COMPARATIVE QUANTITATIVE GENETICS OF PROTEIN STRUCTURES: A COMPOSITE APPROACH TO PROTEIN STRUCTURE EVOLUTION

by

Jose Sergio Hleap

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

at

Dalhousie University
Halifax, Nova Scotia
December 2015

To my parents for all the support and the drive they instilled on me

To my girlfriend for all the love and support during this process

To my supervisor for the guidance provided

And last but not least,

To whomever had the patience to read this document

Table of Contents

Abstract

Structural biology has been long concerned about the emergence of protein structures and the convergence to particular folds. It can be said that protein structures are the realization of genetic information given thermodynamical and biological constraints. Given these properties, let's refer to a structure as a phenotype. As such, protein structures can be analysed as shapes within a geometric morphometrics framework, and as a phenotype in a quantitative genetics framework. Here, I present a robust way to analyse protein structures statistically in either evolutionary or molecular dynamics sampling. I show how General Procrustes Analysis (GPA) can be applied to aligned molecular dynamics snapshots, and provide evidence that the scaling component of GPA is not applicable to protein structures. I also show how analysing protein structures as shapes can give insights into dynamic and evolutionary patterns. Analysing proteins as shapes also gives the possibility to apply known techniques to assess modularity. Traditional techniques have dimensionality limitations. I show how to overcome these limitations and propose a robust way to analyse protein structure modularity. I show how a protein can be partitioned into biologically meaningful clusters, which can be used for description, protein prediction, or analysis of protein dynamics and evolution. The meaning of such modules is discussed further, and a hierarchical model for protein structure modularity is proposed. Also, methods to explore different kinds of modules at different kinds of hierarchy are explored.

Finally, given that protein structures are phenotypes, the potential response to selection can be assessed by means of comparative quantitative genetics. I show that traditional comparative approaches have a heavy computational burden, therefore making the analysis infeasible. Nevertheless, similar approaches are developed to efficiently and accurately generate the estimations when the phenotypic variance is partitioned based on repeated measures, using a pooled-within covariance estimation.

List of Abbreviations Used

GM:	Geometric Morphometrics
GPS:	Generalized Procrustes Superimposition
PCA:	Principal Component Analysis
CVA:	Canonical Variates Analysis
PCoA:	Principal Coordinate Analysis
cMDS:	Classical Multidimensional Scaling
PLS:	Partial Least Square Regression
FM:	Form Matrix
FDM:	Form Difference Matrix
MATT:	Multiple Alignment with Translations and Twists
HOMSTRAD:	HOMologous STRucture Alignment Database
SABmark:	Sequence Alignment Benchmark
SCOP:	Structural Classification of Proteins
RMSD:	Root Mean Square Deviation
I:	Most influential point
FD:	Form Difference/Form deformation
PPA:	Porcine Pancreatic Amylase
PSI-BLAST:	Position-Specific Iterative Basic Local Alignment Search Tool
PDB:	Protein Data Bank

PFAM:	Protein Families database
GH13:	Glycoside Hydrolase Family 13
ET:	Evolutionary Trace
HPA:	Human Pancreatic Amylase
TMA:	*Tenebrio molitor* α-Amylase
AHA:	*Pseudoalteromonas haloplanktis* α-Amylase
Arg:	Arginine; also abbreviated as ARG or R
Lys:	Lysine; Also abbreviated as LYS or L
Asn:	Asparagine; Also abbreviated as ASN or N
EC:	Enzyme Commission
ML:	Maximum Likelihood
GTR:	General Time Reversible
HGT:	Horizontal Gene Transfer
NPC1:	Nieman-Pick Type C-1
MD:	Molecular Dynamics
PVP:	Proportion of Variable with enough Power
PCS	principal components space
LD:	Linear Discriminants
AFU:	Autonomic Folding Units
CATH:	Class, Architecture, Topology, Homology database
LGT:	Lateral Gene Transfer

RIN:	Residue Interaction Networks
PSN:	Protein Structure Networks
PK:	Pyruvate Kinase
RCN:	Residue contact Networks
AN:	All Neighbours
NSN:	Non-Sequential Neighbours
FA:	Factor Analytic
SEM:	Structural Equations Modelling
GLM:	Generalized Linear Models
PMM:	Phylogenetic Mixed Models
REML:	Restricted Maximum Likelihood
MCMC:	Markov Chain Monte Carlo
CPC:	Common Principal Component
RS:	Random Skewers
BGLMM:	Bayesian Generalized Linear Mixed Models

Acknowledgements

I would like to especially thank my supervisor Christian Blouin, for teaching me more than expected, and for making my PhD experience as enjoyable as possible, for the lunch talks and for introducing me to Hive, Go and GURPS!. I would also like to thank Professors Norbert Zeh and Robert Beiko as well as the members of Dr. Beiko's Lab in Dalhousie University for some helpful suggestions throughout the thesis, as well as the laughs, beers, camps, and all the extracurricular activities, to all of you thank you. To Wilson Chan, Alex Safatli, Kyle Nguyen, Simiao Lu, Tyler Brunet, and Jack Ryan for the collaboration in some parts of the thesis, but mainly for the Pepsi's o'clocks, the 60's Mondays and musical weeks, the laughs, and to made working more entertaining. To Liz Mackay for bearing with my writing and help me editing the manuscripts, and to her and her family for those delicious barbecues and suppers I enjoyed so much. I want to thank Jitka Krejci, for all the editing of my thesis, manuscripts, and letters, thank you. To all my Latin friends, for the talks, the "simposios gastronómicos", among much other things we have done during our stay in Halifax. An special mention to Roisin McDevitt, my angel in the department office, for all the help and how she cared!

Finally I want to thank NSERC for funding this project through the grant No. 120504858, and The Departamento Administrativo de Ciencia y Tecnología - Colciencias (Colombia) for supporting me trough the CALDAS scholarship.

Chapter 1

Introduction

It is often assumed that the information required to produce a properly folded polypeptide is mostly contained in the sequence of the encoding gene (Anfinsen and Scheraga, 1975). However, the protein structure universe is smaller than the sequence universe (Soundararajan et al., 2010). Here, universe is defined as the biologically plausible sequences and structures. The mechanisms through which the two universes differ remain unclear (Tiana et al., 2004). There is a fundamental relationship between protein structure and function (Osadchy and Kolodny, 2011), and a significant effect in the biological processes is due to restrictions in the conformational space in tight cellular compartments (Thirumalai et al., 2010). Thus, the study of evolutionary patterns in protein structures is useful to understand the evolution of function and adaptation as a whole.

Despite the literature produced in recent years about protein structure evolution (Orengo et al., 2001; Kinch and Grishin, 2002; Trifonov and Berezovsky, 2003; Xia and Levitt, 2004; Goldstein, 2008; Sternberg et al., 2009; Finn et al., 2010; Lakner et al., 2011; Fleishman and Baker, 2012; Debès, 2013, and references therein), little is known about how the three dimensional (3D) structures of proteins change over time. As more data are gathered, more principles of protein folding are questioned. There is a need for different methods and models to explain the protein folding patterns and processes. However, the exploration of different methods to evaluate protein structure and its evolution, the evaluation of other modularity levels in protein structures other than the domain, the quantification of the genetic contribution (and therefore the response to selection) of protein structure, and the applicability of a quantitative genetic approach to the study of protein structure evolution, remain open non-trivial issues.

The first question can be addressed by means of geometric morphometrics (GM)

(Macleod, 2002; Zelditch et al., 2004; Slice, 2007, and references therein): statistical analysis of shapes borrowed from biological sciences (Adams and Naylor, 2000, 2003). It allows for an efficient comparison of structures abstracted as shapes, an evaluation of different geometric parameters, and discovery of shape patterns across samples. Exploring the shape space, and therefore structure, will allow us to gain insights into protein structure and function, as well as underlying dynamic and evolutionary relationships.

The second question applies to another area of GM: the phenotypic evolution and morphological integration. The degree of co-variation between parts of a structure (its modularity) can be analysed and studied by means of morphometric methods (Klingenberg, 2009). It can be assessed by analysing the covariance between the variables of a given shape. This allows the evaluation of evolutionary and/or dynamic modules, which can be defined as clusters of internally correlated residues within protein structures. Applying appropriate statistical tests and mathematical tools, such co-varying behaviour can be assessed.

The third and fourth questions are tackled in a comparative quantitative genetics framework (Steppan et al., 2002). This framework unifies the fields of macro-evolution and micro-evolution, and allows for the assessment of how natural selection affects traits (Eroukhmanoff and Svensson, 2011). It is also applicable to protein structures since they are themselves phenotypes (Csaba et al., 2005): They evolve according to selective pressures and have underlying genetics that interact with environmental factors to create the actual structure. Analysing the response to selection of a given structure might address evolutionary questions, and hopefully also be applicable to other problems in biology, biotechnology, medicine and/or systems biology.

This thesis is organized in four thematic chapters (excluding the present introduction and the conclusions chapter):

Chapter 2 talks about the geometric morphometrics (GM) tools applied to protein structures. It gives a general background of GM methods: the application of GM to protein structures, the differences between traditional and molecular applications. Finally I will show applications of such methods in two real datasets: the α-amylase and Niemann-Pick disease, type C1 (NPC1) proteins.

Chapter 3 tackles the protein modularity problem. It gives background development of a clustering method and significance testing, and the results of simulation studies. It ends by explaining the application of this method for generalized 3D shapes.

Chapter 4 discusses further the modular architecture of protein structures, from its putative causes for emergence to potential ways to explore the hierarchy. In this chapter, the domain definition and domain as modules are discussed.

Chapter 5 contains background, modelling, and analysis of the comparative quantitative genetics approach applied to protein structures. In the modelling section the classic quantitative genetics models are modified to be applicable within a protein structures framework. These methods are applied to simulations of phenotypic data, as well as simulation of dynamics. It will be concluded by the application of the methods developed to a subset of the α-amylase data referred to above.

Chapter 2

Morphometrics of protein structures: Geometric morphometric approach to protein structure evaluation

Geometric morphometrics is a collection of approaches for the multivariate statistical analysis of Cartesian coordinate data (Slice, 2007). The "geometry" referred to by the word "geometric" refers to the spatial configuration of the estimation of mean shapes and the description of sample variation of shapes using Procrustes distance (Kendall's shape space) (Rohlf, 2002). The multivariate part of geometric morphometrics is usually carried out in a linear tangent space to the non-Euclidean shape space in the vicinity of the mean shape[1]. It is mainly based on landmarks which are "*discrete anatomic loci that can be recognized as the same loci in all specimens of study*" (Zelditch et al., 2004, p.443), and must: 1) be homologous anatomical loci; 2) not alter their topology position relative to other landmarks; 3) provide adequate coverage of the morphology; 4) consistently be found; and 5) lie within the same plane (Zelditch et al., 2004).

Landmark data is more informative than traditional data since its coordinates also contain positional information and thus geometric structure. Once homologous landmarks are assigned, "noisy" factors affecting the dimensionality and degrees of freedom of the possible shape analysis (such as rotation, translation and size) are stripped by means of generalized Procrustes superimposition (GPS). This is done by (Adams et al., 2004; Zelditch et al., 2004):

1. Assigning homologous landmarks to meaningful and descriptive parts of the shape (Figure 2.1B)

2. Centering each configuration of landmarks at the origin by subtracting the coordinates of its centroid from the corresponding (X or Y) coordinates of each

[1] http://life.bio.sunysb.edu/morph/glossary/gloss1.html

landmark: Translating each centroid to the origin, which removes the positional variation (Figure 2.1C).

3. Scaling the landmark configuration to unit centroid size by dividing each coordinate of each landmark by the centroid size configuration (Figure 2.1D).

4. Setting one configuration as reference and rotating the other configurations to minimize the summed squared distances between homolog landmarks, thus removing rotational variation (Figure 2.1E).

The above method can be expressed as (Rohlf and Slice, 1990):

$$X_i = \rho X H + 1\tau \tag{2.1}$$

where matrix X is the original configuration, ρ is the scaling done to X, 1τ is the translation performed to X_i to a reference position, and H is a rotation matrix with an angle of θ of the form:

$$H = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

Once the above is done, the set of all matrices representing the landmark configurations become the shape space, and its dimensions are given by:

$$KM - (M + 1) \tag{2.2}$$

Where K is the number of landmarks, and M the number of dimensions in each landmark. After removing the effects of size, rotation and translation, $2K-3$ dimensions are left for 2D data and $3K - 4$ for 3D data (Zelditch et al., 2004).

The comparison and analysis is based in the the Procrustes distance (D_P). GPS applies the Procrustes analysis method to align a population of shapes instead of only two shape instances (Dryden and Mardia, 1998). The Procrustes distance is the square root of the sum of squared differences between the positions of the landmarks in two optimally superimposed configurations (C_1 and C_2) (Rohlf, 2002):

$$d_P^2 = \sum_{j=1}^{K} [(x_{j1} - x_{j2})^2 + (y_{j1} - y_{j2})^2] \tag{2.3}$$

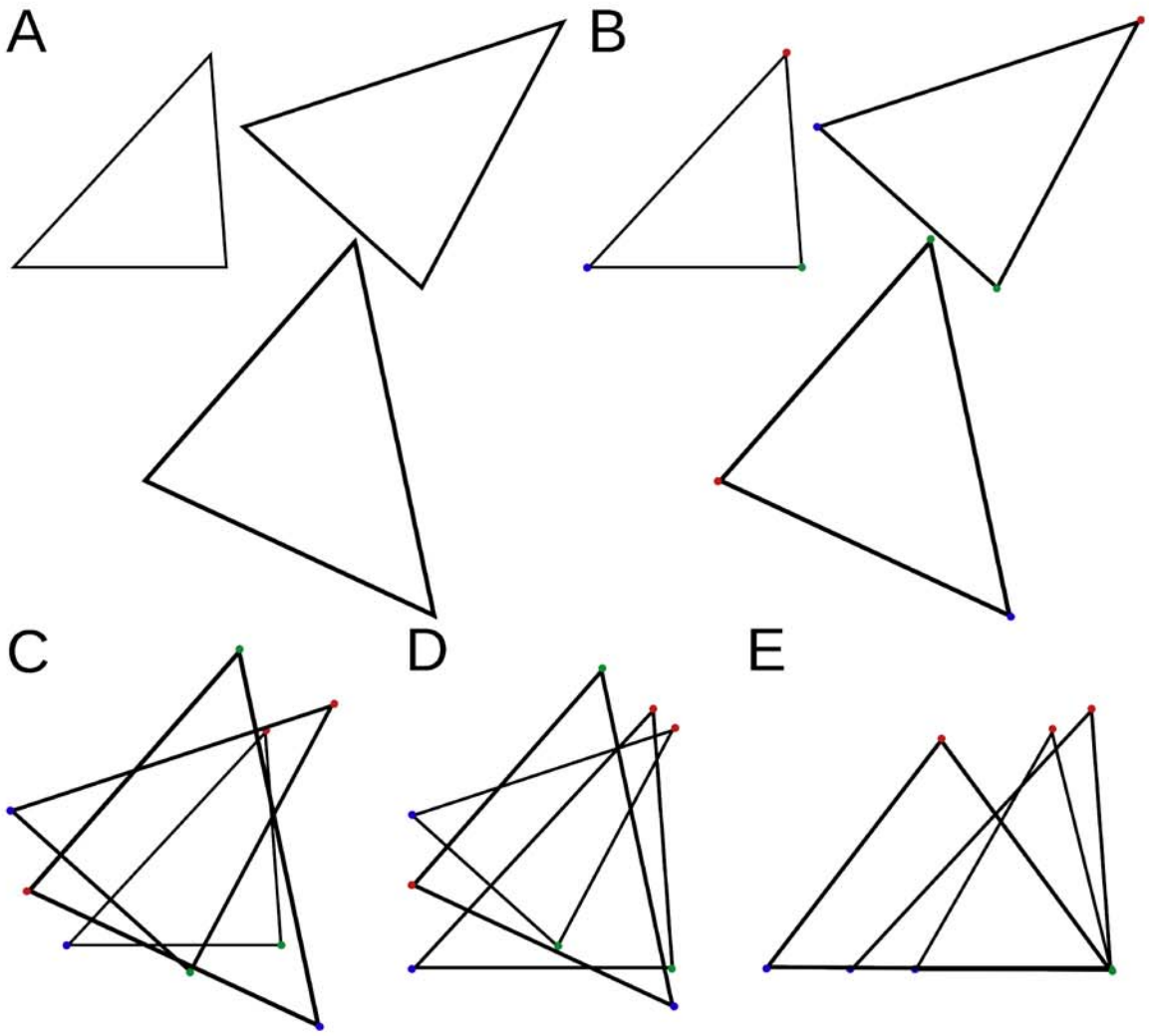


Figure 2.1: Example of landmark superimposition of triangles. A) Original set of shapes; B) Assignment of landmarks by homology (color coded); C) Translating the shapes to its centroid origin; D) Scaling the shapes, and E) Rotating the shapes to minimize the distance between landmarks. The landmarks at each vertex of the triangles are color-coded by homology.

where d_P^2 is the squared Procrustes distance, and K is the number of landmarks. This metric can be then used in several statistical multivariate analysis to attain differences in shapes, clustering, changes in time, test symmetry, etc. Here shapes can be treated as a single point in a multidimensional space, and therefore the information can be summarized in an efficient way. Some traits can also be treated independently in the analyses, extracting information of particular aspects of the shape.

Once comparable variables are set, many multivariate statistics such as Principal

Component Analysis (PCA), Canonical Variates Analysis (CVA), Principal Coordinates Analysis (PCoA) or classical multidimensional scaling (cMDS), Partial Least Squares (PLS), among others, can be used to explore the relationship between observations and between variables.

All these methods can be applied to analyse structures without outliers, but have strong biases when outlying points are included. A sibling field to GM, *Dysmorphometrics* (Claes et al., 2012), can be used to explore the impact of outlier variables. Dysmorphometrics is in summary, "the modeling of morphological abnormalities" (Claes et al., 2012). Such exploration can be performed by means of a corrected maximum likelihood estimates approach (as in Claes et al., 2012) or by means of the Euclidean distance matrix analysis approach (Claude, 2008). In the latter (simpler and less parameterized) approach, an inter-landmark distances matrix (form configuration) is computed using the traditional Euclidean distance for each entry in m dimensions:

$$d(a,b) = \sqrt{\sum_{1}^{m}(a_m - b_m)^2} \tag{2.4}$$

where $d(a,b)$ stands for the Euclidean distance between variables a and b. Therefore the form matrix (FM) is (Claude, 2008):

$$FM = \begin{pmatrix} d_{1,1} & \cdots & d_{i,1} \\ \vdots & \ddots & \vdots \\ d_{1,j} & \cdots & d_{i,j} \end{pmatrix}$$

where i and j are landmarks.

FM is therefore a square symmetric matrix, with zeros in the diagonal and invariant of translation and rotation. With FM computed, different hypotheses can be tested and influential landmarks can be detected (Lele and Richtsmeier, 1992; Lele, 1993; Lele and Cole, 1996). If two forms are identical, they will have the same entries in the FM matrix. We can compute the matrix of differences in form (The form difference matrix or FDM as named by Claude, 2008) between two configurations $S1$ and $S2$ by:

$$FDM_{\frac{S1}{S2}} = FM_{S1} \oslash FM_{S2} \tag{2.5}$$

If by multiplying FM_{S1} by a scalar gives you the second FM_{S2} or *vice versa*, one can tell that both configurations have the same shape, meaning also that all elements of FDM are equal (Claude, 2008). Claude (2008) after the work of Lele and

Richtsmeier (1992), proposed a way to examine the influence of landmarks (variables) in shape difference by calculating the sum of residuals (from the median) for each landmark given the FDM matrix. The landmarks that influence the differences in shape the most would have a higher score which can be mapped to the given shape.

2.1 GM methods in protein structures

2.1.1 Abstracting a protein structure as a shape

A protein fold can be defined as a 3D geometric shape. Sequence analyses help to understand some trends, but explain little about geometry. GM can be used to perform shape analysis from a geometric point of view. It also can be used to give insights into the phylogenetic relationships of the structures rather than the sequences. However, the application of GM to protein structures is not trivial. The scaling component of the Procrustes analysis have no conceptual equivalent for proteins. Since organisms grow, it makes sense to extract the size effect on shape in order to compare young with adults. On the other hand, in proteins the atoms or bonds do not stretch or grow, and therefore the scaling approach (as proposed in Adams and Naylor, 2000, 2003) is not appropriate.

In (Adams and Naylor, 2000) and (Adams and Naylor, 2003) proposals, they:

- Abstract a residue as a landmark
- Evaluate its homology throughout the samples, using ClustalW (Thompson et al., 1994)
- Delete gapped columns
- Perform morphometric analyses

The use of a sequence alignment without structural information to infer structural homology is not appropriate, since the amount of gaps that can be allowed in a loop region can be different than in other regions of the protein (Kann et al., 2005; Kjer et al., 2007), and therefore the definition of structural homology can be different as well. Moreover, since structures are more conserved than sequences, the alignment

based on the structures has more reliable information of equivalent (homologous) residues in more distant clades (Wohlers et al., 2012).

In contrast, I used protein structural alignments which has been worked on extensively (Kolodny et al., 2005; Hasegawa and Holm, 2009; Poleksic, 2011; Joseph et al., 2011; Shibberu et al., 2012). In particular, I used a flexible structure alignment method (MATT; Menke et al., 2008). This strips out rotational and translational information as well as the variability induced by flexible hinges, therefore guaranteeing better fit and homology of the aligned residues.

The abstraction of the residues and landmarks is similar to that in Adams and Naylor (2000) and Adams and Naylor (2003). However, those papers do not fully describe the way the abstraction is made. Here I assign a landmark to the residues' centroids defined by (x,y,z):

$$(\frac{1}{A}\sum_{j=1}^{A} X_j, \frac{1}{A}\sum_{j=1}^{A} Y_j, \frac{1}{A}\sum_{j=1}^{A} Z_j) \tag{2.6}$$

where A will be the number of heavy atoms (C, O, N) that constitute the side chain of a residue including the alpha carbon (C_α). This procedure takes into account only the homologous residues. It captures the variance of both the backbone and the side chain. In the case of glycine, the centroid is the C_α itself.

2.1.2 GPS vs Protein structure alignment

Both GM methods and traditional protein alignment methods strip out the rotation and translation of the configuration. One might ask what are the differences between the two methods. There might be an effect of the scaling process in GPS, the two methods could possibly be used interchangeably, and if so, what are implications of using one versus the other. To test this, the GM methods and traditional analyses have to be done independently but comparatively using the same variables. However, GPS requires the homology information regarding the landmarks. As mentioned before, this homology is estimated by the multiple structural alignment. The GPS is an alignment itself, so providing a starting alignment might bias the result of the superimposition. To cope with this:

1. The protein structures are aligned using MATT software (Menke et al., 2008)

for both approaches.

2. The homologous residues' index for each structure are to be recorded.

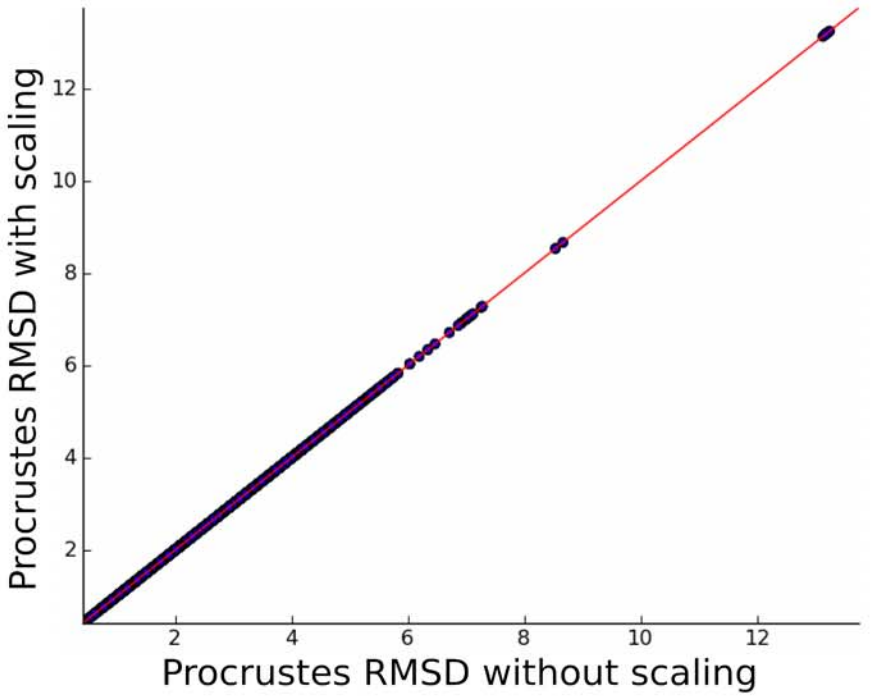
3. For the GPS, the coordinates of the residues corresponding to the recorded indices are used. This process is performed for each independent structure.

The effect of scaling in GPS: Insights from HOMSTRAD and SABmark superfamily databases

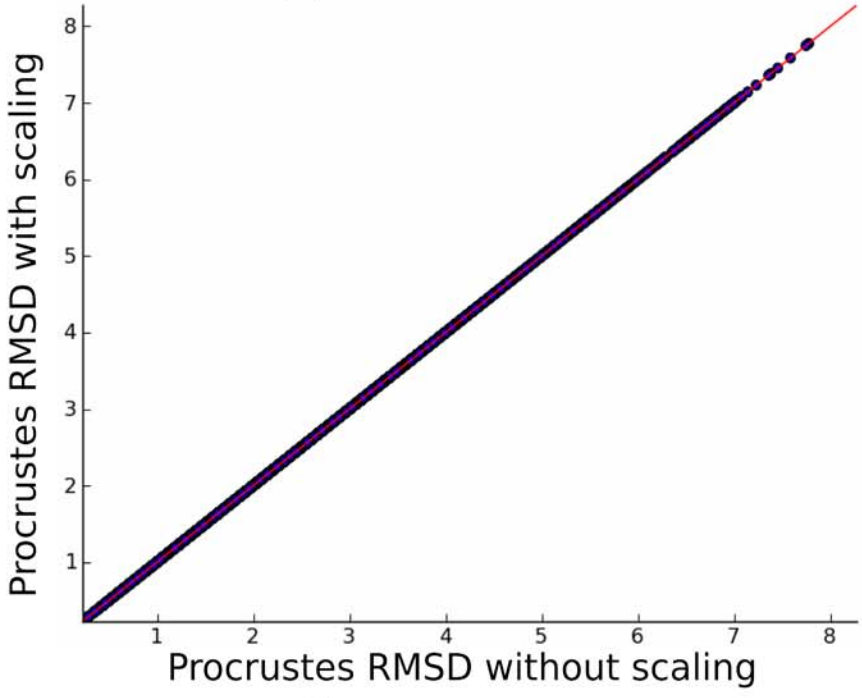
For protein structures it is expected that the scaling does not play a major role in the alignment of structures. To test this assumption, I analysed the HOMSTRAD (386 datasets) and SABmark (425 datasets) superfamily subsets reported in MATT's paper (Menke et al., 2008). The Homstrad database (Mizuguchi et al., 1998) includes structures that are manually curated, guaranteeing their homology and avoiding redundant structures. The Sequence Alignment Benchmark (SABmark) (Walle et al., 2005) includes structures that cover the entire known fold space using SCOP (Murzin et al., 1995) criteria, with the inclusion of un-alignable but apparently similar sequence (Walle et al., 2005). The former database was designed to store structures based on the quality of the X-ray analysis and accuracy of the structure, while the latter database was devised to test multiple alignment problems. Using MATT-reported alignment (Menke et al., 2008, http://groups.csail.mit.edu/cb/matt/) the residue homology was defined. Once the residue index was defined as homologous, the coordinates of the corresponding centroid (see section 2.1.1) for that residue were computed and stored. GPS with and without scaling was performed to the resulting centroid's coordinates. Figure 2.2 shows the results of this comparison.

All datasets (Figs. 2.2a and 2.2b) showed the same behaviour with and without scaling, hence confirming the expectation given the fixed lengths of atomic bonds. From this point further, all GPS analyses made here will be referred to as non-scaled GPS.

At this point it is important to state that given the lack of scaling to a unit size, we remain in the Euclidean space. In GM analysis of 3D shapes, once rotation, translation, and scaling have been stripped of the data, the latter are placed into a new space called Kendall space or shape space (Zelditch et al., 2004). In this space,



(a) Homstrad Dataset



(b) SABmark dataset

Figure 2.2: Effect of scaling in GPA using the HOMSTRAD subsets (2.2a) and the SABmark super family subset (2.2b). X axis corresponds to the Root Mean Square Deviation (RMSD) using GPS with no scaling, while the Y axis corresponds to the scaled GPS.

a set of configurations (group of shapes) have the following dimensions:

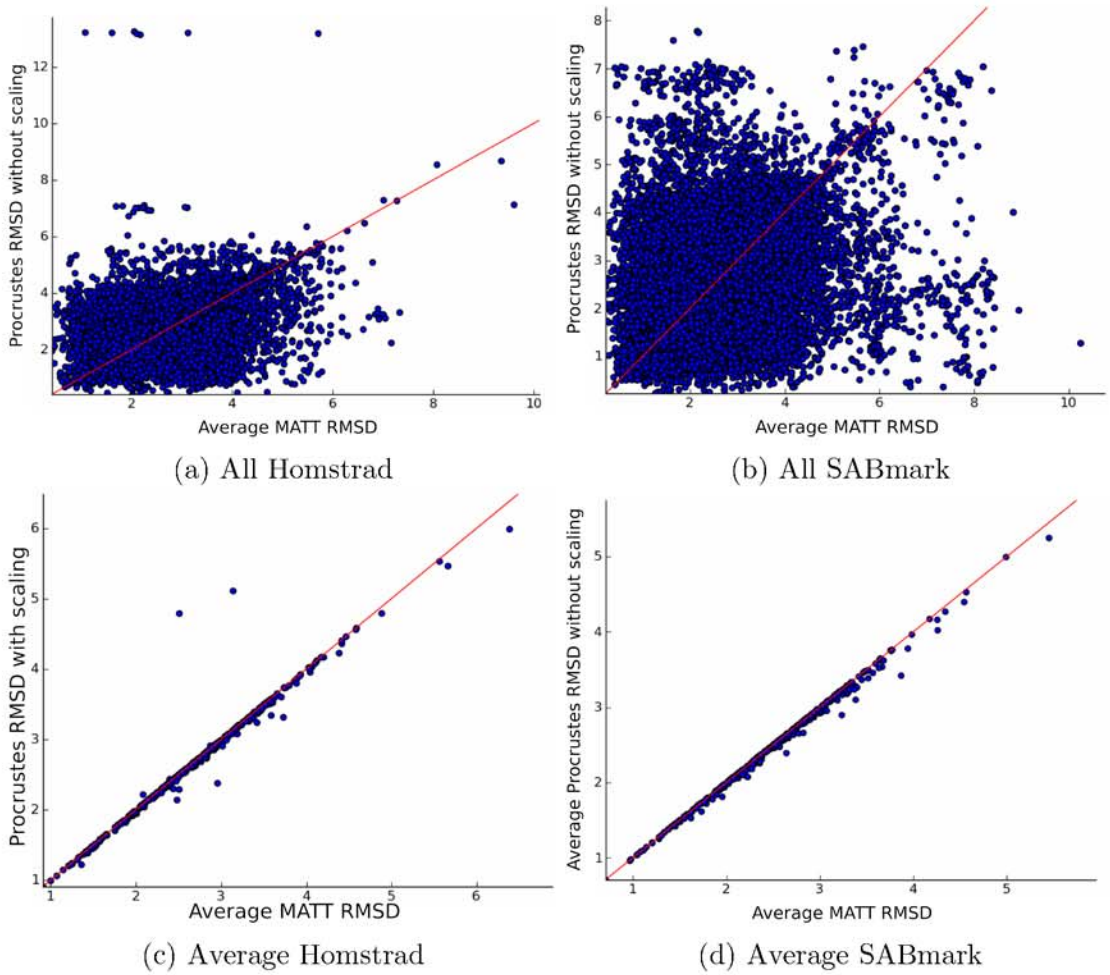
$$D = KM - \frac{M(M-1)}{2} \tag{2.7}$$

K being the number of landmarks and M the number of coordinates per landmark. In the 3D framework three degrees of freedom are lost but seven dimensions are removed (one in size, three in the translation, and three in rotation), thus 3K - 7 are the dimensions for the GM analysis. However, if the scaling is not performed, and the space remains Euclidean, the true dimensionality will be $3K - 6$.

Comparing MATT flexible alignment and GPS: Results from the HOMSTRAD and SABmark super family databases

Since GPS aligns the structures based only on rotation and translation, it is logical to think of it as having a rigid body super imposition alignment. GPS does not allow any deformation of the structure (shape), unlike software, such as MATT, which allows for flexibility. Therefore, it is plausible to hypothesize that a flexible alignment will perform better than GPS, with GPS being similar to a non-flexible alignment (Menke et al., 2008; Konc and Janežič, 2010; Nguyen et al., 2011; Daniluk and Lesyng, 2011; Joseph et al., 2012; Shah and Sahinidis, 2012, among others). Other flexible structural alignment software have shown a slightly better performance than MATT (e.g. Joseph et al., 2012); however, MATT is the only software that reports the alignment on their website (http://groups.csail.mit.edu/cb/matt/). This saves time since protein structure alignment can be time consuming for big datasets. Also, the improvements are not significant (See Joseph et al., 2012) and MATT returns more core residues than most of its competitors, as well as a statistical test of the "goodness" of the alignment (Menke et al., 2008).

There is not a particular trend towards either of the alignment methods evaluated, which differs from the expectations (Figs. 2.3a, and 2.3b). In spite of some outliers, the average RMSDs per dataset in each of the databases are very similar (Figs. 2.3c, and 2.3d). However, the SABmark dataset (Fig. 2.3d) shows a slightly better fit with GPS than with MATT. This is a striking observation since the general belief (some evidence gathered) is that if a protein structure is allowed to bend, the fit will be better. These results can be explained in the following cases:



(a) All Homstrad (b) All SABmark

(c) Average Homstrad (d) Average SABmark

Figure 2.3: Comparison of the alignments using MATT and GPS. 2.3a all by all pairwise RMSDs in the HOMSTRAD subsets. 2.3b all by all pairwise RMSDs in the SABmark super family subset. 2.3c Comparison of average RMSDs per dataset in HOMSTRAD subset. 2.3d Comparison of average RMSDs per dataset in SABmark super family subset.

1. Most of the analysed datasets comprise single domain proteins and therefore there is a smaller probability that the proteins include hinges.

2. Since GPS algorithm uses an iterative fit to the hypothetical mean shape, and given that I am analysing centroids of residues, this approach can give results as good as (or sometimes better on average than) the structure alignment. However, analysis on alpha carbon showed almost identical results (Figure B.1).

3. GPS does not behave as a rigid-body superimposition, and its optimization algorithm performs better without distorting the shape.

4. Since GPS depends on the definition of homology created by MATT, GPS will be functioning as a secondary alignment and an extra optimization.

Giventhis lack of difference on average, I will keep using MATT in the rest of this thesis, since it has to be used in the definition of homology for the GPS.

2.1.3 Form Difference in protein structures

Once aligned, the protein structures abstracted as shapes can be analysed by means of dysmorphometrics. To detect influential residues in a geometric perspective, the form difference matrix (FDM; see equation 2.5) can be analysed. In the context of geometric morphometrics, the FDM accounts for landmarks that have an excessive variation among a pair of shapes. This "excessive variation" will skew the statistical analysis towards the influential point in an effect called "Pinocchio effect". Adding the sum of the differences from the median value per column (variable) and ranking the positions, will yield the most influential point (I) in the data. That could be summarized as:

$$I = max_c(\sum_i |FDM - median(FDM)|_{ic}) \tag{2.8}$$

c being the number of columns and FDM, the form difference matrix.

However, as explained in equation 2.5, this FDM is the representation of the difference between two shapes. We can generalize this by summing the residuals of all shapes versus a hypothetical mean shape, which for simplicity can be calculated as the per-variable per-dimension average. That is, the average of each dimension of each landmark. This approach will then return a Form Difference (FD) value per landmark; however, this value is not bounded and it is difficult to interpret. Also, in the GM context, only the maximum value is important, while the extremely conserved points are of importance for protein datasets. For this reason I scaled the resulting FD vector ($\vec{FD}$) such that it is bounded from -1 (least variation) to 1 (highest variation) by:

$$FD_s = \left(\frac{\vec{FD} - min(\vec{FD})}{max(\vec{FD}) - min(\vec{FD})} * 2 \right) - 1 \tag{2.9}$$

To illustrate how this works, a simulation of 500 hexagons was performed (Figure 2.4). Given an initial shape, for each point and each dimension in the point, a

distribution of random normal numbers is created. This distribution has a given standard deviation (0.05 for regular points, 0.005 for a low variation point and 0.2 for a highly variable point) and the mean was set to 0. This distribution per point and per dimension was then added to the original shape, creating the simulated dataset with controlled variation.

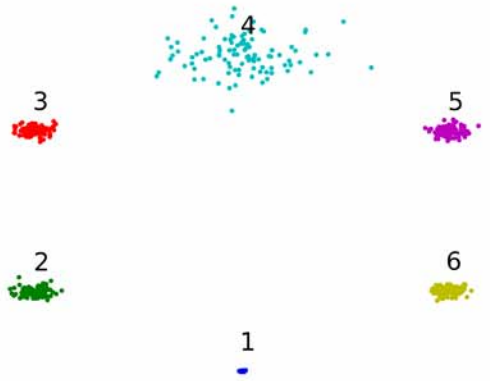


Figure 2.4: Hexagon simulation. Points 2, 3, 5, and 6 have a standard deviation of 0.05, while point 1 (low) has 0.005 and point 4 (high) has an standard deviation of 0.2. This plot was performed using the Python library Matplotlib (Hunter, 2007).

Here, a point with high variation (point 4 in Figure 2.4) and a point with minimal variation (point 1 in Figure 2.4) are introduced, along with four other points exhibiting an average variation without Pinocchio effect of approximately a standard deviation of 0.05. In the context of geometric morphometrics, the Pinocchio effect is biased due to "the distribution landmark error randomly across the configuration, thereby minimizing the overall error by reducing the residual variation around imprecise landmarks and increasing the variation around highly precise landmarks" (von Cramon-Taubadel et al., 2007).

Table 2.1 shows the results for the simulation in Figure 2.4.

As can be seen, the FD_s represents the overall influence of a point in the shape. Negative FD_s suggests that the point is more conserved, with -1 being the most conserved. On the other hand, the more positive the point, the more influence it will

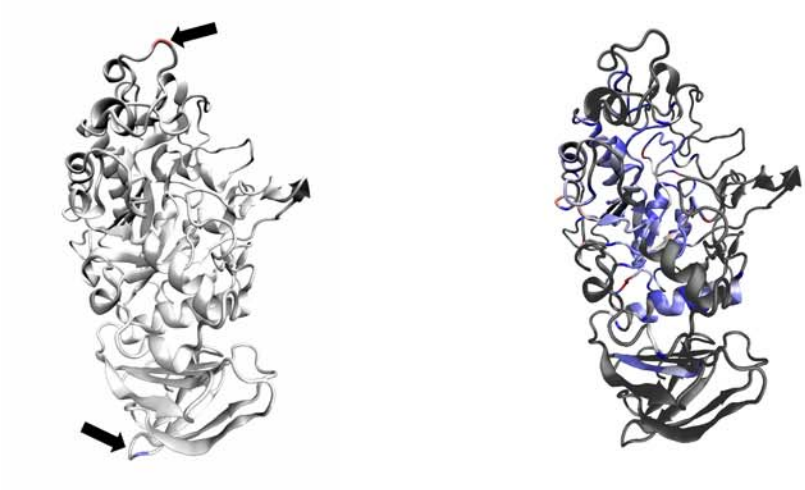
Table 2.1: Scaled FD values for the simulation illustrated in Figure 2.4.

Landmark	Standard deviation	Scaled FD
1	0.005	-1
2	0.05	-0.78
3	0.05	-0.33
4	0.2	1
5	0.05	-0.28
6	0.05	-0.83

have on the shape, with +1 being the most variable. From Table 2.1, one can also see that the "average" points are closer to the least variation than to the highest one, thus displaying a negative tendency. At first sight it might seem trivial to use the FD_s since the standard deviation (sd) seems to correlate with it. However, FD_s represents the influence of a landmark relative to the overall shape, as opposed to the sd which represents the variation at a single variable level. Variables with high sds in a medium-sd neighbourhood will have very high FD_s, while very low-sd points will have low FD_s. This suggests that the sd can be used as proxy for FD. However, in a setting where all the points have high sd, the relation between sd and FD is not as direct. Moreover, for the relationship between sd and FD_s to be proportional, a model of isotropic variation in all dimensions is needed. Such model assumes an equal amount of variation at each landmark and at each dimension in each landmark. It also assumes that landmarks are independent, which is not a fair assumption in most shape analysis (Klingenberg, 2003).

Non-scaled FD can also be used to screen a set of points in a shape to look for Pinocchio outliers, using statistical tests as the Dixon's Q test or the Grubbs' test for outliers.

In protein structures one can compute the FD_s and visualize it in the protein structure coloring. To explore a more complex system than the hexagon, a protein simulation was also performed in the same fashion as the one done for the 2D case. In this case, more extreme points were used for visualization purposes, with 0.0005 being the lowest variation and 0.5 the highest. The protein used for the simulation and visualization was the Porcine Pancreatic Amylase (PPA) with pdb code 1PPI (Figure 2.5a).



(a) Simulation (b) 135 homologs

Figure 2.5: FD_s values mapped in the porcine pancreatic Amylase structure (PDB code: 1PPI). Red represents the highly variable, while blue the least variable. Figure 2.5a is a simulation of the values (The locations of the points are selected at random and do not represent any biological meaning). Here the color scale was offset by 0.5 and the midpoint was set at 0.01 for visualization purposes. Figure 2.5b shows the FD_s for a dataset of 135 structures, gathered with a PSI-BLAST seeded with a PFAM seed alignment. The visualization structure also corresponds to the porcine pancreatic amylase (PDB code: 1PPI). The grey chain corresponds to the non homologous section of the 1PPI with respect to the alignment. Both figures were rendered with VMD v1.91 (Humphrey et al., 1996).

Here we can see which are the residues that contribute the most and the least to the overall shape, as well as their relative position.

This relationship with the geometric variability might not be completely related with sequence variability despite the fact that the centroid is correlated with the size/complexity of the residue it represents. To explore this, I performed a protein structure sampling, seeding a PSI-BLAST (Altschul et al., 1997) search with a PFAM (Finn et al., 2010) seed alignment. The PSI-BLAST search was restricted to structures available at the protein data bank (http://www.rcsb.org/pdb/). There were 135 structures gathered in total (Table A.1 in the appendix) for which homology and membership to the α-amylase family (the Glycoside Hydrolase Family 13, GH13) was guaranteed.

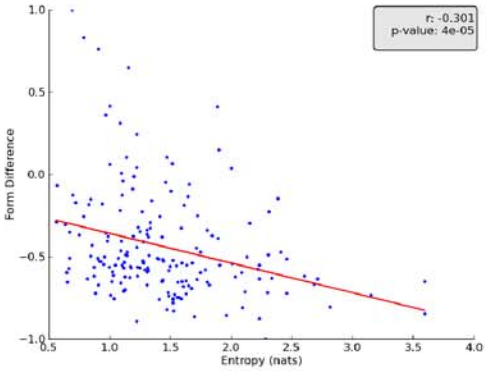
Those 135 structures were aligned using the algorithm proposed by Hleap et al. (2013a) that modifies the pairwise MATT flexible structure aligner (Menke et al., 2008) to complete the multiple structure alignment. This procedure is performed

because the multiple structural alignment version of MATT cannot process this many structures reliably or it requires a prolonged amount of time to finish (Hleap et al., 2013a).

