penalty.

- Total score (log scale) of an alignment:
total score of aligned pairs + total gap penalty.
- We consider alignment of proteins using both 1-dimensional (amino acid sequence order and perhaps type) and 3-dimensional (C_α atomic coordinates) information.

- Gap penalty prior (Rodriguez and Schmidler, 2014) for the matching matrix M :

$$p(M; g, h) = Z(g, h) \exp \{ -gS(M) - hL(M) \},$$

where $S(M)$ is the number of gaps, r_i is the length of the i th gap, $L(M) = \sum_{i=1}^{S(M)} (r_i - 1)$ is the total gap extensions and $Z(g, h)$ is a normalising constant.

- Previous prior for M was

$$p(M; \kappa) \propto \kappa^L,$$

where L is the number of matches given by M .

- Gap penalty prior (Rodriguez and Schmidler, 2014) for the matching matrix M :

$$p(M; g, h) = Z(g, h) \exp \{ -gS(M) - hL(M) \}.$$

- This ignores amino acid *type* information, and just penalises the gap component of the alignment. (Though type information can easily be included.)
- For example, the alignment

S^x		H	E	A	G	A	W	G	H	E	-	E
S^y		-	P	-	-	A	W	-	H	E	A	E

gives $S = 4$, and $r = 1, 1, 2, 1$.

- Write

$$U(M; g, h) = gS(M) + hL(M),$$

the total gap penalty.

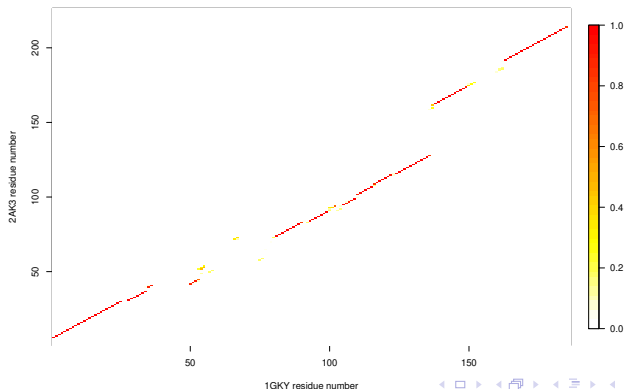
- The joint posterior distribution is

$$p(M, A, \delta, \sigma, x, y) \propto p(A)p(\delta)p(\sigma)v^L\sigma^{-Ld}\exp\{-U(M; g, h)\} \\ \times \exp\left\{-\frac{1}{4\sigma^2}\sum_{j,k:M_{jk}=1}\|(x_j - Ay_k - \delta)\|^2\right\}.$$

- Matching matrix M updated by proposing small perturbations whilst preserving sequence order.

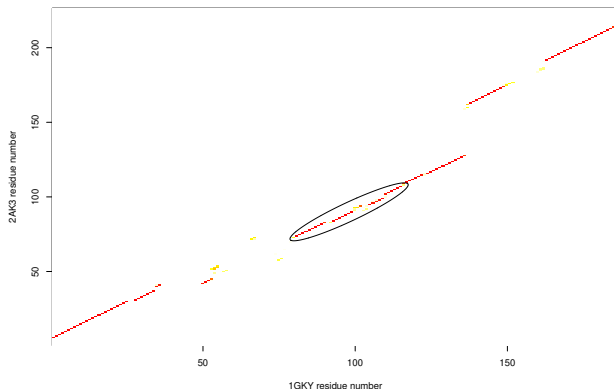
Example

- A guanylate kinase, 1GKY, and an adenylate kinase, 2AK3. “Twilight zone” — low sequence identity ($< 20\%$ of matched pairs are of same type).
- Use $g = 4$, $h = 0.1$, so higher penalty for gap openings than extensions.



