Point Pattern Analysis of Protein Adaptation: incorporating protein structures into selection detection methods using spatial statistics

Hanon Mcshea

Final for BIOS221/STAT366: Modern Statistics for Modern Biology

08/17/2019

0 All data files and code for this paper are available at github.com/cyclase/mutationmap

1 Introduction of problem

Distinguishing signals of adaptation and neutral evolution in sequence datasets is an important problem in the field of molecular evolution. Metrics like the McDonald-Kreitman (MK) test (McDonald and Kreitman, 1991) and branch-site test (Yang and Dos Reis, 2010) have been developed for evaluating the evolutionary signal in protein-coding DNA sequence datasets that have an underlying phylogenetic structure. Both metrics define an adaptative evolutionary regime as deviation from a neutral model. The present study takes up this definition of adaptation and proposes a method for detecting adaptive evolution that incorporates information from three-dimensional protein structures. The method is presented as a workflow in R, with the intention of soliciting feedback as it is developed further.

I investigate whether mutations that occur along long branches in a protein's phylogenetic tree are clustered in its 3D structure, relative to a neutral evolution model. The MK test contends that adaptive signals in one-dimensional sequence data look like an uptick in nonsynonymous substitutions (those that change the sequence of the protein encoded by the DNA) relative to synonymous substitutions. The assumption underlying this contention is that at the level of the protein, nonsynonymous mutations that accrue under adaptation alter its function in response to selective pressure. Given this assumption, I predict that adaptation in response to selective pressure (functional change at the protein level) will produce a clustered pattern of amino acid mutations in the protein's 3D structure, while neutral evolution will produce a random pattern of mutations (no functional change at the protein level). Put another way, the method I explore in this paper uses the definition of adaptation accepted in the molecular evolution literature, but measures departure from neutrality in 3D protein structural space rather than in 1D sequence space.

The advantage of using protein structures is that they introduce biologically-meaningful constraint and context to adaptation testing, making it feasible for ancient proteins where second- and third-codon site saturation in DNA sequences has occurred, and by allowing biochemically-informed interpretations of adaptive signals. To my knowledge, although *post hoc* analysis of mutation clustering in protein structures has been useful in several studies (e.g. Enard et al., 2016; Kacar et al., 2017), no naïve method for detecting adaptation signals using protein structures exists in the literature. This paper explores such a method, using the squalene-hopene cyclase (SHC) group of the triterpenoid cyclase family of proteins as a dataset.

2 Approach

The method maps the mutations that occur along a branch in a tree to their position in the three-dimensional structure of the protein, and then assesses the degree of mutational clustering. Clustering analysis is performed naively with respect to biochemistry ("are mutations clustered in the protein's three-dimensional structure?"), which will be underpowered but will minimize Type I errors. Later I intend to implement a biochemically-informed clustering test ("are mutations clustered in this region of the protein that has been functionally characterized in the biochemistry literature, relative to the rest of the protein?"), which will minimize Type II errors and have greater power, but be limited by the previous work on the protein. Branches with clusters of mutations discovered by either of these methods can be hypothesized to be undergoing adaptive evolution/directional selection in comparison to a neutral or nearly-neutral process, where mutations are expected to be distributed randomly outside of the protein domains that are strongly conserved/under purifying selection (Ohta, 1992).