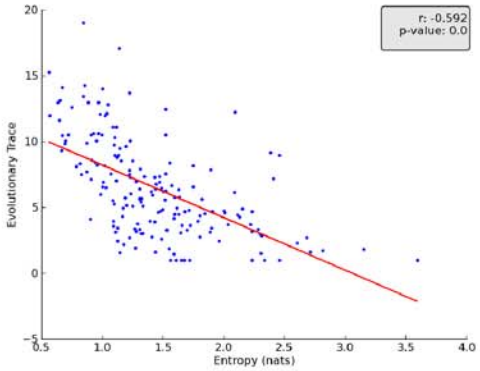
After the alignment, the FD_s was calculated as before. Figure 2.5b shows the mapped FD_s values in the Porcine Pancreatic Amylase (PPA) structure. As expected, most of the highly variable values (and therefore with positive FD_s) are in loops, with the residue 10Arg being the most variable.. This latter residue is also the beginning of the chain after the signal peptide, and therefore its high variability might not disrupt the protein greatly despite being conserved among amylases when analyzed with PDBsum (de Beer et al., 2013). The residue with the least variation was found to be 136Phe which, surprisingly, was also found in a loop. The 136Phe residue is not reported to bind to ligands or to have any catalytic activity. However, this residue is within 15 Å (in a $C_\alpha - C_\alpha$ perspective) of metal and ligand binding residues and it is also highly conserved.

A correlation analysis between the FD_s, entropy, $\overline{\Delta\Delta G}$, and evolutionary trace was performed to test the relationship between the FD_s and relative measures of residue importance. The entropy was computed using the program WebLogo (Crooks et al., 2004) with default parameters. The $\overline{\Delta\Delta G}$ was performed by mutating all residues to Alanine and computing the average difference in energy from the original residue. This procedure was performed using the program FoldX (Schymkowitz et al., 2005), with the default options for the exhaustive replacement of residues to Alanine. The evolutionary trace (ET) estimates a functionality score for each residue important residues based on sequence conservation patterns and their mapping onto the protein surface, generating functional clusters to generate clusters identifying functional interfaces (Lichtarge et al., 1996). The ET scores were computed with the program Evolutionary Trace (Wilkins et al., 2012). The input maximum likelihood tree was previously inferred using FastTree (Price et al., 2010) with WAG as a substitution matrix, and the alignment was provided as stated above.

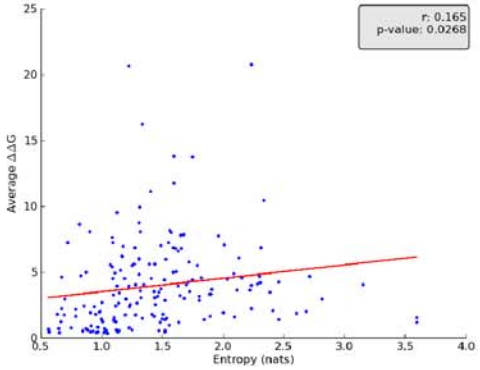
Figure 2.6 shows the relationship between some measurements of residue importance and the FD_s, as well as their co-relationship. There is a weak ($|r| < 0.3$) relationship between the FD_s with entropy and with the evolutionary trace and no relationship with the $\overline{\Delta\Delta G}$. The relationship with the entropy (Figure 2.6a) was



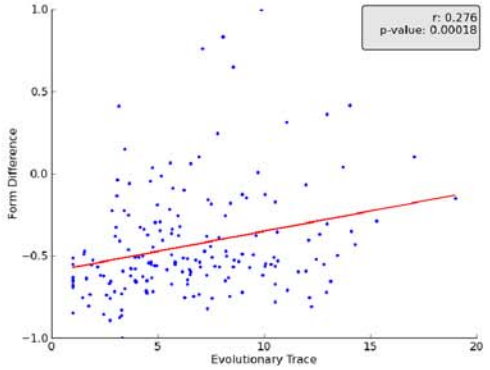
(a) Entropy vs FD



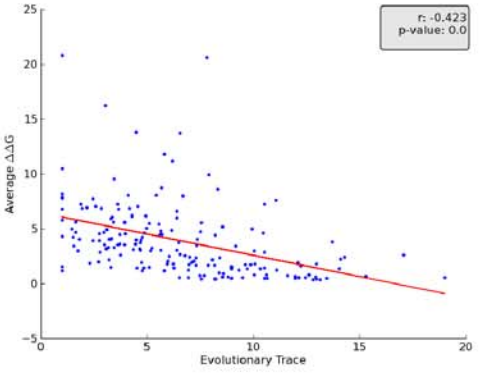
(b) Entropy vs Evolutionary Trace



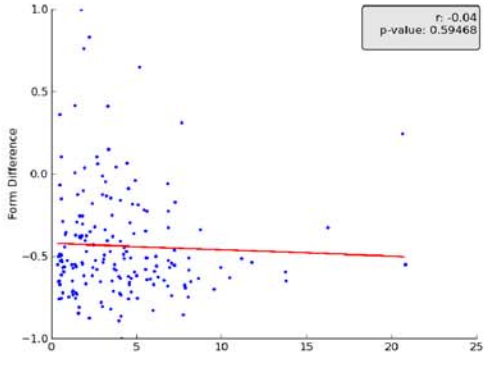
(c) Entropy vs $\overline{\Delta\Delta G}$



(d) Evolutionary trace vs FD



(e) Evolutionary Trace vs $\overline{\Delta\Delta G}$



(f) $\overline{\Delta\Delta G}$ vs FD

Figure 2.6: Correlations between 2.6a Form difference (FD) and entropy; 2.6b Evolutionary trace (ET) with entropy; 2.6c $\overline{\Delta\Delta G}$ (as computed by FoldX) versus entropy; 2.6d FD and evolutionary trace ; 2.6e Evolutionary Trace and $\overline{\Delta\Delta G}$; and 2.6f $\overline{\Delta\Delta G}$ with FD. The correlation values and its p-values are shown in each plot.

expected since there is a correlation between the centroid and the sequence. It was also expected because the FD_s depends on the degree of variability (entropy) in the structure as shown in figures 2.4 and 2.5. However, the weakness of this relationship was not expected. This could be explained by the fact that the abstraction of the structure as a shape (e.g. landmark extraction) includes some information from the sequence, but some structural information might have been confounded. However, I believe that a more plausible explanation is the fact that the structure deformation is following a more complex process than the one that can be explained by $\overline{\Delta\Delta G}$, ET, and entropy themselves.

In the case of the relationship between FD_s and the evolutionary trace (Figure 2.6d), a stronger signal was expected, since the evolutionary trace (ET) measures the functional importance of sites. However, the weak correlation might also be due to the zero-bound values of ET, as can also be seen in its lowered correlation with entropy (Figure 2.6b), where one can see that the bound of ET causes a sub-estimation of the correlation. That is, given that the points tend to aggregate towards the zero boundary, the correlation estimation is affected by an artificial zero slope aggregation towards that boundary. Nonetheless, this under-estimation does not fully explain the low correlation. Another explanation is that functionally important sites will be more geometrically conserved, but there might be higher geometric conservation in other sites related to the geometry of functionally important sites, burying the FD_s value under higher conserved (geometrically) sites.

2.2 Applicability of GM-like methods

So far I have explained how to extract the shape of a protein and enumerate some possible analyses that can be made with this type of data. However, I have not gone in depth into the biological meaning of such analysis. This section will go over two test sets: the homologous alignment of the α-Amylase family and a molecular dynamic simulation of the NPC1 protein.

2.2.1 Statistical analysis of the α-Amylase evolutionary variation

The α-Amylase-like family catalyzes the hydrolysis of α-(1,4) glycosidic bonds of polysaccharides, therefore being classified as glycoside hydrolases (Davies and Henrissat, 1995) in the family 13 (Svensson and Janeček, 2015). It is a multi-reaction catalytic family since its members can catalyze different reactions (hydrolysis, transglycosylation, condensation and cyclization) (Ben Ali et al., 2006). All members of this family share a highly symmetrical TIM-barrel ($(\beta/\alpha)_8$) catalytic domain (Svensson, 1994) (Figure 2.7), including those without any catalytic activity (Fort et al., 2007).

Figure 2.7: Structure of the catalytic domain of the α-Amylase. The TIM-barrel is highlighted. The image was rendered using VMD (Humphrey et al., 1996) and POVray (www.povray.org). The structure used to vizualize this is the PDB 1BF2 chain A from *P. amyloderamosa*.

The TIM-barrel fold is highly versatile and widespread among the structurally characterized enzymes, being present in almost 10% of them (Farber, 1993; Höcker et al., 2001; Wierenga, 2001; Gerlt and Raushel, 2003). There has been a debate about whether the type of evolution that this fold has been through is convergent, divergent or a mixture of both (Farber, 1993), but there is some evidence suggesting the divergent evolution hypothesis is the most likely (Höcker et al., 2001). The catalytic activity and substrate binding residues occur at the C-termini of β-strands and in loops that extend from these strands (Svensson, 1994). The catalytic site includes aspartate as a catalytic nucleophile, glutamate as an acid/base, and a second aspartate for stabilization of the transition state (Uitdehaag et al., 1999). The

catalytic triad plus an arginine residue are totally conserved in this family across all catalysis-active members (Svensson and Janeček, 2015).

The sampling of homologous structures was made as shown in section 2.1.3.

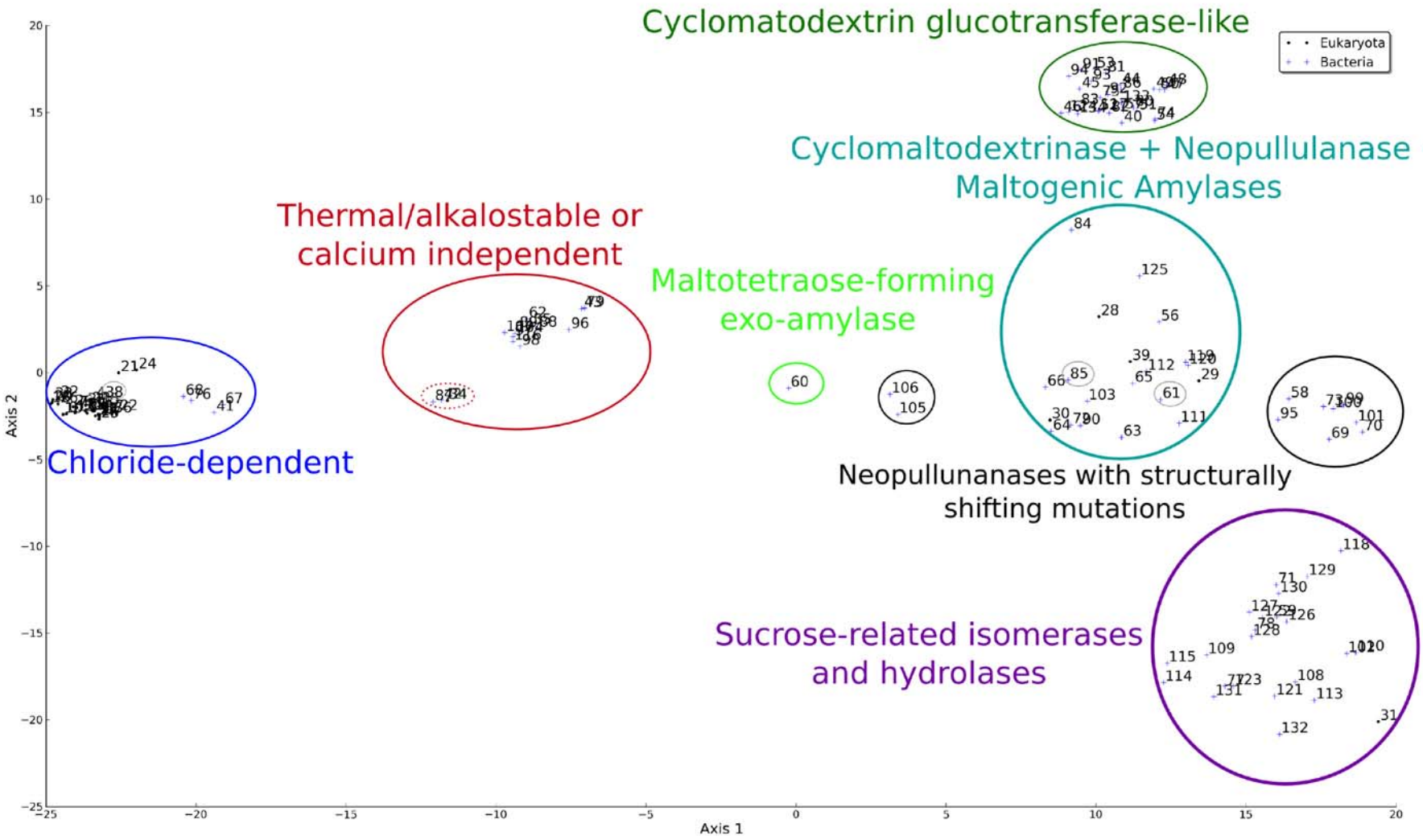


Figure 2.8: Principal Co-ordinates Analysis (PCoA) of 135 protein crystals of the α-Amylase. The circled groups show clusters of structural similarity. The PCoA was performed in R (R, 2011) and modified with a Python script using Matplotlib (Hunter, 2007) library.

Geometry, function, and classification: insights from the principal coordinate analysis

After aligning the structures and applying the methods described in section 2.1, a principal coordinate analysis (PCoA) was performed to the resulting landmark data (Figure 2.8). Analysing the geometry of the protein structures using a PCoA can give us insights into the relationships of such shapes. This procedure tests for differences in the structures being compared, and will show patterns of clustering based on their geometric similarity which in turn might be highly correlated with the functional similarity (Wright and Dyson, 1999).

The PCoA of the multiple structure alignment (Figure 2.8) showed seven distinct and tightly clustered groups:

Chloride-dependent α-Amylases

The first group corresponds to the Chloride-dependent α-Amylases (with amylase function or EC # 3.2.1.1). The similarity among these α-amylases is not a new observation. D'Amico et al. (2000) identified potential chloride-dependent amylases, based on the chloride allosteric activation positives: A) PPA or porcine (*Sus scrofa*) pancreatic α-amylase; B) HPA or human (*Homo sapiens*) pancreatic α-amylase; C) TMA or *Tenebrio molitor* (mealworm) α-amylase; and D) AHA or *Pseudoalteromonas haloplanktis* (before classified as *Alteromonas*) α-amylase. They showed that the side chains of residues Arg195, Asn298 and Arg/Lys337 (PPA numbering) are related to chloride ion binding capabilities (Da Lage et al., 2004).

Thermal/alkalostable or calcium independent α-Amylases

The next tightly defined group in Figure 2.8 are structures that show higher stability in extreme pH and/or thermal conditions or are calcium independent. As shown in Figure 2.8 there is a subgroup of mutants (dotted red oval) with higher structural shift from the main group. In this sub-cluster, three thermo-stable α-amylases (EC # 3.2.1.1) mutants from the genus *Bacillus* can be found. In two of the three cases (3DC0, Rahimzadeh et al., 2012; 1BF2, Fujimoto et al., 1998), directed mutagenesis was performed to increase thermal stability. The 1UA7 represents a mutant of the catalytic site that is not supposed to change

stability or function with respect to the wild type (Kagawa et al., 2003). However, this structure was modeled using 1BF2, and the clustering observed in its structure suggests a higher performance or thermal-stability than other non-chloride binding bacterial amylases. The rest of the group includes α-amylases that exhibits higher thermal/alkaline stability or enhanced efficiency with respect to other amylases of similar function (α-1,4-glucan-4-glucanohydrolase, EC # 3.2.1.1) (Shimi et al., 2008). Most of these structures were created by directed mutagenesis to enhance their industrial applicability by either increasing their thermal or alkaline resistance (Hwang et al., 1997; Machius et al., 1998; Brzozowski et al., 2000; Machius et al., 2003; Lyhne-Iversen et al., 2006; Shirai et al., 2007; Shimi et al., 2008; Alikhajeh et al., 2010) or to make them calcium independent (Prakash and Jaiswal, 2010). There is also a structure with a different enzymatic classification, the maltohexaosidase from *Bacillus licheniformis* (1WP6; glucan 1,4-α-maltohexaosidase or EC # 3.2.1.98). Despite catalysing a slightly different reaction, its native state exhibits higher alkaline stability than other native amylases (Kanai et al., 2004a).

Cyclomaltodextrinase-like α-Amylases

The Cyclomaltodextrinase + Neopullulanase + Maltogenic Amylases group (Figure 2.8) includes enzymes classified in seemingly different functional groups (Cyclomaltodextrinases EC # 3.2.1.54; maltogenic amylases, EC # 3.2.1.133; neopullulanases EC # 3.2.1.135) that can hydrolyze cyclomaltodextrins efficiently (Park et al., 2000) but cannot hydrolyse starch and pullulan as efficiently (Lee et al., 2002). However, Lee et al. (2002) have shown that despite their different enzyme codes, there are no thoroughly documented differences in the literature about their function or structure. They proposed to unify this group under the same enzyme number and the same name (Cyclomaltodextrinases). The result, shown in Figure 2.8, suggests that this is the case given our clustering based on shape. It is important to mention that this Cyclomaltodextrinase group has to be distinguished from the Cyclomaltodextrin glucotransferase group, since those are extracellular enzymes whereas the Cyclomaltodextrinase-like α-Amylases are intracellular (Lee et al., 2002).

Cyclomaltodextrinase-like α-Amylases with structural shifts

The Neopullunanases with structurally shifting mutations groups is a subset of the Cyclomaltodextrinases described above. They carry the same functions (mainly Neopullunanse; EC # 3.2.1.135), but have been subjected to mutagenesis either for binding studies (i.e. 2FH8 and 2FHB ; Mikami et al., 2006) or to inactivate the enzyme using site-directed mutagenesis (Ohtaki et al., 2001; Yokota et al., 2001; Ohtaki et al., 2004; Mizuno et al., 2005). As can be seen in Figure 2.8, even a small number of substitutions cause structural shifts that can be identified by means of a PCoA.

Cyclomaltodextrin glucotransferases-like α-Amylases

This group is composed entirely of bacterial (mainly from the genus *Bacillus*) α-Amylases that catalyze the conversion of starch to cyclodextrins (EC # 2.4.1.19) (Kanai et al., 2004b). As can be seen in Figure 2.8, it is a tightly defined group markedly different from the rest. These differences can be explained by the presence of four aromatic residues that are not present in other amylases and are strongly associated with the protein function (Tonkova, 1998; Kanai et al., 2004b).

Maltotetraose-forming exo-amylase

This is a singleton group, containing the structure 1GCY (Mezaki et al., 2001) from *Pseudomonas stutzeri*. It is a glucan 1,4-alpha-maltotetraohydrolase (EC # 3.2.1.60) that works hydrolyzing amylaceous polysaccharides and removing successive maltotetraose residues from the non-reducing chain ends (Fleischmann et al., 2004). It behaves as an exo-amylase and structural differences with respect to endo-amylases were expected. These differences allow the removal of the residues at the end of the chain instead of just breaking the 1-4 glycosidic linkages. The PCoA in Figure 2.8 expresses these differences by showing a distance between this structure with the "endo-amylases".

Sucrose-related isomerases and hydrolases

This group contains the structures mainly classified as Sucrose glucosylmutases (or Isomaltulose synthase EC # 5.4.99.11)(Fleischmann et al., 2004). However, it also contains three structures (namely 2ZIC, 2ZID, and 4AIE) with Glucan

1,6-alpha-glucosidase (EC # 3.2.1.70) function (Kim et al., 2005; Hondoh et al., 2008; Møller et al., 2012). Of these three structures, 2ZIC and 2ZID have been subjected to directed mutagenesis to improve their catalytic efficiency. This group also harbors an α-glucosidase (2ZE0; Shirai et al., 2008) and an Oligo-1,6-glucosidase (Isomaltase; 1AXH) mutant (Yamamoto et al., 2011). In the rest of structures there are Isomaltulose synthases (EC # 5.4.99.11), including three (4GIN,4GI6, and 4H2C) misannotations: inexistent EC # 5.4.11.99 instead of EC # 5.4.99.11. Despite the somewhat disparity in function they all are classified in the GH13 family (Cantarel et al., 2009; Svensson and Janeček, 2015), and the results shown in Figure 2.8 suggest a high structural similarity.

As can be seen, the principal coordinate analysis of protein structures are tightly correlated with function, and this might give some insights into misannotations or potential functional discoveries. It can be useful for the classification of proteins. This clustering scheme might show an apparent correlation with phylogeny. However, this approach was shown to be sensitive to structural changes. It identified even mutants from the wild type if a structural shift has occurred. This may suggest that this approach is capturing more structural similarities than only phylogenetic ones. It would be interesting to explore phylogenetic signal free variables to test such a hypothesis. This approach seems to be robust to find functional/structural groups.

Phylogeny, function and structural similarity

To support and compare the results from the PCoA (Figure 2.8), a Maximum Likelihood (ML) analysis was performed using `RAxML` (Stamatakis, 2006). One hundred rapid bootstrap replicates (Stamatakis, 2006) were computed to estimate the reliability of the resulting phylogeny. The substitution matrix used was WAG with the GAMMA model of rate heterogeneity. The alpha-parameter was estimated and the empirical amino acid frequencies were used. Figure 2.9 shows the result from the ML search, colored by the EC numbers gathered from the PDB files.

Most of the phylogeny correlates to the function (according to the Enzyme Commission classification) with the exception of the genus *Pseudoalteromonas*. This genus clusters robustly (100% bootstrap) with the animal clade and some bacteria

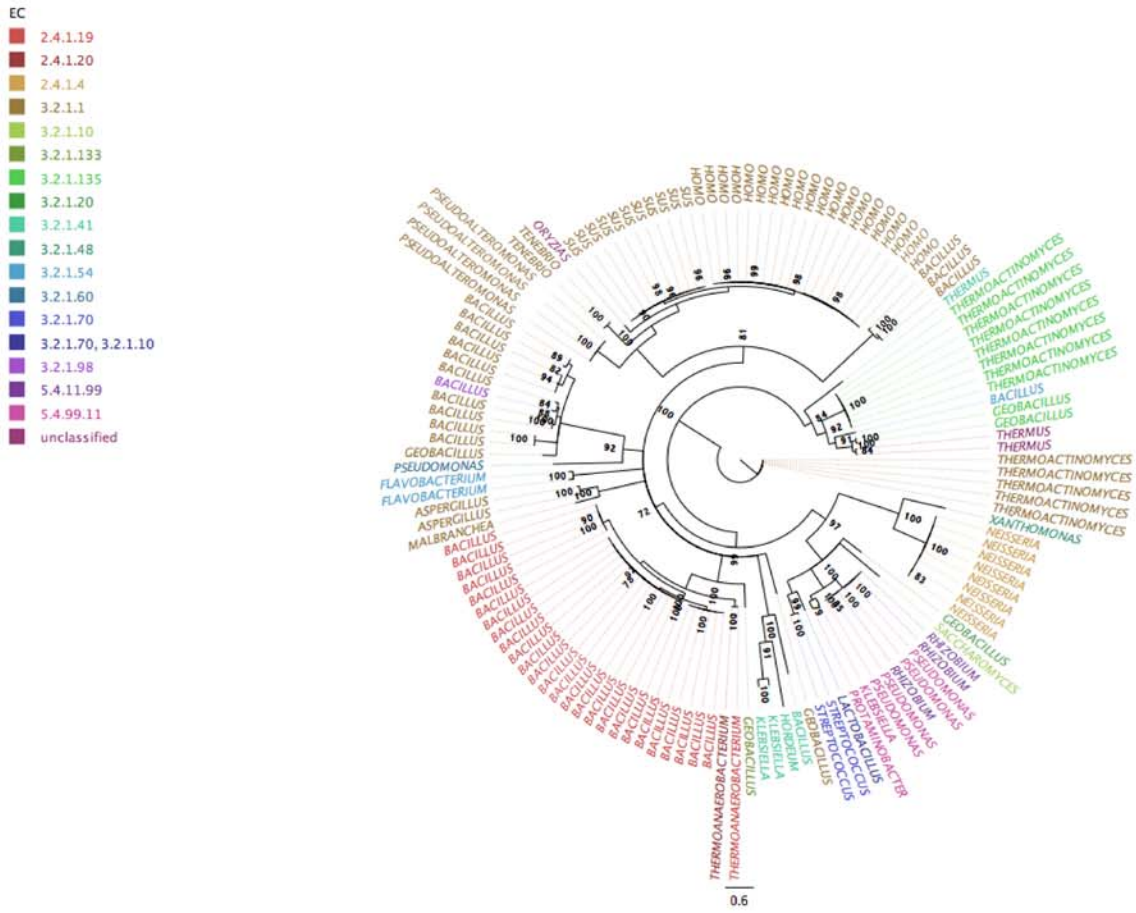


Figure 2.9: 70% consensus Maximum likelihood tree of the amino acid sequences of the *alpha*-amylase dataset, inferred using `RAxML` with 100 rapid bootstraps (Stamatakis, 2006). The matrix of substitutions used was WAG, and the gamma parameters were estimated. The colors represent the Enzyme Commission number (EC) as a proxy for function. This tree was vizualized using FigTree (http://tree.bio.ed.ac.uk/software/figtree/)

that appear in different places on the tree. Despite this, the ML tree topology is highly correlated with the PCoA and all of the groups in Figure 2.8 are represented as monophyletic clades in the tree. Those clades that differed from the PCoA, had bootstrap values under 70%, and therefore appear as collapsed clades in Figure 2.9, and could potentially be placed according the PCoA grouping.

The observation of an Antarctic bacterium (*Pseudoalteromonas*) having a chloride dependent amylase is interesting but it is not new (Feller et al., 1994; Aghajari et al., 1998b), and has been discussed extensively (Aghajari et al., 1998a; Nielsen and Borchert, 2000; Park et al., 2000; D'Amico et al., 2000; Da Lage et al., 2004;

D'Amico et al., 2006; Feller, 2010; MacGregor et al., 2001; Janeček et al., 2013).

Da Lage et al. (2004) proposed that there should be an ancient horizontal gene transfer between a bacterium and a putative animal host. However, *P. haloplanktis* is an Arctic free-living bacterium, and therefore an HGT event is unlikely to happen. However, both structurally and in sequence, the α-Amylase of the psychrophilic bacteria is similar to the animal α-Amylase, supporting the possibility of an ancient HGT. Nonetheless, it is also known that functional and/or structural constraints allow only a restricted range of substitutions at each site. This effectively limits the feasible mutation space and produces different rates at each position (Brocchieri, 2001) and therefore allows different sequences to share high sequence similarity. To test this further, the protein sequences were used as query against the nucleotide database using tBLASTn algorithm (Altschul et al., 1997; Gertz et al., 2006) with Blosum90 substitution matrix to gather closely related sequences available at the GenBank. The sequences were aligned and pruned with TranslatorX (Abascal et al., 2010) (using default options), and only codon positions 1 and 2 were taken into account to avoid mutation saturation noise. A ML tree was inferred with `RAxML` (Stamatakis, 2006) using 100 rapid bootstraps and under the GTR model, estimating the gamma parameter and proportion of invariant sites. This approach is performed under the assumption that the nucleotide sequences will show relatedness between more closely related taxonomic units, since the amino acid sequence might have a higher bias towards the structural constraints and less information available to fully resolve the hypothesis. If the so-called bacterial animal-like α-amylases still branch together with the animal clade, more evidence will support an ancient HGT. My result showed that even taking into account only positions 1 and 2, the mutation saturation is too high to discern the phylogeny for the taxa sampled. With this I cannot disprove the hypothesis that an ancient HGT occurred between an ancestor of the *Pseudoalteromonas* and an animal. More over, my results give some structural evidence supporting this theory.

2.2.2 Statistical analysis of the NPC1 protein simulation

The Niemann-Pick Type C-1 protein (NPC1) binds cholesterol and oxysterols (Infante et al., 2008a) and has an important role in metabolism of cholesterol and some

other lipids. Defects in NPC1 cause malfunction of the cholesterol, sphingolipids, phospholipids, and glycolipids pathways. It is a 1278 residue protein, with 13 membrane helices and three large loops that project to the lumen of lysosomes (Infante et al., 2008a). The first luminal domain is the N-terminal domain, which comprises approximately 240 amino acids. This is a lumen domain (therefore not in the transmembrane region of the protein) and the cholesterol bound to it has an opposite orientation of cholesterol bound to NPC2 (Kwon et al., 2009).

Figure 2.10: Structure of the N-Terminal Domain of the NPC1 protein (PDB code: 3GKH) with the cholesterol bound to it.

The Niemann-Pick type C1 N-terminal domain (NPC1; Fig. 2.10) was simulated in solution using the software `GROMACS 4` (Hess et al., 2008). The force field modes used for the simulations were GROMOS96 for the protein, and the SPCE for the water molecules. Data were collected every two picoseconds for 100 nanoseconds, discarding the first 10 nanoseconds of simulation to achieve stability. This process was performed using a workstation with 24 CPU cores and an NVIDIA TESLA^{TM} GPU. Each sample is treated as an individual observation for the subsequent analyses, and the data are extracted and processed as explained in section 2.1. To avoid biases in rotation and translation in the simulation a superimposition was made. Given that each atom in each snapshot is positionally and structurally homologous throughout the dataset, GPS can be performed without any other structural alignment to achieve homology. For this, the GPS was performed using each atom coordinate as a landmark.

Two simulations were performed: with Cholesterol bound to the structure, and another one without the ligand.

PCA of MD simulations: Insights into the simulation trajectory

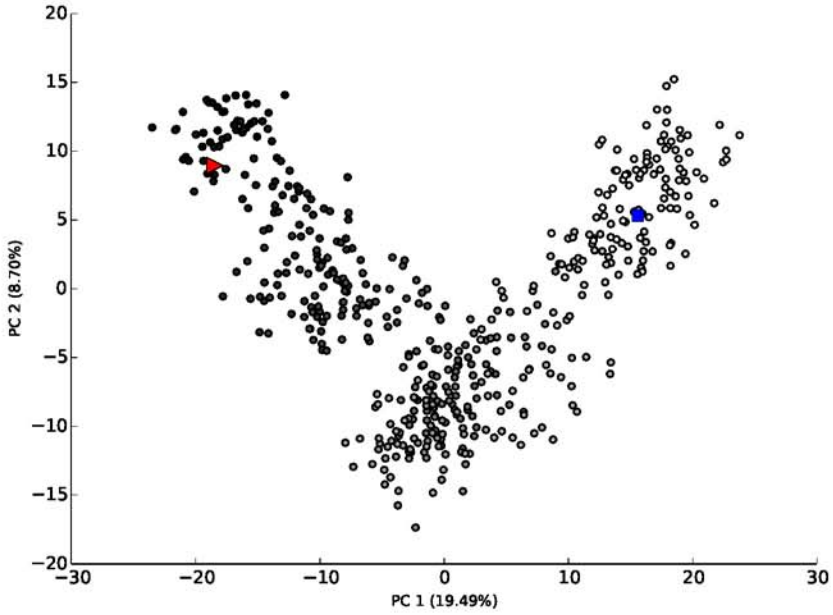
The NPC1 N-terminal domain dataset was explored using a principal component analysis (PCA; Figures 2.11a and 2.11b) to analyse the trajectories as a composite measure of overall structures (principal components).

In both cases it seems that there is more than one state being sampled, and in neither of the cases is the final state close to the starting state. This can be explained by the short length of the simulation or by a short burn-in period. Despite this, with the cholesterol bound to it (figure 2.11b), the structure seems to explore a narrower structural space than the non-bound version (Figure 2.11a). The space exploration here is defined by the movement width in the principal components space (PCS). When the PCS is wider, it implies higher degrees of structural changes. In Figure 2.11a a greater variance in the points can be seen in each conformation state, here denoted as a denser cloud of points. The principal component analysis also showed three somewhat defined conformational states. The simulation with cholesterol bound to the structure, Figure 2.11b, show at least two major conformation spaces and a transition between them. However, in the latter simulation, the variation components are smaller.

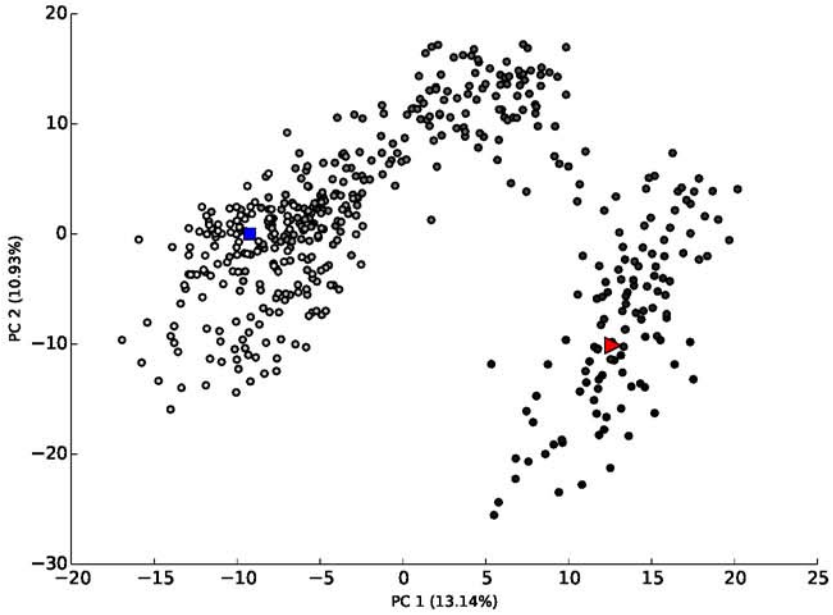
FDM analysis of the NPC1: Insights into cholesterol binding

To check the overall contributions of each of the residues to the deformations during the simulation, the equation 2.5 was applied and the result mapped in the PDB. Figure 2.12 shows that once the cholesterol is bound to the NPC1 (Figure 2.12b) most of the highly movable (and therefore higher FD_s) residues are not contributing to the deformation, seemingly to be held in place by interactions with the ligand.

In Figure 2.12a the opposite behaviour can be seen. When cholesterol is not bound to the structure, the residues in charge of the cholesterol intake/outtake are highly movable and therefore responsible for most of the deformation found in this protein during the simulation.

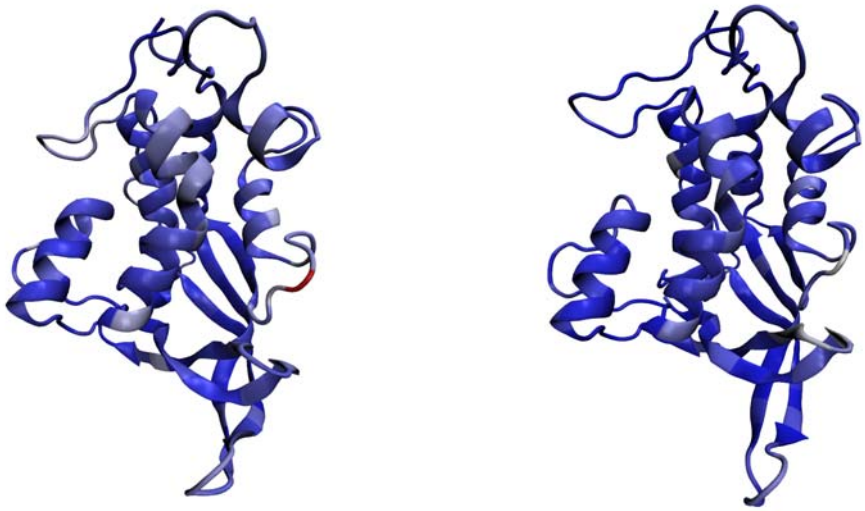


(a) No cholesterol bound.



(b) Cholesterol bound.

Figure 2.11: Principal component analysis of the NPC1 N-terminal domain simulated in solution without (2.11a) and with (2.11b) cholesterol bound to it. The gray scale is proportional to the time of the simulation getting progressively lighter as the simulation develops. The right red triangle depicts the start of the simulation, while the blue square depicts the end.



(a) Whithout Cholesterol (b) With Cholesterol

Figure 2.12: FDM analysis of the NPC1 N-terminal domain (PDB code 3GKH) with 2.12b and without 2.12a ligand. The color represents the FD_s, red being higher score and blue the lowest. The midpoint of the gradient was set at 0.01, and offset by 0.5 for better visualization. In this particular case, the negative FD_s values (therefore the least variable) dominate the scale and the most influential is shown in white, and fades with other positive values.

These results show that the GM-like methods, such as FDM analysis, can be used to identify possible interesting sites for either protein engineering or biomedical applications.

Chapter 3

Defining structural and evolutionary modules in proteins: A community detection approach to explore sub-domain architecture

Jacob (1977) stated that "*Nature is a tinkerer and not an inventor*". Proteins are no exception to this rule, since domains are accepted as the evolutionary modules of proteins, and modular reuse has been demonstrated in all domains of life (Voigt et al., 2002). This modularity confers protein structures with enhanced flexibility (Del Sol et al., 2007) and might influence their ability to respond to selection. This ability is a main concern for evolutionary biology and is related to the robustness of a system (Pigliucci, 2008; Wagner, 2008). Robustness is the ability of a system to maintain its function under perturbations. In a robust system heritable phenotypic variation neither increases nor decreases under disturbances (Wagner, 2008; Rorick and Wagner, 2011). In the protein world, the phenotype is the structure and the phenotypic variance is given by slight variations in protein structure.

In organismal biology, the coordination of subunits within a whole, e.g. mammalian limb bones, floral and leaf traits, parts of wings, individual organs, etc., has long been known as morphological integration (Cheverud, 1996b), which was renamed by evolutionary developmental biologists as modularity (Klingenberg, 2009, and references therein). The modularity of a system is a property that is closely related with both evolvability and robustness (Rorick and Wagner, 2011; Rorick, 2012). This property allows a system to increase its evolvability by diminishing adaptative constraints as well as giving the system the possibility for plasticity and

A modified version of this chapter has been published in BMC structural biology: "Hleap, J.S., Susko, E. & Blouin, C. 2013. BMC Structural Biology 13:20, DOI: http://dx.doi.org/10.1186/1472-6807-13-20". The simulations of protein structures exposed here are not included in the article and some real datasets are used instead.

the emergence of novel functions by rearranging the modules (Rorick, 2012). As stated by Klingenberg (2009), integration and modularity are concerned with the degree of covariation between parts of a structure. It is important, from an evolutionary viewpoint, to determine whether a structure is a single unit or consists of several modules. In molecular biology, the modularity of systems has been used to an extent, but more work has been done in systems biology (Kitano, 2002; Gavin et al., 2006; Popescu and Popescu, 2011; Fraser et al., 2013) including analyses of metabolic networks (Holme, 2011; Yamada and Bork, 2009; Takemoto and Borjigin, 2011; Zhou and Nakhleh, 2012), cell signaling networks (Sudol and Harvey, 2010; Pan et al., 2012; Tran et al., 2013), and protein interaction networks (Taylor et al., 2009; Kim and Tan, 2010; Seebacher and Gavin, 2011; Taylor and Wrana, 2012; Di Paola et al., 2013). In the context of protein architecture, modularity has been used to refer to modules of exon shuffling (Patthy, 1999; Xing and Lee, 2005), and complexes of enzymatic machineries (Gavin et al., 2006). Some approaches to protein structure modularity have also been explored (Gherardini et al., 2010; Rorick and Wagner, 2011; Rorick, 2012), showing modules as domains (Murzin et al., 1995; Andreeva et al., 2008) and also as sub-domain components (Berezovsky and Trifonov, 2001; Fedorov et al., 2001; Gelly et al., 2006; Ahnert et al., 2010). However, the criterion to define protein modules depends on the definition of a proper quantitative treatment, which is not a trivial problem (Rorick, 2012).

There have been different attempts to identify modules in protein structures (Rorick, 2012) such as highly conserved close loops (Sobolevsky et al., 2007), foldons, and autonomous folding units (Haglund et al., 2012). Some of the aforementioned modules can only be identified experimentally and/or in single proteins. Another particularly robust way is to perform modular decomposition by using community detection algorithms (Girvan and Newman, 2002). This approach has been applied extensively in systems biology (Kitano, 2002; Gavin et al., 2006; Holme, 2011; Taylor et al., 2009; Yamada and Bork, 2009; Kim and Tan, 2010; Sudol and Harvey, 2010; Popescu and Popescu, 2011; Seebacher and Gavin, 2011; Takemoto and Borjigin, 2011; Taylor and Wrana, 2012; Pan et al., 2012; Zhou and Nakhleh, 2012; Di Paola et al., 2013; Fraser et al., 2013; Tran et al., 2013) as well as to the protein structure modularity identification problem (Del Sol et al., 2007; Feldman, 2012). However, most of these

attempts only consider the contact matrix (Del Sol et al., 2007; Feldman, 2012). This approach bears no evolutionary information and depends exclusively in the definition of contact between residues (Rorick, 2012). Here, I postulate that correlation information across a group of homologous structures, or a group of snapshots from a molecular dynamics simulation, is more relevant than molecular contact alone.

The analysis of graphs has become crucial to understanding the features of different systems (Fortunato, 2010) such as community structure (Clauset et al., 2004). Several clustering algorithms have been developed (for a review on such algorithms see Fortunato, 2010) and applied successfully to different kinds of networks, such as networks of email messages (Tyler et al., 2003), biological, and social networks (Girvan and Newman, 2002; Del Sol et al., 2007; Novák et al., 2010; Feldman, 2012). However, all clustering techniques, including the graph-based ones, lack a statistical framework to determine the significance of the inferred clusters. This may lead to results that may not be biologically meaningful. In this chapter I present a graph theory-based clustering method that includes a test of statistical significance, a power test, and a test for the accuracy of the estimates given the sample size (i.e. bootstrap). To do this, a permutation-based t-test to assess statistical significance and a power test based on Cohen (1988) to assess the reliability of the estimates are proposed. Also, a bootstrap test and a power analysis to infer cluster robustness are developed. These tests are applied to coordinate data, but can be generalized to other applications. Here, a module is defined as any group of residues that has significant correlation within the group, i.e. among residues within the group. The correlation within the group also has to be significantly higher than the one obtained when correlating these residues with residues of other groups in the dataset.

To explore the relationship among residues in a protein, a correlation graph can be built and its properties can be used. To do this I will need to:

1. Extract the coordinate information as landmarks (Section 2.1.1), and estimate the residue contact matrix (Section 3.2).

2. Create a graph where each landmark is a node and these nodes are connected if significant correlation among them is found (Section 3.3).

3. Test if the partition of the data (grouping of residues) is statistically significant

(Section 3.4).

4. Test for statistical power of each partition (Section 3.5).

5. Test for the stability of the partition to sample size: Bootstrapping (Section 3.6).

Each of these steps are explained further.

3.1 Landmark definition

As defined in section 2.1.1, a landmark is a point, vertex, site or control point in a shape object (protein or simulation object in my case) that can be found repeatedly, and consistently, in a group of such objects (Dryden and Mardia, 1998). Here, we define a landmark as the centroid of homologous residues in a multiple structure alignment. The residue centroid is used to include both sequence (residue side chain) and geometry, as opposed to only the geometry of the backbone. Equation 2.6 shows the computation of the landmarks for a 3D shape such as protein structures.

3.2 Contact definition

3.2.1 In 2D simulation datasets

In structured (shape-defined) datasets, a contact matrix can be inferred. Each landmark in a given configuration is said to be in contact with any other landmark in the dataset if the distance between a given pair of landmarks is not greater than one unit plus the standard deviation of the simulation. This holds true only if the shape being constructed lays on a grid of one unit per tick.