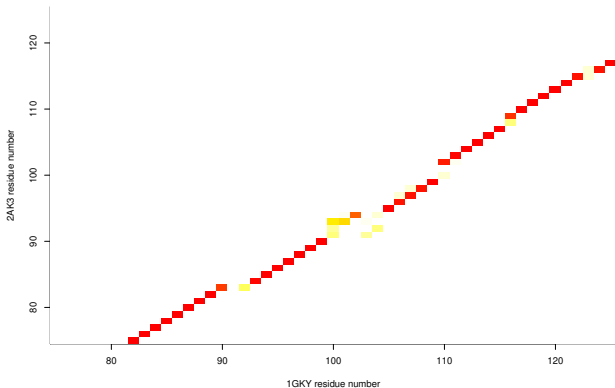
Example

- A guanylate kinase, 1GKY, and an adenylate kinase, 2AK3.
- Sample from posterior distribution of alignments, highlighting areas of uncertainty in the alignment and regions of high structural conservation.



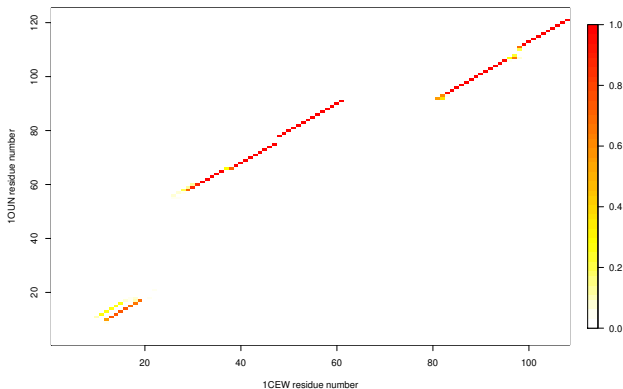
Example

- A guanylate kinase, 1GKY, and an adenylate kinase, 2AK3.
- Samples posterior distribution of alignments, highlighting areas of uncertainty in the alignment and regions of high conservation.



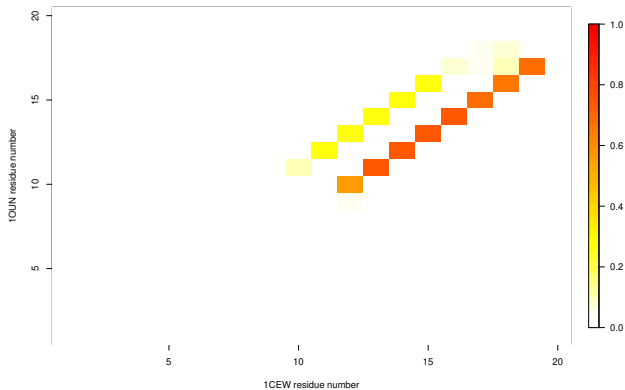
Example

- Pair 1CEW — 1OUN. Two alternative alignments of first helix.
- Assess relative merit by posterior probabilities.



Example

- Alternative alignments of helix.



A general form of penalty


- More generally, we¹ consider priors of the form

$$p(M; \theta) = Z(\theta) \exp \left\{ - \sum_i f(\text{penalty for gap } i; \theta) \right\}.$$

- Implementation (MCMC) can still proceed in the same way – just change in penalty required given a proposal M' .
- Motivation: For previous gap penalty prior, given L matched points and S “gap instances” (blocks of consecutive gaps, over both sequences), the j and k indices forming the matched points are independent *a-priori*.
- In fact,

$$U(M; g, h) = (g - h)S + (m + n - 2L),$$

so penalty depends only on S and L .

¹Fallaize, Green, Mardia and Barber (2019). *arXiv*: 1404.1556. 

Example (Green, 2015.)²

- Suppose $m = 9, n = 15$ and $L = 3$. Two possible alignments are given by the sets of indices

Xindex		0	4	5	9	10
Yindex		0	7	8	12	16

and

Xindex		0	4	5	9	10
Yindex		0	7	11	12	16

- Both give $S = 5$ and $L = 3$, and hence are equally preferable under the prior.
- At least intuitively, we might prefer the first alignment.
- In fact, if the X indices are $(1, 2, 3)$ (ignoring endpoints), then any set of y indices of the form $(2, k, 14)$, $k = 4, \dots, 12$ would be equally preferable.

²In *Geometry Driven Statistics*. Dryden I.L. and Kent, J.T. (eds).