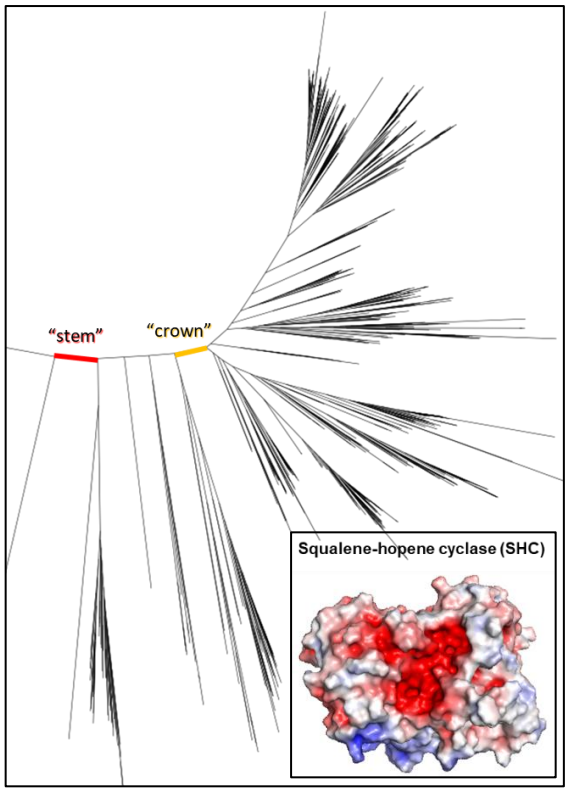

Figure 1: the source of the two mutation point patterns examined in this study. The tree is a maximum-likelihood phylogeny of the triterpenoid cyclase proteins, with the squalene-hopene cyclase (SHC) clade shown (to which other triterpenoid cyclases are the outgroup). The "stem" is the branch between the SHC group and the outgroup, and the "crown" is a branch of similar length within the SHC crown group. Inset is SHC structure used as an observation window in this paper, generated by x-ray crystallography (right; PDB ID = 1SQC; Wendt, et al., 1999) Electrostatic calculations were performed using the APBS PyMol plugin (Jurrus, et al., 2018) and structures were projected in PyMol (DeLano, et al., 2014).

To measure spatial clustering of mutations, I borrowed tools from the field of spatial statistics. Spatial statistics has developed methods for analyzing the distribution of points within an observation window. A dataset of spatial points within an observation window, called a "point pattern," can be characterized by its intensity and by the dependence between points, and can be compared to generative models, called "point processes" (Baddeley et al., 2015). Here, the **point pattern** is the set of mutations accrued along a branch in a protein phylogeny (estimated using ancestral sequence reconstruction) and their positions in the 3D structure of the protein, the **observation window** is defined by the surface of the protein, and the **null hypothesis** is a Poisson point process with the same number of points as the mutation pattern observed. The distribution of residues in the protein (which is inhomogeneous across different secondary and tertiary folds), is supplied as a **covariate**.

I used the R package spatstat (Baddeley et al., 2015) to investigate dependence on the protein structure covariate and dependence between points in two mutation point patterns (MPPs) extracted from the triterpenoid cyclase tree and reconstructed ancestral sequences. In 3D, I calculated the empty-space function E (distance to nearest point from an empty point within the window), the nearest-neighbor function G (distance from a point to the nearest other point), and the pairwise distance function K (distance between points in the pattern). In 2D, I calculated these functions and performed other tests of homogeneity, independence, and complete spatial randomness (CSR). All code for this method is in the rmd supplement Mutation_map.rmd.

3 Methods

3.1 Data wrangling

Evolutionary analysis of a protein [family] using this method requires (1) a phylogenetic tree with internal nodes labeled, (2) reconstructed ancestral protein sequences at those nodes, and (3) a 3D structure of the protein (obtained from x-ray crystallography, NM, cryo-EM, etc). The tree is not loaded into R but can be consulted to identify branches of interest. An alignment of reconstructed ancestral protein sequences (estimated with a program like PAML; Yang, 1997), as well as extant sequences, are accessed in R as a dataframe where columns are sequence identifiers and rows are sites in the alignment. An additional column is created with the non-gap positions of the sequence that corresponds to the structure (in this paper, the *Alicyclobacillus acidocaldarius* SHC). Another column is added for each branch of interest, populated with logical values indicating whether the sequences at either end of the branch are identical at that alignment site (in this paper, the stem of the SHC clade is investigated, as well as a long branch within the SHC crown group; Figure 1). The protein's 3D structure is read from a Protein Data Bank (PDB) file using the package Rpdb, and a dataframe is created that contains the sequence positions and calculated centers of geometry for each amino acid residue. This dataframe is merged with the alignment dataframe (removing sites that do not align to the protein structure) to form an object that contains all information necessary for analysis of MPPs. Here I use spatstat for the entire analysis of this object, but propose the development of more specialized methods in section 6.

3.2 Three-dimensional analysis

For analysis of the 3D point pattern of mutations, I used untransformed coordinates of the centers of geometry of each mutated residue. I defined the observational window as the bounding rectangular prism of the protein because spatstat requires a box-shaped window in 3D (later versions of the method will allow an approximation of the protein surface to be used as the observation window). Within a box-shaped observation window, the globular protein (supplied as a covariate) looks clustered, but MPPs can still be analyzed by comparison to it.

3.3 Two-dimensional analysis

To perform a more extensive analysis than spatstat currently allows in three dimensions, I projected the mutational point pattern into two dimensions. Although a 2D representation of

protein structure necessarily loses biophysical and biochemical information, I undertook the analysis to better understand the implementation and interpretation of spatial statistics methods.