3.2.2 In protein structures

Inter-residue contact maps are a widely used approach to analyze protein structures (Faure et al., 2008). They are also important to understand protein folding and stability (Punta and Rost, 2005), and to identify residues playing structural and/or functional roles (Faure et al., 2008). Despite this, and the advances in the contact

definition (Faure et al., 2008, and references therein), accurate contact map predictions are still mainly unsolved. There are some proposed tests (Faure et al., 2008) and software (Yuan et al., 2012), but they are mainly using C_α-C_α or C_β-C_β distances with a threshold of about 7 to 8 Å (Faure et al., 2008; Yuan et al., 2012). However, these types of contacts are a mere approximation to true contacts. Here, I defined a contact between any two residues if the distance between them is equal or less than 4.5 Å in an all-atom (all side chain atoms) contact analysis. The all-atom approach is more accurate since it takes into account the distance between each possible pair of atoms in two side chains. This approach is recommended in real datasets, since it reduces the number of edges in the graph and is a more naturally plausible definition of subsets in protein structures.

3.3 Graph construction

Assume that one has a dataset made of n observed protein structures. For each of these structures the input data matrix is composed of k landmarks. Here, a landmark is defined as the Cartesian coordinates in three dimensions of the centroid of a residue. This centroid is calculated using the residue's side-chains (see section 3.1). To deal with dimensionality, the original data matrix is split into its components (X,Y,Z) and, for each dimension, a correlation matrix between landmarks is computed. For each entry in each dimension, I test the significance of the correlation coefficient. This coefficient is set to 0 if it meets the following criteria:

$$\frac{1}{2}\log\left(\frac{1+r}{1-r}\right) < \frac{Z_\alpha}{\sqrt{(n-3)}} \tag{3.1}$$

where the left-hand side of the equation 3.1 is the Fisher transformation of the estimated correlation r. The right-hand side of the equation 3.1 is the critical value for an alpha-level test of the null hypothesis that the correlation is 0. There, the Z_α is the standard score which allows the calculation of the probability of a value occurring within our normal distribution and compare scores from different distributions. In this thesis the α value used was 0.05.

This step is done to simplify the graph building process such that insignificant correlations are ignored.

The overall magnitude of the correlation vector is calculated as:

$$\Xi_{ij} = \sqrt{\sum_{p=1}^{3} P_p^2} \tag{3.2}$$

where the value for the pth dimension, P_p, is either r or 0 depending on the result of equation 3.1. The Ξ_{ij} is obtained for each pair of landmarks and assigned to the edges of an undirected graph S, using the `Python-igraph` library (Csardi and Nepusz, 2006).

The summation in equation 3.2 is performed to agglomerate the dimensions (dimension reduction minimizing information loss). Since r is not additive and r^2 is, the sum of r^2 is the appropriate way to add the correlations without violating non-additivity. Also, Ξ is the correlation vector magnitude that guarantees that if there is any correlation in any of the dimensions, Ξ_{ij} will include it, regardless of the vector direction. Let us assume that a given residue is highly and significantly correlated in the X axis, but poorly and/or not significantly correlated in Y or Z axes. Ξ_{ij} will reflect such correlation since the residues must behave completely independent for Ξ_{ij} to be zero or close to zero. By doing this, a sensitivity to big rotations might be created. However, the input data is a set of aligned structures (with a structural aligner, such as MATT in the case of real datasets) and therefore any rotation, translation, and (if the alignment method allow twists) natural deformations are dealt with.

3.3.1 Graph abstraction

Let $S = (N, f)$ be an undirected graph, where N is a list of nodes (landmarks), and f is a function $f : N \times N \to \Bbbk$ that assigns an edge weight to each landmark pair. An edge E_{ij} is assigned only if $\Xi_{ij} > 0$, *and if* the residues are in contact (if the restriction is enforced). The edge weight value is set to Ξ_{ij}.

3.3.2 Community structure or clustering optimization

With the defined graph, the community structure is assessed using a fast-greedy approach, since it is an efficient way to detect clusters (Clauset et al., 2004). Clusters are defined by finding the partition of landmarks that maximizes the modularity

index (Q) (Newman, 2004):

$$Q = \frac{1}{2m} \sum_{vw} \left[A_{vw} - \frac{\sum_w A_{vw} \sum_v A_{vw}}{2m} \right] \delta(C_v, C_w) \tag{3.3}$$

where m is the number of edges in the graph, A_{vw} represents the weight of the edge between vertices v and w, $\sum_w A_{vw}$ and $\sum_v A_{vw}$ are the weighted degree of a vertex (v or w), defined to be the sum of the edge weights of the adjacent edges for each vertex. C_v and C_w are communities to which the vectors v and w belong to, and δ is a binary function where $\delta(C_v, C_w)$ is 1 if $C_v = C_w$ and 0 otherwise.

The modularity index (Q) is then the proportion of edges shared within groups minus the expected proportion if edges were distributed at random. For a given partition, Q indicates the density of nodes within groups when compared against a random distribution of connections regardless of the partition. Q ranges between -1 and 1. If positive, there are more connections inside the module than expected by chance and therefore a possible community structure exists (Newman, 2004, 2006) (i.e. partition or clustering of the data). In my case, a partition made by the optimization of Q is a group of residues that correlate in space (i.e. move together) given the sample. If the sample is across homologous proteins, a cluster or partition represents a concerted movement in the evolution of the protein. Sampling across molecular dynamic simulation snapshots represents parts of the protein that are moving together in solution.

The output is a membership vector that corresponds to the community structure (partition or clustering) in the graph of landmarks. It is interpreted as a set of clusters which number is given by the optimization procedure and therefore there is no need for an *a priori* determination of the number of clusters to be obtained. Each cluster is assumed to be a putative module, but this membership vector provides no support or information about its statistical robustness and significance.

3.4 Statistical significance test of clusters: Controlling the false positives

Despite the usefulness and ubiquity of tests using similar algorithms, the question of significance of clusters is critical since there is no support for the obtained clusters,

and therefore its validity is questionable. To test if each cluster is significant, a permutation t-test (Good, 2000), as implemented in `R` (R, 2011; Maindonald and Braun, 2011), is applied.

The rationale for the test is based on the definition of cluster as an entity where the distribution of correlations of the elements inside the cluster (*intracorrelation*) is significantly distinct from the distribution of correlations with elements from other clusters (*intercorrelation*). This test is applied for each possible pair of clusters defined by a membership vector. For a given pair of clusters, we compare the distribution of the intracorrelation for that cluster with the distribution of intercorrelations for this pair. If one cluster is artificially broken down by the clustering algorithm, there should be no significant differences between the distribution of intra and intercorrelations.

Because the test is performed for a number of pairs, multiple comparisons are made. Let $M(A)$ and $M(B)$ be the mean intracorrelations for two clusters A and B, found by the community detection algorithm. Let $M(AB)$ be the mean intercorrelation. The null hypothesis is then $H_0 : M(A) = M(AB)$. With more than two clusters the number of comparisons (K_C) will be $K(K-1)$, K being the number of clusters. If a single-inference procedure is used, a false increased significance can result, which I correct for using the Benjamini-Hochberg False Discovery Rate correction (FDRc) procedure (Benjamini and Hochberg, 1995).

For example, a given set of homologous proteins is analyzed with this method and a possible partition is obtained. This will give different pieces of the protein that correspond to groups of residues that are correlating (moving together) more within each cluster than among clusters. We use the correlations inside a given group and test against the correlation that exist between that group and other groups. If there is no significant difference, both entities are moving together and therefore should be merged.

3.4.1 Refinement of the membership vector

The results of the significance testing are summarized into a new graph. Let graph $S = (C, E)$ be a directed graph, where C is a list of inferred clusters, and E a list of assigned edges. There will be a directed edge from cluster C_u to cluster C_v if the

hypothesis that $M(u)$ is distinct from $M(uv)$ cannot be rejected. If C_u and C_v are connected by a bi-directional edge, they are merged into a single cluster. The process is iterated until no clusters can be merged.

Following the example in the previous section, let's assume that the protein dataset analyzed was partitioned into 4 groups of residues (A,B,C, and D). Each of those groups will be the vertices (nodes) in a new graph. I will draw an arrow if there are no significant differences between a given group and another group (e.g. correlations within A are not significantly different than the correlations between residues in A and residues in B). If this is reciprocal (e.g. correlations within B are not significantly different than the correlations between residues in B and residues in A), both groups of residues are merged.

3.5 Statistical power test of clusters: Acknowledging the false negatives probability

The above statistical test assesses false positives (Type I error). It is important, as well, to assess the strength of association between members of a cluster. To determine the minimum resolvable correlation for a given sample size, and for a given significance and power, let ρ_{res} be the correlation that can be resolved with a power of $1-\beta$, and a significance level of α. Given the number of observations n, as suggested by (Cohen, 1988) and implemented in the `R` package `PWR` (R, 2011; Champely, 2009), let γ be a function of i and j:

$$\gamma(i,j) = \begin{cases} 1 & \text{if } r_{ij} \geq \rho_{res} \\ 0 & \text{Otherwise} \end{cases} \tag{3.4}$$

where $r_{i,j}$ is the correlation coefficient between landmarks i and j. To assess the power of a candidate cluster C with c elements, we estimate the proportion of correlation values between landmarks of C that are larger than ρ_{res}. For each C the proportion of variables with enough power (PVP) is thus:

$$PVP_C = 2\left(\frac{\sum_1^p \gamma(i,j)}{(c^2-c)}\right) \tag{3.5}$$

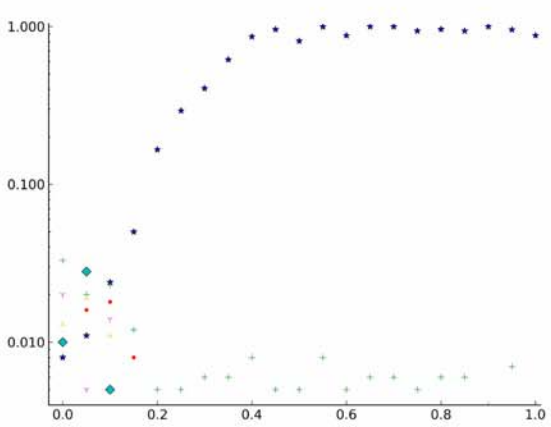
where p is the number of pairs i,j in cluster C.

Here, PVP_C is the estimated PVP which should be distinguished from the true PVP, that arises when the estimated r_{ij} in equation 3.4 is replaced by the actual ρ_{ij}. PVP_C provides qualitative information to help interpret the results given the used sample size. Figure 3.1 shows the behavior of the PVP in the intracorrelations evaluated for 85 (Figure 3.1a), 1000 (Figure 3.1b), and 5000 (Figure 3.1c) observations. Even in simulated data, PVP deviates from the possible values of 0.0 and 1.0 when the number of observations is small.

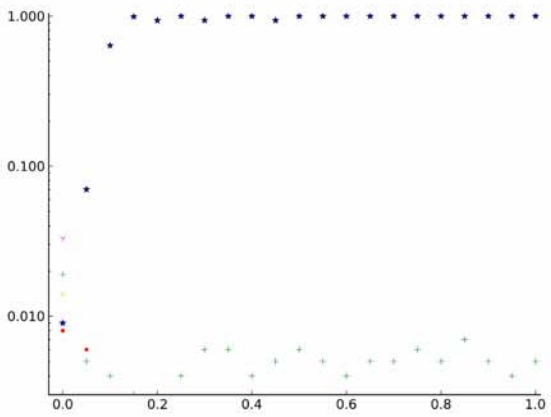
For instance, take a cluster (group of residues from the previous example) A that contains 10 elements, and 45 entries in the upper triangle of its correlation matrix. Assume that A was inferred with 100 observations (protein structures from the example). With that sample size, ρ_{res} will be approximately 0.28 with a power of 0.8 and a significance level of 0.05. If two thirds of the entries in the upper triangle of the correlation matrix of A are below ρ_{res}, PVP_A will be equal to 0.66. In other words, for 30 entries of the correlation matrix I estimate that there was a power of 0.8 or greater. If there are clusters created by optimizing the modularity score (Q) using weakly correlated landmarks, this cluster's PVPs will tend to be close to 0. This test is post-hoc, and is only to inform about the robustness of the partition created.

3.6 Bootstrapping: Measuring the accuracy of sample estimates

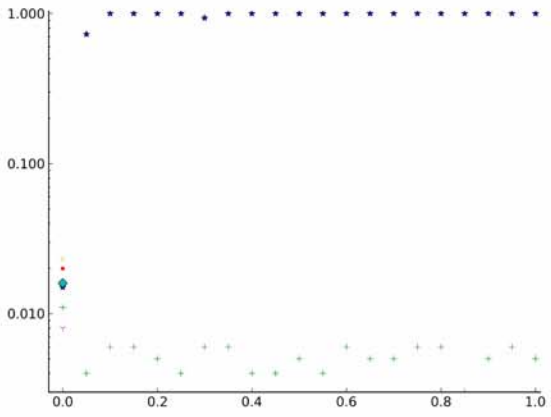
The previous tests evaluate the probability of false positives (Type I errors) and false negatives (Type II errors). However, the sensitivity to sampling error in each estimated cluster can be tested using bootstrapping techniques. The clusters for any set of n samples can be represented as a set of K bipartitions, $b_1, \ldots, b_K$, where $b_{ji} = 1$ or 0 according to whether the ith landmark was in the cluster j or not. Thus, $b_1, \ldots, b_K$ are a series of binary vectors. The bootstrap approach repeatedly generates sets of n samples with replacement from the j^{th} original data. For each of these sets of n samples, I obtain a membership vector as with the original data. All of the bipartitions from all bootstrap sets are then aggregated. The bootstrap percentage for an inferred cluster in the original dataset is calculated as the proportion of bipartitions in the aggregate set showing no conflicts with that cluster. This proportion is reported as the bootstrap value which evaluates the cluster's robustness (Figure



(a) 85 observations



(b) 1000 observations



(c) 5000 observations

Figure 3.1: Behavior of the estimated PVP in different sample sizes. PVP values (y axis) against intracorrelations (x axis) for the simulated data. The star represents the true cluster, the cross represents the grouped singletons, and the rest of the markers represent other singleton clusters recovered as false positives (and therefore low PVP values)

3.2).

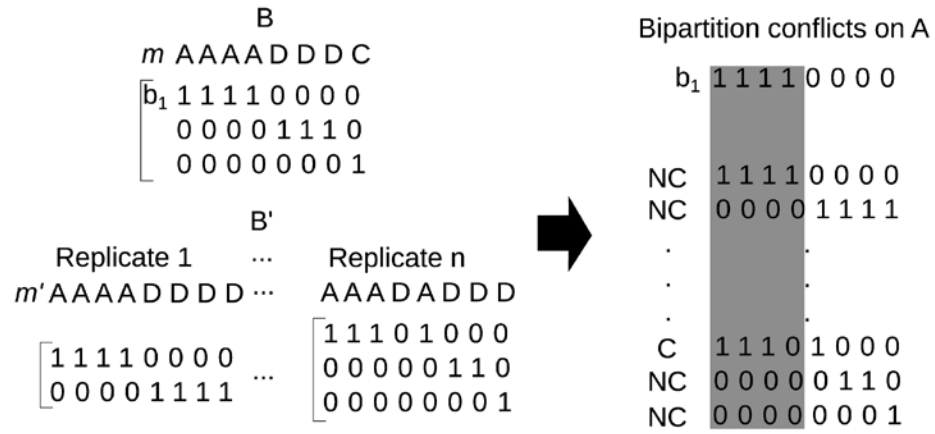


Figure 3.2: Bootstrapping: Defining conflicts. After the clustering method is applied to a given data set, a membership vector m is inferred (3 clusters, A, D, C, in this example) from which a set of bipartitions B can be deduced. By re-sampling with replacement (bootstrapping) the data set, I obtain n replicates which results in a set of bipartitions B' (left-hand side). When a single bipartition from B is compared against B', a conflict (C) or non-conflict (NC) is inferred (right-hand side). This procedure is repeated for all clusters in the original m, giving support for each cluster as the percentage of non-conflicts for that cluster.

From the example, 100 protein structures correspond to the original data from which I have the bipartitions (as shown in Figure 3.2). I created N new replicates of size n by sampling the original n samples with replacement. In some occasions, the same protein structure will be picked. The bootstrap resampling evaluates the effect of possible missing data. However, this measure of reliability is highly conservative, since a slight change in the membership vector might create big variances on the bootstrap value.

Overall, the computation can be summarized in Algorithm 1, where the input is a set of aligned structures, and the output a membership vector of the modules inferred.

Algorithm 1 Pseudo-code for the modularity estimation and testing. This algorithm is a high-level pseudo-code where not all details are expressed.

Input: $D \leftarrow$ A dataset of aligned PDB structures

Output: $V \leftarrow$ a membership vector of clustered residues

Output: P, T, Files with the power and significance of each cluster

procedure MODULER(D, P, T, V)

$X, Y, Z \leftarrow$ Compute matrix of residues' centroids M

$C_X, C_Y, C_Z \leftarrow$ Estimate significant (Fisher transformed equation 3.1) correlation matrices

$LM \leftarrow$ correlation vector magnitude matrix (Apply equation 3.2 to C_X, C_Y, C_Z)

$S(nodes) \leftarrow$ Create a graph with column labels of LM as *nodes*

for $n \in S$ **do**

for $m \in S$ **do**

if $LM_{nm} > 0$ **then**

Connect n and m with edge of length LM_{nm}

end if

end for

end for

membership $\leftarrow$ Find the best partition that optimize the modularity score (equation 3.3)

if contacts are used **then** Refine(membership,LDA) ▷ Merge linear discriminants collisions

end if

for $c \in$ membership **do**

$P \leftarrow$ Test for statistical power in each c ▷ Power to resolve the correlations $\in c$

for $d \in$ membership **do**

$T \leftarrow$ Test for statistically significant differences between c and d.

end for

end for

while membership $\neq$ Membership **do**

Membership $\leftarrow$ Refine(membership, T) ▷ Merge clusters when non-significant

end while

end procedure

3.7 Simulations

To test the method in known modular entities, two types of simulations were performed using Cholesky decomposition. The first one (Multivariate normal simulation) correlates variables in a matrix of random values given a known correlation matrix. The second type of simulation (Protein shape simulation) starts with a real protein shape and then simulates samples with a given correlation.

3.7.1 Multivariate normal simulation

First a multivariate normal random vector is generated as Ly, where y is a vector of independent $N(0,1)$ variates. To simulate multivariate normal vectors, a Cholesky decomposition for the covariance matrix of interest was obtained:

$$LL^T = \begin{pmatrix} 1 & \cdots & \rho & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \rho & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 1 \end{pmatrix}$$

A vector, y, of independent N(0,1) variates was generated so that Ly has the covariance matrix above. The result is a matrix with a set of correlated variables (cluster), surrounded by random (uncorrelated) variables. Cluster intracorrelations ranged from 0 to 1 in increments of 0.05. The first 60 entries (accounting for a cluster with 30 elements with X and Y coordinates) have a given correlation, while 140 entries (accounting for 70 landmarks) are uncorrelated.

A simulation to evaluate the effectiveness in solving the boundaries between two modules can also be performed. In that case, the correlation matrix was:

$$LL^T = \begin{pmatrix} 1 & \cdots & \rho_1 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \rho_1 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & \rho_2 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \rho_2 & \cdots & 1 \end{pmatrix}$$

The resulting matrix contains two clusters whose intracorrelations ranged from 0 to 1 in increments of 0.05.

The output of the simulation is a set of coordinates for a given number of samples to which the method (explained in previous sections) will be applied.

Estimated correlations: Precision of the simulations

Accurate estimates of correlations were gathered for the simulation performed. However, precision varied substantially with sample size.

Table 3.1: Precision of the simulations with one module in background noise. Precision of the simulation of one module on background noise with 100, 500, and 1000 samples. The quantiles describe the distribution of values in the lower triangle of the correlation matrix for the full simulated data set. The background quantiles represent the distribution of values in the rest of the matrix as background noise.

Sample	Intra-cluster	Intracorrelation Quantiles					Background Correlation Quantiles				
Size	correlation	0%	25%	50%	75%	100%	0%	25%	50%	75%	100%
100	0.2	-0.087	0.145	0.206	0.274	0.545	-0.348	-0.058	0.009	0.077	0.304
	0.4	0.061	0.288	0.341	0.395	0.589	-0.314	-0.075	-0.0064	0.059	0.331
	0.6	0.334	0.542	0.578	0.619	0.765	-0.361	-0.072	-0.002	0.066	0.296
	0.8	0.651	0.771	0.791	0.811	0.875	-0.365	-0.073	0.002	0.067	0.291
	1.0	1.0	1.0	1.0	1.0	1.0	-0.209	-0.069	-0.019	0.051	0.232
500	0.2	0.061	0.168	0.193	0.220	0.354	-0.169	-0.029	0.001	0.029	0.139
	0.4	0.261	0.365	0.389	0.411	0.494	-0.172	-0.029	0.002	0.0332	0.204
	0.6	0.485	0.562	0.580	0.597	0.656	-0.152	-0.031	2×10^{-5}	0.031	0.149
	0.8	0.743	0.776	0.786	0.795	0.824	-0.161	-0.041	-0.011	0.019	0.144
	1.0	1.0	1.0	1.0	1.0	1.0	-0.142	-0.0350	1.7×10^{-4}	0.029	0.087
1000	0.2	0.087	0.170	0.190	0.209	0.308	-0.131	-0.021	7.2×10^{-4}	0.023	0.116
	0.4	0.338	0.391	0.407	0.422	0.4745	-0.108	-0.016	0.005	0.0267	0.126
	0.6	0.535	0.580	0.591	0.603	0.647	-0.103	-0.018	0.002	0.023	0.129
	0.8	0.759	0.791	0.798	0.804	0.827	-0.131	-0.024	-0.005	0.012	0.083
	1.0	1.0	1.0	1.0	1.0	1.0	-0.038	-0.014	0.009	0.024	0.068

Despite this, even with low sample sizes the median correlations were close to the true values (Tables 3.1 and 3.2).

As can be seen, some variance was allowed to make the simulation more realistic.

Performance of the the method

In noisy data, this method is able to correctly identify and assign the membership vector at very low modular intracorrelations (Figure 3.3) when the sample size is sufficient. Even for intracorrelations as low as 0.05, the method identifies the true clusters if more than 3000 observations are used.

Table 3.3 shows the results of the significance tests, power analysis, and bootstrapping. The significance test controls the Type I error and therefore the false positives. Here it is reported for an α (false positives or Type I error probability) of

Table 3.2: Precision of the simulations with two modules. Precision of the simulation of two modules with 100, 500 and 1000 samples. The quantiles describe the distribution of values in the lower triangle of the correlation matrix for the full simulated data set. The intercorrelation quantiles represent the distribution of the sub-matrix corresponding to the correlation between the two modules.

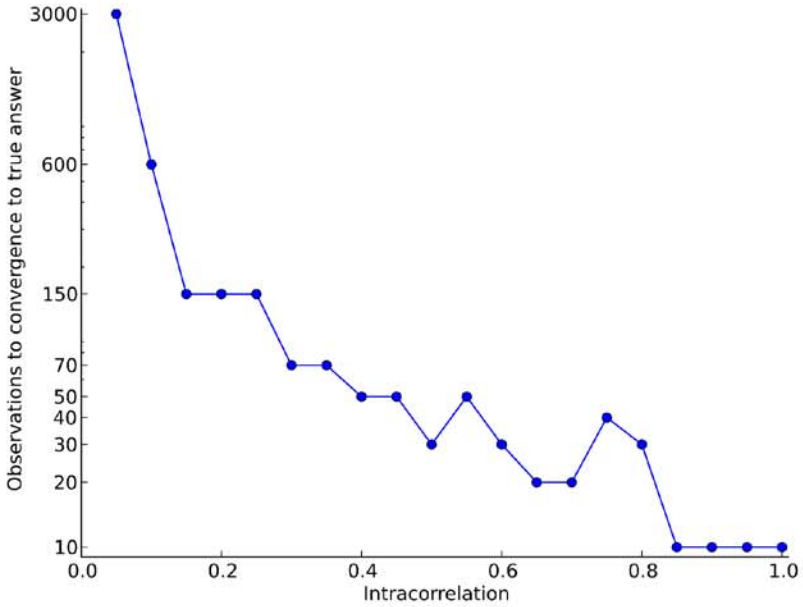
Sample	Intra-cluster	Intracorrelation Quantiles					Intercorrelation Quantiles				
Size	correlation	0%	25%	50%	75%	100%	0%	25%	50%	75%	100%
100	0.2	-0.12	0.112	0.178	0.247	0.532	-0.349	-0.061	0.009	0.081	0.357
	0.4	0.135	0.347	0.401	0.451	0.627	-0.372	-0.075	-0.014	0.049	0.314
	0.6	0.405	0.581	0.617	0.653	0.758	-0.468	-0.225	-0.165	-0.104	0.116
	0.8	0.716	0.776	0.795	0.813	0.879	-0.256	-0.087	-0.049	-0.011	0.147
	1.0	1.0	1.0	1.0	1.0	1.0	-0.018	-0.018	-0.018	-0.018	-0.018
500	0.2	0.059	0.172	0.2	0.229	0.337	-0.177	-0.036	-0.008	0.023	0.151
	0.4	0.286	0.371	0.393	0.414	0.495	-0.099	0.013	0.04	0.066	0.183
	0.6	0.516	0.587	0.602	0.618	0.67	-0.099	0.005	0.029	0.051	0.153
	0.8	0.748	0.793	0.801	0.808	0.838	-0.069	0.014	0.031	0.05	0.131
	1.0	1.0	1.0	1.0	1.0	1.0	-0.012	-0.012	-0.012	-0.012	-0.012
1000	0.2	0.109	0.182	0.201	0.221	0.301	-0.116	-0.023	-0.001	0.021	0.128
	0.4	0.321	0.388	0.405	0.421	0.488	-0.133	-0.03	-0.011	0.008	0.089
	0.6	0.531	0.582	0.593	0.605	0.647	-0.077	-0.008	0.008	0.023	0.091
	0.8	0.764	0.793	0.799	0.805	0.825	-0.052	0.007	0.021	0.034	0.082
	1.0	1.0	1.0	1.0	1.0	1.0	0.026	0.026	0.026	0.026	0.026

0.05. However, the permutation test is not able to deal with the false negatives or Type II error (Table 3.3).

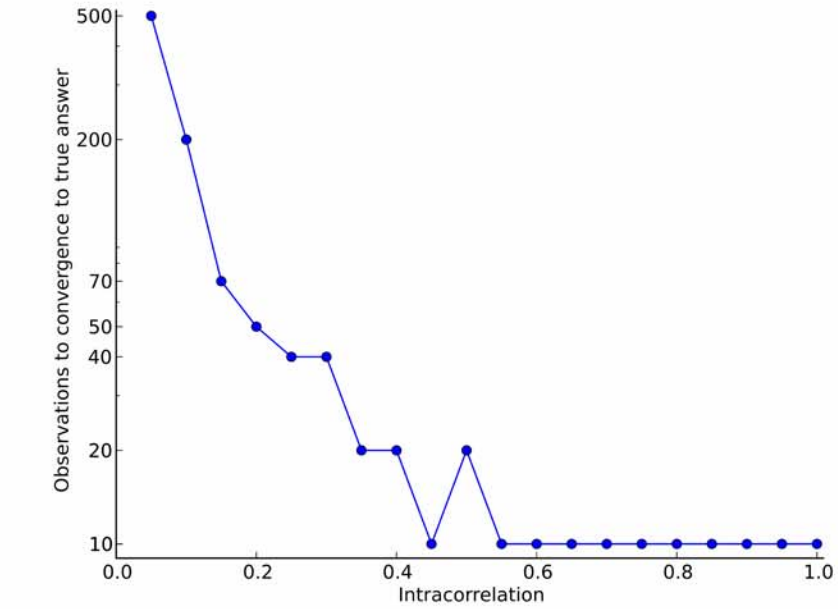
Table 3.3: Clusters, significance, PVP and Bootstrap support for the simulated data. The proportion of pairs that were judged not to be in the same cluster after the permutation test (Significance), the estimated PVP for the simulated data set and the bootstrap value in the simulated datasets with 0.35 intracorrelation and 85 observations. PVP is the proportion of variables in the cluster with enough power (with a β of 0.2 and an α of 0.05) to be resolved. The significance test critical value was corrected using the False Discovery Rate correction.

Clusters	Significance	PVP	Bootstrap
One module on background noise			
A	< 0.0375	0.617	93%
Singletons	0.134	0.006	9%
Two modules			
A	< 0.0125	0.631	100%
B	< 0.0125	0.640	100%

In simulations with correlations of 0.35, the method was able to identify the



(a) One module on background noise



(b) Two modules

Figure 3.3: Performance of the method in simulated data. Performance of the clustering method by number of observations and intracorrelation in multivariate normal simulations for each intracorrelation evaluated.

"true" cluster (Table 3.3) with a relatively low number of observations (85 in this case), and the power analysis gives an estimation of robustness. In the simulation, only cluster A has enough power (0.617 estimated PVP) to resolve almost two thirds of the components of the cluster. The other "module" is a collection of singletons which has an estimated PVP close to zero. Similarly, the bootstrap value highly supports the "true" cluster, while the group of singletons is ruled out. Here it is shown that the significance test efficiently deals with false positives, the PVP gives information about the strength of clustering, and the bootstrap gives information about the repeatability of the clusters when re-sampling.

3.7.2 Structured simulation

So far the method seems to behave very well on un-structured data. However, structured shapes have an inherent co-variance structure that have to be dealt with. To test this I created a second set of simulations with a given shape (Figure 3.4)

Figure 3.4: Starting shape for simulation. Letters and colors represent the true clustering.

To create the true clustering shown in Figure 3.4, the same approach as in section 3.7.1 will be applied. However to retain the shape, the following procedure will be done:

1. Create an original shape coordinates vector.

2. Create a matrix of n samples repeating the shape coordinates of the previous step.

Section 3.7.2 was not included in the BMC paper. This section along with a broader application in networks have been published in PLOS ONE: "Hleap JS, Blouin C (2014) Inferring Meaningful Communities from Topology-Constrained Correlation Networks. PLoS ONE 9(11): e113438. DOI: http://dx.doi.org/10.1371/journal.pone.0113438".

3. Obtain a multivariate normal (as in section 3.7.1) matrix with the same dimensions of that one in the previous step.

4. For the equivalent entries of each true module, perform the Cholesky decomposition on the random matrix as explained in section 3.7.1

5. Sum the random (now correlated) and shape matrices

Given that previous sections showed that 500 samples were enough to resolve most of the correlations (Figure 3.3), in this section I will only use that amount of samples.

In the case of structured samples, it is also useful to include contact information in the same manner that would be advisable in protein datasets. This procedure leads to fewer edges between no modular components and therefore an easier clustering scheme. For this reason, the structured simulation was constructed in such a way that each module's (each colored "H" in Figure 3.4) intracorrelation was varied in 0.2 increments. Also, inference with and without contacts were performed. The measure of accuracy was performed using the F-score measure. F-score corresponds to the harmonic mean of the precision and recall. The former correspond to the number of correct positive results divided by the number of all positive results. The latter, recall, express the number of correct positive results divided by the number of positive results.

Table 3.4: Performance of the structured simulation. Each entry corresponds to the F-Score of the pair intracorrelations. Cor. = Intracorrelation.

		Cluster A											
		No contacts						With contacts					
	Corr.	0.00	0.20	0.40	0.60	0.80	1.00	0.00	0.20	0.40	0.60	0.80	1.00
Cluster B	0.00	0.99	0.73	0.90	0.91	0.91	0.89	0.06	0.86	0.82	0.86	0.84	1.00
	0.20	0.66	1.00	1.00	1.00	0.86	1.00	1.00	1.00	0.83	0.86	0.92	1.00
	0.40	0.91	1.00	1.00	1.00	1.00	1.00	0.87	0.81	1.00	1.00	0.81	0.92
	0.60	0.89	1.00	1.00	1.00	1.00	1.00	0.81	1.00	0.74	1.00	0.76	0.83
	0.80	0.93	1.00	1.00	1.00	1.00	1.00	0.82	0.83	0.86	0.82	0.83	0.83
	1.00	0.89	0.86	1.00	1.00	1.00	1.00	0.99	1.00	0.86	1.00	1.00	1.00

Table 3.4 shows the results of the performance test of the method. Here the performance was measured by means of the F-score.

As can be seen, despite good performance in most of the cases with no contacts, some poorly resolved cases remain (e.g. 0.2 vs 0.0 in Table 3.4). Also, the contact filtering has a worse performance than the non-contact case. In this case, an over-fragmentation was expected, but the method was not able to re-merge artificially broken down clusters.

Despite this result, the use of contacts will be necessary when dealing with real datasets, since modules in proteins are defined as residues being in contact. To deal with this "under performance", I propose a pre-filtering of the membership vector using Linear Discriminants (LD). LD is a multivariate statistical technique that allows the visualization of high-dimensional data by projecting it onto a line while maximizing the distance between the means of the two groups and minimizing the intragroup variance (McLachlan, 2004). This technique has been used extensively in pattern recognition (Jain et al., 1999; McLachlan, 2004) to help cluster difficult datasets. To do so:

1. Clean the membership vector outputted by the optimization of the modularity score (section 3.3.2) by removing all singletons (components in the graphs made of a single value/landmark).

2. Use the cleaned membership vector as a classifier for an LD analysis.

3. Perfom an LD analysis on the correlation magnitude vector matrix (as explained in equation 3.2).

4. Given the first two LDs per classifier, compute the 95% confidence ellipse of the class.

5. Evaluate collisions between the ellipses. That is, checking for overlaps between classes, as delimited by the confidence ellipses.

6. Merge overlapping classes and re-label the membership vector.

7. Continue with the approach explained in sections 3.4, 3.5, and 3.6, using the newly labeled vector as input.

The LD analysis will get rid of most of the possible noise. It is important to state that performing a LD analysis in correlation magnitude matrices will violate

the assumption of independence since most variables will be collinear, and in some cases the normality assumption. However, it has been shown that LD analysis is somewhat robust against violations of these assumptions (Dziuda, 2010). These violations affect mostly the coefficients of the linear discriminant function, but the classification is not greatly affected. Given that my goal does not imply classification of other datasets with a particular inferred linear discriminant function, this violation can be omitted in the pre-filtering. The LD analysis can be performed with the library `MASS` (Venables and Ripley, 2002) and the 95% confidence ellipses with the package `ellipse` (Murdoch and Chow, 2013) available in `R` (R, 2011). The full membership refinement is done with an `R` script that can be coupled with the `Python` script for modularity testing.

Table 3.5: Performance of the structured simulation with LD pre-filtering. Each entry corresponds to the F-Score of the pair intracorrelations. Cor. = Intracorrelation.

		Cluster A											
		No contacts						With contacts					
	Corr.	0.00	0.20	0.40	0.60	0.80	1.00	0.00	0.20	0.40	0.60	0.80	1.00
Cluster B	0.00	1.00	0.95	0.91	0.91	0.93	0.89	0.68	1.00	1.00	1.00	0.99	1.00
	0.20	0.96	0.67	0.67	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	0.40	0.92	0.67	1.00	0.67	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	0.60	0.92	1.00	1.00	0.67	0.67	1.00	1.00	0.87	1.00	1.00	1.00	1.00
	0.80	0.92	1.00	0.67	0.67	0.67	0.67	1.00	1.00	1.00	1.00	1.00	1.00
	1.00	0.89	1.00	1.00	0.67	1.00	0.67	0.99	1.00	1.00	1.00	1.00	0.67

Table 3.5 shows the results of the modularity inference when the LD pre filtering is used. In this case, the problem shown in Table 3.4 is inverted. When there are no contacts involved in the graph construction the performance decreases. This behavior is mainly due to the over merging when too many edges are drawn and therefore a more constricted membership vector is outputted. The contact case, on the other hand, significantly improves the performance of the method, converging to the true answer in most cases. When intracorrelation is equal to one, both approaches present glitches. This is due to the fact that to perfectly correlate two modules with a given structure, both modules will share a similar distribution, and therefore will tend to merge. It is also important to point out that perfect correlation within subsets is not a realistic setup, and therefore it is not going to be common on real datasets. Also, here I am working with an acceptable 95% confidence (an α of 0.05) for the statistical

tests, so the actual performance can be improved by making it more stringent (in the case of simulations) or perhaps more lenient (for real and more variable datasets).

Given these results, I will recommend to use the LD prefilter only in cases when contacts are used, while avoiding its use in the opposite case. With this rule in mind, most datasets will be correctly clustered.

3.7.3 Protein shape simulation

To test modular architecture in more complex shapes, a protein shape simulation in 3D is performed. To do so the landmarks from a real structure are extracted (see section 3.1). With these coordinates as a starting point and a desired membership vector, the Cholesky decomposition explained in the section 3.7.1 is applied. As before, a multivariate normal random vector is generated as Ly, where y is a vector of independent $N(0, 1)$ variates with the desired length (e.g number of samples to generate). The matrix with the desired covariance LL^T is then created by Cholesky decomposition with the correlation matrix that follows the structure of the membership vector with controlled intracorrelation in 0.2 increments. To the resulting correlated matrix, the vector of coordinates of the original shape is added. As result, a given number of samples with the desired correlation between putative residues is created. The correlation matrix is then dependent on the membership vector but follows the overall structure of the previous section. The protein Pyruvate kinase 1PKM (Figure 3.5) is here established as a reference structure given that it is a widely studied protein and its multidomain (and therefore modular) architecture is known.

It contains three CATH domains and a single chain (Greene et al., 2007). The creation of a membership vector corresponding to the CATH domain segments allowed the generation of a visual comparison of the CATH domains and modules, as well as to the simulation of such architecture. To test the convergence to the true cluster, the intracorrelation was also controlled (instead of the uniform random values), from 0.00 to 1.0 in 0.2 increments. The sample size will be kept at 500 for comparative purposes with the other simulations. This sample size is also enough to solve most correlation and is still plausible to find in real datasets (at least by modelling sequences).

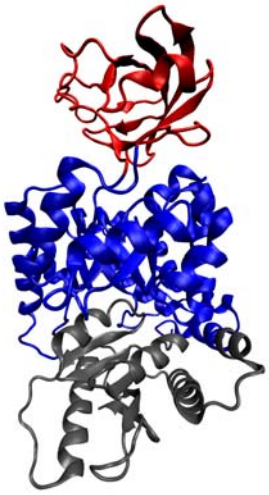


Figure 3.5: Structure of the protein Pyruvate kinase (PDB code: 1PKM) color-coded by CATH domains.

Table 3.6: Performance of the Piruvate Kinase simulation with LD pre-filtering. Each entry corresponds to the F-Score of the pair intracorrelations. Cor. = Intracorrelation.

Cor.	0.00	0.20	0.40	0.60	0.80	1.00
F-Score	0.62	1.00	1.00	1.00	1.00	1.00

Table 3.6 shows the result for the protein simulation. The method developed here is recovering the true cluster in almost all cases. The case of the 0.0 correlation is due to the fact that mostly singletons are found. The F-Score for this case is then computed as if the test vector is a single string of one particular label and therefore the low, yet not zero, F-score.

3.8 Exploration of other real datasets

In this section I will present the results of two real datasets. First a set of snapshots from a molecular dynamics simulation (MD) of the NPC1 N-terminal domain are analysed to provide insights into the modular architecture of dynamic data. That is, group of residues that move together in solution.

Then a set of homologous structures of the *alpha*-amylase catalytic domain are use to test the sub-domain architecture at the evolutionary level. A module here (different than the MD modules) refers to a group of residues that are moving together in the evolution of the structure.

Dynamic modules of the Niemann Pick C1 protein N-terminal domain

The Niemann-Pick disease type C (NPC) is an autosomal recessive disease, expressed when there is an error in the exogenous cholesterol trafficking and as result a lysosomal accumulation of it (Patterson et al., 2006). This disease is caused by a mutation in either of the two NPC proteins (NPC1 and NPC2) (Kwon et al., 2009). The Niemann-Pick C1 (NPC1) protein regulates the lysosomal cholesterol transport to other intracellular compartments (Garver et al., 2002). NPC1 contains 13 (13-16 according to (Patterson et al., 2006)) membrane domains and 3 other domains that are in the lumen of the lysosomes (Davies and Ioannou, 2000). One of these luminal domains is the N-terminal domain which bears the cholesterol binding site (Infante et al., 2008b), and has eight α-helices flanked by three β-sheets (Figure 3.6) and its sequence is highly conserved (Watari et al., 1999). NPC1 N-terminal domain (unlike the NPC2 protein) can bind with the oxygenated derivatives of the cholesterol (Kwon et al., 2009) making it an interesting domain to study dynamic properties.

Figure 3.6 shows the modules gathered when the module identification is applied to the molecular dynamics simulation of the NPC1 N-terminal domain snapshots. All these modules showed a bootstrap above 66.7% and a PVP over 0.96. Interestingly, all modules are related with the binding pocket, surrounding the cholesterol molecule.

The first module (Figure 3.6A) encloses three cholesterol binding residues, and another binding residue to the N-Acetyl-D-Glucosamine (NAG). It also spans a residue associated with the development of the NPC1 disease in adulthood (Fancello et al., 2009). All other residues correlating with these seem to give structural support to the back of the cholesterol binding pocket, as well as serving as receptacles for both ligands. This region also encompasses four residues containing single nucleotide polymorphisms (SNPs) for the human gene (Karchin et al., 2005).

In Figure 3.6B, a module that comprises more than half of the residues that make the sterol pocket is shown. From these residues, this module is the only one that includes the non-hydrophobic ones, being of importance in the direct protein - 3β-hydroxyl interactions, as well as the water-mediated interaction with such groups.

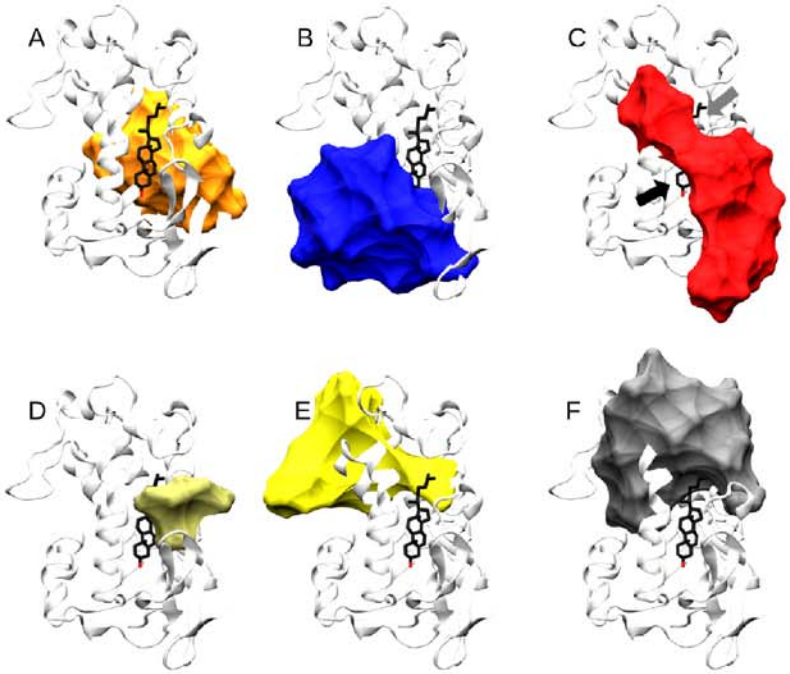


Figure 3.6: Modules recovered by the method in a molecular dynamics simulation of the The Niemann-Pick disease, type C1 (NPC1; PDB code: 3GKH) protein with cholesterol (Licorice-type structure) bound. The arrows in sub-figure C point to the water (black arrow) and sterol (gray arrow) openings, described in (Kwon et al., 2009). The images were rendered using VMD (Humphrey et al., 1996) and POVray (www.povray.org).