New penalty

- We consider adding an additional penalty term to discourage this sort of situation.
- Consider the matches defined by the triples (j_1, j_2, j_3) and (k_1, k_2, k_3) . Then given j_1, j_3, k_1, k_3 , we encourage matches between j_2 and k_2 which preserve “proportionality”.
- Specifically, we introduce

$$\gamma(q; \nu) = \frac{\nu q^2}{2},$$

where

$$q = \log \left[\frac{(j_2 - j_1)/(j_3 - j_2)}{(k_2 - k_1)/(k_3 - k_2)} \right],$$

and total penalty contribution for these indices is $\gamma(q)$ + any gap opening/extension penalties as previously.

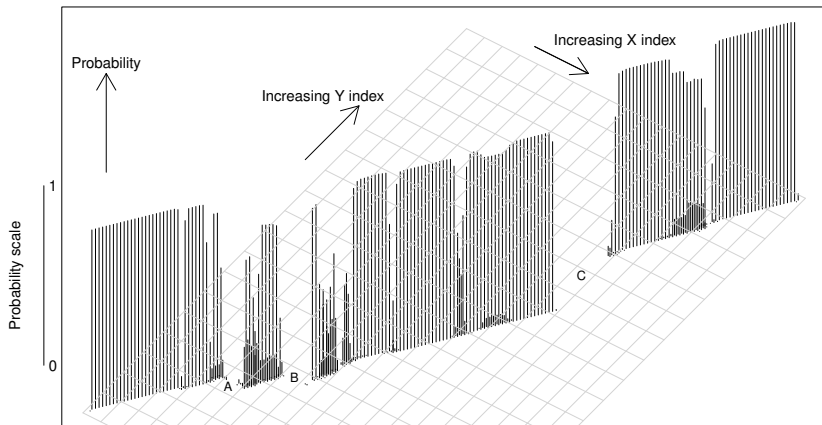
- Total overall penalty is still just a sum over successive pairs and triples of the matching indices.

- In general, given L matches, we have L triples of matching indices in the X sequence, given by $(j_0, j_1, j_2), (j_1, j_2, j_3), \dots, (j_{L-1}, j_L, j_{L+1})$.
- Similarly, in the Y sequence we have the L triples $(k_0, k_1, k_2), (k_1, k_2, k_3), \dots, (k_{L-1}, k_L, k_{L+1})$.
- The total penalty function is

$$U(M; g, h, \nu) = gS(M) + hL(M) + \sum_{i=1}^L \gamma(q_i; \nu).$$

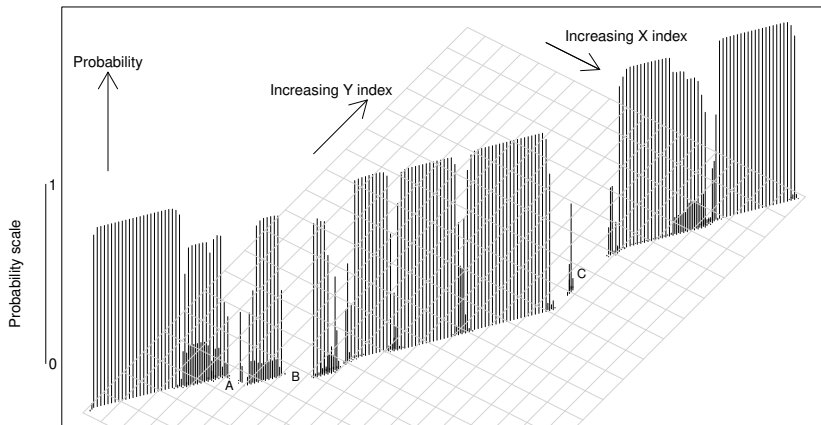
- Letting $\nu = 0$, we obtain the original penalty.
- Many other possibilities