For 2D point patterns, spatstat allows a polygonal observation window. A polygon represents the boundary of a protein better than a rectangular box does. Future versions of this method will define the observation window as the scaled alpha-shape of the protein, computed with the package alphahull (Pateiro-Lopez et al., 2019) in 2D or ashape3D (Lafarge and Pateiro-Lopez 2017) in 3D. The alpha-shape is easier to work with in R than the 3D mesh of solvent-accessible surface area that is typical of molecular dynamics investigations, but retains an intuitive and biophysically-relevant notion of "shape." Scaling the alpha shape by the average length of one amino acid side-chain places all points in the pattern within the observation window, rather than on its edge. I generated a 3D alpha-shape of an SHC for future use (Figure 2; compare to inset crystal structure in Figure 1); it can be examined in an rgl viewing device (Adler et al., 2019) by opening the rmd supplement to this paper in R/RStudio.

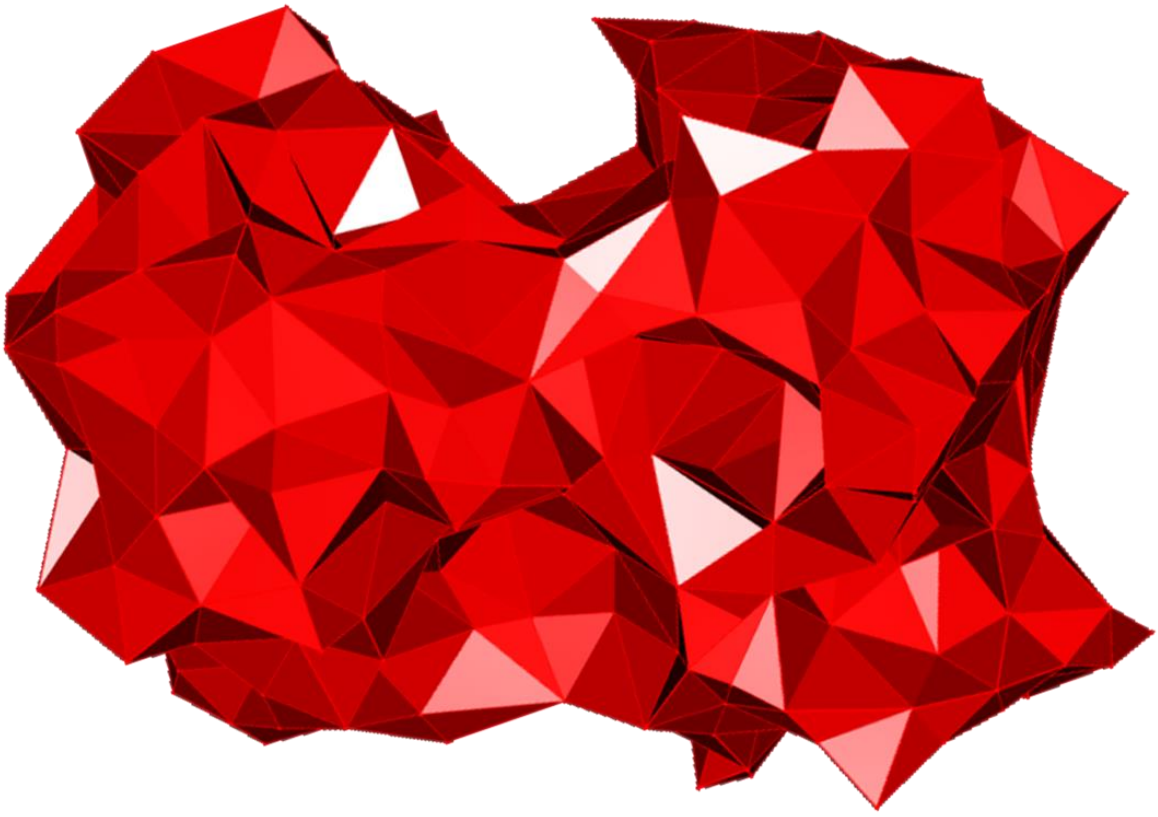

Figure 2: extremely shiny alpha hull of the SHC protein structure.

I performed the workflow on two 2D projections of MPPs: an unscaled principal component analysis (PCA) and a t-distributed Stochastic Neighbor Embedding (t-SNE). These projections were selected, respectively, as less- and more-processed planar representations of

multidimensional data. Because the x, y, and z coordinates of the protein point pattern are already represented in the same units (Angstrom, equivalent to 0.1 nm), I did not perform the data scaling that is necessary for typical uses of PCA. Thus the long axis of the protein (which is about 150 x 50 x 50 A) explains most of the variance in residue coordinates, and the first two PCs explain 98% of the variance. t-SNE does not preserve a recognizable "shadow" of the protein in 2D, but does preserve meaningful short-range spatial relationships between residues. PCA was implemented in the R package ade4 (Dray and Dufour 2004) and t-SNE was implemented in Rtsne (Krijthe 2015).