This helps in the stabilization of the bounded sterols and giving a particular stereo-specificity(Kwon et al., 2009). This module is located in the bottom part of the binding pocket and can be seen as a "glue" for the secondary structures in contact with such pockets, and therefore maintaining the shape of the structure in its less movable part. This also supports the model in (Kwon et al., 2009), where the sterol opening needs to move in order to uptake the cholesterol from NPC2. This module also contains three SNPs found in the human gene (Karchin et al., 2005).

The module shown in figure 3.6C, shows the residues responsible for the water (black arrow) and sterol (gray arrow) openings described in (Kwon et al., 2009) as being of functional importance to the cholesterol uptake and the retention of it in the binding site. If some residues within this module are mutated, the cholesterol might not be taken by the protein and the Niemann-Pick disease is expressed (Kwon et al., 2009). This module also includes two cholesterol binding sites, a residue shown to be related to the development of the disease in infantile stages (Millat et al., 2005),

and a SNP (Karchin et al., 2005).

The module in figure 3.6D shows a small module that coincides with functionally important residues involved in the affinity for cholesterol binding (Kwon et al., 2009). These thus may be related to the expression of the NPC disease. Giving that these modules are analyzed in the light of dynamics, the module in Figure 3.6D shows that the affinity for the cholesterol mediated by these residues is given by geometric constraints induced by cholesterol binding.

Figure 3.6E shows a module that encloses two binding residues to NAG. It has also been shown that two residues are important in the development of a late infantile NPC1 disease (Millat et al., 2005; Fancello et al., 2009), and one SNP is also enclosed. It seems to be also of structural support for the cholesterol binding pocket in the top(E), creating a pocket that receives the ligand.

The module shown in figure 3.6F encloses the α-helices 3, 7 and 8, that have been shown to play an important role in the access and release of cholesterol, since its movement controls the enlargement of the sterol opening (Kwon et al., 2009). This module also contains some of the residues that decrease the cholesterol transfer to the liposomes if mutated (Kwon et al., 2009), as well as four SNPs (Karchin et al., 2005). The module shown in figure 3.6F is therefore of functional importance for the intake and outtake of cholesterol.

Since there are disease-related mutations in all of the modules, it would be important to further study the relationship between modules and protein function. The correlation within modules is large enough to think of them as units, and therefore it is probable that the residues exposed in (Millat et al., 2005; Kwon et al., 2009; Karchin et al., 2005) are not the only major contributors to the disease. Further confirmation of the effects of mutations within these modules is needed.

Evolutionary modules in the α-amylase catalytic domain

Starch is the main storage of carbohydrates in plants. Processing it and discovering novel poly and oligosaccharides is important for biotechnological and chemo industrial applications (Svensson, 1994). Most starch-related enzymes are classified within the α-amylase family. This family catalyzes the hydrolysis of α-(1,4) glycosidic bonds of polysaccharides, and therefore is classified as glycoside hydrolases

(Davies and Henrissat, 1995). This a multi-reaction catalytic family, since its members can catalyze different reactions (hydrolysis, transglycosylation, condensation and cyclization) (Ben Ali et al., 2006). Industrially, some α-amylases are used in the production of ethanol (Bothast and Schlicher, 2005), high-fructose corn syrup(Visuri and Klibanov, 1987), and other oligosaccharides. It is therefore of industrial and biological importance. It has a highly symmetrical TIM-barrel ($(\beta/\alpha)_8$) catalytic domain (Svensson, 1994). This fold is highly versatile and widespread among the structurally characterized enzymes, being present in almost 10% of them (Farber, 1993; Höcker et al., 2001; Wierenga, 2001; Gerlt and Raushel, 2003). There has been a debate about the type of evolution that this fold has been through: convergent, divergent, or both(Farber, 1993). However, there is evidence supporting the divergent evolution hypothesis (Höcker et al., 2001). The catalytic activity and substrate binding residues occur at the C-termini of β-strands and in loops that extend from these strands (Svensson, 1994).

Four modules are identified in the α-amylase sub-domain architecture (Table 3.7 and Figure 3.7). In Figure 3.7, most of the modules span the surface to the TIM-barrel (β-sheets of the TIM-barrel are highlighted in Figure 3.7A). This behavior is due to the interaction of the protein and its catalytic pocket, with the ions calcium and sodium received by this structure mainly on its surface. Modules shown in figures 5B, D and E span regions where these ions are frequently found among the homologs, and the residues in charge of the ligation of the three metal ions (Machius et al., 1998) as co-factors for the hydrolysis.

Table 3.7: Clusters, significance, PVP and Bootstrap support for the α-amylase data set. The proportion of pairs that were judged not to be in the same cluster after the permutation test (Significance), the estimated PVP and the bootstrap value in the α-amylase. The significance test critical value was corrected using the False Discovery Rate correction

Modules	Significance	PVP	Bootstrap
B	< 0.001	0.479	31.3%
C	0.005	0.440	42.9%
D	< 0.001	0.503	39.9%
E	< 0.001	0.580	51.7%

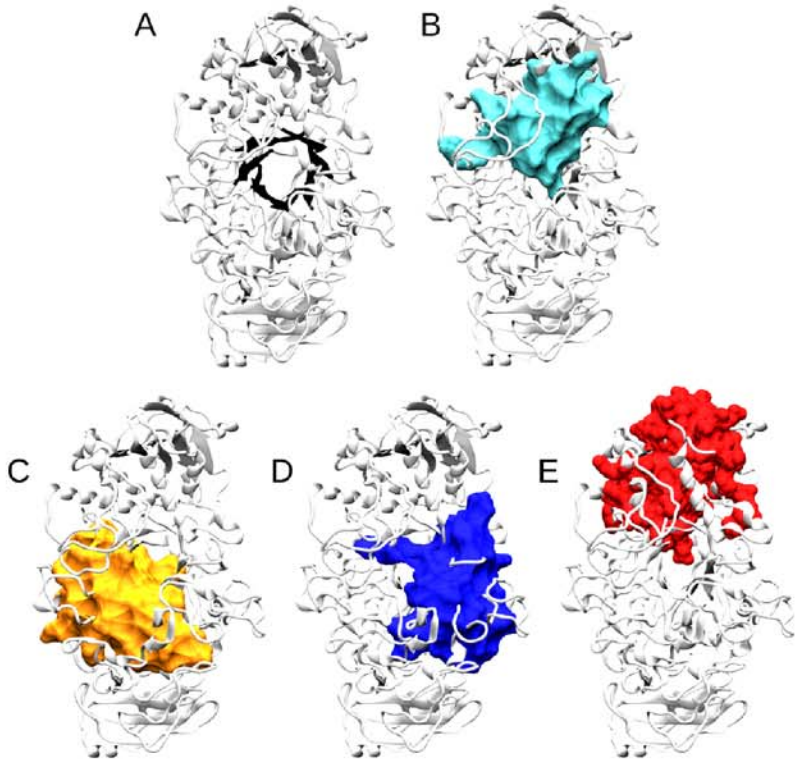


Figure 3.7: Modules recovered by the method in the homologous structures of the α-amylase, its catalytic domain. This analysis was performed using the 85 redundant structures available at the protein data bank (http://www.rcsb.org/pdb/). In the sub-figure A the TIM barrel is highlighted, and sub-figures B-E show the different modules obtained. The images were rendered using VMD (Humphrey et al., 1996) and POVray (www.povray.org). The structure used to visualize the modules is the PDB 1BF2.A from *P. amyloderamosa*.

The module in Figure 3.7B also comprises two residues that mutational studies have shown as important for the cleavage site (MacGregor et al., 2001) and have been reported as substrate binding and catalytic residues. The module shown in Figure 3.7D also spans important catalytic residues. This includes a proton donor, a catalytic nucleophile (MacGregor et al., 2001), and five substrate binding residues (Svensson, 1994). The module in Figure 3.7C comprises a substrate binding residue (Svensson, 1994). Furthermore, module 5C seem to span most of the smallest active sub-domain of a TIM-barrel fold, as shown by (Ben Ali et al., 2011) in *Bacillus stearothermophilus*, comprising almost all of the $\beta_2\alpha_2$ domain (Figure 3.8). No known catalytic residue was found in this module, however (Ben Ali et al., 2011) showed that this module retains its catalytic activity. The second domain showed by (Ben Ali

et al., 2011) (Figure 3.8), was not homologous throughout our sampling (i.e. was not present in all the sampled structures), and therefore, no information was available about this domain.

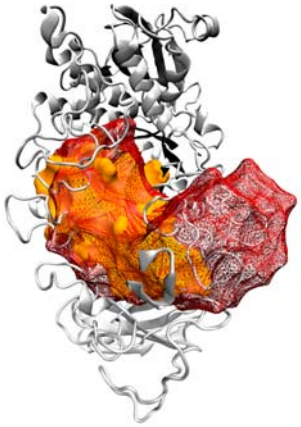


Figure 3.8: Evolutionary module in α-amylase spanning the smallest active sub-domain of a TIM-barrel fold. Expansion of the module shown in Figure 5C with a superimposition of the AmyTM structure reported in Ben Ali et al. (2011) (wire-frame structure). The images were rendered using VMD Humphrey et al. (1996) and POVray (www.povray.org).

3.9 Concluding remarks

Most biological data are typically highly multivariate and multidimensional in nature. Many tools have been developed to deal with such dimensionality (Muirhead, 2008, and references therein). However, the variable selection and dimensionality reduction used in such methods (aiming to reduce matrix complexity) may compromise information conservation (Berge et al., 2010), or require a larger sample size than is possible for protein data. To overcome these drawbacks, I introduce a community detection-based clustering method. Community detection-based approaches do not need *a priori* knowledge of the number of clusters (Mishra and Pandey, 2011), are not heavily parametrized, and can handle multivariate and multidimensional data without dimensionality reduction. Here I propose a graph based method to explore protein structure modularity, where:

1. A graph is built where the vertices are the centroids of residues. The correlation between coordinates is set as edge weight if it is significant (see equations 3.1

and 3.2), and if the two residues are in contact (See section 3.2).

2. The community structure in the graph is inferred by fast-greedy (evaluating and selecting the best result at each step, as opposed to maximizing at the end of the scoring process) optimization of a modularity score (Q; see equation 3.3).

3. The membership vector is pre-filtered using a Linear Discriminant Analysis, with the community structure of the previous step as classifier.

4. The statistical support for each cluster is obtained.

5. The solution is refined based on this statistical support.

6. The statistical power to resolve each partition with respect to the size of the dataset is estimated (equations 3.4 and 3.5).

7. The stability of the estimates with respect to the sampling error is measured using bootstrapping (Figure 3.2).

3.9.1 Putative meaning of the sub-domain architecture

So far I have shown the significant partitions of a domain. But what is the probable meaning of such modules? One might think that these modules can represent autonomous folding units (AFU); however, my data showed discontinuous amino acid sequences (in one dimension, since they are in contact in 3D space) per module. Also, comparative analysis with the dataset analysed by Fischer and Marqusee (2000) showed no relationship with the groups obtained here. Another plausible hypothesis could be to assign modules to close loops, but the same continuity argument can be made. Furthermore, the α-amylase subdomains identified by this method span several of the TIM-barrel close loops exposed by Frenkel and Trifonov (2005) with no particular pattern. These discrepancies are expected, since the definition of foldons, AUFs and close loops have little or no meaning in an evolutionary perspective. These concepts are derived from the analysis of single structures and their internal interactions (i.e. contact matrix, physical interactions, length, distance) and therefore the

non-evolutionary approaches for sub-domain determination will identify a different kind of module than an evolutionary approach.

On a more related framework, Dutheil and Galtier (2007) developed a method to test co-evolving sites. When tested on the α-amylase dataset used in Chapter 2, no pattern correlating the two methods was found. Moreover, the largest significant grouping of co-evolving residues with Dutheil and Galtier (2007) method span only 10 residues of the protein. This discrepancy can be attributed to the fact that Dutheil and Galtier (2007) are testing co-evolution in a sequence based perspective. That is, giving a phylogenetic tree and its source alignment, which residues have significant mutual information. This method disregards completely the geometry of protein structures, therefore answering a different question than the approach exposed here.

So what is the possible meaning of the sub-domains? Although more work (both bioinformatic and experimental) is needed to clearly address this question, the sub-domain architecture here represented is probably co-evolving geometric units (in the case of homologous sampling) and semi-rigid components (in the dynamic perspective) of proteins. The partitions shown here can be depicting a level of modularity out of a possible hierarchical architecture of protein modularity. This concept will be covered in chapter 4.

Chapter 4

The semantics of the modular architecture of protein structures

Protein structures are normally inherently flexible and can be shaped to perform a wide spectrum of functions. This property of proteins allows for the opportunism of evolutionary tinkering (Jacob, 1977), that is, the development of a new system or function by re-engineering an existing one or by combining existing systems. That tinkering and progressive integration of sub-parts in evolutionary time has shaped most protein structures as modular systems by epistasis, linkage, and co-evolution between residues (Schlosser and Wagner, 2008). The epistatic effect refers not only to directly compensatory effects but the fact that a group of changes can have different effects than each one of them individually, and therefore creating a structuring in the protein that tends to be modular. This epistasis does not necessarily require linkage, but if selected, the linkage can come into play thus creating a stronger modularity and heritability of such modules. All these processes do not necessarily mean co-evolving residues, however if epistasis and linkage confer the structure with selective advantages, co-evolution between residues and modularity might arise.

Modularity makes structures evolvable, or capable to cope with selective pressures, by diminishing the number of constraints (Caetano-Anollés et al., 2013; Harms and Thornton, 2013). Modular systems also promote the emergence of novel functions by rearrangements and are therefore said to be plastic (Del Sol et al., 2007; Bridgham et al., 2010; Rorick, 2012; Bornberg-Bauer and Albà, 2013). Plasticity and resilience lead to robust systems (Pigliucci, 2008; Wagner, 2008). Robustness is the ability of a system to maintain its function under perturbations. A robust

A modified version of this chapter has been accepted for publication in the first issue of Current Protein & Peptide Science in 2016; Hleap, J.S. & Blouin, C. 2016. CPPS DOI: 10.2174/1389203716666150923104720.

biological system tends to neither increase nor decrease its heritable phenotypic variation (Wagner, 2008; Rorick and Wagner, 2011), allowing it "to undergo innovative modification without losing functionality" (Rorick and Wagner, 2011). This does not mean that a robust system cannot be present in different environments, it just allows it to accept changes that create phenotypic variation. In proteins, as proposed in Chapter 2, let the structure be considered a phenotype. Phenotypic variance is given by slight variations in the shape of the protein structure (Hleap et al., 2013b; Chapter 3). Keeping the robustness of the protein shape allows for higher degrees of structural variances. This plasticity and the emergence of novel functions was shown in different systems by Wagner (2008). He has shown that phenotypic robustness contributes to evolvability while genotypic robustness diminish it. Here, I focus on the former, since it is the one that relates to the protein shape, its structure.
Modularity in protein structures can be viewed as the linkage within functional units (Schlosser and Wagner, 2008). Here, a functional unit is defined as a group of residues that are required for thermodynamic stability or any particular function. Modularity normally assumes a tight integration of elements within, and a complete independence among elements between modules (Rorick, 2012). However, the assumption of independence is not always valid (Hleap et al., 2013b): a marginal dependence might exist where the integration between modules is incomplete (Mittenthal et al., 2012) or the independence is not under selection. Properly formulated, molecular modularity can be modelled in a graph theoretic framework and explored using Newman and Girvan method (Newman and Girvan, 2004). This approach, which allows to explore the modularity of a protein structure by standard graph theoretic approaches (Rorick, 2012; Hleap et al., 2013b; Chapter 3), will be discussed further later on.

4.1 The emergence of modularity: natural selection and the self-organizing nature of proteins

It has been proposed that modularity arises spontaneously from the duplication process in cellular networks instead of by selection (Solé and Valverde, 2008). However, the authors also support the idea of tinkering as a driving force for the emergence and

retention of modularity. In protein structures, the concept of expansion by duplication is not as clear as for cellular networks. Solé and Valverde (2008) acknowledge the possibility that selection might drive the "deletion" or modification of interactions within the system, and therefore affect its modularity. The question of the emergence and retention of modularity remains an open issue. The hierarchical and modular nature of biological systems have been shown to be widespread (Lorenz et al., 2011). A more general pattern related to the origin and evolution of biological systems heavily relies on this concept. Lorenz et al. (2011) also hypothesized that modularity arises spontaneously in a rugged fitness landscape, a changing environment, and when Lateral Gene Transfer (LGT) is present. These arguments are compelling, since LGT events will greatly benefit modular systems because of tinkering. Also, a changing environment creates the need for a greater resilience where modularity spontaneously emerges. Finally, a rough fitness landscape creates opportunities for the dynamics of epistatic interactions: Flexibility in interactions is likelier in modular systems (Schlosser and Wagner, 2008). However, the causality of those three cases with respect to modularity is not clear, since the modularity itself can facilitate LGT events, shape the current evolvability of the system, and can modify the system behaviour in a given fitness landscape. This does not mean that modularity is an abstract property. It means that there is a synergistic behaviour between these factors and modularity. Along with its emergence, modularity further contributes to the development of these properties. At a lower level, self-organization can explain all the scenarios mentioned before and the emergence of modularity. If protein structures are self-organizing systems, the appearance of emerging properties, such as modularity, are expected. Self-organization is a pattern-forming process, where the information for the organization mainly comes from the interaction within the system without intervention of external influences. A self-organized system normally exhibits the following characteristics (Camazine et al., 2002):

1. *Dynamicity*: It requires continuous and dynamic interactions among elements within the system.

2. *Exhibits emergent properties*: Through interaction between elements within the system, new properties emerge that cannot be explained by the sum of individual contributions of the elements.

3. *Oscillation*: Self-organized systems normally are explained by parameters that can be tuned by feedback (positive or negative), and that parameter tuning causes an oscillation between a pattern-forming system and a chaotic one.

4. *Multi-stability*: Multiple possible states are often exhibited by self-organizing systems, and the transition between states can be due to the oscillation parameters.

However, the diversity and nature of all protein structures cannot be explained exclusively by self-organization since it is obvious that the genetics and cellular machinery provide a template and a set of instructions. Nevertheless, self-organization is playing an important role in the folding of proteins. That is why most proteins can fold spontaneously in a fraction of the time that it would be theoretically expected as posed in the Levinthal's paradox (Zwanzig et al., 1992; Rooman et al., 2002). If we conceptualize protein structures as a hybrid template/blueprint-self-organized system, the paradox is resolved. Given that the only information each of the residues is taking into account are the local interactions at a given time, the degrees of freedom or "conformational space" reduces significantly. This allows for the explanation not only of the folding by close loops modules (as seems to occur in globular proteins) (Berezovsky and Trifonov, 2002), but virtually of any folding pattern. The hybrid template/blueprint-self-organized system would also help to explain the difference between the number protein sequences versus the number of protein structures they code, since a great number of interactions are actually redundant (Davies et al., 2013). This hybrid non-self/self-organizing behaviour in proteins is also in line with current knowledge of energy landscape theory (Wolynes, 2005; Clementi, 2008; Schafer et al., 2013). Self-organization in protein structures has been measured in a variety of ways such as energy frustration (Fernández and Berry, 2000), hydrophobic interactions (Gerstman and Chapagain, 2005), and folding index (Lundgren et al., 2013). This self-organizing nature of protein structures and protein folding does not ignore other folding catalyzers. Its is known that other players such as ligands, specific isomerases, heat-shock proteins, and chaperones might be involved in different degrees in the production of a folded protein (Jaenicke, 1991). In those cases, the hybridism mentioned earlier can account for the protein-protein interactions that might occur during the folding of the peptide.

In summary, protein structures are context-aware self-organizing systems, where parameter tuning is provided by the context of the protein folding environment and natural selection, shifting/modifying the energy landscape of the system (i.e. Crowding) (Minton, 2000).

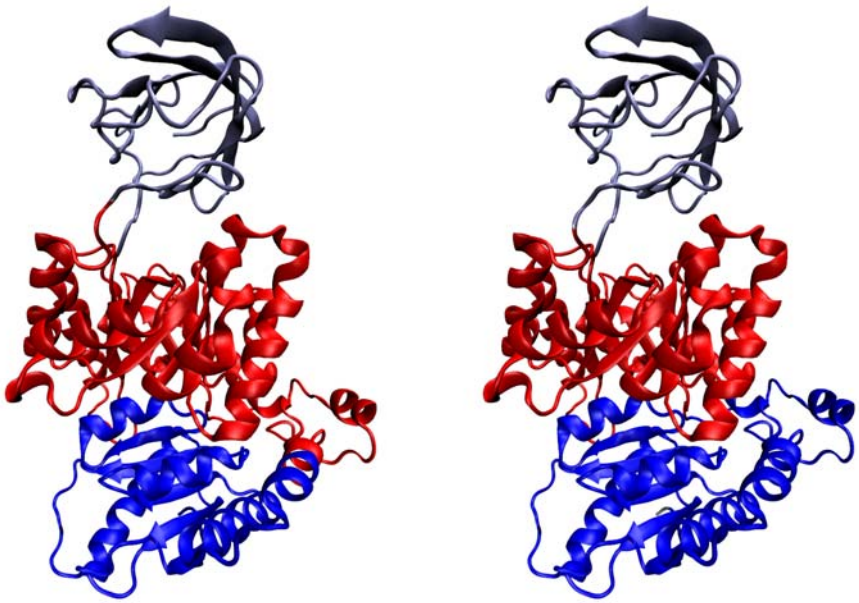
Tinkering, self-organization, and natural selection provide the conditions for the emergence of modularity in proteins. Modularity is an emerging property of a selected self-organization. It provides a model to explain many structural phenomena in protein structures. An example is the more or less hierarchical packing of residues within the protein. Modularity emerges when by tinkering and selection, some interactions are favoured and some other selected against. In Solé and Valverde (2008) terms: *"the rich gets richer and the network will be organized around hubs"*.

Although the emergence of modularity can be intuitively explained, a sound methodology to understand its mechanisms and implications for protein structural biology remains to be explored.

4.2 Domains as modules

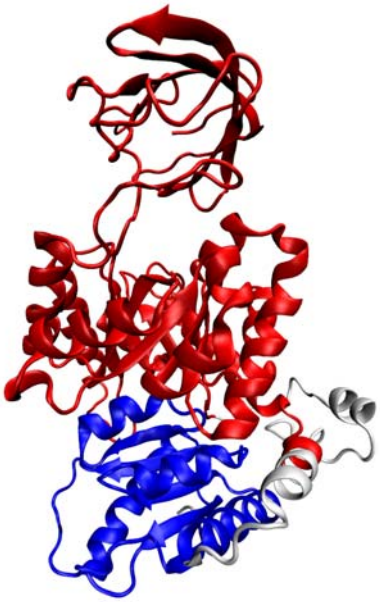
Domains are accepted as the protein structure's evolutionary or functional units (Ponting and Russell, 2002). It has been widely demonstrated that emerging properties in protein function may arise by reusing, rearranging, and shuffling domains (Ponting and Russell, 2002; Voigt et al., 2002; Bridgham et al., 2010). Despite the widespread use of the term domain, the concept is not fully nor universally defined. There are differences in domain classifications even in widely used data sources such as CATH (Greene et al., 2007), SCOP (Murzin et al., 1995), and PFAM (Punta et al., 2012) (Figure 4.1). In general, we can posit that a domain is a *"spatially distinct structure that could conceivably fold and function in isolation"* (Ponting and Russell, 2002). It is also assumed to be the unit of protein evolution (Marchler-Bauer et al., 2005; Koehl, 2006; Jin et al., 2009; Forslund and Sonnhammer, 2012), and an atomic concept in structure and function prediction and annotation (Ezkurdia and Tress, 2011; Xue et al., 2013). Therefore studying and classifying protein domains is an area of intense research.

Discrepancies still exist leading researchers to even propose a consensus solution combining different protein domain assignments (Day et al., 2003; Holland et al.,



(a) SCOP

(b) CATH



(c) PFAM

Figure 4.1: Domain definition the multidomain protein Pyruvate kinase (PK) from rabbit muscle (pdb code: 1PKN). a) SCOP domain definition of the PK enzyme. The catalytic domain (red) is composed by two non-sequential sections; b) CATH domain definition of the PK enzyme. The catalytic domain (red) is composed by two non-sequential sections. c) PFAM domain definition of the PK enzyme, derived from HMM search on the sequence of the 1PKN protein. In the PDB, three domains are reported. Colored cartoon refers to a domain definition and white are residues not assigned to a domain

2006; Alden et al., 2010; Schaeffer et al., 2011). As an example, consider the rabbit Pyruvate kinase enzyme (PDB code:1PKN ; Figure 4.1), a three domain enzyme, with a more or less clear domain definition. Figure 4.1 shows discrepancies in the domain assignment in SCOP, CATH, and PFAM. It is understandable that databases such as SCOP and CATH give similar results, since domain assignment in these has a manual curation step. Despite this, the discrepancies show that every domain classification method is constrained by its definition of a protein domain (Alden et al., 2010). In this light, if what we consider a domain is an evolutionary conserved unit, it will most likely differ from that of a domain as a structurally independent unit. In the latter, discontinuity in sequence is not allowed otherwise the domain cannot behave as a autonomous folding unit (AFU). Assignments like the ones in the catalytic domain (in red) of the Pyruvate kinase in figures 4.1a and 4.1b, for example, would not represent structural domains that can fold autonomously since they are composed of two non-sequential sections interrupted by the PK domain (purple). It has been shown that the "consensus" domain definition among important domain classification databases (i.e. SCOP and CATH) do not correspond with AFUs (Day et al., 2003; Schaeffer et al., 2011). A big fraction of the domains are discontinuous along the chain (Redfern et al., 2007; Kolodny et al., 2013). Despite that, structurally, the lack of sequentiality cannot be easily interpreted; evolutionarily and functionally discontinuity can be explained by different evolutionary scenarios. For example, a domain could be interrupted by an acquisition of a newer domain or flanking regions that serve a functional purpose. These alternative domain definitions have been exploited in the Multidom database (Majumdar et al., 2009; http://prodata.swmed.edu/multidom/) providing insightful information into domain architecture. However, the lack of a unique answer causes difficulties in the automation of the procedure, the interpretation of results, and the possible generalizations of the assignment.

In general, protein domain boundaries are inconsistently defined (Kolodny et al., 2013) and there is a myriad of techniques to try to assess this (Table 4.1).

Not even manually curated domain assignments are in full agreement (Veretnik et al., 2004). This lack of agreement is primarily due to the complexity of the classification. It has been pointed out that domain classification is not tree-like as it was thought before. It contains a more complex hierarchy, leading to a reassessment

Table 4.1: Some of the Domain classification methods available. Auto refers to automatic assignment, that is, domain classification based on an algorithm (i.e. Hidden markov models, modularity optimization). Manual refers to a manually curation step made by experts. Mixed refers to the usage of different resources to build a consensus. BOC refers to basis of classification: Sequence (A) or structure (St). F: Functional; E: Evolutionary; S: Structural.

Method	Assignment	Domain type	Tool	BOC	Reference
Armadillo	Auto	S	Web App.	A	(Dumontier et al., 2005)
CATH	Auto/manual	F/E	Database	St	(Greene et al., 2007)
CDD	Auto	F/E	Database	A	(Marchler-Bauer et al., 2011)
COGS	Auto	F/E	Database	A	(Tatusov et al., 2000, 2001)
CHOPnet	Auto	E	Application	A	(Liu and Rost, 2004a,b)
DOMAC	Auto	E/F	Web App.	A	(Cheng, 2007)
DoBo	Auto	E	Application	A	(Eickholt et al., 2011)
DomainDiscovery	Auto	S/E	Application	A	(Sikder and Zomaya, 2006)
Domain Fishing	Mixed	F/E	Web App.	A	(Contreras-Moreira and Bates, 2002)
DomainParser	Auto	S	Multiple	St	(Xu et al., 2000)
DomCut	Auto	S/E	Web App.	A	(Suyama and Ohara, 2003)
DomNet	Auto	S/E	Application	A	(Yoo et al., 2008)
DomPred	Mixed	F/E	Web App.	A	(Bryson et al., 2005)
DOMpro	Auto	S/E	Multiple	A	(Cheng et al., 2006)
DomSSEA	Mixed	S	Application	A	(Marsden et al., 2002)
Galzitskaya's	Auto	S	Method	A	(Galzitskaya and Melnik, 2003)
Ginzu	Mixed	S/E	Multiple	A	(Chivian et al., 2003)
InterPro	Mixed	F/E	Database	A/St	(Hunter et al., 2009)
Li et.al.	Auto	E	Method	A	(Li et al., 2012)
Nagarajan's	Mixed	S	Web App.	A	(Nagarajan and Yona, 2004)
PDP	Auto	S	Application	St	(Alexandrov and Shindyalov, 2003)
PFAM	Auto/manual	F	Database	A	(Finn et al., 2010)
PIRSF	Mixed	S/E	Web App.	A	(Nikolskaya et al., 2006; Wu et al., 2004)
PRODO	Auto	E	Application	A	(Sim et al., 2005)
ProDom	Auto	F	Database	A	(Servant et al., 2002)
SBASE	Auto/manual	F	Web App.	A	(Dhir et al., 2010)
SCOP2	Auto/manual	F/E	Database	St	(Andreeva et al., 2014)
SCOP	Auto/manual	F/E	Database	St	(Murzin et al., 1995)
Sistla et. al.	Auto	F/E	Method	St	(Sistla et al., 2005)
SMART	Auto	F/E	Database	A	(Schultz et al., 1998; Letunic et al., 2012)
SnapDRAGON	Auto	S	Application	A	(George and Heringa, 2002)
SSEP-Domain	Auto	S/E	Web App.	A	(Gewehr and Zimmer, 2006)
ThreaDom	Auto	S/E	Web App.	A	(Xue et al., 2013)
Yalamanchili's	Auto	S/E	Method	St	(Yalamanchili and Parekh, 2009)

of the classes in some databases (Andreeva et al., 2014) and a call for a review of the concept of domain (Yegambaram et al., 2013). Let's consider here that there are many kinds of domains and that each one of them contains a given architecture. It is important to differentiate between structural, evolutionary, and functional domains. The former speaks more about the folding and packing of the protein structure. It is therefore more related to autonomic folding units (AFUs). It is also the concept that is probably the most relevant for protein structure prediction. The evolutionary domain is a partition of a structure based on conservation and homology in evolutionary time, and therefore primary structure contiguity is not required. In these kind of modules, the evolutionary events in protein structure (i.e. domain translocation, recombination, shuffling, splicing,etc.) can be traced. This type of domain is the one that matters the most for studies of natural history of the structures, but it can also be useful for protein prediction (i.e. assignment of homologs), drug design, and engineering. The latter type of domain, the functional domain, will be sometimes tightly related with the evolutionary one, since one of the main drives for the selection of structures in evolutionary time is function. In this kind of domain, the classification of residues' groups are based on the interaction with the catalytic site (in the case of enzymes) or any functional property that the structure might have. The latter domain type is probably the most important for drug design and protein engineering.

Largely because of the distinction about domains explained above, defining domain boundaries is still a hard problem. One of the reasons for this is that there are many forces shaping protein structures (i.e. physical, chemical, and biological). Also, it seems that there is an architecture below the level of domain that may be shaping residues interactions (Hleap et al., 2013b; Chapter 3). That lower-level hierarchy in protein structure architecture will be explored further in the next section.

4.2.1 Proteins as networks: Identifying domains by graph theory

Since a protein structure can be summarized as a chain of amino-acids interacting in 3D, a natural way of abstracting it is as residue interaction networks (RINs)(Doncheva et al., 2011) or protein structure networks (PSN)(Vishveshwara et al., 2009). By using edges as proxies for contacts between residues as nodes, a protein structure will

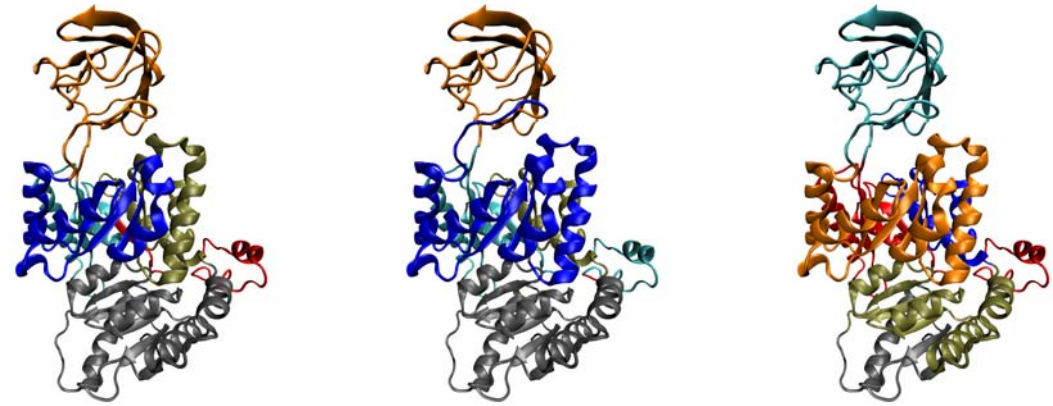
be shown as a graph or network. Let $S = (N, f)$ be an undirected graph representing a protein structure with N residues abstracted as nodes. f will be a function $f : N \times N \to \Bbbk$ that assigns an edge (and weight if necessary) to each residue pair. An edge E_{ij} is assigned only if $\exists(N_i, N_j)$, that is if residues N_i and N_j are in contact. This strategy can use many contact definitions (i.e. all atom, C_α, C_β; (for a test on contact definitions see Yuan et al., 2012)), and can have some other constrains. It can also be constructed as a weighted network, where the edge weight is a given property (i.e. hydrophobicity difference, van der Waals energy, etc.), including more information into the abstraction. These approaches have been already used for solving the domain boundary definition (Xu et al., 2000; Sistla et al., 2005; Yalamanchili and Parekh, 2009), and other applications in protein structures (Vishveshwara et al., 2002).

With this definition of protein topology, many graph theory applications can be used. In my case, the modularity of the protein can be explored as explained in Chapter 3 and equation 3.3. Here I will focus on the structural definition of domain. To do so, I will constrain he Q optimization to clusters following a contiguity criteria. That is, a given residue can only belong to a group if there is sequential connection of that residue with at one other member of the group. This optimization was performed with a genetic algorithm, for wich Q represented the fitness of the populations.

By optimizing the modularity score (Q; Equation 3.3) we can obtain the partition in the protein that contains more contacts within than in between clusters. It is therefore a proxy for the domain architecture of the protein, at least at the structural definition level.

Figure 4.2 shows the results of a graph in which partition has been constrained by contiguity, that is, a graph with unweighted edges representing the contacts in the structure, and optimized keeping the sequentiality. For the PK example here, let's restrict ourselves to interactions based on contacts. As can be seen, there is an over-fragmentation of the anticipated domains. This particular behaviour is sensitive to a number of issues:

1. Protein crystal is a rigid body: While the crystal is a rigid body showing some contacts, the protein in a functional environment is flexible, and therefore the contacts abstracted might not represent true contacts but a crystallization bias.



(a) 4.5Å all neighbours (b) 4.5Å NSN (c) 5.0Å AN

(d) 5.0Å NSN (e) 6.0Å AN (f) 6.0Å NSN

Figure 4.2: Partition of the rabbit's PK (PDB code: 1PKN) by means of modularity optimization of the contact map with different contact thresholds. Clusters are colored differentiating the membership of each residue. Clusters in the Pyruvate kinase (PK) were inferred by means of modularity (Q) optimization of the structure abstracted as a residue contact network (RCN). The contact was defined as an all-atom interaction less than a threshold. Figures 4.2a and 4.2b are based in residues within 4.5Å; Figures 4.2c and 4.2d with 5Å; and Figures 4.2e and 4.2f within 6Å. Figures 4.2a, 4.2c, and 4.2c show full sequentiality while figures 4.2b, 4.2b and 4.2f do not include immediate neighbors up to the second position in the sequence. The optimization was constrained to sequentiality by means of a Genetic Algorithm for the optimization of Q. **AN:** All neighbours. **NSN:** no sequential neighbours.

2. Too many parameters to tune: Using only the contact map can give you approximate answers (Figure 4.2). However, to render such accuracies, different parameters have to be tuned, such as modularity normalization parameter, number of contiguous neighbours to be disregarded, and again, the contact definition/threshold (Duarte et al., 2010; Yuan et al., 2012).

3. The domain definition: Depending on what you are optimizing for, your domain definitions will likely have different answers.

4. A hierarchical architecture of protein structures: Modularity solves for one optimal partition and disregards hierarchies of organization.

In the particular example in Figure 4.2, despite the over-fragmentation, the inferred modules are somewhat sensible in figures 4.2b, 4.2d, and 4.2f. The partition of the N-terminal domain (mid domain; red in Figure 4.1) was expected since this section is composed by two non-sequential fragments. However, based on the agreement on most contact definitions, there seems to be more than one contact-based domain, implying either a domain definition bias or a subdomain architecture of the N-terminal domain. However, the disparity in results remains, and the parameter tuning can become *ad hoc*.

There are software packages available that make use of contact graphs and render reasonable results (Xu et al., 2000; DomainParser). However, such programs do not rely entirely on the contact map, and normally include other variables and normalization procedures to achieve a visually reasonable Domain-level resolution. Also, the accuracy reported in such software is measured in a given domain definition that might not be the one being sought in a given problem (i.e. SCOP domain definition in DomainParser software). Another issue with the software development is that given a growing amount of citations of databases like SCOP, most programs optimize their algorithms to fit such domain. Such approaches try to bypass the interesting properties of protein structure in an attempt to reproduce manual assignment.

4.3 Sub-domain architecture

The hierarchical architecture of proteins can be explored in many ways. The contact maps explored in the previous section showed a clear example of how a level smaller

than the domain affects the way we can predict higher level modules. Even in lower levels of architecture, the definition of modules (sub-domains) is potentially interesting. In Figure 4.2 we can see that by constraining the modularity (Q) partition by sequence contiguity, structural sub-domains can be found. This sub-domain might be including known elements such as close-loops (i.e. the tan module in figures 4.2a and 4.2b) and AFUs (i.e. most other clusters in Figure 4.2 can behave as such). Identifying such structural modules is of great importance in the development of new protein structure prediction software. Protein structure prediction currently is based on the domain boundary prediction with high variability in the accuracies obtained (Wang et al., 2014). Given that protein structure prediction is more related to folding thermodynamics, partition of the structure into its AFUs and close loops might give future developers better methodological and conceptual tools for protein structure prediction.

In the evolutionary/functional definition of modules, I have shown in Chapter 3 that by obtaining correlation information of the coordinates among residues across homologs, a graph theoretic approach can be employed to explore evolutionary (if samples are homologs) or dynamic modules (if the samples are snapshots of a molecular dynamics simulation). A quick protein BLAST (Altschul et al., 1997) restricted to the PDB database can be used in the PK protein example to gather homologous structures. With a strict cut-off ($e - value \leq 10^{-100}$) 42 protein structures were sampled, ranging from bacteria to human. Applying the method proposed in Hleap et al. (2013a) and Chapter 3, 5 significant modules were inferred (Figure 4.3).

It can be noticed that both structural modules and evolutionary ones, converge in the separation of the SCOP N-terminal domain into sub-domains. It seems that such separation is an evolutionary constraint in the TIM-barrel evolution, since the results I have shown in Chapter 3 and in Hleap et al. (2013a), the α-Amylase TIM-barrel have a similar pattern. This does not mean that the two domains are necessarily homologous, but does stress the need of selection to keep such fold.

These different kinds of data and modules support the idea of a hierarchical organization of protein structure architecture. (Andreeva et al., 2014) acknowledging this, created a new SCOP database (SCOP2) in which those motifs, sub-folds and other kind of classifications, are now included.

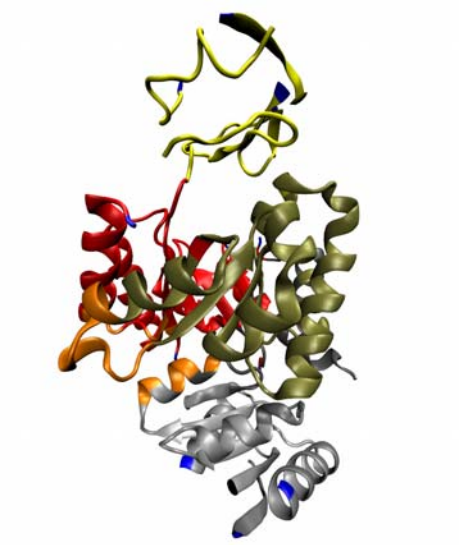


Figure 4.3: Evolutionary modules of the Pyruvate kinase rendered in the rabbit's PK (PDB code: 1PKN). After a protein BLAST restricted to the PDB database, 42 structures where gathered. By using the method in Hleap et al. (2013a), 5 modules were inferred. Non-homologous sections of the protein are stripped off, but its connections are highlighted in blue.

An unsolved issue about these sub-levels of architecture is the question of how many these are. In Chapter 3, I have shown that there might be more than one level in the evolutionary definition of sub-domain. This question is important in order to discover new patterns that can lead to the functional role of groups of residues (Csermely, 2008; Vishveshwara et al., 2009), drug discovery (Csermely et al., 2013), exploration of protein-protein interaction interfaces (Doncheva et al., 2011) and binding sites (Reichmann et al., 2007), protein-ligand interaction (Sathyapriya et al., 2008; Ozbek et al., 2010; Liu and Hu, 2011), or protein folding (Khor, 2012). Also, it is important to assess the contribution that scaffolding residues are having with respect to the architecture, since they seem to cluster together often. Luckily, there are tools such as RINalyzer (http://www.rinalyzer.de) (Doncheva et al., 2011) and protocols (Doncheva et al., 2012) to explore a more fine grain detail about the protein residue networks. However, these tools are a post-hoc solution for the domain and sub-domain boundary prediction and might serve as the exploration/validation of results.

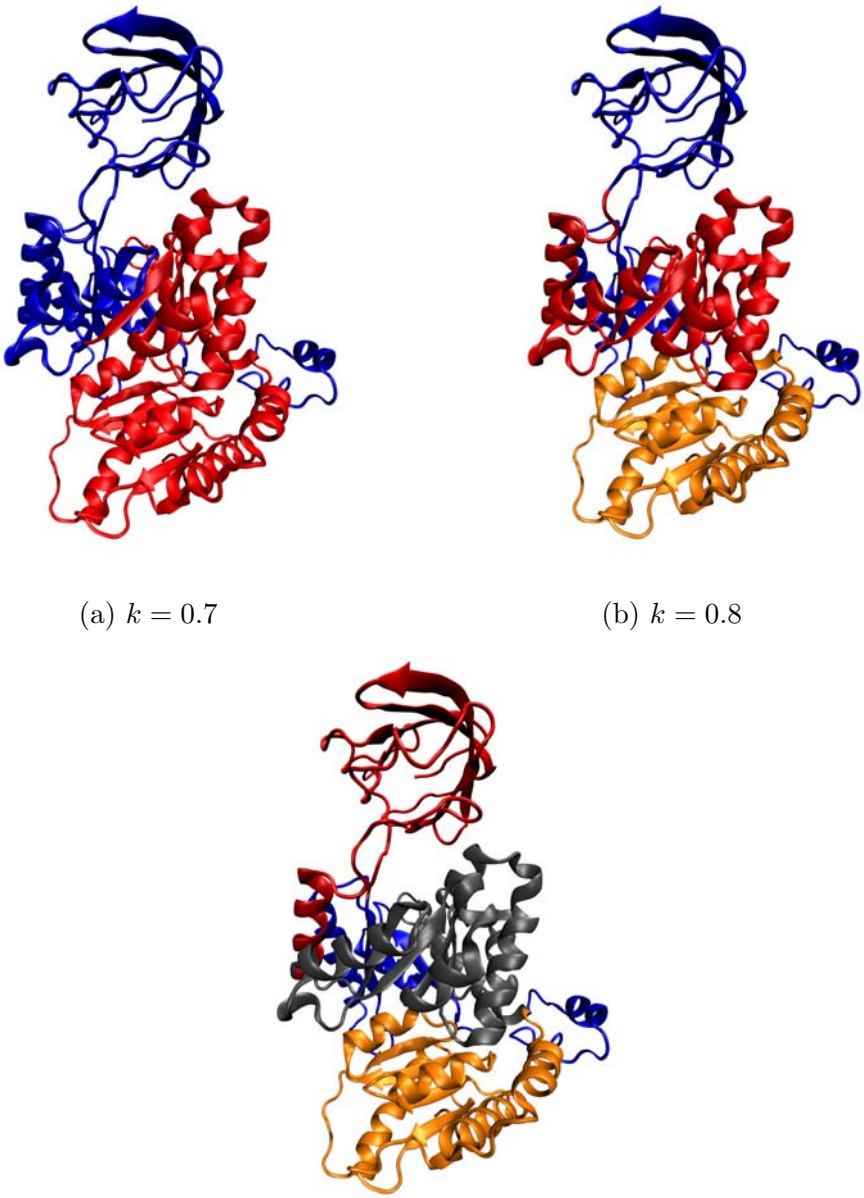
4.4 Exploring the hierarchy of protein structure architecture: Perspectives

In the previous section I have shown that there is a hierarchy in protein structure architecture. One possible way to explore this hierarchy relies on Newman-Girvan modularity (Equation 3.3) and related optimization algorithms. If we constrain the optimization of Q in such a way that it avoids weaker connections to be included within the resulting cluster, we might be able to "control" the level of exploration. Figure 4.2 shows how a constraint on the sequentiality applied to the modularity optimization (here performed with a genetic algorithm) will yield different levels of modularity. We can also optimize $k^m \times Q$, where k is the degree of constraint and m is the number of modules. With this constraint, optimizing $k^m \times Q$ also implies minimizing m, or the number of modules. This means that a partition is penalized if not enough edges are shared, and the inverse of k can be interpreted as the "strength" or degree of improvement needed to further break the protein into modules. That lack of information for a given level can be controlled by the degree of constraint. Figure 4.4 shows the results of such strategy for three levels of k: 0.70, 0.80, and 0.90.