Application: 1GKY – 2AK3



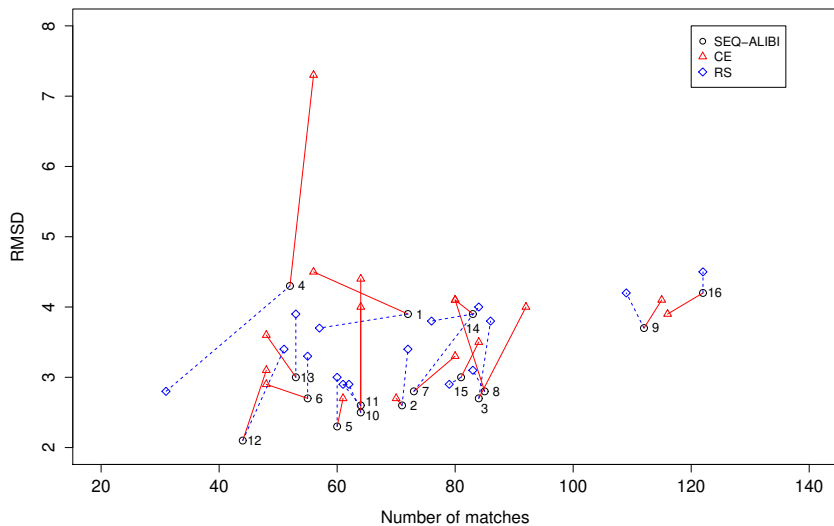
$$\nu = 0.25.$$

Application: 1GKY – 2AK3



$$\nu = 4.0.$$

Application: 16 “challenging” protein pairs.



- Prior is

$$p(M; \theta) = Z(\theta) \exp \{-U(M; \theta)\}.$$

- Could treat θ as unknown, adding an extra layer to the hierarchy.
- Standard MCMC requires knowledge of

$$Z(\theta)^{-1} = \sum_{M'} \exp \{-U(M'; \theta)\}.$$

- Methods which avoid this require ability to simulate from the distribution.
- For the standard gap penalty, there are efficient recursions for both computation of constant and simulation (mimicking the standard forward/backward algorithm in sequence alignment).
- This algorithm doesn't seem feasible computationally for general penalty.

Incorporating amino acid types

- We can also incorporate amino acid type information.
- Sequence of X is S^x , with elements $s_j^x \in \mathcal{S}, j = 1, \dots, m$ and \mathcal{S} is the set of integers 1 – 20 representing each of the 20 amino acid types.
- Similar definition for the sequence of Y , S^y .
- The sequence likelihood is

$$p(S^x, S^y | M, \Psi^l) = \prod_{j,k: M_{jk}=1} \psi_{s_j^x s_k^y}^l \prod_{j=1}^m q_{s_j^x} \prod_{k=1}^n q_{s_k^y},$$

where Ψ^l is a 20×20 PAM matrix for scoring each pair of amino acid types, accounting for an evolutionary distance l .

- q_s is the background proportion of an amino acid of type s in all proteins.

PAM matrices

- PAM — “point accepted mutations”.
- The elements of Ψ^l are

$$\psi_{ab}^l = \frac{p_{ab}^{(l)}}{q_a q_b}, \quad a, b = 1, \dots, 20,$$

where $p_{ab}^{(l)}$ is the probability of an amino acid of type a being substituted into an amino acid of type b over an evolutionary distance of l , and q_a, q_b are the relative proportions of amino acid types a and b in all proteins.

- One-step transitions $p_{ab}^{(1)}$ estimated from alignments of closely related proteins, rescaled so that probability of a substitution to a *different* amino acid type at any one site over one “evolutionary unit” is 0.01.
- For PAM- l matrix, $l\%$ “point accepted mutations”. The larger the value of l , the greater the tolerance to substitutions, implying a longer evolutionary distance.
- As $l \rightarrow \infty$, $p_{ab}^{(l)}$ tends to product of background probabilities.

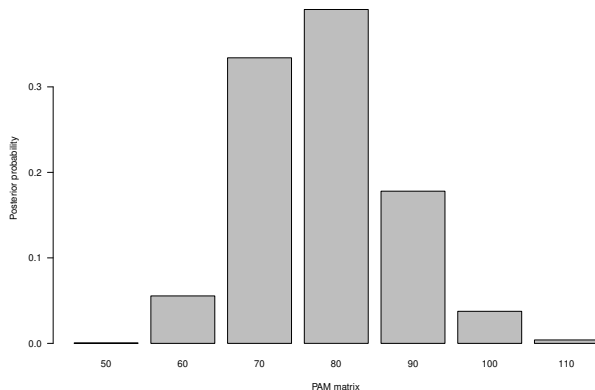
- The joint posterior distribution is now

$$\begin{aligned}
 p(M, A, \delta, \sigma, x, y, S^x, S^y) &\propto p(A)p(\delta)p(\sigma)v^L \exp\{-U(M; \theta)\} \\
 &\times \prod_{j,k:M_{j,k}=1} \frac{\psi_{s_j^x s_k^y}^l \phi\{(x_j - Ay_k - \delta)/(\sigma\sqrt{2})\}}{(\sigma\sqrt{2})^d} \\
 &\times \prod_{j=1}^m q_{s_j^x} \prod_{k=1}^n q_{s_k^y}.
 \end{aligned}$$