4 Results

4.1 Dependence between points in 3D

The whole protein point pattern, the stem mutation point pattern, and the crown group point pattern, as well as their F, G, and K functions, are show in Figure 3. The interpretation of these functions (and of the same functions for 2D processes in section 4.2) is as follows: an F function value less than that of a Poisson point process at any radius R indicates "clustering," where the nearest point to an empty space is greater than in a Poisson process, and an F function value greater that of the Poisson point process indicates overdispersion or a "regular" generative process;

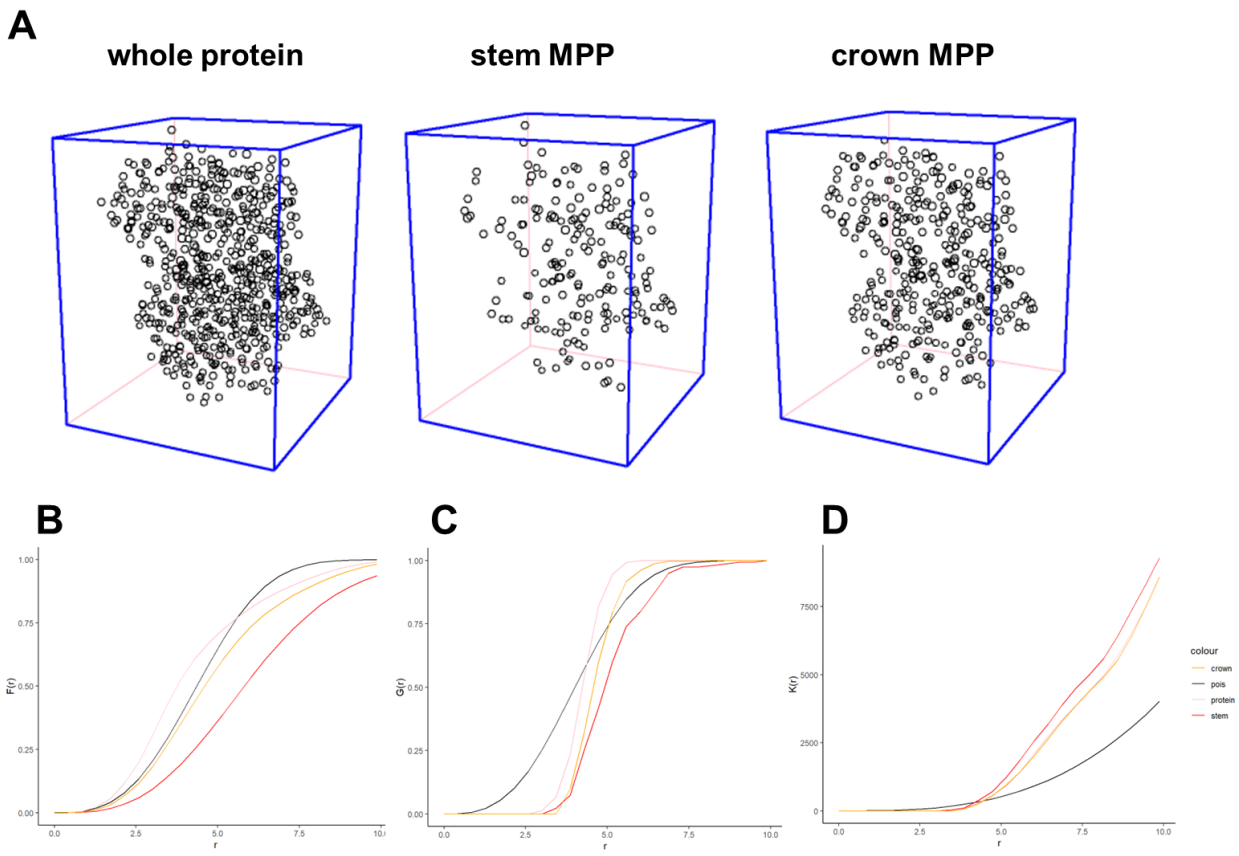

Figure 3: 3D point patterns visualization and clustering analysis (A) point patterns; (B) F function of the three point patterns as labeled in comparison to a Poisson point process; (C) G function of point patterns and Poisson process; (D) K function of point patterns and Poisson process.