With this approach different levels of modularity on a structural domain definition can be seen, as well as control numerical issues that can arise in a sparse graph. In Figure 4.4a only two modules are found that, when inspected, might reflect a higher level organization of folding. In Figure 4.4b the partitions of the structure are already similar to the SCOP C-terminal domain (orange), half of the N-terminal domain, and the rest, showing how a second level of interaction reveals a hierarchical architecture. Finally, in Figure 4.4c a partition into the SCOP domains can already be seen, plus the partition (by sequentiality) of the N-terminal/TIM-barrel domain. These results suggest the sub-domain architecture is a real phenomenon, and that this modularity has a hierarchical nature. It also shows that by means of optimization of $k^m \times Q$, this hierarchical sub-domain architecture can be explored.

Within a graph theoretic framework, there are other approaches that can also be used to explore the hierarchical nature of protein structure architecture. Most graph software use a dendrogram to partition the network. Each level of the dendrogram

can be scored for Q in the different levels, and therefore get an idea of the modular and hierarchical architecture of the structure. Other approaches use statistical tests to assess the quality of hierarchical partitions. An example of such algorithm is the



(a) $k = 0.7$ (b) $k = 0.8$

(c) $k = 0.9$

Figure 4.4: Weighted modularity optimization. Optimization of the Newman modularity score Q constrained by a weight and sequentiality. The weight is the Q multiplier k^m, where k is the degree of contrain and m the number of modules generated. The constrained optimization was performed using a genetic algorithm.

OSLOM (Order Statistics Local Optimization Method) (Lancichinetti et al., 2009, 2011) which simultaneously evaluates the hierarchical and the overlapping community structure of graphs. This method includes a fitness function with an α parameter, similar to the weighted Q presented here.

Approaches like these should be explored further in the future in order to fully understand the protein structure architecture.

Chapter 5

The response to selection in protein structures: A comparative quantitative genetics approach

Let's assume that protein structures are phenotypes and that their 3D structures respond to both genetics and environment. Because of this, protein structures' coordinates can be treated as phenotypic traits and can be analysed within a quantitative genetics framework. In this framework, the observed variance of a phenotype is (Falconer and Mackay, 1996):

$$V_P = V_G + V_{GE} + V_E \tag{5.1}$$

where V_G is the genetic variance, V_{GE} is the variance of the interaction between environment and the genetic component, and V_E is the variance caused by the environment. Each variance component can be decomposed into finer components (Falconer and Mackay, 1996):

$$V_G = V_A + V_D + V_I \tag{5.2}$$

where V_A is the additive genetic variance, V_D is the dominance variance, and V_I is the epistatic interactions variance. In evolution, the additive variance is the most relevant from the equations above. This is due to the fact that is the additive genetic component which is "adding" to the phenotype by genes. It is also the only one that can be selected since the environmental, dominance, and interaction effects are not encoded in genes. These other components are therefore not affected by directional selection but might affect estimations. The dominance effects, for example, can bias the estimation of the genetic component upward. This dominance variance can be selected only if there is no fixation in the genes (heterozygotes still exist), and if the dominance shifts in the phenotypes have a selective pressure, otherwise they tend to zero. Despite that, selection can affect the dominance effects and their contribution to the genetic component, however their intensity is typically small.

In protein structures, the dominance effect is not present in the model since the phenotype being analysed comes from a unique sequence, thus it can be regarded as an haplotypic phenotype.

Decomposing the phenotype and the corresponding quantitative analysis can be performed by analysing the phenotype under the light of kinship in a generalized linear model (GLM) (Falconer and Mackay, 1996):

$$y_i = \mu + a_i + e_i \tag{5.3}$$

where y_i is a phenotypic trait being measured in individual i, μ is the population mean, a_i is the breeding value or the genetic effect on the mean value, and e_i is a residual value.

One can generalize the model of equation 5.3 and include other effects as fixed effects for which we have sampled all intended levels. One example of the usage of fixed effects is to extract sexual dimorphism from the analysis. With this generalization commonly named *the animal model*, both random (a_i) and fixed effects (b) can be estimated (Wilson et al., 2010):

$$y_i = \mu + b_i + a_i + e_i \tag{5.4}$$

The estimation can be done with regression analysis given filial designs (i.e. mother-offspring, half-siblings, full-siblings, etc ...). The filial design is needed since it is the one that allows to partition the phenotype into relatedness components. A pedigree is therefore required for this kind of model.

All above background is applied to multigenic single traits. That is, a particular trait encoded by a number of different genes that needs to interact in order to "form" the trait (i.e. length of a femur, number of abdominal bristles in an insect, milk production in cattle, etc).

Proteins and other shapes are highly multivariate in nature (Klingenberg and Leamy, 2001). To comply with such nature a multivariate equivalent of the models expressed above can be formulated as (Thompson, 2008):

$$y = Xb + Za + e \tag{5.5}$$

where y represents a vector of observations of multiple traits, X and Z represent design matrices for the fixed and random effects in vectors b and a respectively, and e is the residual component that cannot be explained by the model.

The additive variance (V_A) is the actual contribution from genes to the phenotype. V_A is also the portion of the genetic variance that can be passed from parents to offspring and be selected. The direct relationship between V_A and selection is expressed in the *breeder's equation* (Falconer and Mackay, 1996):

$$R = \frac{V_A}{V_P} * S \tag{5.6}$$

where R is the response to selection of a population of a given trait, S is the selection differential, and $\frac{V_A}{V_P}$ is also called narrow-sense heritability (h^2). Lande and Arnold (1983) generalized this for a multivariate case as:

$$\Delta \bar{z} = G\beta \tag{5.7}$$

where $\Delta \bar{z}$ is a vector of changes in traits, G is the genetic variance-covariance matrix (G-matrix), and β is a vector of selection gradients.

As explained in Chapter 2, protein structures can be abstracted as shapes and it is known that shapes are multivariate traits (Klingenberg and Leamy, 2001). Monteiro et al. (2002) proposed the use of an univariate approach for multivariate traits, particularly shapes, to estimate the heritability. However, Klingenberg (2003) demonstrated that Monteiro et al. (2002) approach only works when all variation in each landmark is uniform in all orientations. It is intuitive that this is hardly the case for most shapes. Therefore an estimation of the G-matrix is required (Klingenberg, 2003); this task is not trivial. To deal with the fact that the number of samples is limited, that inversion of matrices require expensive computation, and that an eigen decomposition is required, the restricted maximum likelihood (REML) approach has been traditionally employed to carry out the variance decomposition. When applied to univariate data, it is more accurate since it deals better with missing data (i.e. unknown parents, arbitrary breeding designs, etc) and can account for selection processes. However, REML has good properties only asymptotically. The reliability of the estimates is questionable when data is scarce. One way to deal with complex cases that might bias the REML estimates is to use Bayesian inference of the animal

model. This approach uses Markov chain Monte Carlo simulations and is a more robust estimation than REML, with equivalent results in less complex cases, however (Blasco, 2001). This robustness assumes that the Bayesian model has enough information in the prior probability distribution. This can considerably affect the estimation of the variance components. In particular, uninformative priors, such as flat priors, can lead to biases in the estimation. Another approach that can be used when the number of variables is large is factor-analytic (FA) models and structural equations modelling (SEM). These strategies have been proven to have an equivalent robustness in the estimation of genotype effects to more traditional models (for a review on such models refer to Smith et al., 2005; Piepho et al., 2008). FA models allow a higher dimensionality by partitioning the estimations into latent factors (Meyer, 2009; Runcie and Mukherjee, 2013). Latent factors are variables that are not directly observed but inferred through a model, using the directly observed variables to estimate them. The definition of the number of latent factors to use (defined by the researcher) is the downside of FA models. A wrong number of latent factor can significantly bias the estimations of parameters. FA models also assume that the between-factors covariance is a single value. This can warp the response to selection. Also, FA models have an increased square root of mean square error when estimating the genetic component as the rank of the matrix get smaller (Meyer and Kirkpatrick, 2008).

5.1 Lynch's comparative quantitative genetic model: Applications in protein structures

GLM estimation of quantitative genetics parameters rely on the filial design. Normally, filial design needs a pedigree to express the relationship within families and then the variance can be decomposed. Such pedigree can be summarized in an inbreeding or relationship matrix, and incorporated into the GLM.

In protein structures this type of relationship is not retained since most of the sampling is performed in different species. Lynch (1991) developed a wider generalization of the linear mixed models, in what he called *phylogenetic mixed model (PMM)*. In this model, the correlation of phylogenetically heritable components is

the time to the shared common ancestor (length of the path from the most recent common ancestor among two species and the root) in the phylogeny (Housworth et al., 2004; Hadfield and Nakagawa, 2010). The PMM can be described as (Lynch, 1991):

$$\bar{z}_{ci} = \mu_c + a_{ci} + e_{ci} \tag{5.8}$$

where $\bar{z}_{ci}$ is the observed mean of the trait, μ_c is the grand mean of the cth character over the phylogeny, a_{ci} is the additive (phylogenetic) value the cth character in the ith taxon. Finally, e_{ci} is the residual deviation or Cheverud's specific effect (Cheverud et al., 1985) if within species data has been sampled.

An assumption of the model is that μ_c is shared among all taxa in the phylogeny. This is a sensible assumption to make when analysing truly homologous protein structures, since the mean effect on the phenotype is shared by common ancestry. This also means that $\mu_c + a_{ci}$ can be interpreted as the heritable component of the mean phenotype for the ith taxon (Lynch, 1991).

The univariate model in equation 5.8 can be generalized to any number of characters by (Lynch, 1991):

$$\bar{z} = X\mu + a + e \tag{5.9}$$

where X is an $np \times p$ incidence matrix, p being the number of traits and n the number of observations.

As can be seen, equations 5.4 and 5.5 are almost identical to equations 5.8 and 5.9. The differences are in the definition of the effects, equating the additive effects to be phylogenetic. Also, in Lynch's multivariate equation the fixed effect is equated to the mean.

Here, the phylogenetic effects are the portion of the variation that has been inherited from ancestral species (Cheverud et al., 1985). It does not only contain the genetic component, but also some environmental contributions given the shared evolutionary history of the taxa (Housworth et al., 2004). In PMM, analogous to traditional quantitative genetics, the ratio between the additive component and the total variance is the heritability (h^2) in an univariate approach. Housworth et al. (2004) point out also that an univariate h^2 in a PMM is actually equivalent to Freckleton et al. (2002)'s and Pagel (1999)'s phylogenetic correlation (λ).

Martins et al. (2002) showed that the PMM is comparable to the phylogenetic generalized least squares (PGLS), spatial auto-regression, and phylogenetic eigenvector regression. They all yielded good statistical performance, regardless of the evolutionary model and even when some of the assumptions where violated.

Despite the robustness of the models, the REML technique has two major drawbacks: assumption of normality, and restrictions in the sample size . It is widely known that REML would poorly estimate the genetic correlation when overparameterized (multi-trait inference), the sample size is small (Martins, personal communication), and when the normality assumption is violated (Hadfield and Nakagawa, 2010). These violations can be handled in a Bayesian framework using Markov Chain Monte Carlo techniques. In such techniques, the higher complexity of the joint probability calculation needed for the likelihood estimation can be broken down in lower dimensional conditionals. From those conditionals the MCMC sampling can be performed and marginal distributions can be extracted (Hadfield and Nakagawa, 2010). The discussion of the usage of Bayesian MCMC techniques is beyond the scope of this thesis. I refer the interested reader to Sorensen and Gianola (2002) for a better description of likelihood and Bayesian methods in quantitative genetics.

Despite its strengths, the Bayesian framework also has weaknesses. The most important one is that it requires proper and informative priors. Too vague or uninformative priors lead to important biases with high variation in results. This is why the sensitivity to the choice of prior distribution should always be assessed (Lambert et al., 2005). Given that in evolutionary biology datasets the amount of knowledge on the estimator is scarce, well informed priors are normally not available and by informing priors with partial information, the estimation can become ill-conditioned.

5.1.1 Computational infeasibility of the full comparative approach

In a multivariate scenario, both REML and Bayesian methods have a computational chokepoint. This chokepoint is in the matrix operations needed to estimate a large number of parameters. As the matrix grows bigger in both number of individuals (n) and number of traits (p), the computation becomes untractable in terms of memory and time. For p traits, $p(p+1)/2$ covariance components need to be estimated per random effect, and thus the estimation burden increases quadratically (Meyer and

Kirkpatrick, 2008; Houle, 2010).

To test how feasible the methods are, I developed a simulation where I constructed the phenotype from its components. Following equation 5.9, the phylogenetic effects (a) were simulated using the `rbv` function in the `MCMCglmm R` package (Hadfield, 2010) to create a matrix of randomly generated multivariate normal phylogenetic effects. The generation of phylogenetic effects was constrained with a known $p \times p$ genetic matrix (G) that was drawn from an inverse Wishart distribution with an scale matrix with one in the diagonal and 0.5 in the off diagonal entries. Also, a random tree was generated by the functions `compute.brtime` and `rtree` of the `ape R` package (Paradis et al., 2004). The error component (e) was built by performing a Cholesky factorization of a known $p \times p$ covariance matrix (E) and multiplying the decomposed matrix with a matrix of random values of shape $n \times p$. The known E matrix was also drawn from an inverse Wishart distribution, and its scale matrix contained ones in the diagonal and 0.1 in the off diagonal. Summarizing, the simulation was performed by:

1. Create the known phylogenetic (G) and error (E) covariance matrices
2. Using G to constrain the simulation of the phylogenetic effects (a), simulate the $n \times p$ matrix a given a random tree
3. Incorporate the desired covariation E into a $n \times p$ random matrix
4. Add the error and phylogenetic effects

Table 5.1 shows the time and memory spent in a simulation fixing n to 100 and varying p. The estimation was performed using Lynch's PMM model as implemented in the `R` package `ape` (Paradis et al., 2004), where many matrix inversions and Kronecker products are involved (see Lynch, 1991; for details). This computation was performed in a PC Intel®Xeon®CPU E5-2620 v2 @ 2.10GHz Intel i3 3.10GHz 128 Gb 2x hexa core processor.

Table 5.1 and Figure C.1 show that at 16 traits it required over 300 Mb of memory and over 9.4 hours to compute. The behaviour with 32 traits is erratic. In the current simulation the computation cannot be performed since the memory requirements are too high. The problem scales up quickly not only with the number of traits, but also

Table 5.1: Feasibility of the phylogenetic mixed model (PMM). Memory, time and accuracy of the PMM using Lynch's and Bayesian approaches. RS_A correspond to the random skewer test for the phylogentic covariance and RS_E for the residual. Bold values indicate correlation of random skewers greater than 0.9 and significant.

Method	**Traits**	**Time (secs)**	**Memory (Mb)**	RS_A		RS_E	
				p-val	ρ	p-val	ρ
Lynch	2	113.44	13.9	0.109	0.955	0.036	**0.997**
	4	1132.34	27	0.004	**0.952**	0.000	**0.995**
	8	1780.78	77.7	0.003	0.878	0.000	**0.997**
	16	34159.34	276.8	0.000	0.888	0.000	**0.998**
BGLMM	2	35.94	50.5	0.103	0.948	0.024	**0.997**
	4	110.36	62.2	0.002	**0.971**	0.001	**0.997**
	8	624.88	107.8	0.000	**0.940**	0.000	**0.994**
	16	2508.96	310	0.000	0.897	0.000	**0.995**
	32	9777.48	1110	0.000	**0.904**	0.000	**0.995**
Algorithm 2	2	37	47.4	0.062	0.975	0.026	**0.996**
	4	227.96	60.5	0.018	0.890	0.000	**0.996**
	8	1714.86	74.5	0.002	**0.911**	0.000	**0.992**
	16	6066.8	127	0.000	0.871	0.000	**0.994**
	32	17254.76	406.3	0.000	0.878	0.000	**0.993**
	64	70207.5	427.5	0.000	0.858	0.000	**0.992**
	128	317766.02	74.3	0.000	0.866	0.000	**0.993**

with the number of individuals. This trend is due to computation of the inverse of the relationship and the identity matrices. The computation of these matrices take the most time (≈94%) and memory (≈60%). Also, the time and memory required to compute the Kronecker product of these very large matrices scale up quickly with the dimensions of the input matrix. It can take up almost a third of the spent memory, showing the dependency between complexity and both n and p. The reliability of the estimates is also affected. The sample size needs to be increased in order to estimate the covariance matrices with confidence. However, by increasing the sample size the computation becomes more complex. To test the extent of the bias Table 5.1 shows the mean correlation and corresponding p-values of the Cheverud's Random Skewer (RS) test (Cheverud, 1996a; Cheverud and Marroig, 2007) implemented in the `R` package `phytools` (Revell, 2012). Despite that works such as Bégin et al. (2004) have contentions about any given covariance matrix comparison methods, Cheverud's

test is better suited for my framework. It introduces random vectors of change and compares the correlation of resulting vectors. This is in line with the quantitative genetic framework as in equation 5.7. It is also better suited for comparative studies than other tests of equality such as Anderson's maximum likelihood test of equality of covariances (Anderson, 1958) or the common principal component (CPC) analysis (Phillips and Arnold, 1999). Those tests have big biases given the sample size and the number of traits. Steppan (1997) showed a positive relationship between the number of traits and and the likelihood of rejecting equality.

Surprisingly, despite the expected instability of the matrix estimation given the sample size, most of the estimation were highly correlated (>0.70) and significant. It is important to state that the (RS) test does not evaluate equality of the covariances (in fact most covariances were very dissimilar). However, the overall response to a vector of selection is highly correlated. For the purposes of this thesis, the matrix equality in response to disturbances is more relevant than the exact match between the covariance matrices' values. This is because the response to disturbances expressed in the RS test, follows the same framework as the expression detailed in equation 5.7.

Bayesian solution to the memory requirements

Given that Bayesian generalized linear mixed models (BGLMM) use Markov chain Monte Carlo (MCMC) simulations and usually Gibbs sampling, the memory requirements should lower significantly. However, Table 5.1 show the same trend on a different scale when using the `R` package for BGLMM `MCMCglmm` (Hadfield, 2010).

I simulated up to 32 traits using approximately 2 hours in the bigger dataset when the sample size and the number of MCMC iterations are held constant. The results show that it was not the memory but the time that benefited from the Bayesian approach since over 1Gb was used (Table 5.1 and Figure C.3). However, this memory requirement can be lowered if fewer MCMC iterations are performed. Nevertheless, lowering the number of iterations can only be done on per case basis since the convergence has to be guaranteed.

Data in Table 5.1 describe the accuracy of this approach. With these particular data, it behaves better or similar to Lynch's approach while being faster when the

number of traits is high. With the given dataset, when more than 32 variables were analysed, `MCMCglmm` (Hadfield, 2010) reported ill-conditioned priors when completely flat ones (identity matrices) were used. With this in mind, and the fact that on 32 traits the memory requirements remain over 1Gb, a new approach was required.

5.1.2 Dealing with computational constraints: An approximation to the G-matrix

Dealing with time constraints is easier than dealing with memory constraints. For this reason, a naïve approach can be developed by exploring all pairwise estimations (Algorithm 2). This approach is time consuming, but theoretically memory efficient.

With this approach, $\binom{p}{2}$ estimations are computed. This means that many repetitions of estimated variances are explored. To choose the best value among the repeats, one can keep track of each estimation and its Deviance Information Criterion (DIC), and select the best estimation of each pair to reconstruct the full matrix. The `R` package `MCMCglmm` was also used for the pairwise estimations.

The computational cost of this approach is shown in Table 5.1 and Figure C.5. There it can be seen the high time constraint but the memory feasibility of this method. Despite the time cost this method can be easily parallelized. This approach cannot currently be applied with the full-matrix method. If parallelized, Algorithm 2 could in theory take the same amount of time to estimate the parameters as a 2 trait full-matrix analysis. This is true, independently of how many variables are being analysed, if the number of processors are equal or greater than the number of pairs explored.

Algorithm 2 computes only an approximation to the true matrix, since this method does not account for the covariance of other pairs in the estimation. To test this, I ran this algorithm and used the RS test the response to random vectors between the real matrix and the pairwise-re-constructed matrix (Table 5.1 and Figure C.6).

Despite its feasibility, algorithm 2 showed a lower accuracy obtained in the correlation of the response to selection vectors. This can be due to an overly simplistic pairwise approach, but also by the limited sample size.

Algorithm 2 Pseudo-code for the partial G-matrix estimation. This algorithm is a high-level pseudo-code where not all details are expressed.

Input: $gm \leftarrow$ A dataset of homologous coordinates.

$P \leftarrow gm$ Extract the information as phenotypic matrix P

$N \leftarrow$ Extract the column labels of P as trait labels

$S \leftarrow$ Create an array of all the pair combinations of N

for $s \in S$ **do**

$G_s \leftarrow$ Estimate the pair variance-covariance matrix

$M \leftarrow G_s$ Store variance-covariance matrix in an array

$D \leftarrow D_s$ Compute and store the DIC for the matrix G_s

end for

G Prepare a matrix of $p \times p$ dimensions

for $i \in G$ **do**

$d \leftarrow$ minimize D where $i \in s$ and store variance with minimum DIC

$G_{ii} \leftarrow d$

for $j \in G$ **do**

if $i == j$ **then** continue ▷ Do not compute the same variance twice

end if

$e \leftarrow$ minimize D where $j \in s$ and store variance with minimum DIC

$G_{jj} \leftarrow e$

$G_{ij} \leftarrow M_{ij}$

end for

end for

Fill G_{ji} with G_{ij} **return** G

Sample size effect in the estimation

It is known that the sample size is also an is issue in the estimations of G specially when the number of traits increases. To test this assumption, I simulated and measured time and memory usage for the estimations following the same strategy used in section 5.1.2. In this case, the simulation has a fixed p equal to 8 and varying levels of n: low (16 observations), medium (64 observations), and high (256 observations). Table 5.2 shows the requirements in time and memory, as well as the average accuracy and the standard deviation for 10 replicates. The estimated mean is slightly biased resulting in a decrease in the mean. These trends can be seen in Appendix C.7

Table 5.2: Effect of sample size in the phylogenetic mixed model(PMM). Average (± standard deviation) memory, time and accuracy of the PMM using Lynch's and Bayesian approaches. RS_A corresponds to the random skewer test for the phylogenetic covariance and RS_E for the residual.

Method	**Samples**	**Time (secs.)**	**Memory(Mb)**	RS_A	RS_E
Lynch	16	43.76±53.77	15.77±0.05	0.72±0.26	0.85±0.26
	64	2943.23±2746.06	37±0	0.72±0.24	0.98±0.04
	256	111986.44±53576.95	444.63±0.05	0.71±0.26	1±0
BGLMM	16	70.41±7.8	100.62±1.96	0.67±0.2	0.86±0.25
	64	343.31±65.44	102.82±9.73	0.7±0.12	0.9±0.31
	256	1198.12±100.73	123.06±37.9	0.78±0.15	1±0
Algorithm 2	16	235.32±22.64	57.01±5.3	0.63±0.3	0.93±0.09
	64	973.11±158.28	65.88±3.41	0.73±0.17	0.99±0.02
	256	3002.19±333.75	158.04±114.79	0.66±0.26	1±0

It is expected that accuracy increases with sample size. However, Table 5.2 shows that this is only significantly true for the residual matrices. This phenomenon might be attributable to the differences in the scaling structure in the original simulated matrices.

Data from Table 5.2 also suggest that the cost in time and memory can be prohibitive. However, a greater problem is that obtaining sample sizes of over 200 protein structures is not always possible. One possibility to increase the sample size is to include snapshot structures from molecular dynamic simulations of each of the

homologs. This approach also introduces an extra component to the variance: the within group (within homolog/species) component. Given that this component is actually of interest, it is worth to assess it separately. Section 5.1.3 explores this issue.

5.1.3 Beyond the OTUs: partitioning the variance within taxonomic units

We know by equation 5.1 that the phenotypic variance can be explained by the genetic, environmental, and interaction components. It is also known that if repeated measures of a trait are available, the variance can be further partitioned into a third component. In general comparative evolutionary biology, such components may include differences among populations, phenotypic plasticity, sampling variation, instrument-related error, physiological state, variation related to age, sex, season, or time of day, among others (Ives et al., 2007). All these sources of variations greatly depend on the way the sampling was done and the trait in hand. By including a within-group, within individual in traditional quantitative genetics or within species in comparative studies, nuisance parameters can be dealt with. Many studies have shown the importance of dealing with measurement errors and deviations from the between-group analyses (Harmon and Losos, 2005; Ives et al., 2007; Felsenstein, 2008; Hadfield and Nakagawa, 2010; Garamszegi and Møller, 2010; Silvestro et al., 2015). Therefore the within-group analysis is ideal. In protein structures, repeated measures can be taken as snapshots of molecular dynamic simulations of a protein in solution (as performed in section 2.2.2). Thereby adding a partition to the variance. In this set up, another variable is added to the model:

$$\bar{z} = Xb + Za + e + m \tag{5.10}$$

where m is the matrix of individual effects or effects of the dynamics of a protein.

This approach has an application in structural biology, since it allows partitioning the structural variation into:

1. Phylogenetic component: This component will provide information about the evolutionary constraint in the protein structure. Therefore, it can be used in

informing decisions of protein engineering, structural constraints in bioinformatics, etc.

2. Dynamic component: This component provides information about the thermodynamic constraints in a set of proteins. It can be use as before to guide within species protein engineering and on the analysis of the dynamics of a given fold.

3. Residual component: This will encompass all other components of the phenotype including noise, different sources of error inclusive of measurement error, and other environmental factors.

This approach is ideal since it allows to dissociate the phylogenetic (evolutionary) variability from dynamic variability. Liu and Bahar (2012) show a correlation between sequence evolution and dynamics. However, in their approach both entities can be confounded by permanent effects created by descent. Ideally, this bias should be stripped to avoid an artificially increased estimation of the correlation. Using the linear mixed model to estimate this components implies the inclusion of extra parameters to estimate. To test the computational complexity and the accuracy, the same approach as in section 5.1.1 was used, but including a dynamic term in the simulation and the model. This term was included by:

1. Generating a Multivariate normal $p \times r$ matrix O, where r are the number of repetitions.

2. Correlating the variables with a known covariance matrix M (drawn from an inverse Wishart distribution), by means of Cholesky decomposition.

3. Populating a MD matrix $(n * r) \times p$, with the correlated matrix by repeating it n times.

4. Generating a multivariate normal $(n * r) \times p$ matrix of independent effects.

5. Adding the MD matrix with the independent effects.

The results can be seen in Table 5.3.

With the inclusion of extra parameters there is a significant increase in the time and memory required for the estimation. However, the high accuracy in the phylogenetic covariance matrix estimation is surprising since a more unstable estimation

Table 5.3: Accuracy and feasibility of the two random effect PMM. Memory (Mb), time (sec) and accuracy (random skewer correlation) of the PMM using a Bayesian approach with two random effects. RS_A corresponds to the random skewer test for the phylogenetic covariance, RS_M for the dynamic component, and RS_E for the residual.

Method	**Traits**	**Time (hours)**	**Memory (Mb)**	RS_A		RS_M		RS_E	
				p-val	ρ	p-val	ρ	p-val	ρ
BGLMM	2	1.497	185.1	0.033	0.992	0.077	0.963	0.227	0.758
	4	3.635	255.1	0.005	0.968	0.003	0.958	0.095	0.681
	8	11.1946	457.4	0.000	0.947	0.001	0.868	0.060	0.554
	16	40.562	1184.4	0.000	0.931	0.000	0.835	0.029	0.450
Algorithm 2	2	1.624	213.4	0.052	0.990	0.090	0.958	0.222	0.758
	4	8.663	233	0.004	0.977	0.012	0.957	0.104	0.701
	8	36.974	252	0.000	0.950	0.001	0.868	0.046	0.592
	16	149.193	245.9	0.000	0.946	0.000	0.842	0.011	0.583

was expected. The accuracy was anticipated to diminish quickly with the number of variance-covariance components to be estimated. For 3 random effects, there are $\frac{3q(q+1)}{2}$ variance-covariance components that have to be estimated. This complexity makes the estimation of parameters unstable and intractable. However, data in Table 5.3 show high correlation values for the phylogenetic and acceptable values for the dynamic component. In the case of the residual value, most correlations were low and non significant. This is also an interesting observation since the trend seen in previous sections showed that an accurate estimation of the residual matrix was more feasible than the phylogenetic one.

Despite an improvement in accuracy, it can be seen that the time and memory required for the computation are still prohibitive; therefore, when q and n are high, another approximation is needed.

5.2 Overcoming over-parametrization: Approaching the G-matrix by means of the P-matrix

Given the results in section 5.1, the estimation of the G matrix within the Lynch's PMM is infeasible. This is not a new observation since in comparative evolutionary biology it is widely known that accurate measures of G are difficult or impossible to obtain (Marroig and Cheverud, 2001). This pattern is even more evident when

dimensionality is high. On average, protein structures are composed of over 200 residues in a three-dimensional system, which means over 600 variables. Also, the sample size at the species level is typically small. Because of these reasons, a full and stable estimation of the G-matrix is not possible. However, as referred in section 5.1.3, an increased number of samples can be achieved by means of molecular dynamic simulations. This increases n considerably depending on the length of the simulation. I have shown the infeasibility of the GLMM to deal with the dimensionality and very large sample size. However, it has been shown that phenotypic (P) matrices can be estimated with more confidence with large sample sizes (Cheverud, 1988, 1996a; Roff, 1995). It is also shown that in some cases, P can be used as surrogate for G when the two are proportional (Marroig and Cheverud, 2001; Revell et al., 2007). To test this, the same simulation scheme as in section 5.1.3 was used. In this section, the simulation was performed with 500 replicates as molecular dynamics snapshots, 100 taxa, and the traits were varied from 2 to 1024 in a geometric series increase. Since the within-homolog matrix structure is known, a pooled-within covariance matrix (W) was computed as:

$$W = \frac{1}{n-S}\sum_{s=1}^{S}\left(\sum_{\omega:f(\omega)=s}[x_i(\omega)-\bar{x}_{i,s}]\times[x_j(\omega)-\bar{x}_{j,s}]\right)_{i,j=1,\cdots,p} \tag{5.11}$$

where S is a the number of categorical variables describing the groups or species, ω is an instance, were $f(\omega)$ correspond to the class value of the instance, and $\bar{x}_{i,s}$ is the mean of the variable i for individuals belonging to s. n is the sample size.

Here, W contains the covariance matrix of the within-homolog (i.e. Molecular dynamic data). To estimate the evolutionary component of P, the between structures/species covariance matrix (B) has to be taken into account. B will be simply the difference between the P and W.

Table 5.4 and Figure C.8 show the feasibility and accuracy of the pooled-within species covariance estimation method. Even with highly multivariate data (1024 traits), the memory requirement is manageable (less than 2 Gb). The evaluation is done in under an hour. The accuracy of the estimation is high, with the estimated G matrix being almost identical to the simulated one, and the estimated MD having

Table 5.4: Accuracy and feasibility of the pooled-within covariance estimation. Memory (Mb), time (sec) and accuracy (random skewer correlation) of the pooled-within covariance estimation approach. RS_B corresponds to the random skewer test for the phylogenetic covariance and RS_W to the dynamic component.

Traits	**Time (secs.)**	**Memory (Mb)**	RS_B		RS_W	
			p-val	ρ	p-val	ρ
2	0.60	182.9	0.002	1.000	0.021	0.999
4	0.80	238.2	0.000	0.999	0.007	0.952
8	1.00	387.6	0.000	0.998	0.000	0.983
16	1.82	407.5	0.000	0.998	0.000	0.963
32	6.08	428.5	0.000	0.998	0.000	0.966
64	20.32	465.9	0.000	0.999	0.000	0.953
128	91.14	539.4	0.000	0.999	0.000	0.947
256	341.90	686.8	0.000	0.999	0.000	0.950
512	1342.36	982.2	0.000	0.999	0.000	0.938
1024	5268.82	1843.7	0.000	0.999	0.000	0.937

over 0.97 correlated response to random vectors to the actual MD. This is a surprising result since this method cannot separate completely the error terms from the genetic and the dynamic component. However, it seems that the split of the error term between the two other components make the error terms negligible. Moreover, it seems that error does not affect significantly the structure of G and MD, allowing them to behave almost identically than the simulated counterparts. Given these results, and the fact that the application to real datasets can only be made with this approach, it seems reasonable to keep using this from this point forward. However, the biological and evolutionary meaning of this approach is less clear than in the other methods since there is no explicit use of a phylogeny.

5.2.1 The meaning of the pooled within-structure covariance matrix

It has been widely described that P-matrices can be used as surrogates of G-matrices in cases were they are proportional or sufficiently similar (Cheverud, 1996b; Marroig and Cheverud, 2001; House and Simmons, 2005; de Oliveira et al., 2009; Marroig et al., 2009; Porto et al., 2013). Prôa et al. (2013) showed that this assumption can be relaxed if the correlation between G and P is higher than 0.6. In protein structures,

we can assume that given the strong selective pressures and long divergence times that they have been subjected to, the relationship between P and G could potentially be standardized. Assuming that this is true in protein structures, the estimated pooled variance-covariance (V/CV) matrices in real datasets might have specific biological meaning. This has been described in Haber (2015) for morphological integration in mammals. Following Haber's (2015) logic, the translation of the pooled V/CV matrix could be as follows:

1. The within-structure/species (i.e. thermodynamic V/CV matrix) refers to integration of residues in a thermodynamical and functional manner, and also contains information about environmental factors affecting the physical-chemistry of the structure. Haber (2015) includes a genetic component for his within population estimation, since populations follow a filial design. My data, on the other hand, have a controlled amount of genetic component given that the sampling is done in a time series instead of a static population. My approach would be more related to an estimation of within repeated measures design.

2. The among-structure/species (i.e additive or evolutionary V/CV matrix) refers to the concerted evolution of traits given integration and selection (Haber, 2015).

Therefore the pooled within-structure covariance matrix approach is not only the one possible to compute for protein structures, but also biologically meaningful.

5.3 Response to selection: The case of α-Amylase

As portrayed in equations 5.6 and 5.7, the response to selection of a phenotype depends on the within-species change in mean due to selection, the correlation between different traits, and the amount of heritable component of the shape. The first component can be referred as $\beta = P^{-1}S$, and also known as the vector of selection gradients (Rausher, 1992) or directional selection gradient. The second and third elements are summarized in the G matrix. As expressed in equation 5.7, this covariance matrix represents the genetic component of the variation in the diagonal, and the correlated response of every trait to each other in the off-diagonal.

Another extension from equation 5.7 is to compute the long-term selection gradient assuming that G is more or less constant over long periods of time:

$$\beta_\lambda = G^{-1}\Delta\bar{z} \tag{5.12}$$

Here $\Delta\bar{z}$ would be proportional to the differences in mean between two diverging populations.

It is important to stress the relationship between these concepts and fitness. Given that fitness (w) is directly related to selection, its mathematical relationship can be expressed as (Blows and Walsh, 2009):

$$f = a + \sum_{i=1}^{n} \beta_i z_i + e_i \tag{5.13}$$

and so it behaves as the weights of a multiple regression of f on the vector of phenotypes z. Fitness is intuitive in organismal biology and can be represented by the count of the offspring of a given phenotype after an event of selection. In proteins, however, the definition is not as straightforward. We can portray fitness in many different ways, depending on the hypothesis being tested. If the analysis is done comparatively (i.e. across different protein structures from different sources), a fitness analysis including exclusively structural measures such as Gibbs free energy (ΔG) can be misleading. The fitness surface that can arise from this data would only represent departures from every individual native state. Nevertheless, ΔG and the energy of unfolding (ΔG°), are important measures to determine the stability of the protein which is important for the fitness of a protein structure. To improve this fitness landscape, f can be defined by ΔG° coupled with a functional measure. Since in proteins function is the main selective trait, including a term accounting for this would create a more realistic fitness surface. In enzymes this can be achieved by using the efficiency or K_{cat}/K_M of each of the enzymes for a given substrate. The fitness function (F) can be expressed as:

$$F(i,s) = \Delta G_i^\circ \frac{K_{cat}^{i,s}}{K_M^{i,s}} \tag{5.14}$$

where ΔG_i° is the free energy of unfolding of the structure i, $K_{cat}^{i,s}$ is the turnover number for structure i in substrate s, and $K_M^{i,s}$ is the the Michaelis constant of protein i working on substrate s.

As can be seen this is a relative fitness, and its relativity depends on the substrate in which is being computed.

In the case of the α-amylase family (GH13), one might try to apply the framework developed in previous sections and try to estimate the response to selection of a subset of them. However, for this dataset equation, 5.14 cannot be applied since the information of the relative efficiency given a common substrate is not consistently available. For this reason I am going to work exclusively with $\Delta G^{\circ}_{unfold}$, but knowing the caveats that this only speaks about structural stability and it has been shown that $\Delta G_{equilibrium}$ or $\Delta G^{\circ}_{unfold}$ are not optimized for during evolution (Alfaro, 2014).

5.3.1 Estimating dynamic and genetic variance-covariance matrices in the α-Amylase dataset

Given that molecular dynamic simulations are very time consuming, I subset the dataset presented in chapters 2 and 3, taking randomly one quarter of the 135 structures to a total of 34 protein structures (Table 5.5).

Following the methods depicted in section 2.2.2, 30 nanoseconds were simulated and up to 500 snapshots were sampled per simulation. The estimation of the pooled-within covariance matrix was performed as follows:

1. Align every model within each MD simulation using GPS: Remove extra rotations and translations that could occur during MD simulation.

2. Select an ambassador structure that is closest to the mean structure.

3. Align all ambassadors using MATT flexible structure aligner to identify homologous sites: Multiple structure alignment to identify structural homology.

4. Extract the centroid of the fully homologous sites (gapless columns): Identify shared information among all structures.

5. Concatenate the centroid information for all trajectories

6. Perform a GPS on the entire set of shapes: Bring all pre-aligned structures into the same reference plane.

7. Apply method described in section 5.2: Estimate Variance-Covariance (VCV) matrices.

Parallel to this, I computed the $\Delta G^{\circ}_{unfold}$ on each model for each protein using the command line version of `FoldX` (Schymkowitz et al., 2005). It is important to notice that the computed $\Delta G^{\circ}_{unfold}$ is not comparable in proteins of different size, therefore I computed the average $\Delta G^{\circ}_{unfold}$ per residue as $\Delta G^{\hat{\circ}}_{unfold} = \frac{\Delta G^{\circ}_{unfold}}{n}$, n being the number of residues. With this $\Delta G^{\hat{\circ}}_{unfold}$ as proxy for fitness we can try to explore the fitness surface. To do this, I used the first two principal components of

Table 5.5: Subset of the α-Amylase dataset. PDB codes, taxonomic information, Enzyme comission code and reference of the 34 structures used to estimate the pooled-within covariance matrix

PDB code	**Species**	**EC code**	**Mutation**	**Reference**
1CGY	*Bacillus circulans*	2.4.1.19	Y195W	Penninga et al. (1995)
1E3X	*Bacillus amyloliquefaciens*	2.2.1.1	Chimeric	Brzozowski et al. (2000)
1G5A	*Neisseria polysaccharea*	2.4.1.4	None	Skov et al. (2001)
1GVI	*Thermus* sp.	3.2.1.54*	None	Lee et al. (2002)
1J0H	*Geobacillus stearothermophilus* TRS40	3.2.1.135	None	Hondoh et al. (2003)
1KB3	*Homo sapiens*	3.2.1.1	R195A	Numao et al. (2002)
1KXH	*Pseudoalteromonas haloplanctis*	3.2.1.1	D174N†	Aghajari et al. (2002)
1M53	*Klebsiella* sp. LX3	5.4.99.11	None	Zhang et al. (2003)
1SMA	*Thermus* sp. IM6501	3.2.1.133	None	Kim et al. (1999)
1TMQ	*Tenebrio molitor*	3.2.1.1	None	Strobl et al. (1998)
1UA7	*Bacillus subtilis*	3.2.1	N356Q	(Kagawa et al., 2003)
1UD3	*Bacillus* sp. KSM-K38	3.2.1.1	N289H	Nonaka et al. (2003)
1VJS	*Bacillus licheniformis*	3.2.1.1	None‡	Hwang et al. (1997)
1W9X	*Bacillus halmapalus*	3.2.1.1	None	Davies et al. (2005)
1WZL	*Thermoactinomyces vulgaris* R-47	3.2.1.135	R469L	Mizuno et al. (2005)
1ZJA	*Pseudomonas mesoacidophila*	5.4.99.11	None	Ravaud et al. (2007)
2DIE	*Bacillus* sp. KSM-1378	3.2.1.1	None	Shirai et al. (2007)
2FH8	*Enterobacter aerogenes*	3.2.1.41	G680L/V882L	Mikami et al. (2006)
2TAA	*Aspergillus oryzae*	3.2.1.1	None	Matsuura et al. (1984)
2WAN	*Bacillus acidopullulyticus*	3.2.1.41	None$^{\text{II}}$	Turkenburg et al. (2009)
2Y4S	*Hordeum vulgare*	3.2.1.41	None	Vester-Christensen et al. (2010)
2Z1K	*Thermus thermophilus HB8*	3.2.1.41	NA	NA
2ZE0	*Geobacillus* sp. HTA-462	3.2.1.20	None	Shirai et al. (2008)
2ZIC	*Streptococcus mutans*	3.2.1.70	N536L	Hondoh et al. (2008)
3AXH	*Saccharomyces cerevisiae*	2.1.1.64/3.2.1.10	E277A	(Yamamoto et al., 2011)
3CZK	*Xanthomonas axonopodis* pv. glycines	3.2.1.48	E322Q	(Kim et al., 2008)
3DC0	*Bacillus* sp. KR-8104	3.2.1.1	NA	NA
3EDE	*Flavobacterium* sp. 92	3.2.1.54	T49P	Buedenbender and Schulz (2009)
3GBD	*Serratia plymuthica*	5.4.99.11	None	(Ravaud et al., 2009)
3UEQ	*Neisseria polysaccharea*	2.4.1.4	None	(Guérin et al., 2012)
3VM5	*Oryzias latipes*	3.2.1.1*	None	Mizutani et al. (2012)
3VM7	*Malbranchea cinnamomea*	3.2.1.73*	None	Han et al. (2013)
4E2O	*Geobacillus thermoleovorans* CCB_US3_UF5	3.2.1.1	None$^{\text{Ш}}$	Mok et al. (2013)
4GI6	*Rhizobium* sp. MX-45	5.4.99.11	F164L	Lipski et al. (2013)

* EC number derived from enzyme name using BRENDA (Schomburg et al., 2004)
† Inactive mutant
‡ Thermostable α-amylase
II A588 CYS Modelled as oxidised CYS (CSX)
Ш Truncated
NA Not provided by the RCSB PDB or secondary sources

a PCA analysis of the shapes as X and Y axes; $\Delta\hat{G^{\circ}}_{unfold}$ in the Z axis. Figure 5.1 shows the result for the 16006 models of the 34 MD simulations.