- We can consider l to be fixed (use a fixed PAM matrix) or include it as an unknown in the model and obtain its marginal posterior.
- This framework allows a natural measure of the evolutionary distance between two proteins.
- For convenience, we consider a discrete set of possible values for l .

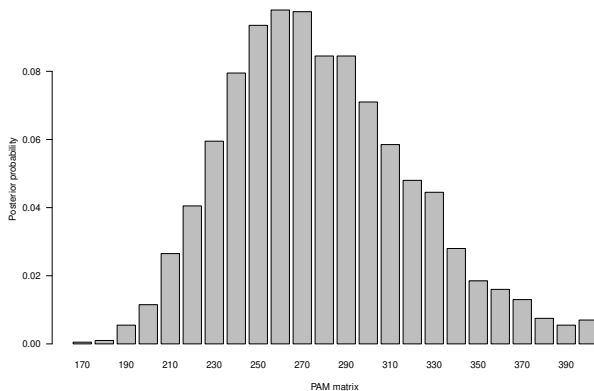
Example

- Example: Guanylate kinase pair 1GKY-1LVG. Closely related ($\approx 52\%$ sequence identity).
- Posterior mode of l is 80.



Example

- Example: The pair 1GKY-2AK3 revisited.
- Posterior mode of l is 260, indicating a longer evolutionary distance.



- Fully Bayesian model allows joint inference for matching and transformation/alignment.
- Flexibility to incorporate various forms of prior information.
- Biologically-meaningful results.
- Future work:
 - Large-scale assessment of improvement in alignments using general penalty functions.
 - Inference for θ in general penalty functions.
 - Changes to model/prior to allow e.g. protein flexibility (“twists”), non-sequential matching (domain swaps).
 - Incorporate additional information, e.g. hydrogen bonding, electrostatic potentials.
 - Multiple configurations – alignment, clustering, structure classification.

- 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.
- Fallaize, C.J., Green, P.J., Mardia, K.V. and Barber, S. (2019). Bayesian protein sequence and structure alignment. *arXiv*: 1404.1556.
- Forbes, P.G.M and Lauritzen, S. (2013). Fingerprint Analysis using Bayesian Alignment. In *Proceedings of LASR 2013*, 81–84.
- Gold, N.D. and Jackson, R.M. (2006). SitesBase: a database for structure-based protein-ligand binding site comparisons. *Nucleic Acids Research*, 34, D231–D234.
- Green, P.J. and Mardia, K.V. (2006). Bayesian alignment using hierarchical models, with applications in protein bioinformatics. *Biometrika*, 93, 234–254.
- Kent, J.T., Mardia, K.V. and Taylor, C.C. (2010). Matching unlabelled configurations and protein bioinformatics. Technical report, University of Leeds.

- Mardia, K.V., Fallaize, C.J., Barber, S., Jackson, R.M. and Theobald, D.L. (2013). Bayesian alignment of similarity shapes. *Annals of Applied Statistics*, 989–1009.
- Mardia, K.V., Nyirongo, V.B., Fallaize, C.J., Barber, S. and Jackson, R.M. (2011). Hierarchical Bayesian modeling of pharmacophores in bioinformatics. *Biometrics*, 67, 611—619.
- Rodriguez, A. and Schmidler, S.C. (2014). Bayesian protein structure alignment. *Annals of Applied Statistics*, 8, 2068–2095.
- Ruffieux, Y. and Green, P.J. (2009). Alignment of multiple configurations using hierarchical models. *Journal of Computational and Graphical Statistics*, 18, 756—773.
- Schmidler, S.C. (2007). Fast Bayesian shape matching using geometric algorithms. In *Bayesian Statistics 8*, 471—490.