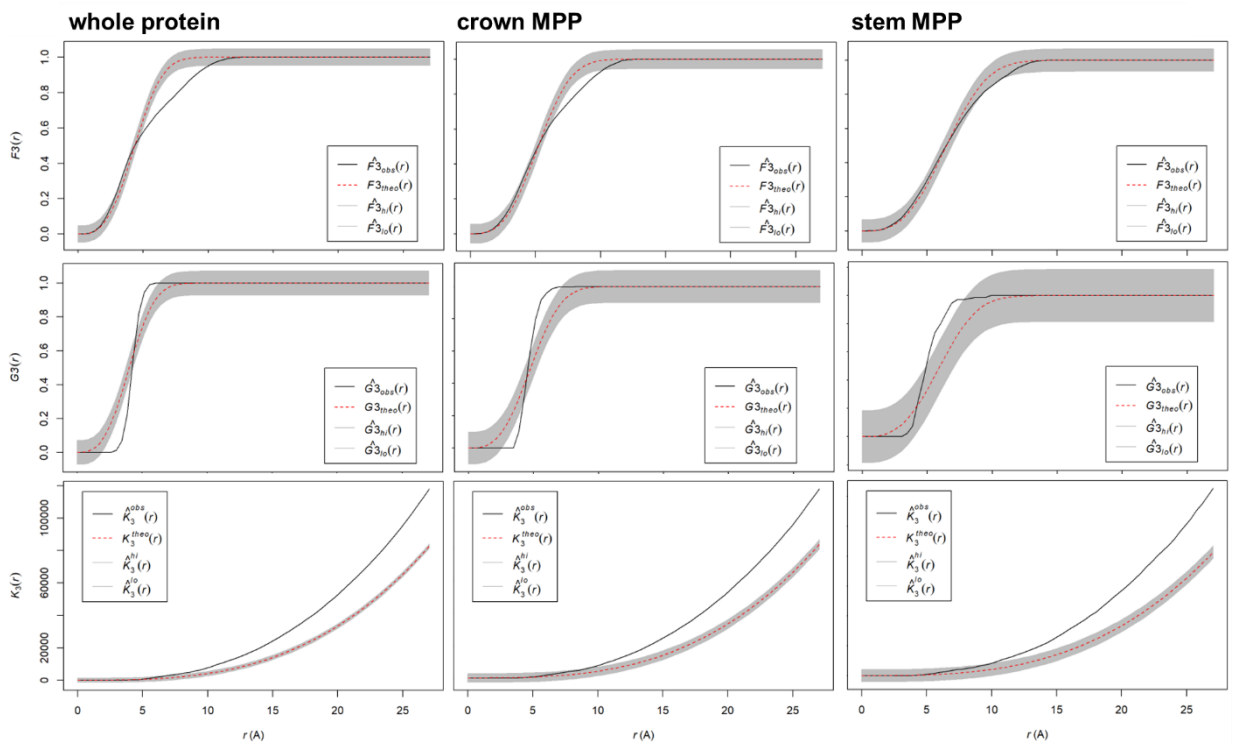

Figure 4: global Monte Carlo simulation envelopes for the F (first row), G (second row), and K (third row) functions of the 3D point patterns as labeled. 39 replicates were performed to create a 95% confidence interval.

inversely $G>G_{pois}$ and $K>K_{pois}$ indicate clustering, and $G<G_{pois}$ and $K<K_{pois}$ indicate overdispersion. The F (empty-space) function suggests that the whole-protein and the crown mutation point process do not differ much from a Poisson process in the same window, while the stem mutation point process is more clustered than a Poisson process. However, the stem MPP is never clustered according to the G (nearest-neighbor) function, whereas the whole-protein and crown MPPs become clustered around 4 and 4.5 Å, respectively. In the K function, the stem MPP transitions from overdispersed to clustered at a slightly shorter radius than the crown and whole-protein MPPs.

I performed Monte Carlo simulation to capture the variability of the Poisson process and plotted critical bands around the Poisson process to determine significance of deviations from the Poisson process with 95% confidence. In these plots (Figure 4), the whole-protein MPP is clustered at length scales of 4-10 Å according to the F function, at length scales of 5-6 Å according to the G function, and at length scales >9 Å according to the K function; it is overdispersed according to the G function at 2-5 Å. The stem MPP is clustered at length scales of 10-11 Å according to the F function, at length scales of 5-7 Å according to the G function, and at length scales >12 Å according to the K function. The crown MPP is clustered at length scales of 6-9 Å according to the F function, at length scales of 5-7 Å according to the G function, and at length scales >10 Å according to the K function. The stem MPP is never significantly overdispersed and the crown MPP is overdispersed at length scales of 3-4 Å according only to the G function.

4.2 2D

4.2.1 Dependence between points

For the 2D protein structure projection made with PCA, Monte Carlo simulation of Poisson process critical bands showed that for the whole-protein pattern and the stem and crown MPPs, neither the F nor G functions differed significantly from a Poisson process (Figures in rmd supplement). The K function for each did diverge from a Poisson process, showing clustering at longer length scales (becoming significant at around 5 units in PCA space for the whole protein and the stem MPP, and around 8 units in PCA space for the crown MPP).