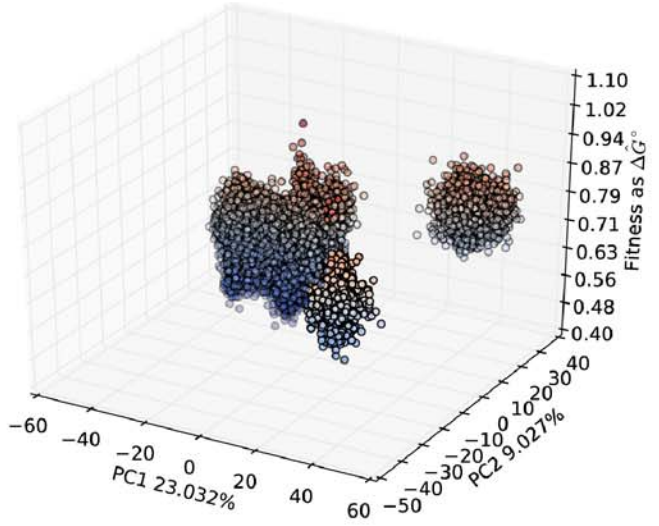


Figure 5.1: Fitness surface of the MD simulations in the 34 structures dataset. Fitness in the Z axis is defined as $\Delta\hat{G}^{\circ}$

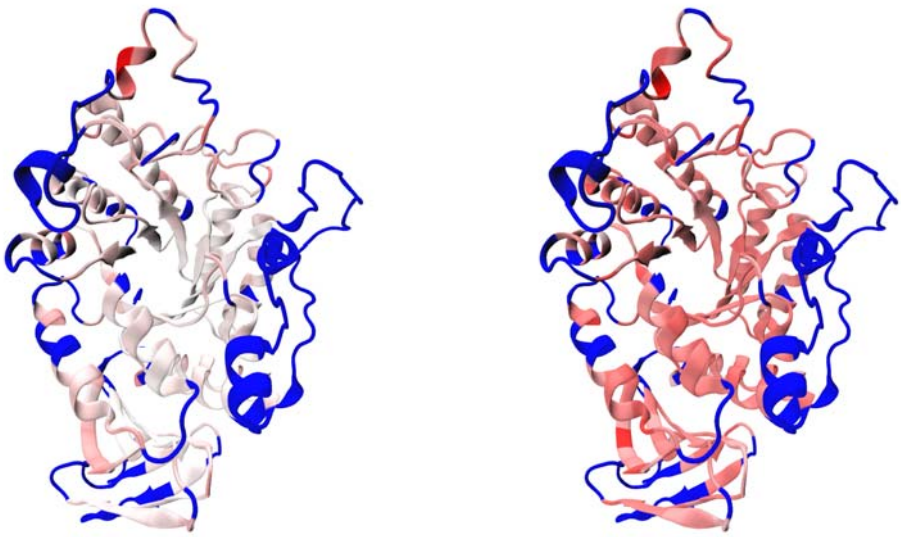
The structure depicted with the higher fitness was the model 1 of structure 2TAA, from *Aspergillus oryzae*. Here I am going to assume that evolution optimizes for $\Delta\hat{G}^{\circ}$, despite that the evidence shows that this might not be the case. However, a more comprehensive fitness function can be obtained by experimentally determining enzyme efficiency in a given substrate and by applying equation 5.14. The model 1 of structure 2TAA can be assumed to be the result of the goal of selection. The realized response to selection $\Delta\bar{z}_{\varpi}$ can be defined as:

$$\Delta\bar{z}_{\varpi} = \mu_{\oplus} - \mu_0 \tag{5.15}$$

where $\mu_{\oplus}$ is the target or after-selection mean structure and μ_0 is the starting structure or before selection structure. To estimate $\Delta\bar{z}_{\varpi}$ it is essential to have the fitness defined based on the questions to be asked, giving that the interpretation of the realized response to selection depends on it.

In an engineering perspective, let's assume that $\mu_{\oplus}$ is the mean of a population

of structures with the desired stability. On the other hand, μ_0 is the mean of a population of structures created by a desired vector. One might ask the question of how does μ_0 have to change towards the stability of $\mu_\oplus$. This can be achieved by computing β_λ (equation 5.12), and replacing $\Delta\bar{z}$ by $\Delta\bar{z}_\varpi$. In the particular case of the GH13 dataset, let's assume that the model 1 of the structure 2TAA is the desired phenotype (with the higher fitness in Figure 5.1), and the model 643 of the structure 4E2O from *Geobacillus thermoleovorans* CCB_US3_UF5 (with the lower fitness in Figure 5.1) corresponds to the source phenotype. β_λ would have a length corresponding to the dimensions of the shape. In the GH13 case 297 homologous residues were identified, which means that these shapes have a dimensionality of 891 traits. This dimension-per-dimension output is important since it tells us the amount of pressure in each dimension per each residue. However, it makes the visualization more difficult. For the sake of visualization simplicity, Figure 5.2 shows the absolute value of the sum of β_λ per residue, standardized from 0 to 1.



(a) Selection gradient on G (b) Dynamic gradient on M

Figure 5.2: $\sum_{i=x,y,z} |\beta_{\lambda_i}|$ rendered in the source structure 4E2O. White represents the lowest magnitude (0), while red the highest (1). Blue depics the non-homologous residues.

Figure 5.2a shows the selection gradient using the estimated G. Not surprisingly, the selection gradient for the TIM-barrel is very low. This means that there is not

much directional selection on this sub-structure. However, it is somewhat surprising that there is not any purifying selection either. This can be explained by the fixation of the trait in the evolution. Since the TIM-barrel is a widespread sub-structure that has been strongly selected during evolution, it might have reached a point of fixation of its geometry. Therefore, the G matrix shows little covariation among these residues since the geometric variability is also low. It is important to stress here that the phenotype measured is the geometry of the structure more than the sequence. Therefore, despite some variation may have occurred at the sequence level, it might not have meaningfully affected the positional information.

However, one must be cautious with the approach employed in Figure 5.2 since the signs are missed, thereby ignoring the direction of selection and the correlated response to selection. Nevertheless, this approach allows for a coarse-grained visual exploration of β_{λ_i}. Individual instances identified by this method should be analysed afterwards in each dimension. Table 5.6 shows the actual values of β_λ for the top 5 positive values (directional selection) and top 5 negative values (purifying selection).

Table 5.6: Selection gradient in the top 5 residues. Top panel shows the residues were at least one of its coordinates is under directional selection and the sum of their absolute values is the highest. Bottom panel contains the information of residues where at least one of its coordinates is under purifying selection, and the sum of the raw values are the lowest.

ResIndex	**Residue**	β_X	β_Y	β_z	$\Delta\bar{z}_X$	$\Delta\bar{z}_Y$	$\Delta\bar{z}_Z$
			Directional				
112	TYR	-5.225	1.082	11.138	-5.106	2.043	10.248
122	LYS	12.333	-2.321	-0.964	12.452	-1.360	-1.854
124	ASP	14.28	-6.963	-10.036	14.399	-6.002	-10.926
125	TRP	18.001	-0.984	0.336	18.121	-0.022	-0.554
126	PHE	11.53	-0.833	3.253	11.650	0.128	2.363
			Purifying				
80	HIS	-5.580	-2.148	4.023	-5.461	-1.187	3.13
121	THR	2.508	-4.644	-5.731	2.627	-3.683	-6.621
223	TYR	-0.010	-7.631	-7.634	0.110	-6.670	-8.524
358	SER	-8.647	-3.461	1.963	-8.527	-2.500	1.073
394	GLU	-4.561	-0.449	-4.002	-4.442	0.512	-4.892

Figure 5.2b and Table 5.7 show the mean difference between target and source

when effects of correlated dynamic differentials are removed. Given that effectively in equation 5.12, G acts as a rotation matrix to remove the selection differentials, one may posit that the same can be achieved with the dynamic (M) matrix. This concept is more difficult to interpret than the actual response to selection. Once G is replaced by M in equation 5.12, we might call it *dynamic gradient* to differentiate it from the selection gradient already explained. In this case, if the gradient is zero for a given trait, this can be interpreted as that the dynamic component of the phenotype does not contribute significantly to the difference in shape for that particular trait. In the case of non-zero gradients, these can be interpreted as contributions of the dynamics to the differential, either towards the target (positive gradient) or away from the target (negative gradient).

Table 5.7: Dynamics gradient in the top 5 residues. Top panel shows the residues where at least one of its coordinates is under positive gradient. Bottom panel cointains the information of residues where at least one of its coordinates is under negative gradient.

ResIndex	**Residue**	β_X	β_Y	β_z	$\Delta\bar{z}_X$	$\Delta\bar{z}_Y$	$\Delta\bar{z}_Z$
			Directional				
117	LEU	13.028	37.149	11.848	2.130	3.521	4.437
125	TRP	29.019	33.605	6.857	18.121	-0.022	-0.554
126	PHE	22.548	33.755	9.774	11.650	0.128	2.363
262	LYS	12.972	38.081	11.412	2.073	4.454	4.001
367	LEU	13.590	34.561	15.609	2.692	0.933	8.197
			Purifying				
124	ASP	25.297	27.625	-3.515	14.399	-6.002	-10.926
223	TYR	11.008	26.958	-1.113	0.110	-6.670	-8.524

In the GH13 subset, most dynamic gradients were positive having only two residues that had at one coordinate under negative gradient (Table 5.7). This can also be inferred by Figure 5.2b. The values of the dynamic gradient are high but sensible given the definition of fitness. Since I defined fitness as the energy of unfolding (ΔG°), most of the information used to select the target and source structures comes from stability, and therefore thermodynamic information. The results depicted in Table 5.7 and Figure 5.2b suggest that most of the variation that explains the difference in phenotype between the structure 4E2O and 2TAA, is contained within the molecular

dynamic component rather than the approximation to the phylogenetic component.

Orientation of G

Arnold (1992) showed that, despite high additive variances, G might not be aligned with the fitness surface. This implies that even though β_λ can be non-zero, the response to selection might send the phenotype in a different direction than the fitness surface. To test this, I used the GH13 dataset and applied Blows et al. (2004) matrix subspace projection approach. This approach assumes that it is usually the case that most of G eigenvalues account for almost no variation. Therefore a submatrix A can be created by choosing k eigenvectors e_i such that $A = (e_1, e_2, \ldots, e_k)$. To choose k I used Minka (2000) probabilistic PCA model, implemented in `Scikit-learn Python` package (Pedregosa et al., 2011).

We can determine what is the closest vector (projection P_{ro}) of G onto β_λ, by projecting the A into the subspace as :

$$P_{ro} = A(AA^T)^{-1}A^T \tag{5.16}$$

Then, the projection that is closest to beta (P_β) can be calculated as (Blows et al., 2004):

$$P_\beta = P_{ro}\beta_\lambda \tag{5.17}$$

The angle of the direction of optimal response (θ) can be estimated by (Walsh and Blows, 2009):

$$\theta = cos^{-1}\left(\frac{P_\beta^T \beta_\lambda}{\sqrt{P_\beta P_\beta^T}\sqrt{\beta_\lambda \beta_\lambda^T}}\right) \tag{5.18}$$

The GH13 θ was 1.4 degrees, which means that the direction of optimal response is 1.4 degrees away from the total genetic variation of 99% explained by the projection. According to this, the *Geobacillus thermoleovorans* structure is susceptible to the selection in the actual direction of the fitness landscape towards the structure of *Aspergillus oryzae* to achieve maximum stability. The extend of such change is given

by $\Delta\bar{z}$, which means that the centroid position of the residue i should be displaced by $\vec{v} = (\Delta\bar{z}_{ix}, \Delta\bar{z}_{iy}, \Delta\bar{z}_{iz})$.

In the case of the dynamics, the same approach can be taken. Here, θ_M was 1.5 degrees which means that the optimal dynamic response is 1.5 degrees away from the optimal response. This can be interpreted in a similar way than the regular θ. However, manipulating the structure along the dynamics gradient is not feasible, and thus is more applicable to thermodynamic theory of protein structures.

With this approach the spaces of G or M are not considered in their entirety (Walsh and Blows, 2009). Blows and Walsh (2009) and Hansen and Houle (2008) develop a similar approach which measures the angle between β and the predicted response to selection from the multivariate breeders equation, $\Delta\bar{z}$ as:

$$\theta_{\Delta\bar{z}\text{-}\beta} = cos^{-1}\left(\frac{\Delta\bar{z}^T\beta_\lambda}{\sqrt{\Delta\bar{z}\Delta\bar{z}^T}\sqrt{\beta_\lambda\beta_\lambda^T}}\right) \tag{5.19}$$

$\theta_{\Delta\bar{z}\text{-}\beta}$ would be zero when there is no genetic constraint, whereas an angle of 90° would represent an absolute constraint (Walsh and Blows, 2009).

The GH13 dataset $\theta_{\Delta\bar{z}\text{-}\beta}$ was 0.3. This means that the genetic constraints on 4E2O are not affecting the direction of selection. This posits the possibility that a strong directional selection selection will drive the source structure towards the target one. The same happens when this approach is applied to M. $\theta^M_{\Delta\bar{z}\text{-}\beta}$ is 1.46 degrees, which is almost identical to θ_M. Thus there is almost no within-variation or dynamic constraints to the vector of response given the dynamic gradient.

5.3.2 Concluding remarks

In this section I have shown the application of the framework described in section 5.2 in a subset of the GH13 proteins. I have demonstrated that this approach is feasible and gives sensible results given the definition of fitness. This definition is essential in the interpretation of the results since it is the one that gives polarity to R_ϖ. Therefore, all conclusions about the response to selection and the selection gradient itself must be analysed under this light.

The usage of M in the determination of the dynamic gradient could be controversial. This is due to the fact that, in the partition of the phenotypic variance, M

is expected to be the environmental variance plus an error term. However, since the source data for the estimation of G and M comes from repeated measures by MD, M contains information about the thermodynamics and folding stability of the protein. It is therefore also contributing to selection.

It is important to stress the fact that this is an approximation to the true G and true M, since I have shown in previous sections that these cannot be estimated given the dimensionality of the phenotype (Section 5.1). However, I have shown in section 5.2 that the pooled-within group approach gives consistent results.

In this section I have also shown that in a stability perspective, the TIM-barrel show a small phylogenetic/genetic component to the selection gradient when a less stable structure (4E2O) is analysed with respect to a more stable one (2TAA). In an engineering perspective, this means that most of the changes in shape come from the dynamics. Nevertheless, the small $\theta_{\Delta\bar{z}\text{-}\beta}$ show that most of the changes applied to 4E2O would result directly into increasing the stability towards the one expressed by 2TAA. 4E2O is a truncated protein, and therefore some loss of stability is expected. It seems that residues 112Y, 122K, 124D, 125W, and 126P, are good candidates to increase the stability of the molecule giving their $\Delta\hat{z}$s. In these cases, the goal will be to shift the position of their centroids given the resulting vector of the three dimensions. Testing this is beyond the scope of this thesis.

Chapter 6

Conclusion

Studying protein structure variation if not a trivial issue. It requires different ways to tackle the issues that arise when your data is highly multivariate and the sample size is small.

In chapter 2 I have shown a sound methodology by applying geometric morphometric (GM) principles into protein structure analysis. By aligning protein structures with standard structural alignment software, I demonstrated that a meaningfully abstraction can be made by computing the geometric center (centroid) of each residue. I also have shown how this abstracted shape can be analysed by standard multivariate statistics techniques, giving insights into protein structure and function. Particularly, I found that a classic multidimensional scaling allowed me to explore the shape space of the α-amylase dataset, giving insights into the relationship between structure and function. This same technique allowed me to show how certain mutations can create significant structural shifts. In this chapter, I also showed how principal components analysis can be used to explore the trajectory of molecular dynamic simulations allowing to focus attention into possible lower energy states that are being sampled more often than the transitions. I applied a form difference estimation to both kinds of data and showed how they can be used to identify interesting residues. My results suggest that despite that the most influential point is not a catalytic or binding site, it is always close to the catalytic pocket or other functional scaffolds. In the case of the α-amylase, it seemed to be important in the protein's ability to bind to the substrate and to bind metal ions given the proximity ($\approx 15\mathring{A}$) to catalytic and metal binding residues. In the NPC dataset, my results suggest that the deformation of the structure is affected by the presence of the ligand. When it it present, it decreases the amount of deformation allowed in the protein. Overall, chapter 2 effectively showed how GM-like methods can be useful in structural biology.

In Chapter 3, I have developed a robust methodology to explore a structural architecture under the domain level. By abstracting the shape correlation into a correlation graph, I was able to infer modules. Further statistical significance tests were performed, along with a bootstrap test of robustness of each of the inferred modules. Given that the sample size is of concern, a power test also informs the user about the confidence in the particular partition. In simulation data, I found that the accuracy of the expected modules was high and highly dependant on the intracorrelation and the sample-size needed to resolve such correlation. However, I found that when a constraint is imposed over the graph (such as contacts), an over-fragmentation is likely to occur. However, I developed a LDA-based pre-filtering approach to cope with such fragmentation, yielding high accuracies for most of the simulated data. I have developed a sound and robust method to analyse the geometric co-variability among residues giving an evolutionary or molecular dynamic sampling.

Chapter 4 expands on the semantics of protein structure architecture. I have shown a plausible scenario for the emergence of protein structure modularity. I have also discussed the need for a better definition of domain specifying at least three types of domains. This need can be seen in the lack of convergence among domain definitions. This problem arises because the domain boundary prediction is a hard problem. In this chapter, I provided evidence to suggest that the hardness of this problem is related to a hierarchical architecture, as well as problems in the definition. I have also proposed a method to explore such hierarchy in structural domains, where one constraint is the sequentiality. I have shown that, by using this method, the estimated modules resemble AFUs and it might be related to the folding process.

Finally, in Chapter 5, I have developed a a method to estimate the selection gradient on protein structures. I have shown how traditional quantitative genetics approach are not feasibly applied to protein structures given computational and sample size constraints. However, I also have shown how to overcome these issues by approximation using molecular dynamic simulations along with homolog sampling. My results suggest that this approach gives more accurate results than traditional methods. The framework developed in this chapter is, however, dependent on the

definition of fitness. For simplicity I used ΔG°, but I suggest more accurate representation of fitness by incorporating substrate specific experimental data.

This thesis has set a framework to analyse protein structures at different levels. It provides a framework that can be used in many fields of structural biology which can give insights into a wide variety of problems.

Appendix A

Table of PDB codes and plot (Figure 2.8) equivalences

Table A.1: PDB codes of the α-amylase homologues, the species from which it was crystallized, and its corresponding equivalence number in plot 2.8. Only the chain A, corresponding to the catalytic domain, was used.

Species	PDB code	Plot equivalences	PDB code	Plot equivalences	PDB code	Plot equivalences	PDB code	Plot equivalences	PDB code	Plot equivalences
Pseudomonas mesoacidophila	1ZJA	102	2PWE	108	2PWF	109	2PWG	110		
Bacillus sp. Ksm-k38	1UD2	88	1UD3	89						
Geobacillus thermoleovorans	4E2O	125								
Malbranchea cinnamomea	3VM7	39								
Klebsiella aerogenes	2FH8	105	2FHB	106						
Thermus thermophilus	2Z1K	112								
Flavobacterium sp. 92	3EDE	119	3EDF	120						
Pseudomonas stutzeri	1GCY	60								
Lactobacillus acidophilus ncfm	4AIE	123								
Neisseria polysaccharea	1G5A	59	1JGI	71	1MVY	78	3UEQ	122	4FLO	126
	4FLQ	127	4FLR	128	4FLS	129				
Sus scrofa	1BVN	1	1DHK	4	1HX0	5	1JFH	6	1KXQ	15
	1OSE	17	1PIF	18	1PPI	19	1UA3	22	1VAH	23
	3L2L	36								
Hordeum vulgare	2Y4S	30								
Bacillus sp.	1D7F	53	1EA9	56	1PAM	81	1UKS	91	1UKT	92

Table A.1: (continued)

Species	PDB code	Plot equivalences	PDB code	Plot equivalence	PDB code	Plot equivalence	PDB code	Plot equivalence	PDB code	Plot equivalence
	1V3J	93	1V3K	94	1WP6	98	2DIE	104		
Bacillus subtilis	1BAG	42	1UA7	87						
Thermoactinomyces vulgaris	1G1Y	58	1IZJ	63	1IZK	64	1JF5	69	1JF6	70
	1JI1	72	1JL8	73	1UH2	90	1VFM	95	1WZK	99
	1WZL	100	1WZM	101	2D0F	103				
Bacillus licheniformis	1BLI	43	1OB0	79	1VJS	96				
Streptococcus mutans	2ZIC	114	2ZID	115						
Thermus sp.	1GVI	61								
Homo sapiens	1B2Y	0	1C8Q	2	1CPU	3	1JXJ	7	1JXK	8
	1KB3	9	1KBB	10	1KBK	11	1KGU	12	1KGW	13
	1KGX	14	1NM9	16	1Q4N	20	1XGZ	25	1Z32	26
	2CPU	27	3BLK	32	3BLP	33	3DHP	35	3OLD	37
Geobacillus stearothermophilus	1HVX	62	1J0H	65	1J0J	66	1QHO	84		
Thermus sp. Im6501	1SMA	85								
Bacillus circulans	1CDG	44	1CGT	45	1CGU	46	1CGV	47	1CGW	48
	1CGX	49	1CGY	50	1CXK	51	1CXL	52	1DTU	54
	1EO5	57	1KCK	74	1KCL	75	1OT1	80	1PEZ	82
	1PJ9	83	1TCM	86	4CGT	124	6CGT	133	8CGT	134
Bacillus acidopullulyticus	2WAN	111								
Bacillus halmapalus	1W9X	97	2GJP	107						
Thermoanaerobacterium thermosulfurigenes	1A47	40	3BMV	117						
Protaminobacter rubrum	3GBD	121								

Table A.1: (continued)

Species	PDB code	Plot equivalences	PDB code	Plot equivalence	PDB code	Plot equivalence	PDB code	Plot equivalence	PDB code	Plot equivalence
Oryzias latipes	3VM5	38								
Tenebrio molitor	1TMQ	21	1VIW	24						
Geobacillus sp.	2ZE0	113								
Pseudoalteromonas haloplanktis	1B0I	41	1JD7	67	1JD9	68	1KXH	76		
Bacillus amyloliquefaciens	1E3X	55	3BH4	116						
Aspergillus oryzae	2GUY	28	2TAA	29						
Klebsiella sp. Lx3	1M53	77								
Bacillus sp. Kr8104	3DC0	34								
Saccharomyces cerevisiae	3AXH	31								
Rhizobium	4GI6	130	4GIN	131	4H2C	132				
Xanthomonas axonopodis pv. Glycines	3CZK	118								

Appendix B

PCoA using C_α as landmark

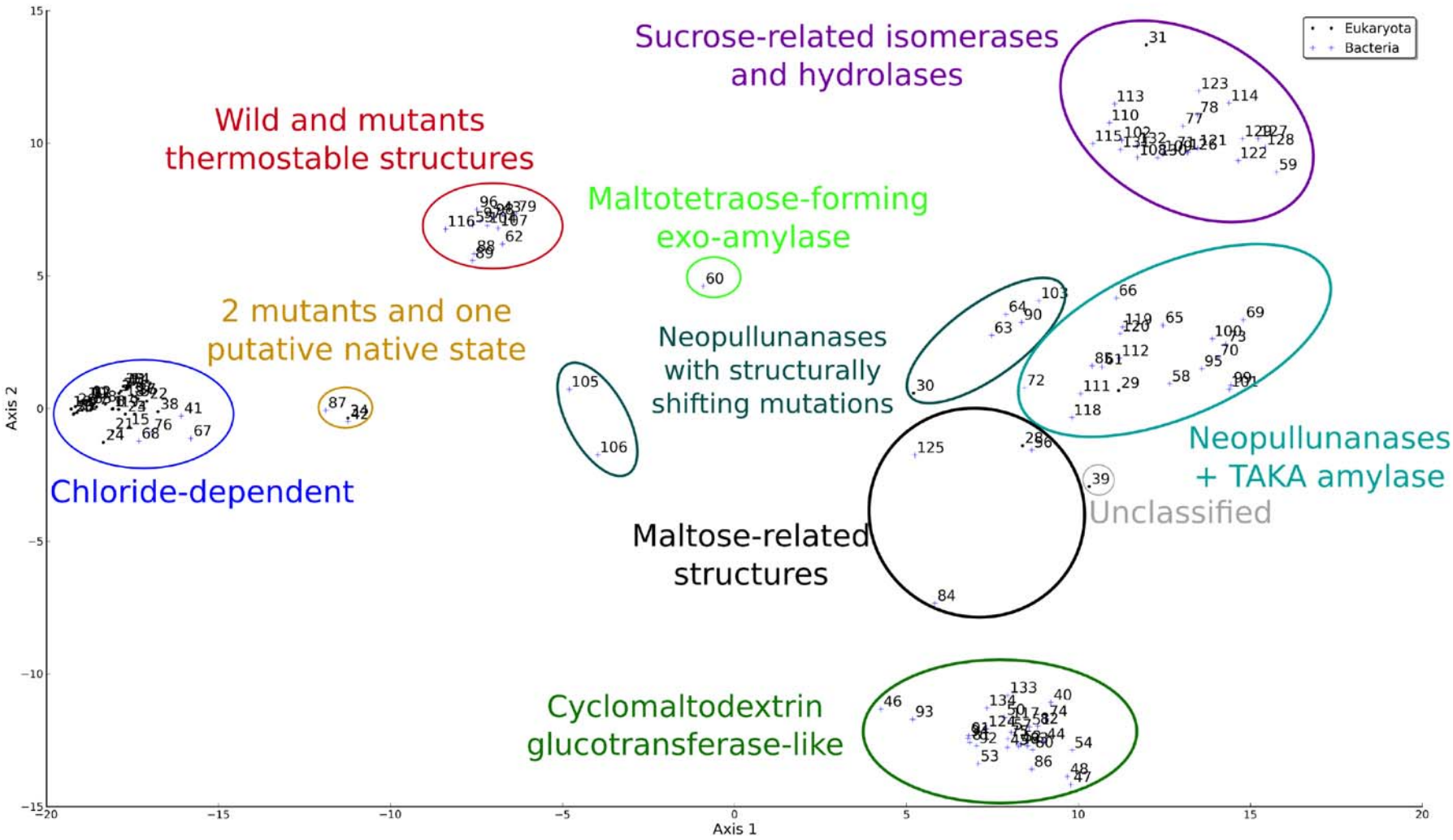


Figure B.1: Principal coordinates analysis on the GH13 dataset using C_α as landmark

Appendix C

Feasibility of different phylogenetic mixed model implementations

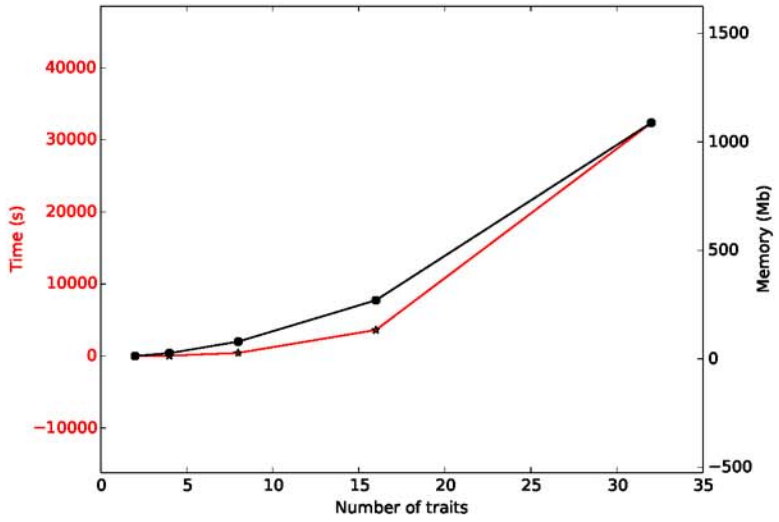


Figure C.1: Computational cost in memory (red) and time (black) of the Lynch's PMM implementation in the R package ape. The x axis represents the number of traits evaluated in a constant number of observations (100)

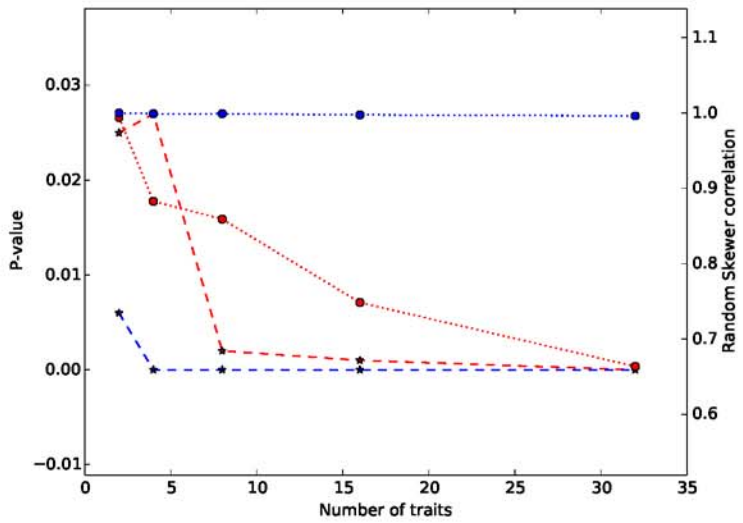


Figure C.2: Random Skewer test of equality of covariances. Correlation (dots and dotted line) and significance (stars and dashed line) of the response to random skewers (1000 simulations), between the simulated and estimated matrices using the Lynch's PMM implementation in the R package `ape`. The color red corresponds to the test on the additive covariance matrices, while the blue on the residual covariance matrices.

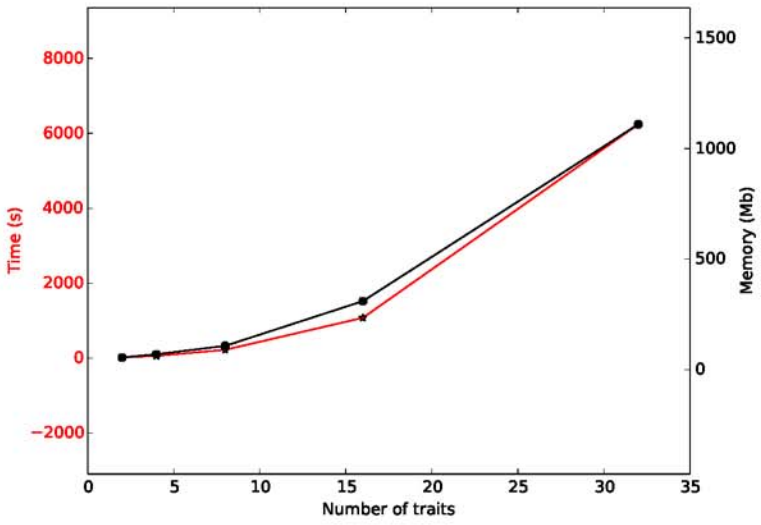


Figure C.3: Computational cost in memory (red) and time(blue) of the PMM implementation in the R package `MCMCglmm`. The x axis represents the number of traits evaluated in a constant number of observations (100). The iterations for the MCMC sampling are also set constant to 100000 with a burnin of 10000, and a thinning interval of 5. The prior was set flat to a diagonal matrix of the size of the estimated matrix.

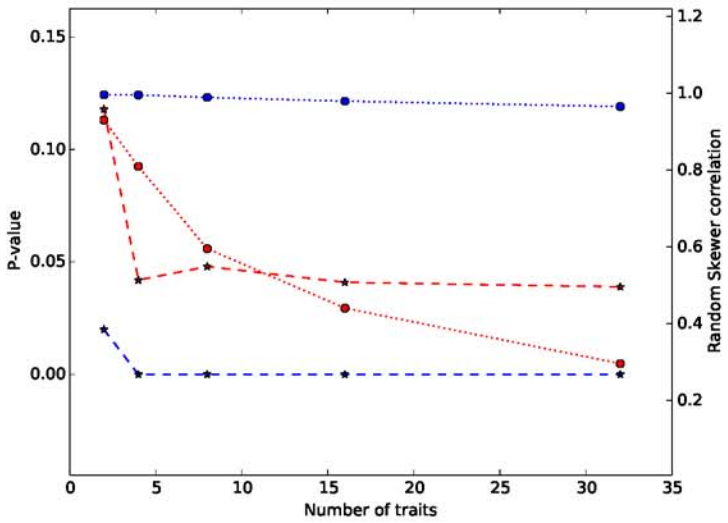


Figure C.4: Random Skewer test of equality of covariances for the Bayesian approach. Correlation (dots and dotted line) and significance (stars and dashed line) of the response to random skewers (1000 simulations), between the simulated and estimated matrices using the the Bayesian Linear Mixed Model R package `MCMCglmm`. Red corresponds to the test of the additive covariance matrices, while blue to the residual covariance matrices.

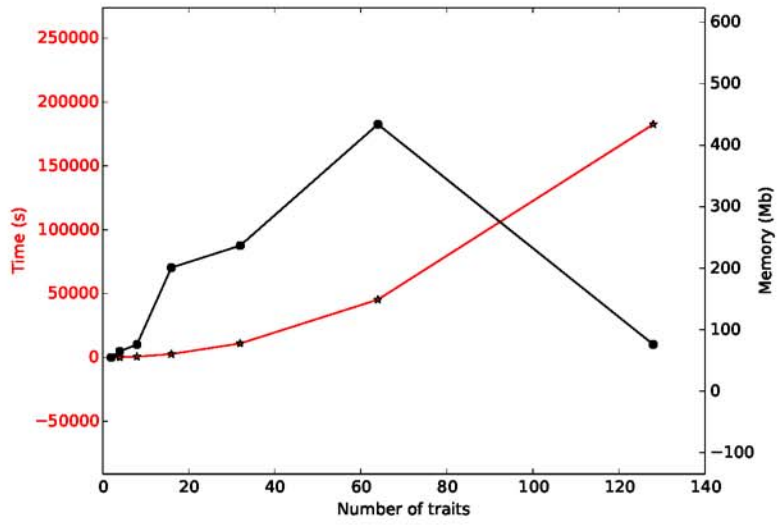


Figure C.5: Computational cost in memory (red) and time (black) of the algorithm 2 implementation. The x axis represents the number of traits evaluated in a constant number of observations (100). The iterations for the MCMC sampling are also set constant to 100000 with a burnin of 10000, and a thinning interval of 5. The prior was set flat to a diagonal matrix of the size of the estimated matrix.

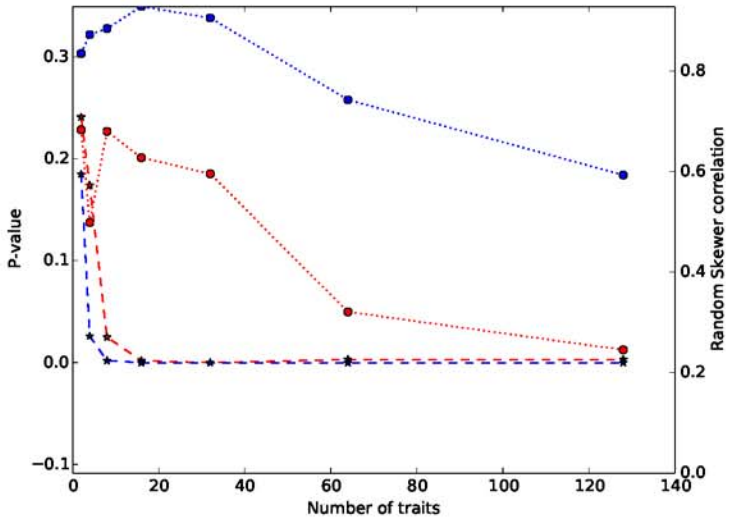
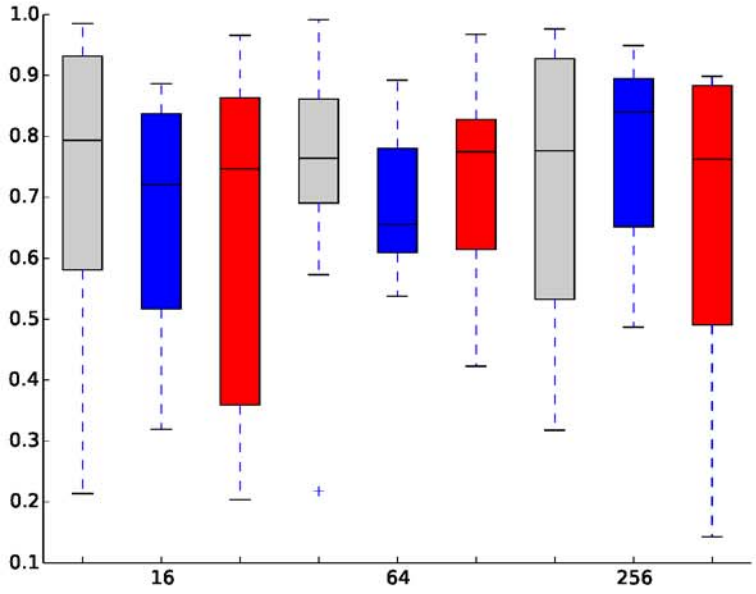
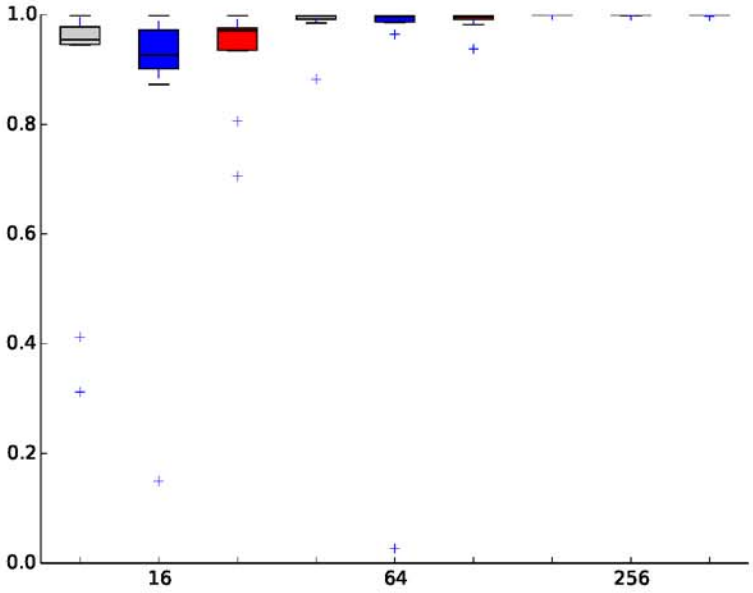


Figure C.6: Random Skewer test of equality of covariances for algorithm 2. Correlation (dots and dotted line) and significance (stars and dashed line) of the response to random skewers (1000 simulations), between the simulated and estimated matrices using the the Bayesian Linear Mixed Model R package `MCMCglmm`. Red corresponds to the test of the additive covariance matrices, while blue to the residual covariance matrices.

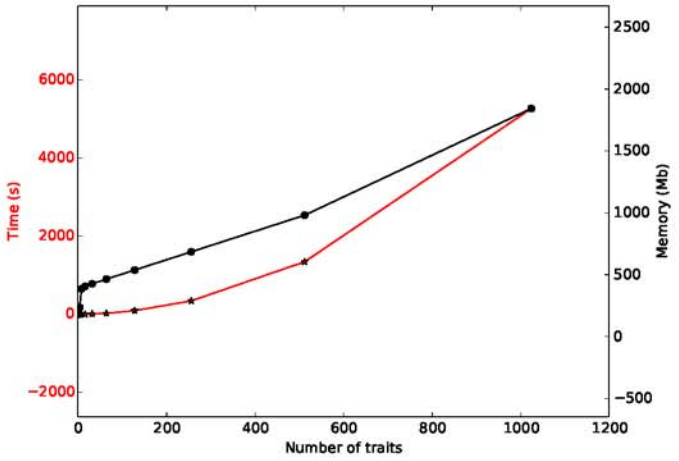


(a) Additive VCV matrices

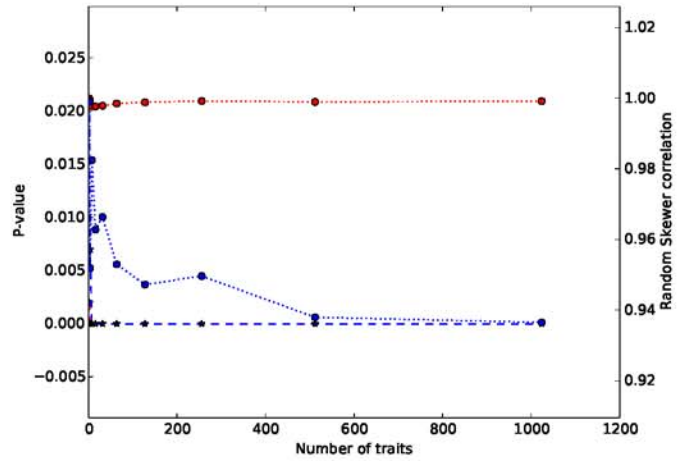


(b) Residual VCV matrices

Figure C.7: Box-plot of the Random Skewer test of equality of covariances in three levels of sample size. Panel C.7a correspond to the estimated additive VCV, and panel C.7b the resduals. Boxes are colored by algorithm: Lynch is depicted as grey, BGLMM as blue, and Algorithm 2 red.



(a) Memory and time profile



(b) Random skewers

Figure C.8: Feasibility and Random Skewer test of equality of covariances for the pooled-within covariance estimation model. Panel C.8a is the feasibility in memory (red dotted line is the correlation and the dashed line represents the p-value) and time(blue dotted line is the correlation and the dashed line represents the p-value) for the pooled within species covariance estimation of G and MD. Panel C.8b shows the accuracy as vector correlation of random skewers.

Appendix D

BioMed Central license agreement

In submitting an article to any of the journals published by BioMed Central I certify that:

1. I am authorized by my co-authors to enter into these arrangements.

2. I warrant, on behalf of myself and my co-authors, that:

the article is original, has not been formally published in any other peer-reviewed journal, is not under consideration by any other journal and does not infringe any existing copyright or any other third party rights; I am/we are the sole author(s) of the article and have full authority to enter into this agreement and in granting rights to BioMed Central are not in breach of any other obligation; the article contains nothing that is unlawful, libellous, or which would, if published, constitute a breach of contract or of confidence or of commitment given to secrecy; I/we have taken due care to ensure the integrity of the article. To my/our - and currently accepted scientific - knowledge all statements contained in it purporting to be facts are true and any formula or instruction contained in the article will not, if followed accurately, cause any injury, illness or damage to the user. 3. I, and all co-authors, agree that the article, if editorially accepted for publication, shall be licensed under the Creative Commons Attribution License 4.0. In line with BioMed Central's Open Data Policy, data included in the article shall be made available under the Creative Commons 1.0 Public Domain Dedication waiver, unless otherwise stated. If the law requires that the article be published in the public domain, I/we will notify BioMed Central at the time of submission, and in such cases not only the data but also the article shall be released under the Creative Commons 1.0 Public Domain Dedication waiver. For the avoidance of doubt it is stated that sections 1 and 2 of this license agreement shall apply and prevail regardless of whether the article is published under

This is the licence agreement that can be found in http://www.biomedcentral.com/authors/license

Creative Commons Attribution License 4.0 or the Creative Commons 1.0 Public Domain Dedication waiver.