In 2D, spatstat can calculate the F, G, and K functions without edge corrections. Edge corrections are performed in spatial analysis to account for points beyond the boundary of the observation window, but are not relevant for MPPs, which are biophysically confined to their observation window. Thus I recalculated the F, G, and K functions of the PCA-projected point patterns without edge correction, and found that they did not differ in shape or significance from the same functions with edge correction applied (rmd supplement).

For the 2D protein structure projection made with t-SNE, the whole-protein and crown MPP F functions showed significant clustering beginning around 0.6 units in t-SNE space, and the stem MPP F function became clustered at longer length scales (1.2 units in t-SNE space). The G function showed overdispersion at intermediate length scales for the whole protein and the crown MPP, but no divergence from Poisson for the stem tSNE. The K function conversely showed clustering at intermediate length scales for the whole protein and the crown MPP, but no divergence from Poisson for the stem tSNE.

4.2.2 Dependence on covariate

A basic test of complete spatial randomness is a chi-squared test of the counts of points in subsections ("quadrats") of the observation window. The PCA projections of each of the MPPs, as well as the whole-protein point process, differed significantly with $p<0.05$ in a test with 15 quadrats. However, the test is very sensitive to quadrat size, and loses significance at finer and coarser scales. The interval for which it is significant cannot be interpreted because the PCA units are not physically meaningful. A Kolmogorov-Smirnov test of dependence on a spatial covariate – a pixel representation of the density of the whole-protein pattern – suggested that neither the stem nor crown MPPs depended on it ($p=6.657e-11$ and 9.381e-12, respectively). A Berman test of the same dependence, which compares means of observed and expected cumulative probability functions of MPP intensity over the covariate, was not significant for either MPP. The Kolmogorov-Smirnov test and Berman test results were the same (K-S test $p=0.0240$ and $p=2.53e-12$, respectively; Berman test not significant) for the t-SNE projections.

I plotted the intensity of points in each pattern as a smoothed function of covariate value for the PCA and t-SNE projections (Figure 5). This gives an alternate visualization of length scales at which point pattern intensity appears independent of protein density. The relative distribution estimate for the PCA projection of the crown MPP had no major intensity peaks, while that of the PCA-projected stem MPP had a peak at longer length scales. For the t-SNE projections, the relative distribution estimates show that the crown MPP has slight intensity peaks away from the greatest density of the covariate, while the stem MPP intensity has a more pronounced peak at covariate densities lower than the covariate mode.

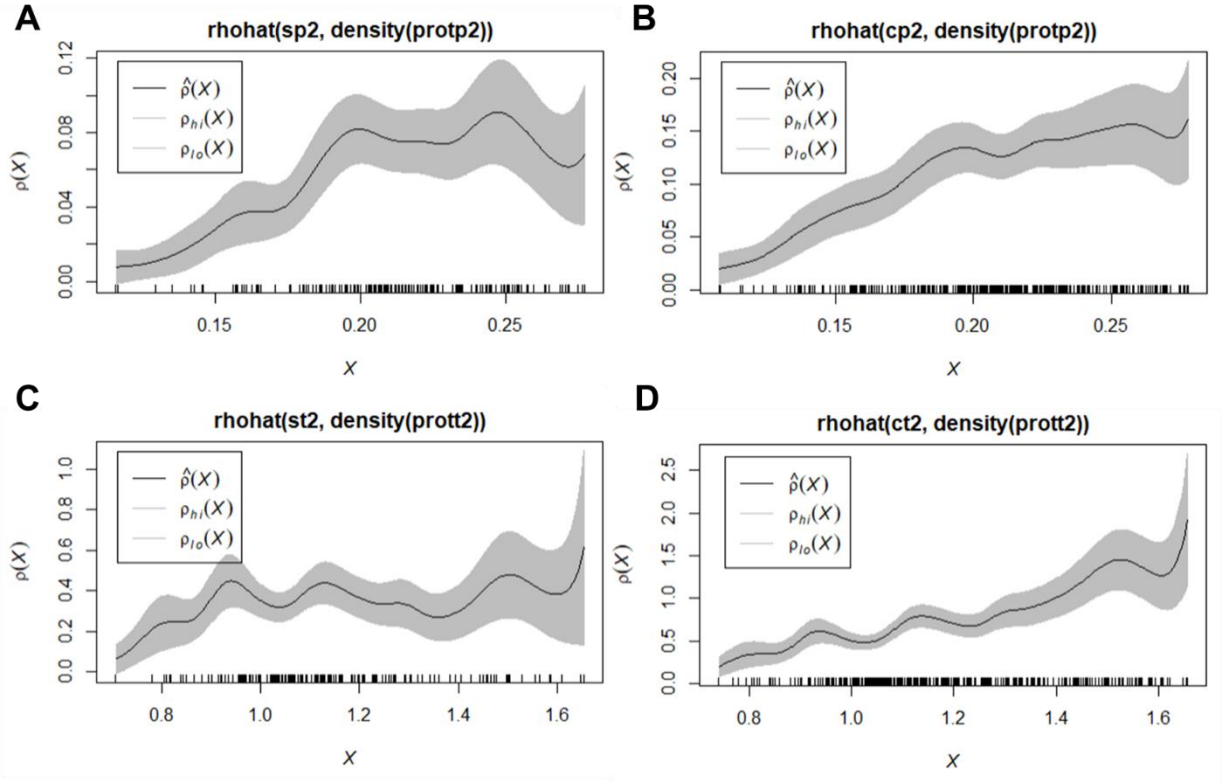

Figure 5: distribution estimates for the intensity (A-B) PCA-projected stem and crown MPPs and (C-D) t-SNE-projected stem and crown MPPs, relative to the density of the relevant projection of the whole protein.

5 Discussion

5.1 In 3D, possible clustering in stem but not in crown

In three dimensions, the stem MPP appears clustered in relation to a single Poisson process for empty-space and pairwise distance functions (and not clustered for the nearest-neighbor function; Figure 2), but the F and K clustering is not significant when Monte Carlo simulation of Poisson processes is performed. Clustering appears in its G function (Figure3). The lack of concordance in the F function could be due to the box-shaped observation window, which introduces a large amount of biologically-meaningless empty space. If this is the case (which could be tested by re-analysis in an alpha-hull-shaped observation window), then the G function clustering at 5-8 A may

be an adaptive signal. It is important to emphasize that the evolutionary interpretations of MPP properties are all guesses until we have a positive control (see section 6).

The crown MPP looks consistently similar to the point process of the whole protein. The appearance of clustering in the whole protein is an artifact of the empty space in the box-shaped window and/or an intrinsic property of its structure, so clustering in MPPs identical to those in the whole protein cannot be differentiated from artifacts. In a sense the protein structure provides an alternative null model to the Poisson process – under neutral evolution, we expect clustering curves with the same shape as those of the protein point pattern. The F, G, and K curves for the crown MPP do not differentiate it from the protein point pattern. However, this does not rule out adaptation-driven clustering in the crown MPP that is masked by method artifacts.

5.2 2D

5.2.1 No clear signal of dependence between points

The only clustering signal in the PCA-projected MPPs was in the K function, which measures the average number of other points within distance *r* of a point in the pattern, and which always reports clustering at long length scales for these protein point patterns and MPPs. In the t-SNE-projected point patterns, the whole-protein pattern and the crown MPP showed G function clustering, but it is hard to know whether to interpret the latter as tentatively neutral (as I suggested for the same phenomenon above) because the curves are different shapes/diverge from the Poisson at different length scales. Further, *r* values are not physically meaningful in either 2D projection, and cannot be interpreted as linear in the t-SNE projection (van der Maaten and Hinton, 2008), further confusing interpretation of the projections into 2D. The main result from this exercise is support for the necessity of implementing a reasonable observation window in 3D, to avoid projection altogether.

5.2.1 No chance of dependence on protein structure

Although the quadrat count exercise was not robust enough to interpret, the Kolmogorov-Smirnoff test strongly suggested that the intensity of the MPPs is not dependent on the protein structure density. This holds even though the mean intensities are not significantly different from those predicted by the structure density, and these two results together might be interpreted as constraint from the structure on some parameters of the cumulative intensity function, but not to a deterministic extent on the shape of the function. Lack of MPP dependence on protein structure suggests that deviations from the null Poisson model, where observed, are not driven by protein structure and may instead be due to dependence between points. A next step in covariate dependence interpretation is building the analysis capability in 3D (not currently possible in spatstat), where the possible clustering signal was observed in this work, and where any clustering signal will be more reliably interpretable than a projected one.

6 Future work

Future versions of the method will ideally manage the uncertainty generated in tree estimation and ancestral sequence reconstruction by computing whether mutation clusters are robust to tree perturbation. The version presented here took a single phylogeny and set of ancestral sequences as working hypotheses (Pauling and Zuckerkandl, 1963) and did not integrate their inhering uncertainty. I analyzed a very small data set here (two point patterns from two branches in a tree with thousands of branches), and hope to improve the workflow to assess mutation clustering across all branches in a tree or all branches above some length threshold. Doing so would allow comparison of MPPs across the tree, which might reveal trends in intensity or fitted model parameters even where patterns cannot be distinguished from Poisson processes.