[End of BioMed Centrals license agreement]

Explanatory notes regarding BioMed Centrals license agreement

As an aid to our authors, the following paragraphs provide some brief explanations concerning the Creative Commons licenses that apply to the articles published in BioMed Central-published journals and the rationale for why we have chosen these licenses.

The Creative Commons Attribution License (CC BY), of which CC BY 4.0 is the most recent version, was developed to facilitate open access as defined in the founding documents of the movement, such as the 2003 Berlin Declaration. Open access content has to be freely available online, and through licensing their work under CC BY authors grant users the right to unrestricted dissemination and re-use of the work, with only the one proviso that proper attribution is given to authors. This liberal licensing is best suited to facilitate the transfer and growth of scientific knowledge. The Open Access Scholarly Publishers Association (OASPA) therefore strongly recommends the use of CC BY for the open access publication of research literature, and many research funders worldwide either recommend or mandate that research they have supported be published under CC BY. Examples for such policies include funders as diverse as the Wellcome Trust, the Australian Governments, the European Commissions Horizon 2020 framework programme, or the Bill & Melinda Gates Foundation.

The default use of the Creative Commons 1.0 Public Domain Dedication waiver (CC0 or CC zero) for data published within articles follows the same logic: facilitating maximum benefit and the widest possible re-use of knowledge. It is also the case that in some jurisdictions copyright does not apply to data. CC0 waives all potential copyrights, to the extent legally possible, as well as the attribution requirement. The waiver applies to data, not to the presentation of data. If, for instance, a table or figure displaying research data is reproduced, CC BY and the requirement to attribute applies. Increasingly, however, new insights are possible through the use of big data techniques, such as data mining, that harness the entire corpus of digital

data. In such cases attribution is often technically infeasible due to the sheer mass of the data mined, making CC0 the most suitable licensing tool for research outputs generated from such innovative techniques.

It is important to differentiate between legal requirements and community norms. It is first and foremost a community norm, not a law, that within the scientific community attribution mostly takes the form of citation. It is also a community norm that researchers are expected to refer to their sources, which usually takes the form of citation. Across all cases of research reuse (including data, code, etc), community norms will apply as is appropriate for the situation: researchers will cite their sources where it is feasible, regardless of the applicable license. CC0 therefore covers those instances that lie beyond long-established community norms. The overall effect, then, of CC0 for data is to enable further use, without any loss of citations. For further explanation, we recommend you refer to our Open Data FAQ.

In the following, we provide the licenses summaries as they can be found on the Creative Commons website:

The Creative Commons Attribution License 4.0 provides the following summary (where you equals the user): You are free to:

* Share: copy and redistribute the material in any medium or format

* Adapt: remix, transform, and build upon the material

for any purpose, even commercially. The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:

Attribution You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

No additional restrictions You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

Notices:

You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable exception or limitation.

No warranties are given. The license may not give you all of the permissions necessary for your intended use. For example, other rights such as publicity, privacy, or moral rights may limit how you use the material.

Please note: For the terms set in italics in the summary above further details are provided on the Creative Commons web page from which the summary is taken (http://creativecommons.org/licenses/by/4.0/).

The Creative Commons 1.0 Public Domain Dedication waiver provides the following summary:

No Copyright

The person who associated a work with this deed has dedicated the work to the public domain by waiving all of his or her rights to the work worldwide under copyright law, including all related and neighbouring rights, to the extent allowed by law.

You can copy, modify, distribute and perform the work, even for commercial purposes, all without asking permission. See Other Information below.

Other Information

In no way are the patent or trademark rights of any person affected by CC0, nor are the rights that other persons may have in the work or in how the work is used, such as publicity or privacy rights.

Unless expressly stated otherwise, the person who associated a work with this deed makes no warranties about the work, and disclaims liability for all uses of the work, to the fullest extent permitted by applicable law.

When using or citing the work, you should not imply endorsement by the author or the affirmer.

Please note: For the terms set in italics in the summary above further details are provided on the Creative Commons web page from which the summary is taken (http://creativecommons.org/publicdomain/zero/1.0/).

Appendix E

PLOS license

The following policy applies to all of PLOS journals, unless otherwise noted.

PLOS applies the Creative Commons Attribution (CC BY) license to works we publish. This license was developed to facilitate open access namely, free immediate access to, and unrestricted reuse of, original works of all types.

Under this license, authors agree to make articles legally available for reuse, without permission or fees, for virtually any purpose. Anyone may copy, distribute or reuse these articles, as long as the author and original source are properly cited.

Using PLOS Content

No permission is required from the authors or the publishers to reuse or repurpose PLOS content provided the original article is cited. In most cases, appropriate attribution can be provided by simply citing the original article.

Example citation:

Kaltenbach LS et al. (2007) Huntingtin Interacting Proteins Are Genetic Modifiers of Neurodegeneration. PLOS Genet 3(5): e82. doi:10.1371/journal.pgen.0030082.

If the item you plan to reuse is not part of a published article (e.g., a featured issue image), then indicate the originator of the work, and the volume, issue, and date of the journal in which the item appeared.

For any reuse or redistribution of a work, you must also make clear the license terms under which the work was published.

This is the licence that can be found in http://journals.plos.org/plosone/s/content-license

Figures, Tables, and Images

Figures, tables, and images are published under the Creative Commons Attribution (CC BY) license.

Data

If any relevant accompanying data is submitted to repositories with stated licensing policies, the policies should not be more restrictive than CC BY.

sectionSubmitting Copyrighted or Proprietary Content

Do not submit any figures, photos, tables, or other works that have been previously copyrighted or that contain proprietary data unless you have and can supply written permission from the copyright holder to use that content. This includes:

> maps and satellite images

> slogans and logos

> social media content.

Appendix F

BENTHAM science Self-archiving policies and copyright agreement

F.1 Self-archiving policies

By signing the Copyright Letter the authors retain the rights of self-archiving. Following are the important features of self-archiving policy of Bentham Science journals:

1. Authors can deposit the first draft of a submitted article on their personal websites, their institutions repositories or any non-commercial repository for personal use, internal institutional use or for permitted scholarly posting.

2. Authors may deposit the ACCEPTED VERSION of the peer-reviewed article on their personal websites, their institutions repository or any non-commercial repository such as PMC, arXiv after 12 MONTHS of publication on the journal website. In addition, an acknowledgement must be given to the original source of publication and a link should be inserted to the published article on the journal's/publishers website.

3. If the research is funded by NIH, Wellcome Trust or any other Open Access Mandate, authors are allowed the archiving of published version of manuscripts in an institutional repository after the mandatory embargo period. Authors should first contact the Editorial Office of the journal for information about depositing a copy of the manuscript to a repository. Consistent with the copyright

This is the Self-archiving policy that can be found in http://benthamscience.com/self-archiving-policies-main.php

agreement, Bentham Science does not allow archiving of FINAL PUBLISHED VERSION of manuscripts.

4. The link to the original source of publication should be provided by inserting the DOI number of the article in the following sentence: The published manuscript is available at EurekaSelect via

   `http://www.eurekaselect.com/openurl/content.php?genre=article&doi=[insert DOI]`

   .

5. There is no embargo on the archiving of articles published under the OPEN ACCESS PLUS category. Authors are allowed deposition of such articles on institutional, non-commercial repositories and personal websites immediately after publication on the journal website.

F.2 Copyright agreement as per electronic mail: Grant of permission

Dear Dr. Hleap,

Thank you for your interest in our copyrighted material, and for requesting permission for its use.

Permission is granted for the following subject to the conditions outlined below:

Hleap, JS & Blouin, C. The Semantics of the Modular Architecture of Protein Structures, Current Protein & Peptide Science, Volume 16 (E-pub ahead of print)

To be used in the following manner:

Bentham Science Publishers grants you the right to reproduce the material indicated above on a one-time, non-exclusive basis, solely for the purpose described. Permission must be requested separately for any future or additional use.

For an article, the copyright notice must be printed on the first page of article or book chapter. For figures, photographs, covers, or tables, the notice may appear with the material, in a footnote, or in the reference list.

Thank you for your patience while your request was being processed. If you wish to contact us further, please use the address below.

Sincerely,

AMBREEN IRSHAD

Permissions & Rights Manager
Bentham Science Publishers
Email: ambreenirshad@benthamscience.org
URL: www.benthamscience.com

Bibliography

Federico Abascal, Rafael Zardoya, and Maximilian J. Telford. Translatorx: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic acids research*, 38(suppl 2):–7, 2010.

Dean C. Adams and Gavin J. P. Naylor. A new method for evaluating the structural similarity of proteins using geometric morphometrics. In S. Miyano, R. Shamir, and T. Takagi, editors, *Currents in Computational Molecular Biology*, volume 30 of *Frontiers Science Series*, pages 120–121, Tokyo, 2000. Universal Academy Press.

Dean C. Adams and Gavin J. P. Naylor. A comparison of methods for assessing the structural similarity of proteins. In *Mathematical Methods for Protein Structure Analysis and Design*, pages 109–115, 2003. doi: 10.1007/978-3-540-44827-3_6.

Dean C. Adams, F. James Rohlf, and Dennis E. Slice. Geometric morphometrics: Ten years of progress following the 'revolution'. *Italian Journal of Zoology*, 71(1): 5–16, 2004. doi: 10.1080/11250000409356545.

Nushin Aghajari, Georges Feller, Charles Gerday, and Richard Haser. Structures of the psychrophilic *Alteromonas haloplanctis* α-amylase give insights into cold adaptation at a molecular level. *Structure*, 6(12):1503–1516, 1998a.

Nushin Aghajari, Richard Haser, Georges Feller, and Charles Gerday. Crystal structures of the psychrophilic α-amylase from alteromonas haloplanctis in its native form and complexed with an inhibitor. *Protein science*, 7(3):564–572, 1998b.

Nushin Aghajari, Michel Roth, and Richard Haser. Crystallographic evidence of a transglycosylation reaction: ternary complexes of a psychrophilic α-amylase. *Biochemistry*, 41(13):4273–4280, 2002.

S. E. Ahnert, I. G. Johnston, T. M. A. Fink, J. P. K. Doye, and A. A. Louis. Self-assembly, modularity, and physical complexity. *Physical Review E*, 82(2):026117, 2010.

Kieran Alden, Stella Veretnik, and Philip E Bourne. dconsensus: a tool for displaying domain assignments by multiple structure-based algorithms and for construction of a consensus assignment. *BMC Bioinformatics*, 11:310, 2010. doi: 10.1186/1471-2105-11-310.

Nickolai Alexandrov and Ilya Shindyalov. Pdp: protein domain parser. *Bioinformatics*, 19(3):429–430, 2003. doi: 10.1093/bioinformatics/btg006.

Javier Antonio Alfaro. *Capturing the dynamics of protein sequence evolution through site-independent structurally constrained phylogenetic models*. PhD thesis, Department of biochemitry & molecular biology, Dalhousie University, Halifax, Canada, 2014.

Jahan Alikhajeh, Khosro Khajeh, Bijan Ranjbar, Hossein Naderi-Manesh, Y. H. Lin, Enhung Liu, H. H. Guan, Y. C. Hsieh, Phimonphan Chuankhayan, Y. C. Huang, et al. Structure of bacillus amyloliquefaciens-amylase at high resolution: implications for thermal stability. *Acta Crystallographica Section F: Structural Biology and Crystallization Communications*, 66(2):121–129, 2010.

S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402, 1997. doi: 10.1093/nar/25.17.3389.

Theodore Wilbur Anderson. *An introduction to multivariate statistical analysis*, volume 2. Wiley New York, 1958.

Antonina Andreeva, Dave Howorth, John-Marc Chandonia, Steven E. Brenner, Tim J. P. Hubbard, Cyrus Chothia, and Alexey G. Murzin. Data growth and its impact on the scop database: new developments. *Nucleic Acids Res.*, 36(Database issue): D419–D425, 2008. doi: 10.1093/nar/gkm993.

Antonina Andreeva, Dave Howorth, Cyrus Chothia, Eugene Kulesha, and Alexey G Murzin. Scop2 prototype: a new approach to protein structure mining. *Nucleic Acids Res.*, 42(Database issue):D310–4, 2014. doi: 10.1093/nar/gkt1242.

C. B. Anfinsen and A. Scheraga. Experimental and theoretical aspects of protein folding. *Advances in Protein Chemistry*, 29:205–300, 1975.

Stevan J. Arnold. Constraints on phenotypic evolution. *American Naturalist*, 61: S85–S107, 1992.

M. Bégin, D. A. Roff, and V. Debat. The effect of temperature and wing morphology on quantitative genetic variation in the cricket gryllus firmus, with an appendix examining the statistical properties of the jackknife–manova method of matrix comparison. *Journal of evolutionary biology*, 17(6):1255–1267, 2004. doi: 10.1111/j.1420-9101.2004.00772.x.

M. Ben Ali, M. Ghram, H. Hmani, B. Khemakhem, R. Haser, and S. Bejar. Toward the smallest active subdomain of a tim-barrel fold: Insights from a truncated α-amylase. *Biochemical and biophysical research communications*, 411(2):265–270, 2011.

Mamdouh Ben Ali, Bassem Khemakhem, Xavier Robert, Richard Haser, and Samir Bejar. Thermostability enhancement and change in starch hydrolysis profile of the maltohexaose-forming amylase of bacillus stearothermophilus us100 strain. *Biochem. J.*, 394(Pt 1):51–6, 2006. doi: 10.1042/BJ20050726.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol*, 57(1):289–300, 1995. ISSN 00359246. doi: 10.2307/2346101.

I. N. Berezovsky and E. N. Trifonov. Loop fold nature of globular proteins. *Protein Engineering*, 14(6):403–407, 2001.

Igor N. Berezovsky and Edward N. Trifonov. Loop fold structure of proteins: resolution of levinthal's paradox. *Journal of Biomolecular Structure and Dynamics*, 20(1):5–6, 2002.

Claude Berge, Nicolas Froloff, Ravi Kiran Reddy Kalathur, Myriam Maumy, Olivier Poch, Wolfgang Raffelsberger, and Nicolas Wicker. Multidimensional fitting for multivariate data analysis. *J. Comput. Biol.*, 17(5):723–32, 2010. doi: 10.1089/cmb.2009.0126.

A. Blasco. The bayesian controversy in animal breeding. *Journal of Animal Science*, 79(8):2023–2046, 2001.

Mark Blows and Bruce Walsh. Spherical cows grazing in flatland: Constraints to selection and adaptation. In Julius van der Werf, Hans-Ulrich Graser, Richard Frankham, and Cedric Gondro, editors, *Adaptation and fitness in animal populations*, pages 83–101. Springer, 2009. ISBN 978-1-4020-9004-2. doi: 10.1007/978-1-4020-9005-9_6. URL http://dx.doi.org/10.1007/978-1-4020-9005-9_6.

Mark W. Blows, Stephen F. Chenoweth, and Emma Hine. Orientation of the genetic variance-covariance matrix and the fitness surface for multiple male sexually selected traits. *The American Naturalist*, 163(3):329–340, 2004. doi: 10.1086/381941.

Erich Bornberg-Bauer and M Mar Albà. Dynamics and adaptive benefits of modular protein evolution. *Curr. Opin. Struct. Biol.*, 23(3):459–66, 2013. doi: 10.1016/j.sbi.2013.02.012.

R J Bothast and M A Schlicher. Biotechnological processes for conversion of corn into ethanol. *Appl. Microbiol. Biotechnol.*, 67(1):19–25, 2005. doi: 10.1007/s00253-004-1819-8.

Jamie T. Bridgham, Geeta N. Eick, Claire Larroux, Kirti Deshpande, Michael J. Harms, Marie EA Gauthier, Eric A. Ortlund, Bernard M. Degnan, and Joseph W. Thornton. Protein evolution by molecular tinkering: diversification of the nuclear receptor superfamily from a ligand-dependent ancestor. *PLoS biology*, 8(10):e1000497, 2010. doi: 10.1371/journal.pbio.1000497.

Luciano Brocchieri. Phylogenetic inferences from molecular sequences: review and critique. *Theoretical population biology*, 59(1):27–40, 2001. doi: 10.1006/tpbi.2000.1485.

Kevin Bryson, Liam J McGuffin, Russell L Marsden, Jonathan J Ward, Jaspreet S Sodhi, and David T Jones. Protein structure prediction servers at university college london. *Nucleic Acids Res.*, 33(Web Server issue):W36–8, 2005. doi: 10.1093/nar/gki410.

Andrzej M. Brzozowski, David M. Lawson, Johan P. Turkenburg, Henrik Bisgaard-Frantzen, Allan Svendsen, Torben V. Borchert, Zbigniew Dauter, Keith S. Wilson, and Gideon J. Davies. Structural analysis of a chimeric bacterial α-amylase. high-resolution analysis of native and ligand complexes. *Biochemistry*, 39(31):9099–9107, 2000. doi: 10.1021/bi0000317.

Stefan Buedenbender and Georg E Schulz. Structural base for enzymatic cyclodextrin hydrolysis. *J. Mol. Biol.*, 385(2):606–17, 2009. doi: 10.1016/j.jmb.2008.10.085.

Gustavo Caetano-Anollés, Minglei Wang, and Derek Caetano-Anollés. Structural phylogenomics retrodicts the origin of the genetic code and uncovers the evolutionary impact of protein flexibility. *PLoS ONE*, 8(8):e72225, 2013. doi: 10.1371/journal.pone.0072225.

Scott Camazine, Jean-Louis Deneubourg, N. Franks, James Sneyd, Guy Theraulaz, and Eric Bonabeau. *Self-organization in biological systems*. Princeton University Press, Princeton, New Jersey, 2002.

Brandi L. Cantarel, Pedro M. Coutinho, Corinne Rancurel, Thomas Bernard, Vincent Lombard, and Bernard Henrissat. The carbohydrate-active enzymes database (cazy): an expert resource for glycogenomics. *Nucleic acids research*, 37(suppl 1): D233–D238, 2009. doi: 10.1093/nar/gkn663.

Stephane Champely. *pwr: Basic functions for power analysis*, 2009. URL http://CRAN.R-project.org/package=pwr. R package version 1.1.1.

Jianlin Cheng. Domac: an accurate, hybrid protein domain prediction server. *Nucleic Acids Res.*, 35(Web Server issue):W354–6, 2007. doi: 10.1093/nar/gkm390.

Jianlin Cheng, Michael J. Sweredoski, and Pierre Baldi. Dompro: protein domain prediction using profiles, secondary structure, relative solvent accessibility, and recursive neural networks. *Data Mining and Knowledge Discovery*, 13(1):1–10, 2006. doi: 10.1007/s10618-005-0023-5.

J. M. Cheverud. Quantitative genetic analysis of cranial morphology in the cotton-top (saguinus oedipus) and saddle-back (s. fuscicollis) tamarins. *Journal of Evolutionary Biology*, 9(1):5–42, 1996a.

James M. Cheverud. A comparison of genetic and phenotypic correlations. *Evolution*, 42(5):958–968, 1988.

James M. Cheverud. Developmental integration and the evolution of pleiotropy. *American Zoologist*, 36(1):44–50, 1996b. doi: 10.1093/icb/36.1.44.

James M. Cheverud and Gabriel Marroig. Comparing covariance matrices: random skewers method compared to the common principal components model. *Genetics and Molecular Biology*, 30(2):461–469, 2007.

James M. Cheverud, Malcolm M. Dow, and Walter Leutenegger. The quantitative assessment of phylogenetic constraints in comparative analyses: sexual dimorphism in body weight among primates. *Evolution*, pages 1335–1351, 1985.

Dylan Chivian, David E. Kim, Lars Malmström, Philip Bradley, Timothy Robertson, Paul Murphy, Charles EM Strauss, Richard Bonneau, Carol A. Rohl, and David Baker. Automated prediction of casp-5 structures using the robetta server. *Proteins: Structure, Function, and Bioinformatics*, 53(S6):524–533, 2003. doi: 10.1002/prot.10529.

Peter Claes, Katleen Daniels, Mark Walters, John Clement, Dirk Vandermeulen, and Paul Suetens. Dysmorphometrics: the modelling of morphological abnormalities. *Theor Biol Med Model*, 9:5, 2012. doi: 10.1186/1742-4682-9-5.

Julien Claude. *Morphometrics with R.* Use R! Springer, 2008. ISBN 9780387777894. URL http://www.worldcat.org/isbn/9780387777894.

Aaron Clauset, M. E. J. Newman, and Cristopher Moore. Finding community structure in very large networks. *PHYS.REV.E*, 70:066111, 2004. doi: 10.1103/PhysRevE.70.066111.

Cecilia Clementi. Coarse-grained models of protein folding: toy models or predictive tools? *Curr. Opin. Struct. Biol.*, 18(1):10–5, 2008. doi: 10.1016/j.sbi.2007.10.005.

Jacob Cohen. *Statistical power analysis for the behavioral sciences : Jacob Cohen.* Lawrence Erlbaum Associates, Hillsdale,NJ, 2 edition, 1988. ISBN 0805802835.

Bruno Contreras-Moreira and Paul A. Bates. Domain fishing: a first step in protein comparative modelling. *Bioinformatics*, 18(8):1141–1142, 2002. doi: 10.1093/bioinformatics/18.8.1141.

Gavin E. Crooks, Gary Hon, John-Marc Chandonia, and Steven E. Brenner. Weblogo: a sequence logo generator. *Genome research*, 14(6):1188–1190, 2004.

Gergely Csaba, Fabian Birzele, and Ralf Zimmer. Protein structure alignment considering phenotypic plasticity. *Bioinformatics*, 24(16):–98, 2005. doi: 10.1093/bioinformatics/btn271.

Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006. URL http://igraph.sf.net.

Peter Csermely. Creative elements: network-based predictions of active centres in proteins and cellular and social networks. *Trends Biochem. Sci.*, 33(12):569–76, 2008. doi: 10.1016/j.tibs.2008.09.006.

Peter Csermely, Tamás Korcsmáros, Huba J M Kiss, Gábor London, and Ruth Nussinov. Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. *Pharmacol. Ther.*, 138(3):333–408, 2013. doi: 10.1016/j.pharmthera.2013.01.016.

J. L. Da Lage, Georges Feller, and Š Janeček. Horizontal gene transfer from eukarya to bacteria and domain shuffling: the α-amylase model. *Cellular and Molecular Life Sciences CMLS*, 61(1):97–109, 2004. doi: 10.1007/s00018-003-3334-y.

S. D'Amico, C. Gerday, and G. Feller. Structural similarities and evolutionary relationships in chloride-dependent α-amylases. *Gene*, 253(1):95–105, 2000. doi: 10.1016/S0378-1119(00)00229-8.

Salvino D'Amico, Jean-Sébastien Sohier, and Georges Feller. Kinetics and energetics of ligand binding determined by microcalorimetry: insights into active site mobility in a psychrophilic α-amylase. *Journal of molecular biology*, 358(5):1296–1304, 2006.

Paweł Daniluk and Bogdan Lesyng. A novel method to compare protein structures using local descriptors. *BMC Bioinformatics*, 12(1):344, 2011. doi: 10.1186/1471-2105-12-344.

G. Davies and B. Henrissat. Structures and mechanisms of glycosyl hydrolases. *Structure*, 3(9):853–859, 1995.

Gideon J. Davies, A. Marek Brzozowski, Zbigniew Dauter, Michael D. Rasmussen, Torben V. Borchert, and Keith S. Wilson. Structure of a bacillus halmapalus family 13 α-amylase, bha, in complex with an acarbose-derived nonasaccharide at 2.1 ă resolution. *Acta Crystallographica Section D: Biological Crystallography*, 61 (2):190–193, 2005.

J. P. Davies and Y. A. Ioannou. Topological analysis of niemann-pick c1 protein reveals that the membrane orientation of the putative sterol-sensing domain is identical to those of 3-hydroxy-3-methylglutaryl-coa reductase and sterol regulatory element binding protein cleavage-activating protein. *Journal of Biological Chemistry*, 275(32):24367–24374, 2000.

Paul CW Davies, Elisabeth Rieper, and Jack A. Tuszynski. Self-organization and entropy reduction in a living cell. *Biosystems*, 111(1):1–10, 2013. doi: 10.1016/j.biosystems.2012.10.005.

Ryan Day, David A C Beck, Roger S Armen, and Valerie Daggett. A consensus view of fold space: combining scop, cath, and the dali domain dictionary. *Protein Sci.*, 12(10):2150–60, 2003. doi: 10.1110/ps.0306803.

Tjaart AP de Beer, Karel Berka, Janet M. Thornton, and Roman A. Laskowski. Pdbsum additions. *Nucleic acids research*, 42(D1):D292–D296, 2013. doi: 10.1093/nar/gkt940.

Felipe Bandoni de Oliveira, Arthur Porto, and Gabriel Marroig. Covariance structure in the skull of catarrhini: a case of pattern stasis and magnitude evolution. *J. Hum. Evol.*, 56(4):417–30, 2009. doi: 10.1016/j.jhevol.2009.01.010.

Cédric Debès. *Physical constraints on protein structure evolution.* PhD thesis, Ruperto-Carola University, Heidelberg, Germany, 2013.

Antonio Del Sol, Marcos J Araúzo-Bravo, Dolors Amoros, and Ruth Nussinov. Modular architecture of protein structures and allosteric communications: potential implications for signaling proteins and regulatory linkages. *Genome Biol.*, 8(5): R92, 2007. doi: 10.1186/gb-2007-8-5-r92.

Somdutta Dhir, Mircea Pacurar, Dino Franklin, Zoltán Gáspári, Attila Kertész-Farkas, András Kocsor, Frank Eisenhaber, and Sándor Pongor. Detecting atypical examples of known domain types by sequence similarity searching: The sbase domain library approach. *Current Protein and Peptide Science*, 11(7):538–549, 2010.

L. Di Paola, M. De Ruvo, P. Paci, D. Santoni, and A. Giuliani. Protein contact networks: An emerging paradigm in chemistry. *Chemical Reviews*, 113(3):1598–1613, 2013. doi: 10.1021/cr3002356.

Nadezhda T Doncheva, Karsten Klein, Francisco S Domingues, and Mario Albrecht. Analyzing and visualizing residue networks of protein structures. *Trends Biochem. Sci.*, 36(4):179–82, 2011. doi: 10.1016/j.tibs.2011.01.002.

Nadezhda T Doncheva, Yassen Assenov, Francisco S Domingues, and Mario Albrecht. Topological analysis and interactive visualization of biological networks and protein structures. *Nat Protoc*, 7(4):670–85, 2012. doi: 10.1038/nprot.2012.004.

I. L. Dryden and Kanti V. Mardia. *Statistical Shape Analysis.* Wiley, Chichester, 1 edition, 1998. ISBN 9780471958161.

Jose M Duarte, Rajagopal Sathyapriya, Henning Stehr, Ioannis Filippis, and Michael Lappe. Optimal contact definition for reconstruction of contact maps. *BMC Bioinformatics*, 11:283, 2010. doi: 10.1186/1471-2105-11-283.

Michel Dumontier, Rong Yao, Howard J Feldman, and Christopher W V Hogue. Armadillo: domain boundary prediction by amino acid composition. *J. Mol. Biol.*, 350(5):1061–73, 2005. doi: 10.1016/j.jmb.2005.05.037.

Julien Dutheil and Nicolas Galtier. Detecting groups of coevolving positions in a molecule: a clustering approach. *BMC evolutionary biology*, 7(1):242, 2007.

Darius M. Dziuda. *Data mining for genomics and proteomics: analysis of gene and protein expression data*, volume 1. John Wiley & Sons, Inc., Hoboken,NJ, 2010.

Jesse Eickholt, Xin Deng, and Jianlin Cheng. Dobo: Protein domain boundary prediction by integrating evolutionary signals and machine learning. *BMC Bioinformatics*, 12:43, 2011. doi: 10.1186/1471-2105-12-43.

F. Eroukhmanoff and E. I. Svensson. Evolution and stability of the g-matrix during the colonization of a novel environment. *J. Evol. Biol.*, 24(6):1363–73, 2011. doi: 10.1111/j.1420-9101.2011.02270.x.

Iakes Ezkurdia and Michael L Tress. Protein structural domains: definition and prediction. *Curr Protoc Protein Sci*, Chapter 2:Unit2.14, 2011. doi: 10.1002/0471140864.ps0214s66.

D. S. Falconer and T. F. C. Mackay. *Introduction to quantitative genetics*. Longman, 1996. ISBN 9780582243026. URL http://books.google.ca/books?id=lrCYQgAACAAJ.

T. Fancello, A. Dardis, C. Rosano, P. Tarugi, B. Tappino, S. Zampieri, E. Pinotti, F. Corsolini, S. Fecarotta, A. D'Amico, et al. Molecular analysis of npc1 and npc2 gene in 34 niemann–pick c italian patients: identification and structural modeling of novel mutations. *neurogenetics*, 10(3):229–239, 2009.

G. K. Farber. An α/β-barrel full of evolutionary trouble. *Current opinion in structural biology*, 3(3):409–412, 1993.

Guilhem Faure, Aurélie Bornot, and Alexandre G de Brevern. Protein contacts, inter-residue interactions and side-chain modelling. *Biochimie*, 90(4):626–39, 2008. doi: 10.1016/j.biochi.2007.11.007.

A. Fedorov, X. Cao, S. Saxonov, S. J. De Souza, S. W. Roy, and W. Gilbert. Intron distribution difference for 276 ancient and 131 modern genes suggests the existence of ancient introns. *Proceedings of the National Academy of Sciences*, 98(23):13177–13182, 2001.

H. J. Feldman. Identifying structural domains of proteins using clustering. *BMC Bioinformatics*, 13(1):286, 2012.

Georges Feller. Protein stability and enzyme activity at extreme biological temperatures. *Journal of Physics: Condensed Matter*, 22(32):323101, 2010.

Georges Feller, Francoise Payan, Fabienne Theys, Minxie Qian, Richard Haser, and Charles Gerday. Stability and structural analysis of α-amylase from the antarctic psychrophile *Alteromonas haloplanctis* a23. *European Journal of Biochemistry*, 222(2):441–447, 1994.

Joseph Felsenstein. Comparative methods with sampling error and within-species variation: contrasts revisited and revised. *The American Naturalist*, 171(6):713–725, 2008.

Ariel Fernández and R. Stephen Berry. Self-organization and mismatch tolerance in protein folding: General theory and an application. *Journal of Chemical Physics, Volume 112, Issue 11, pp. 5212-5222 (2000).*, 112:5212–5222, mar 2000. doi: 10.1063/1.481076.

R. D. Finn, J. Mistry, J. Tate, P. Coggill, A. Heger, J. E. Pollington, O. L. Gavin, P. Gunasekaran, G. Ceric, K. Forslund, Liisa Holm, Erik L. L. Sonnhammer, Sean R. Eddy, and Alex Bateman. The pfam protein families database. *Nucleic acids research*, 38(suppl 1):D211 – D222, 2010. doi: 10.1093/nar/gkp985.

Kael F. Fischer and Susan Marqusee. A rapid test for identification of autonomous folding units in proteins. *Journal of molecular biology*, 302(3):701–712, 2000.

Astrid Fleischmann, Michael Darsow, Kirill Degtyarenko, Wolfgang Fleischmann, Sinéad Boyce, Kristian B. Axelsen, Amos Bairoch, Dietmar Schomburg, Keith F. Tipton, and Rolf Apweiler. Intenz, the integrated relational enzyme database. *Nucleic acids research*, 32(suppl 1):–434, 2004. doi: 10.1093/nar/gkh119.

S. J. Fleishman and D. Baker. Role of the biomolecular energy gap in protein design, structure, and evolution. *Cell*, 149(2):262–273, 2012.

Kristoffer Forslund and Erik L L Sonnhammer. Evolution of protein domain architectures. In Maria Anisimova, editor, *Methods in Molecular Biology*, volume 856, pages 187–216. Humana Press, London, 2012. doi: 10.1007/978-1-61779-585-5_8.

Joana Fort, R. Laura, Hans E. Burghardt, Carles Ferrer-Costa, Javier Turnay, Cristina Ferrer-Orta, Isabel Usón, Antonio Zorzano, Juan Fernández-Recio, Modesto Orozco, María Antonia Lizarbe, and Andres Palacín. The structure of human 4f2hc ectodomain provides a model for homodimerization and electrostatic interaction with plasma membrane. *Journal of Biological Chemistry*, 282(43): 31444–31452, 2007.

Santo Fortunato. Community detection in graphs. *Phys Rep*, 486(3-5):75–174, 2010. ISSN 0370-1573. doi: 10.1016/j.physrep.2009.11.002.

James S. Fraser, John D. Gross, and Nevan J. Krogan. From systems to structure: Bridging networks and mechanism. *Molecular cell*, 49(2):222–231, 2013.

R. P. Freckleton, P. H. Harvey, and M. Pagel. Phylogenetic analysis and comparative data: a test and review of evidence. *The American Naturalist*, 160(6):712–726, 2002.

Zakhar M. Frenkel and Edward N. Trifonov. Closed loops of tim barrel protein fold. *Journal of Biomolecular Structure and Dynamics*, 22(6):643–655, 2005. doi: 10.1080/07391102.2005.10507032.

Zui Fujimoto, Kenji Takase, Nobuko Doui, Mitsuru Momma, Takashi Matsumoto, and Hiroshi Mizuno. Crystal structure of a catalytic-site mutant α-amylase from bacillus subtilis complexed with maltopentaose. *Journal of Molecular Biology*, 277 (2):393–407, 1998. ISSN 0022-2836. doi: 10.1006/jmbi.1997.1599.

Oxana V Galzitskaya and Bogdan S Melnik. Prediction of protein domain boundaries from sequence alone. *Protein Sci.*, 12(4):696–701, 2003. doi: 10.1110/ps.0233103.

László Z Garamszegi and Anders P. Møller. Effects of sample size and intraspecific variation in phylogenetic comparative studies: a meta-analytic review. *Biological Reviews*, 85(4):797–805, 2010.

W. S. Garver, K. Krishnan, J. R. Gallagos, M. Michikawa, G. A. Francis, and R. A. Heidenreich. Niemann-pick c1 protein regulates cholesterol transport to the trans-golgi network and plasma membrane caveolae. *Journal of lipid research*, 43(4): 579–589, 2002.

A. C. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L. J. Jensen, S. Bastuck, B. Dümpelfeld, et al. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440(7084):631–636, 2006.

J. C. Gelly, A. G. De Brevern, and S. Hazout. 'protein peeling': an approach for splitting a 3d protein structure into compact fragments. *Bioinformatics*, 22(2): 129–133, 2006.

Richard A George and Jaap Heringa. Snapdragon: a method to delineate protein structural domains from sequence data. *J. Mol. Biol.*, 316(3):839–51, 2002. doi: 10.1006/jmbi.2001.5387.

J. A. Gerlt and F. M. Raushel. Evolution of function in $(\beta/\alpha)_8$-barrel enzymes. *Current opinion in chemical biology*, 7(2):252–264, 2003. doi: 10.1016/S1367-5931(03)00019-X.

Bernard S. Gerstman and Prem P. Chapagain. Self-organization in protein folding and the hydrophobic interaction. *The Journal of chemical physics*, 123(5):054901, 2005.

E. Michael Gertz, Yi-Kuo Yu, Richa Agarwala, Alejandro A. Schäffer, and Stephen F. Altschul. Composition-based statistics and translated nucleotide searches: improving the tblastn module of blast. *BMC biology*, 4(1):41, 2006.

Jan E Gewehr and Ralf Zimmer. Ssep-domain: protein domain prediction by alignment of secondary structure elements and profiles. *Bioinformatics*, 22(2):181–7, 2006. doi: 10.1093/bioinformatics/bti751.

Pier Federico Gherardini, Gabriele Ausiello, Robert B Russell, and Manuela Helmer-Citterich. Modular architecture of nucleotide-binding pockets. *Nucleic Acids Res.*, 38(11):3809–16, 2010. doi: 10.1093/nar/gkq090.

M Girvan and M E J Newman. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. U.S.A.*, 99(12):7821–7826, 2002. doi: 10.1073/pnas.122653799.

Richard A. Goldstein. The structure of protein evolution and the evolution of protein structure. *Curr. Opin. Struct. Biol.*, 18(2):170–7, 2008. doi: 10.1016/j.sbi.2008.01.006.

Phillip Good. *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. Springer, New York, 2nd edition, 2000. ISBN 038798898X.

Lesley H. Greene, Tony E. Lewis, Sarah Addou, Alison Cuff, Tim Dallman, Mark Dibley, Oliver Redfern, Frances Pearl, Rekha Nambudiry, Adam Reid, Ian Sillitoe, Corin Yeats, Janet M. Thornton, and Christine A. Orengo. The cath domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic Acids Res.*, 35(Database issue):–291, 2007. doi: 10.1093/nar/gkl959.

Frédéric Guérin, Sophie Barbe, Sandra Pizzut-Serin, Gabrielle Potocki-Véronèse, David Guieysse, Valérie Guillet, Pierre Monsan, Lionel Mourey, Magali Remaud-Siméon, Isabelle André, and Samuel Tranier. Structural investigation of the thermostability and product specificity of amylosucrase from the bacterium deinococcus geothermalis. *J. Biol. Chem.*, 287(9):6642–54, 2012. doi: 10.1074/jbc.M111.322917.

Annat Haber. The evolution of morphological integration in the ruminant skull. *Evolutionary Biology*, 42(1):99–114, 2015. doi: 10.1007/s11692-014-9302-7.

J. D. Hadfield and S. Nakagawa. General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. *Journal of evolutionary biology*, 23(3):494–508, 2010.

Jarrod D. Hadfield. Mcmc methods for multi-response generalized linear mixed models: the mcmcglmm r package. *Journal of Statistical Software*, 33(2):1–22, 2010.

Ellinor Haglund, Jens Danielsson, Saraboji Kadhirvel, Magnus O. Lindberg, Derek T. Logan, and Mikael Oliveberg. Trimming down a protein structure to its bare foldons. *Journal of Biological Chemistry*, 287(4):2731–2738, 2012.

Peng Han, Peng Zhou, Songqing Hu, Shaoqing Yang, Qiaojuan Yan, and Zhengqiang Jiang. A novel multifunctional α-amylase from the thermophilic fungus *Malbranchea cinnamomea*: Biochemical characterization and three-dimensional structure. *Applied Biochemistry and Biotechnology*, 2(170):420–435, 2013. doi: 10.1007/s12010-013-0198-y.

Thomas F. Hansen and David Houle. Measuring and comparing evolvability and constraint in multivariate characters. *J. Evol. Biol.*, 21(5):1201–19, 2008. doi: 10.1111/j.1420-9101.2008.01573.x.

Luke J. Harmon and Jonathan B. Losos. The effect of intraspecific sample size on type i and type ii error rates in comparative studies. *Evolution*, 59(12):2705–2710, 2005.

Michael J Harms and Joseph W Thornton. Evolutionary biochemistry: revealing the historical and physical causes of protein properties. *Nat. Rev. Genet.*, 14(8):559–71, 2013. doi: 10.1038/nrg3540.

Hitomi Hasegawa and Liisa Holm. Advances and pitfalls of protein structural alignment. *Current Opinion in Structural Biology*, 19(3):341–348, 2009. ISSN 0959-440X. doi: 10.1016/j.sbi.2009.04.003. URL http://www.sciencedirect.com/science/article/pii/S0959440X09000621.

B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl. Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of chemical theory and computation*, 4(3):435–447, 2008. doi: 10.1021/ct700301q.

Jose Sergio Hleap, Khanh N. Nguyen, Alex Safatli, and Christian Blouin. Reference matters: An efficient and scalable algorithm for large multiple structure alignment. In Fahad Saeed and Bhaskar DasGupta, editors, *Proceedings of the 5th International Conference on Bioinformatics and Computational Biology (BICOB–2013)*, pages 153–158, Winona, MN, USA, 2013a.

Jose Sergio Hleap, Edward Susko, and Christian Blouin. Defining structural and evolutionary modules in proteins: a community detection approach to explore sub-domain architecture. *BMC structural biology*, 13(1):20, 2013b. doi: 10.1186/1472-6807-13-20.

B. Höcker, C. Jürgens, M. Wilmanns, and R. Sterner. Stability, catalytic versatility and evolution of the $(\beta/\alpha)_8$-barrel fold. *Current opinion in biotechnology*, 12(4):376–381, 2001.

Timothy A Holland, Stella Veretnik, Ilya N Shindyalov, and Philip E Bourne. Partitioning protein structures into domains: why is it so difficult? *J. Mol. Biol.*, 361 (3):562–90, 2006. doi: 10.1016/j.jmb.2006.05.060.

Petter Holme. Metabolic robustness and network modularity: a model study. *Plos One*, 6(2):–16605, 2011.

Hironori Hondoh, Takashi Kuriki, and Yoshiki Matsuura. Three-dimensional structure and substrate binding of bacillus stearothermophilus neopullulanase. *Journal of molecular biology*, 326(1):177–188, 2003.

Hironori Hondoh, Wataru Saburi, Haruhide Mori, Masayuki Okuyama, Toshitaka Nakada, Yoshiki Matsuura, and Atsuo Kimura. Substrate recognition mechanism of α-1, 6-glucosidic linkage hydrolyzing enzyme, dextran glucosidase from *Streptococcus mutans*. *Journal of molecular biology*, 378(4):913–922, 2008. doi: 10.1016/j.jmb.2008.03.016.

David Houle. Numbering the hairs on our heads: the shared challenge and promise of phenomics. *Proc. Natl. Acad. Sci. U.S.A.*, 107 Suppl 1:1793–9, 2010. doi: 10.1073/pnas.0906195106.

C M House and L W Simmons. The evolution of male genitalia: patterns of genetic variation and covariation in the genital sclerites of the dung beetle onthophagus taurus. *J. Evol. Biol.*, 18(5):1281–92, 2005. doi: 10.1111/j.1420-9101.2005.00926.x.

Elizabeth A. Housworth, Emília P. Martins, and Michael Lynch. The phylogenetic mixed model. *The American Naturalist*, 163(1):84–96, 2004.

William Humphrey, Andrew Dalke, and Klaus Schulten. Vmd – visual molecular dynamics. *Journal of Molecular Graphics*, 14:33–38, 1996. doi: 10.1016/0263-7855(96)00018-5.

J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing In Science & Engineering*, 9(3):90–95, 2007. doi: 10.1109/MCSE.2007.55.

Sarah Hunter, Rolf Apweiler, Teresa K Attwood, Amos Bairoch, Alex Bateman, David Binns, Peer Bork, Ujjwal Das, Louise Daugherty, Lauranne Duquenne, Robert D Finn, Julian Gough, Daniel Haft, Nicolas Hulo, Daniel Kahn, Elizabeth Kelly, Aurélie Laugraud, Ivica Letunic, David Lonsdale, Rodrigo Lopez, Martin Madera, John Maslen, Craig McAnulla, Jennifer McDowall, Jaina Mistry, Alex Mitchell, Nicola Mulder, Darren Natale, Christine Orengo, Antony F Quinn, Jeremy D Selengut, Christian J A Sigrist, Manjula Thimma, Paul D Thomas, Franck Valentin, Derek Wilson, Cathy H Wu, and Corin Yeats. Interpro: the integrative protein signature database. *Nucleic Acids Res.*, 37(Database issue): D211–5, 2009. doi: 10.1093/nar/gkn785.