I also plan to identify a dataset for which I can perform the MK and branch-site tests, to determine whether the lack of signal in this analysis was due to a true lack of evolutionary signal, or problems with (i) observation window and edge-correction in 3D or (ii) projection into 2D. A group of recent, rapidly-evolving proteins, e.g. those involved in host-virus interactions, would be ideal for this. If I detect adaptation in a clade or along a branch by other metrics but not those attempted in this paper, I would be excited to write script to calculate non-edge-corrected F, G, and K functions in 3D. If mutation clustering were still not detectable by this method, it might merit reappraisal of our assumptions about the dynamics of protein evolution and adaptation.

If the structure-independence of the MPPs holds in three dimensions, then the nearest-neighbor clustering in the stem MPP but not the crown MPP might signal that the population of proteins ancestral to the SHC clade were undergoing adaptation in response to selective pressure, and that later radiation within the SHC crown group was not driven by adaptation at the protein level. The present method does not provide enough evidence to support such a scenario, but it does provide glimmers that clustering may vary across the tree in a way that is worth exploring with a more refined method.

7 Acknowledgments

Thank you to S. Holmes, L. Nguyen, L. Symul, M. Sesia, A.K. Lee, E. Horst, and B. Barros-McShea. I have been wanting to use protein structures for adaptation detection since Fall 2017 but did not have the statistical tools or coding ability until BIOS221/STAT366.

8 References

Adler, D., Murdoch, D., and others (2019). rgl: 3D Visualization Using OpenGL. R package version 0.100.26. https://CRAN.R-project.org/package=rgl

Baddeley, A., Rubak, E., & Turner, R. (2015). *Spatial point patterns: methodology and applications with R*. Chapman and Hall/CRC.

DeLano, W. L. (2014). The PyMOL Molecular Graphics System, Version 1.8. Schrödinger LLC, http://www.pymol.org.

Dray S, Dufour A (2007). "The ade4 Package: Implementing the Duality Diagram for Ecologists." Journal of Statistical Software_, *22*(4), 1-20. doi: 10.18637/jss.v022.i04 (URL:

https://doi.org/10.18637/jss.v022.i04).

Enard, D., Cai, L., Gwennap, C., & Petrov, D. A. (2016). Viruses are a dominant driver of protein adaptation in mammals. Elife, 5, e12469.

Idé, J. (2017). Rpdb: Read, Write, Visualize and Manipulate PDB Files. R package version 2.3. https://CRAN.R-project.org/package=Rpdb

Jurrus E, Engel D, Star K, Monson K, Brandi J, Felberg LE, Brookes DH, Wilson L, Chen J, Liles K, Chun M, Li P, Gohara DW, Dolinsky T, Konecny R, Koes DR, Nielsen JE, Head-Gordon T, Geng W, Krasny R, Wei GW, Holst MJ, McCammon JA, Baker NA. Improvements to the APBS biomolecular solvation software suite. Protein Science, 27, 112-128, 2018.

Kacar, B., Hanson-Smith, V., Adam, Z. R., & Boekelheide, N. (2017). Constraining the timing of the Great Oxidation Event within the Rubisco phylogenetic tree. Geobiology, 15(5), 628-640.

Krijthe, J.H. (2015). Rtsne: T-Distributed Stochastic Neighbor Embedding using a Barnes-Hut Implementation, URL: https://github.com/jkrijthe/Rtsne

Lafarge, T., & Pateiro-Lopez, B. (2017). alphashape3d: Implementation of the 3D Alpha-Shape for the Reconstruction of 3D Sets from a Point Cloud. R package version 1.3. https://CRAN.R-project.org/package=alphashape3d

McDonald, J. H., & Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. Nature, 351(6328), 652.

Ohta, T. (1992). The nearly neutral theory of molecular evolution. Annual review of ecology and systematics, 23(1), 263-286.

Pateiro-Lopez, B., Rodriguez-Casal, A., and others. (2019). alphahull: Generalization of the Convex Hull of a Sample of Points in the Plane. R package version 2.2. https://CRAN.R-project.org/package=alphahull

Pauling, L., & Zuckerkandl, E. (1963). Chemical paleogenetics molecular restoration studies of extinct forms of life. Acta Chemica Scandinavica, 17, 9–16.

van der Maaten, L.J.P. & Hinton, G.E. (2008). Visualizing high-dimensional data using t-SNE. Journal of Machine Learning Research 9(Nov):2579-2605, 2008.

Wendt, K. U., Lenhart, A., & Schulz, G. E. (1999). The structure of the membrane protein squalene-hopene cyclase at 2.0 Å resolution. Journal of Molecular Biology, 286(1), 175–187.

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. Bioinformatics, 13(5), 555-556.

Yang, Z., & Dos Reis, M. (2010). Statistical properties of the branch-site test of positive selection. Molecular biology and evolution, 28(3), 1217-1228.