Kwang Yeon Hwang, Hyun Kyu Song, Changsoo Chang, Jungkyu Lee, S. Y. Lee, K. K. Kim, Senyon Choe, Robert M. Sweet, and S. W. Suh. Crystal structure of thermostable alpha-amylase from bacillus licheniformis refined at 1.7 a resolution. *Molecules and cells*, 7(2):251–258, 1997.

R. E. Infante, A. Radhakrishnan, L. Abi-Mosleh, L. N. Kinch, M. L. Wang, N. V. Grishin, J. L. Goldstein, and M. S. Brown. Purified npc1 protein. *Journal of Biological Chemistry*, 283(2):1064–1075, 2008a.

Rodney E Infante, Arun Radhakrishnan, Lina Abi-Mosleh, Lisa N Kinch, Michael L Wang, Nick V Grishin, Joseph L Goldstein, and Michael S Brown. Purified npc1 protein: Ii. localization of sterol binding to a 240-amino acid soluble luminal loop. *J. Biol. Chem.*, 283(2):1064–75, 2008b. doi: 10.1074/jbc.M707944200.

Anthony R Ives, Peter E Midford, and Theodore Garland. Within-species variation and measurement error in phylogenetic comparative methods. *Syst. Biol.*, 56(2):252–70, 2007. doi: 10.1080/10635150701313830. URL http://sysbio.oxfordjournals.org/content/56/2/252.abstract.

F. Jacob. Evolution and tinkering. *Science*, 196(4295):1161–1166, 1977. doi: 10.1126/science.860134.

Rainer Jaenicke. Protein folding: local structures, domains, subunits, and assemblies. *Biochemistry*, 30(13):3147–3161, 1991.

A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Comput. Surv.*, 31(3):264–323, September 1999. ISSN 0360-0300. doi: http://doi.acm.org/10.1145/331499.331504. URL http://doi.acm.org/10.1145/331499.331504.

Štefan Janeček, Birte Svensson, and E. Ann MacGregor. α-amylase: an enzyme specificity found in various families of glycoside hydrolases. *Cellular and Molecular Life Sciences*, pages 1–22, 2013.

Jing Jin, Xueying Xie, Chen Chen, Jin Gyoon Park, Chris Stark, D Andrew James, Marina Olhovsky, Rune Linding, Yongyi Mao, and Tony Pawson. Eukaryotic protein domains as functional units of cellular evolution. *Sci Signal*, 2(98):ra76, 2009. doi: 10.1126/scisignal.2000546.

Agnel Praveen Joseph, N. Srinivasan, and Alexandre G. de Brevern. Improvement of protein structure comparison using a structural alphabet. *Biochimie*, 93(9): 1434–1445, 2011. ISSN 0300-9084. doi: 10.1016/j.biochi.2011.04.010. URL http://www.sciencedirect.com/science/article/pii/S0300908411001295.

Agnel Praveen Joseph, Narayanaswamy Srinivasan, and Alexandre G. de Brevern. Progressive structure-based alignment of homologous proteins: Adopting sequence comparison strategies. *Biochimie*, 94(9):2025–2034, 2012. ISSN 0300-9084. doi: 10.1016/j.biochi.2012.05.028. URL http://www.sciencedirect.com/science/article/pii/S0300908412002167.

Masayuki Kagawa, Zui Fujimoto, Mitsuru Momma, Kenji Takase, and Hiroshi Mizuno. Crystal structure of *Bacillus subtilis* α-amylase in complex with acarbose. *Journal of Bacteriology*, 185(23):6981–6984, 2003. doi: 10.1128/JB.185.23.6981-6984.2003.

Ryuta Kanai, Keiko Haga, Toshihiko Akiba, Kunio Yamane, and Kazuaki Harata. Biochemical and crystallographic analyses of maltohexaose-producing amylase from alkalophilic bacillus sp. 707. *Biochemistry*, 43(44):14047–14056, 2004a.

Ryuta Kanai, Keiko Haga, Toshihiko Akiba, Kunio Yamane, and Kazuaki Harata. Role of phe283 in enzymatic reaction of cyclodextrin glycosyltransferase from alkalophilic bacillus sp. 1011: Substrate binding and arrangement of the catalytic site. *Protein science*, 13(2):457–465, 2004b. doi: 10.1110/ps.03408504.

Maricel G Kann, Paul A Thiessen, Anna R Panchenko, Alejandro A Schäffer, Stephen F Altschul, and Stephen H Bryant. A structure-based method for protein sequence alignment. *Bioinformatics*, 21(8):1451–6, 2005. doi: 10.1093/bioinformatics/bti233. URL http://bioinformatics.oxfordjournals.org/content/21/8/1451.abstract.

R. Karchin, M. Diekhans, L. Kelly, D. J. Thomas, U. Pieper, N. Eswar, D. Haussler, and A. Sali. Ls-snp: large-scale annotation of coding non-synonymous snps based on multiple information sources. *Bioinformatics*, 21(12):2814–2820, 2005.

Susan Khor. Towards an integrated understanding of the structural characteristics of protein residue networks. *Theory Biosci.*, 131(2):61–75, 2012. doi: 10.1007/s12064-011-0135-y.

H. K. Kim, J. W. Liu, Paul D. Carr, and David L. Ollis. Following directed evolution with crystallography: structural changes observed in changing the substrate specificity of dienelactone hydrolase. *Acta Crystallographica Section D: Biological Crystallography*, 61(7):920–931, 2005.

Jeong-Sun Kim, Sun-Shin Cha, Hyun-Ju Kim, Tae-Jip Kim, Nam-Chul Ha, Sang-Taek Oh, Hyun-Soo Cho, Moon-Ju Cho, Myo-Jeong Kim, Hee-Seob Lee, et al. Crystal structure of a maltogenic amylase provides insights into a catalytic versatility. *Journal of Biological Chemistry*, 274(37):26279–26286, 1999.

Jongkwang Kim and Kai Tan. Discover protein complexes in protein-protein interaction networks using parametric local modularity. *BMC bioinformatics*, 11(1): 521, 2010.

Myung-Il Kim, Hong-Suk Kim, Jin Jung, and Sangkee Rhee. Crystal structures and mutagenesis of sucrose hydrolase from xanthomonas axonopodis pv. glycines: insight into the exclusively hydrolytic amylosucrase fold. *J. Mol. Biol.*, 380(4): 636–47, 2008. doi: 10.1016/j.jmb.2008.05.046.

Lisa N. Kinch and Nick V. Grishin. Evolution of protein structures and functions. *Current Opinion in Structural Biology*, 12(3):400–408, 2002. ISSN 0959-440X. doi: 10.1016/S0959-440X(02)00338-X. URL http://www.sciencedirect.com/science/article/pii/S0959440X0200338X.

H. Kitano. Systems biology: a brief overview. *Science*, 295(5560):1662–1664, 2002.

Karl M Kjer, Joseph J Gillespie, and Karen A Ober. Opinions on multiple sequence alignment, and an empirical comparison of repeatability and accuracy between poy and structural alignment. *Syst. Biol.*, 56(1):133–46, 2007. doi: 10.1080/10635150601156305. URL http://sysbio.oxfordjournals.org/content/56/1/133.short.

Christian Peter Klingenberg. Morphometric integration and modularity in configurations of landmarks: tools for evaluating a priori hypotheses. *Evol Dev*, 11(4): 405–21, 2009. doi: 10.1111/j.1525-142X.2009.00347.x.

Christian Peter Klingenberg and Larry J. Leamy. Quantitative genetics of geometric shape in the mouse mandible. *Evolution*, 55(11):2342–2352, 2001. ISSN 1558-5646. doi: 10.1111/j.0014-3820.2001.tb00747.x. URL http://dx.doi.org/10.1111/j.0014-3820.2001.tb00747.x.

C.P. Klingenberg. Quantitative genetics of geometric shape: heritability and the pitfalls of the univariate approach. *Evolution*, 57(1):191–195, 2003.

Patrice Koehl. Protein structure classification. *Reviews in Computational Chemistry*, 22:1, 2006.

Rachel Kolodny, Patrice Koehl, and Michael Levitt. Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. *J. Mol. Biol.*, 346(4):1173–88, 2005. doi: 10.1016/j.jmb.2004.12.032.

Rachel Kolodny, Leonid Pereyaslavets, Abraham O Samson, and Michael Levitt. On the universe of protein folds. *Annu Rev Biophys*, 42:559–82, 2013. doi: 10.1146/annurev-biophys-083012-130432.

J. Konc and D Janežič. Probis algorithm for detection of structurally similar protein binding sites by local structural alignment. *Bioinformatics*, 26(9):1160–1168, 2010. doi: 10.1093/bioinformatics/btq100.

H. J. Kwon, L. Abi-Mosleh, M. L. Wang, J. Deisenhofer, J. L. Goldstein, M. S. Brown, and R. E. Infante. Structure of n-terminal domain of npc1 reveals distinct subdomains for binding and transfer of cholesterol. *Cell*, 137(7):1213–1224, 2009. doi: 10.1016/j.cell.2009.03.049.

Clemens Lakner, Mark T Holder, Nick Goldman, and Gavin J P Naylor. What's in a likelihood? simple models of protein evolution and the contribution of structurally viable reconstructions to the likelihood. *Syst. Biol.*, 60(2):161–74, 2011. doi: 10.1093/sysbio/syq088.

Paul C. Lambert, Alex J. Sutton, Paul R. Burton, Keith R. Abrams, and David R. Jones. How vague is vague?: A simulation study of the impact of the use of vague prior distributions in mcmc using winbugs. *Statistics in medicine*, 24(15): 2401–2428, 2005.

Andrea Lancichinetti, Santo Fortunato, and János Kertész. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics, Volume 11, Issue 3, id. 033015 (2009).*, 11:3015, mar 2009. doi: 10.1088/1367-2630/11/3/033015.

Andrea Lancichinetti, Filippo Radicchi, José J. Ramasco, and Santo Fortunato. Finding statistically significant communities in networks. *PloS one*, 6(4):e18961, 2011.

Russell Lande and Stevan J. Arnold. The measurement of selection on correlated characters. *Evolution*, pages 1210–1226, 1983.

Hee-Seob Lee, Min-Sung Kim, Hyun-Soo Cho, Jung-In Kim, Tae-Jip Kim, Ji-Hye Choi, Cheonseok Park, Heung-Soo Lee, Byung-Ha Oh, and Kwan-Hwa Park. Cyclomaltodextrinase, neopullulanase, and maltogenic amylase are nearly indistinguishable from each other. *Journal of Biological Chemistry*, 277(24):21891–21897, 2002. doi: 10.1074/jbc.M201623200.

S. Lele and T. M. Cole. A new test for shape differences when variance–covariance matrices are unequal. *Journal of Human Evolution*, 31(3):193–212, 1996. ISSN 0047-2484. doi: 10.1006/jhev.1996.0057. URL http://www.sciencedirect.com/science/article/pii/S0047248496900573.

Subhash Lele. Euclidean distance matrix analysis (edma): Estimation of mean form and mean form difference. *Mathematical Geology*, 25(5):573–602, 1993. ISSN 0882-8121. URL http://dx.doi.org/10.1007/BF00890247. 10.1007/BF00890247.

Subhash Lele and Joan T. Richtsmeier. On comparing biological shapes: Detection of influential landmarks. *American Journal of Physical Anthropology*, 87(1):49–65, 1992. ISSN 1096-8644. doi: 10.1002/ajpa.1330870106. URL http://dx.doi.org/10.1002/ajpa.1330870106.

Ivica Letunic, Tobias Doerks, and Peer Bork. Smart 7: recent updates to the protein domain annotation resource. *Nucleic Acids Res.*, 40(Database issue):D302–5, 2012. doi: 10.1093/nar/gkr931.

Bi-Qing Li, Le-Le Hu, Lei Chen, Kai-Yan Feng, Yu-Dong Cai, and Kuo-Chen Chou. Prediction of protein domain with mrmr feature selection and analysis. *PLoS ONE, Edited by Bin Xue, vol. 7, issue 6, p. e39308*, 7:39308, jun 2012. doi: 10.1371/journal.pone.0039308.

Olivier Lichtarge, Henry R Bourne, and Fred E Cohen. An evolutionary trace method defines binding surfaces common to protein families. *Journal of molecular biology*, 257(2):342–358, 1996.

Alexandra Lipski, Hildegard Watzlawick, Stephanie Ravaud, Xavier Robert, Moez Rhimi, Richard Haser, Ralf Mattes, and Nushin Aghajari. Mutations inducing an active-site aperture in rhizobium sp. sucrose isomerase confer hydrolytic activity. *Acta Crystallographica Section D: Biological Crystallography*, 69(2):298–307, 2013. doi: 10.1107/S0907444912045532.

Jinfeng Liu and Burkhard Rost. Sequence-based prediction of protein domains. *Nucleic acids research*, 32(12):3522–3530, 2004a. doi: 10.1093/nar/gkh684.

Jinfeng Liu and Burkhard Rost. Chop proteins into structural domain-like fragments. *Proteins*, 55(3):678–88, 2004b. doi: 10.1002/prot.20095.

Rong Liu and Jianjun Hu. Computational prediction of heme-binding residues by exploiting residue interaction network. *PLoS ONE*, 6(10):e25560, 2011. doi: 10.1371/journal.pone.0025560.

Ying Liu and Ivet Bahar. Sequence evolution correlates with structural dynamics. *Mol. Biol. Evol.*, 29(9):2253–63, 2012. doi: 10.1093/molbev/mss097.

Dirk M. Lorenz, Alice Jeng, and Michael W. Deem. The emergence of modularity in biological systems. *Physics of Life Reviews*, 8(2):129–160, jun 2011. doi: 10.1016/j.plrev.2011.02.003.

Martin Lundgren, Andrey Krokhotin, and Antti J. Niemi. Topology and structural self-organization in folded proteins. *Physical Review E*, 88(4):042709, 2013. doi: 10.1103/PhysRevE.88.042709.

Louise Lyhne-Iversen, Timothy J. Hobley, Svend G. Kaasgaard, and Pernille Harris. Structure of bacillus halmapalus-amylase crystallized with and without the substrate analogue acarbose and maltose. *Acta Crystallographica Section F: Structural Biology and Crystallization Communications*, 62(9):849–854, 2006.

Michael Lynch. Methods for the analysis of comparative data in evolutionary biology. *Evolution*, pages 1065–1080, 1991.

E. Ann MacGregor, Štefan Janeček, and Birte Svensson. Relationship of sequence and structure to specificity in the α-amylase family of enzymes. *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology*, 1546(1): 1–20, 2001. ISSN 0167-4838. doi: 10.1016/S0167-4838(00)00302-2.

Mischa Machius, Nathalie Declerck, Robert Huber, and Georg Wiegand. Activation of bacillus licheniformis α-amylase through a disorder→order transition of the substrate-binding site mediated by a calcium–sodium–calcium metal triad. *Structure*, 6(3):281–292, 1998. ISSN 0969-2126. doi: 10.1016/S0969-2126(98)00032-X.

Mischa Machius, Nathalie Declerck, Robert Huber, and Georg Wiegand. Kinetic stabilization of bacillus licheniformis α-amylase through introduction of hydrophobic residues at the surface. *Journal of Biological Chemistry*, 278(13):11546–11553, 2003. doi: 10.1074/jbc.M212618200.

N. Macleod. Geometric morphometrics and geological shape-classification systems. *Earth-Science Reviews*, 59(1-4):27–47, nov 2002. doi: 10.1016/S0012-8252(02)00068-5. URL http://dx.doi.org/10.1016/S0012-8252(02)00068-5.

John Maindonald and W. John Braun. *DAAG: Data Analysis And Graphics data and functions*, 2011. URL http://CRAN.R-project.org/package=DAAG. R package version 1.08.

Indraneel Majumdar, Lisa N. Kinch, and Nick V. Grishin. A database of domain definitions for proteins with complex interdomain geometry. *PLoS ONE, Edited by Mark Isalan, vol. 4, issue 4, p. e5084*, 4:5084, apr 2009. doi: 10.1371/journal.pone.0005084.

Aron Marchler-Bauer, John B Anderson, Praveen F Cherukuri, Carol DeWeese-Scott, Lewis Y Geer, Marc Gwadz, Siqian He, David I Hurwitz, John D Jackson, Zhaoxi Ke, Christopher J Lanczycki, Cynthia A Liebert, Chunlei Liu, Fu Lu, Gabriele H Marchler, Mikhail Mullokandov, Benjamin A Shoemaker, Vahan Simonyan, James S Song, Paul A Thiessen, Roxanne A Yamashita, Jodie J Yin, Dachuan Zhang, and Stephen H Bryant. Cdd: a conserved domain database for protein classification. *Nucleic Acids Res.*, 33(Database issue):D192–6, 2005. doi: 10.1093/nar/gki069.

Aron Marchler-Bauer, Shennan Lu, John B Anderson, Farideh Chitsaz, Myra K Derbyshire, Carol DeWeese-Scott, Jessica H Fong, Lewis Y Geer, Renata C Geer, Noreen R Gonzales, Marc Gwadz, David I Hurwitz, John D Jackson, Zhaoxi Ke, Christopher J Lanczycki, Fu Lu, Gabriele H Marchler, Mikhail Mullokandov, Marina V Omelchenko, Cynthia L Robertson, James S Song, Narmada Thanki, Roxanne A Yamashita, Dachuan Zhang, Naigong Zhang, Chanjuan Zheng, and Stephen H Bryant. Cdd: a conserved domain database for the functional annotation of proteins. *Nucleic Acids Res.*, 39(Database issue):D225–9, 2011. doi: 10.1093/nar/gkq1189.

Gabriel Marroig and James M. Cheverud. A comparison of phenotypic variation and covariation patterns and the role of phylogeny, ecology, and ontogeny during cranial evolution of new world monkeys. *Evolution*, 55(12):2576–2600, 2001.

Gabriel Marroig, Leila T. Shirai, Arthur Porto, Felipe B. de Oliveira, and Valderes De Conto. The evolution of modularity in the mammalian skull ii: evolutionary consequences. *Evolutionary Biology*, 36(1):136–148, 2009. doi: 10.1007/s11692-009-9051-1.

Russell L Marsden, Liam J McGuffin, and David T Jones. Rapid protein domain assignment from amino acid sequence using predicted secondary structure. *Protein Sci.*, 11(12):2814–24, 2002. doi: 10.1110/ps.0209902.

Emília P. Martins, José Alexandre F. Diniz-Filho, and Elizabeth A. Housworth. Adaptive constraints and the phylogenetic comparative method: a computer simulation test. *Evolution*, 56(1):1–13, 2002.

Yoshiki Matsuura, Masami KUSUNOKI, Wakako HARADA, and Masao KAKUDO. Structure and possible catalytic residues of taka-amylase a. *Journal of Biochemistry*, 95(3):697–702, 1984.

Geoffrey McLachlan. *Discriminant analysis and statistical pattern recognition*, volume 544. John Wiley & Sons, Inc, Hoboken, NJ, 2004.

Matthew Menke, Bonnie Berger, and Lenore Cowen. Matt: local flexibility aids protein multiple structure alignment. *PLoS Comput. Biol.*, 4(1):e10, 2008. doi: 10.1371/journal.pcbi.0040010.

Karin Meyer. Factor-analytic models for genotype x environment type problems and structured covariance matrices. *Genet. Sel. Evol.*, 41:21, 2009. doi: 10.1186/1297-9686-41-21.

Karin Meyer and Mark Kirkpatrick. Perils of parsimony: properties of reduced-rank estimates of genetic covariance matrices. *Genetics*, 180(2):1153–66, 2008. doi: 10.1534/genetics.108.090159.

Yoshihiro Mezaki, Yoshio Katsuya, Michio Kubota, and Yoshiki Matsuura. Crystallization and structural analysis of intact maltotetraose-forming exo-amylase from pseudomonas stutzeri. *Bioscience, biotechnology, and biochemistry*, 65(1):222–225, 2001. doi: 10.1271/bbb.65.222.

Bunzo Mikami, Hiroyuki Iwamoto, Dominggus Malle, Hye-Jin Yoon, Elif Demirkan-Sarikaya, Yoshihiro Mezaki, and Yoshio Katsuya. Crystal structure of pullulanase: evidence for parallel binding of oligosaccharides in the active site. *Journal of molecular biology*, 359(3):690–707, 2006. doi: 10.1016/j.jmb.2006.03.058.

Gilles Millat, Nathalie Baïlo, Sabine Molinero, Céline Rodriguez, Karim Chikh, and Marie T Vanier. Niemann-pick c disease: use of denaturing high performance liquid chromatography for the detection of npc1 and npc2 genetic variations and impact on management of patients and families. *Mol. Genet. Metab.*, 86(1-2): 220–32, 2005. ISSN 1096-7192. doi: 10.1016/j.ymgme.2005.07.007.

Thomas P. Minka. Automatic choice of dimensionality for pca. In *NIPS*, volume 13, pages 598–604, 2000.

Allen P. Minton. Implications of macromolecular crowding for protein assembly. *Current opinion in structural biology*, 10(1):34–39, 2000.

Pooja Mishra and Paras Nath Pandey. A graph-based clustering method applied to protein sequences. *Bioinformation*, 6(10):372–374, 2011.

Jay Mittenthal, Derek Caetano-Anollés, and Gustavo Caetano-Anollés. Biphasic patterns of diversification and the emergence of modules. *Front Genet*, 3:147, 2012. doi: 10.3389/fgene.2012.00147.

K Mizuguchi, C M Deane, T L Blundell, and J P Overington. Homstrad: a database of protein structure alignments for homologous families. *Protein Sci.*, 7(11):2469–71, 1998. ISSN 1469-896X. doi: 10.1002/pro.5560071126.

Masahiro Mizuno, K.azuhiro Ichikawa, Takashi Tonozuka, Akashi Ohtaki, Yoichiro Shimura, Shigehiro Kamitori, Atsushi. Nishikawa, and Yoshiyuki Sakano. Mutagenesis and structural analysis of thermoactinomyces vulgarius r-47 α-amylase ii (tvaii). *Journal of Applied Glycoscience*, 52(3):225–231, 2005. doi: 10.5458/jag.52.225.

Kimihiko Mizutani, Mayuko Toyoda, Yuichiro Otake, Soshi Yoshioka, Nobuyuki Takahashi, and Bunzo Mikami. Structural and functional characterization of recombinant medaka fish alpha-amylase expressed in yeast pichia pastoris. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, 1824(8):954–962, 2012. doi: 10.1016/j.bbapap.2012.05.005.

Sook-Chen Mok, Aik-Hong Teh, Jennifer A. Saito, Nazalan Najimudin, and Maqsudul Alam. Crystal structure of a compact α-amylase from *Geobacillus thermoleovorans*. *Enzyme and microbial technology*, 53(1):46–54, 2013. doi: 10.1016/j.enzmictec.2013.03.009.

Marie S. Møller, Folmer Fredslund, Avishek Majumder, Hiroyuki Nakai, Jens-Christian N. Poulsen, Leila Lo Leggio, Birte Svensson, and Maher Abou Hachem. Enzymology and structure of the gh13_31 glucan 1, 6-α-glucosidase that confers isomaltooligosaccharide utilization in the probiotic lactobacillus acidophilus ncfm. *Journal of bacteriology*, 194(16):4249–4259, 2012. doi: 10.1128/JB.00622-12.

Leandro R. Monteiro, José Alexcandre F. Diniz-Filho, Sérgio F. Reis, and Edilson D. Araújo. Geometric estimates of heritability in biological shape. *Evolution*, 56(3): 563–572, 2002.

Robb J. Muirhead. *Aspects of Multivariate Statistical Theory*. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2008. ISBN 9780470316559. doi: 10.1002/9780470316559.

Duncan Murdoch and E. D. Chow. *ellipse: Functions for drawing ellipses and ellipse-like confidence regions*, 2013. URL http://CRAN.R-project.org/package=ellipse. R package version 0.3-8.

A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. Scop: a structural classification of proteins database for the investigation of sequences and structures. *Journal of molecular biology*, 247(4):536–540, 1995.

Niranjan Nagarajan and Golan Yona. Automatic prediction of protein domains from sequence information using a hybrid learning system. *Bioinformatics*, 20(9):1335–60, 2004. doi: 10.1093/bioinformatics/bth086.

M. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys Rev E*, 69(2):026113, feb 2004. ISSN 1539-3755. doi: 10.1103/PhysRevE.69.026113. URL http://dx.doi.org/10.1103/PhysRevE.69.026113.

M E J Newman. Analysis of weighted networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, 70(5 Pt 2):056131, 2004. ISSN 1539-3755. doi: 10.1103/PhysRevE.70.056131.

Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.

M. N. Nguyen, K. P. Tan, and M. S. Madhusudhan. Click—topology-independent comparison of biomolecular 3d structures. *Nucleic acids research*, 39(suppl 2):–24, 2011. doi: 10.1093/nar/gkr393.

J. E. Nielsen and T. V. Borchert. Protein engineering of bacterial α-amylases. *Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology*, 1543(2):253–274, 2000.

Anastasia N. Nikolskaya, Cecilia N. Arighi, Hongzhan Huang, Winona C. Barker, and Cathy H. Wu. Pirsf family classification system for protein functional and evolutionary analysis. *Evolutionary bioinformatics online*, 2:197, 2006.

Tsuyoshi Nonaka, Masahiro Fujihashi, Akiko Kita, Hiroshi Hagihara, Katsuya Ozaki, Susumu Ito, and Kunio Miki. Crystal structure of calcium-free α-amylase from bacillus sp. strain ksm-k38 (amyk38) and its sodium ion binding sites. *Journal of Biological Chemistry*, 278(27):24818–24824, 2003.

Petr Novák, Pavel Neumann, and Jirí Macas. Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinformatics*, 11:378, 2010. doi: 10.1186/1471-2105-11-378.

Shin Numao, Robert Maurus, Gary Sidhu, Yili Wang, Christopher M. Overall, Gary D. Brayer, and Stephen G. Withers. Probing the role of the chloride ion in the mechanism of human pancreatic α-amylase. *Biochemistry*, 41(1):215–225, 2002.

Akashi Ohtaki, Shin Kondo, Yoichiro Shimura, Takashi Tonozuka, Yoshiyuki Sakano, and Shigehiro Kamitori. Role of phe286 in the recognition mechanism of cyclomaltooligosaccharides (cyclodextrins) by *Thermoactinomyces vulgaris* r-47 α-amylase 2 (tvaii). x-ray structures of the mutant tvaiis, f286a and f286y, and kinetic analyses of the phe286-replaced mutant tvaiis. *Carbohydrate Research*, 334(4):309–313, 2001. doi: 10.1016/S0008-6215(01)00190-2.

Akashi Ohtaki, Masahiro Mizuno, Takashi Tonozuka, Yoshiyuki Sakano, and Shigehiro Kamitori. Complex structures of thermoactinomyces vulgaris r-47 a-amylase 2 with acarbose and cyclodextrins demonstrate the multiple substrate recognition mechanism. *Journal of Biological Chemistry*, 279(30):31033–31040, 2004. doi: 10.1074/jbc.M404311200.

C. A. Orengo, I. Sillitoe, G. Reeves, and F. M. Pearl. Review: what can structural classifications reveal about protein evolution? *J. Struct. Biol.*, 134(2-3):145–65, 2001. doi: 10.1006/jsbi.2001.4398.

Margarita Osadchy and Rachel Kolodny. Maps of protein structure space reveal a fundamental relationship between protein structure and function. *Proceedings of the National Academy of Sciences*, 108(30):12301–12306, 2011. doi: /10.1073/pnas.1102727108. URL http://www.pnas.org/content/early/2011/06/28/1102727108.abstract.

Pemra Ozbek, Seren Soner, Burak Erman, and Turkan Haliloglu. Dnabindprot: fluctuation-based predictor of dna-binding residues within a network of interacting residues. *Nucleic Acids Res.*, 38(Web Server issue):W417–23, 2010. doi: 10.1093/nar/gkq396.

Mark Pagel. Inferring the historical patterns of biological evolution. *Nature*, 401 (6756):877–884, 1999.

Catherine Qiurong Pan, Marius Sudol, Michael Sheetz, and Boon Chuan Low. Modularity and functional plasticity of scaffold proteins as p(l)acemakers in cell signaling. *Cellular Signalling*, 24(11):2143–2165, 2012.

Emmanuel Paradis, Julien Claude, and Korbinian Strimmer. Ape: Analyses of phylogenetics and evolution in r language. *Bioinformatics*, 20:289–290, 2004. doi: 10.1093/bioinformatics/btg412.

Kwan-Hwa Park, Tae-Jip Kim, Tae-Kyou Cheong, Jung-Wan Kim, Byung-Ha Oh, and Birte Svensson. Structure, specificity and function of cyclomaltodextrinase, a multispecific enzyme of the α-amylase family. *Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology*, 1478(2):165–185, 2000. doi: 10.1016/S0167-4838(00)00041-8.

Marc C. Patterson, Marie T. Vanier, Kinuko Suzuki, Jill A. Morris, Eugene Carstea, Edward B. Neufeld, Joan E. Blanchette-Mackie, and Peter G. Pentchev. *Scriver's OMMBID: The Online Metabolic and Molecular Bases of Inherited Disease*, chapter Chapter 145: Niemann-Pick Disease Type C: A Lipid Trafficking Disorder. Mc-Graw HIll Inc, New York, 2006. doi: 10.1036/ommbid.175.

László Patthy. Genome evolution and the evolution of exon-shuffling — a review. *Gene*, 238(1):103–114, 1999. ISSN 0378-1119. doi: 10.1016/S0378-1119(99)00228-0.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Dirk Penninga, Boris Strokopytov, Henrieette J. Rozeboom, Catherine L. Lawson, Bauke W. Dijkstra, Jack Bergsma, and Lubbert Dijkhuizen. Site-directed mutations in tyrosine 195 of cyclodextrin glycosyltransferase from bacillus circulans strain 251 affect activity and product specificity. *Biochemistry*, 34(10):3368–3376, 1995.

Patrick C. Phillips and Stevan J. Arnold. Hierarchical comparison of genetic variance-covariance matrices. i. using the flury hierarchy. *Evolution*, 53(5):1506–1515, 1999. ISSN 00143820. URL http://www.jstor.org/stable/2640896.

H. P. Piepho, J. Möhring, A. E. Melchinger, and A. Büchse. Blup for phenotypic selection in plant breeding and variety testing. *Euphytica*, 161(1-2):209–228, 2008.

M. Pigliucci. Is evolvability evolvable? *Nature Reviews Genetics*, 9(1):75–82, 2008.

Aleksandar Poleksic. Optimizing a widely used protein structure alignment measure in expected polynomial time. *IEEE/ACM Trans Comput Biol Bioinform*, 8(6):1716–20, nov.-dec. 2011. ISSN 1545-5963. doi: 10.1109/TCBB.2011.122.

Chris P Ponting and Robert R Russell. The natural history of protein domains. *Annu Rev Biophys Biomol Struct*, 31:45–71, 2002. doi: 10.1146/annurev.biophys.31.082901.134314.

G. V. Popescu and S. C. Popescu. *Handbook of Research on Computational and Systems Biology: Interdisciplinary Applications*, volume 1, chapter Complexity and Modularity of MAPK Signaling Networks, pages 355–368. IGI Global, Hershey,PA, 2011. doi: 10.4018/978-1-60960-491-2.ch016.

Arthur Porto, Leila Teruko Shirai, Felipe Bandoni de Oliveira, and Gabriel Marroig. Size variation, growth strategies, and the evolution of modularity in the mammalian skull. *Evolution*, 67(11):3305–22, 2013. doi: 10.1111/evo.12177.

Om Prakash and Nivedita Jaiswal. α-amylase: an ideal representative of thermostable enzymes. *Applied biochemistry and biotechnology*, 160(8):2401–2414, 2010. doi: 10.1007/s12010-009-8735-4.

Morgan N. Price, Paramvir S. Dehal, and Adam P. Arkin. Fasttree 2–approximately maximum-likelihood trees for large alignments. *PloS one*, 5(3):e9490, 2010.

Miguel Prôa, Paul O'Higgins, and Leandro R Monteiro. Type i error rates for testing genetic drift with phenotypic covariance matrices: a simulation study. *Evolution*, 67(1):185–95, 2013. doi: 10.1111/j.1558-5646.2012.01746.x.

M. Punta and B. Rost. Protein folding rates estimated from contact predictions. *Journal of molecular biology*, 348(3):507–512, 2005.

Marco Punta, Penny C Coggill, Ruth Y Eberhardt, Jaina Mistry, John Tate, Chris Boursnell, Ningze Pang, Kristoffer Forslund, Goran Ceric, Jody Clements, Andreas Heger, Liisa Holm, Erik L L Sonnhammer, Sean R Eddy, Alex Bateman, and Robert D Finn. The pfam protein families database. *Nucleic Acids Res.*, 40 (Database issue):D290–301, 2012. doi: 10.1093/nar/gkr1065.

Development Core Team R. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2011. URL http://www.R-project.org/. ISBN 3-900051-07-0.

Mahsa Rahimzadeh, Khosro Khajeh, Manoochehr Mirshahi, Mahmood Khayatian, and Robert Schwarzenbacher. Probing the role of asparagine mutation in thermostability of *Bacillus* kr-8104 α-amylase. *International Journal of Biological Macromolecules*, 50(4):1175–1182, 2012. doi: 10.1016/j.ijbiomac.2011.11.014.

Mark D. Rausher. The measurement of selection on quantitative traits: biases due to environmental covariances between traits and fitness. *Evolution*, pages 616–626, 1992. doi: http://doi.org/10.2307/2409632.

Stéphanie Ravaud, Xavier Robert, Hildegard Watzlawick, Richard Haser, Ralf Mattes, and Nushin Aghajari. Trehalulose synthase native and carbohydrate complexed structures provide insights into sucrose isomerization. *J. Biol. Chem.*, 282 (38):28126–36, 2007. doi: 10.1074/jbc.M704515200.

Stéphanie Ravaud, Xavier Robert, Hildegard Watzlawick, Richard Haser, Ralf Mattes, and Nushin Aghajari. Structural determinants of product specificity of sucrose isomerases. *FEBS Lett.*, 583(12):1964–8, 2009. doi: 10.1016/j.febslet.2009.05.002.

Oliver C. Redfern, Andrew Harrison, Tim Dallman, Frances M. G. Pearl, and Christine A. Orengo. Cathedral: A fast and effective algorithm to predict folds and domain boundaries from multidomain protein structures. *PLoS Computational Biology, vol. 3, issue 11, p. e232*, 3:232, n/a 2007. doi: 10.1371/journal.pcbi.0030232.

Dana Reichmann, Ofer Rahat, Mati Cohen, Hani Neuvirth, and Gideon Schreiber. The molecular architecture of protein-protein binding sites. *Curr. Opin. Struct. Biol.*, 17(1):67–76, 2007. doi: 10.1016/j.sbi.2007.01.004.

Liam J. Revell. phytools: an r package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, 3(2):217–223, 2012.

Liam J. Revell, Luke J. Harmon, R. Brian Langerhans, and Jason J. Kolbe. A phylogenetic approach to determining the importance of constraint on phenotypic evolution in the neotropical lizard anolis cristatellus. *Evolutionary Ecology Research*, 9(2):261–282, 2007.

Derek A. Roff. The estimation of genetic correlations from phenotypic correlations: a test of cheverud's conjecture. *Heredity*, 74(5):481–490, 1995. doi: 10.1038/hdy.1995.68.

F. J. Rohlf. *Morphology, Shape and Phylogeny*, chapter Geometric Morphometrics and Phylogeny, pages 175–193. Taylor and Francis, New York, 2002.

F. James Rohlf and Dennis Slice. Extensions of the procrustes method for the optimal superimposition of landmarks. *Systematic Zoology*, 39(1):40–59, 1990. ISSN 00397989. doi: 10.2307/2992207. URL http://www.jstor.org/stable/2992207.

Marianne Rooman, Yves Dehouck, Jean Marc Kwasigroch, Christophe Biot, and Dimitri Gilis. What is paradoxical about levinthal paradox? *J. Biomol. Struct. Dyn.*, 20(3):327–9, 2002. doi: 10.1080/07391102.2002.10506850.

M. M. Rorick and G. P. Wagner. Protein structural modularity and robustness are associated with evolvability. *Genome biology and evolution*, 3:456, 2011.

Mary Rorick. Quantifying protein modularity and evolvability: A comparison of different techniques. *BioSystems*, 110(1):22–33, 2012. doi: 10.1016/j.biosystems.2012.06.006.

Daniel E Runcie and Sayan Mukherjee. Dissecting high-dimensional phenotypes with bayesian sparse factor analysis of genetic covariance matrices. *Genetics*, 194(3):753–67, 2013. doi: 10.1534/genetics.113.151217.

R Sathyapriya, M S Vijayabaskar, and Saraswathi Vishveshwara. Insights into protein-dna interactions through structure network analysis. *PLoS Comput. Biol.*, 4(9):e1000170, 2008. doi: 10.1371/journal.pcbi.1000170.

R Dustin Schaeffer, Amanda L Jonsson, Andrew M Simms, and Valerie Daggett. Generation of a consensus protein domain dictionary. *Bioinformatics*, 27(1):46–54, 2011. doi: 10.1093/bioinformatics/btq625.

N. P. Schafer, B. L. Kim, W. Zheng, and P. G. Wolynes. Learning to fold proteins using energy landscape theory. *arXiv preprint arXiv:1312.7283*, 2013.

Gerhard Schlosser and Günter P Wagner. A simple model of co-evolutionary dynamics caused by epistatic selection. *J. Theor. Biol.*, 250(1):48–65, 2008. doi: 10.1016/j.jtbi.2007.08.033.

Ida Schomburg, Antje Chang, Christian Ebeling, Marion Gremse, Christian Heldt, Gregor Huhn, and Dietmar Schomburg. Brenda, the enzyme database: updates and major new developments. *Nucleic acids research*, 32(suppl 1):D431–D433, 2004. doi: 10.1093/nar/gkh081.

Jörg Schultz, Frank Milpetz, Peer Bork, and Chris P. Ponting. Smart, a simple modular architecture research tool: identification of signaling domains. *Proceedings of the National Academy of Sciences*, 95(11):5857–5864, 1998.

Joost Schymkowitz, Jesper Borg, Francois Stricher, Robby Nys, Frederic Rousseau, and Luis Serrano. The foldx web server: an online force field. *Nucleic Acids Research*, 33(suppl 2):W382–W388, 2005.

Jan Seebacher and Anne-Claude Gavin. Snapshot: Protein-protein interaction networks. *Cell*, 144(6):1000, 2011.

Florence Servant, Catherine Bru, Sébastien Carrère, Emmanuel Courcelle, Jérôme Gouzy, David Peyruc, and Daniel Kahn. Prodom: Automated clustering of homologous domains. *Briefings in Bioinformatics*, 3(3):246–251, 2002. doi: 10.1093/bib/3.3.246.

S. B. Shah and N. V. Sahinidis. Sas-pro: Simultaneous residue assignment and structure superposition for protein structure alignment. *PloS one*, 7(5):–37493, 2012. doi: 10.1371/journal.pone.0037493.

Yosi Shibberu, Mark Brandt, and Allen Holder. Fundamentals of protein structure alignment. In Yi Pan, Jianxin Wang, and Min Li, editors, *Algorithmic and AI Methods for Protein Bioinformatics*, Wiley Book Series on Bioinformatics. Wiley, 2012.

Tsuyoshi Shimi, Tohru Kobayashi, Susumu Ito, and Koki Horikoshi. *Protein Adaptation in Extremophiles*, chapter Alkaline Adaptation of Proteins, pages 105–141. Nova Publishers, 2008.

Tsuyoshi Shirai, Kazuaki Igarashi, Tadahiro Ozawa, Hiroshi Hagihara, Tohru Kobayashi, Katsuya Ozaki, and Susumu Ito. Ancestral sequence evolutionary trace and crystal structure analyses of alkaline α-amylase from bacillus sp. ksm-1378 to clarify the alkaline adaptation process of proteins. *Proteins: Structure, Function, and Bioinformatics*, 66(3):600–610, 2007. doi: 10.1002/prot.21255.

Tsuyoshi Shirai, Vo Si Hung, Katsuhito Morinaka, Tohru Kobayashi, and Susumu Ito. Crystal structure of gh13 α-glucosidase gsj from one of the deepest sea bacteria. *Proteins: Structure, Function, and Bioinformatics*, 73(1):126–133, 2008. doi: 10.1002/prot.22044.

Abdur R Sikder and Albert Y Zomaya. Improving the performance of domaindiscovery of protein domain boundary assignment using inter-domain linker index. *BMC Bioinformatics*, 7 Suppl 5:S6, 2006. doi: 10.1186/1471-2105-7-S5-S6.

Daniele Silvestro, Anna Kostikova, Glenn Litsios, Peter B. Pearman, and Nicolas Salamin. Measurement errors should always be incorporated in phylogenetic comparative analysis. *Methods in Ecology and Evolution*, 6(3):340–346, 2015. doi: 10.1111/2041-210X.12337.

Jaehyun Sim, Seung-Yeon Kim, and Jooyoung Lee. Pprodo: prediction of protein domain boundaries using neural networks. *Proteins*, 59(3):627–32, 2005. doi: 10.1002/prot.20442.

Ramesh K Sistla, Brinda K V, and Saraswathi Vishveshwara. Identification of domains and domain interface residues in multidomain proteins from graph spectral method. *Proteins*, 59(3):616–26, 2005. doi: 10.1002/prot.20444.

Lars K. Skov, Osman Mirza, Anette Henriksen, Gabrielle Potocki De Montalk, Magali Remaud-Simeon, Patricia Sarçabal, Rene-Marc Willemot, Pierre Monsan, and Michael Gajhede. Amylosucrase, a glucan-synthesizing enzyme from the α-amylase family. *Journal of Biological Chemistry*, 276(27):25273–25278, 2001.

Dennis E. Slice. Geometric morphometrics. *Annual Review of Anthropology*, 36(1): 261–281, 2007. doi: 10.1146/annurev.anthro.34.081804.120613. URL http://www.annualreviews.org/doi/abs/10.1146/annurev.anthro.34.081804.120613.

A. B. Smith, Brian R. Cullis, and R. Thompson. The analysis of crop cultivar breeding and evaluation trials: an overview of current mixed model approaches. *The Journal of Agricultural Science*, 143(06):449–462, 2005. doi: 10.1017/S0021859605005587.

Yehoshua Sobolevsky, Zakharia M. Frenkel, and Edward N. Trifonov. Combinations of ancestral modules in proteins. *Journal of Molecular Evolution*, 65(6):640–650, 2007.

Ricard V. Solé and Sergi Valverde. Spontaneous emergence of modularity in cellular networks. *Journal of The Royal Society Interface*, 5(18):129–133, 2008. doi: 10.1098/rsif.2007.1108.

Daniel Sorensen and Daniel Gianola. *Likelihood, Bayesian and MCMC methods in quantitative genetics*. Springer, 2002.

Venkataramanan Soundararajan, Rahul Raman, S. Raguram, V. Sasisekharan, and Ram Sasisekharan. Atomic interaction networks in the core of protein domains and their native folds. *PLoS ONE*, 5(2):–9391, 2010. doi: 10.1371/journal.pone.0009391.

Alexandros Stamatakis. Raxml-vi-hpc: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688–2690, 2006.

S. J. Steppan, P. C. Philips, and D. Houle. Comparative quantitative genetics: evolution of the g matrix. *Trends in Ecology & Evolution*, 17(7):320–327, 2002.

Scott J. Steppan. Phylogenetic analysis of phenotypic covariance structure. i. contrasting results from matrix correlation and common principal component analysis. *Evolution*, 51(2):571–586, 1997. ISSN 00143820. URL http://www.jstor.org/stable/2411129.

Michael Sternberg, Syed Nabil Ali, Manuela Helmer-Citterich, Pier F. Gherardini, Keiran Fleming, Lawrence A. Kelley, and Mark N. Wass. Evolution of protein structure and function. *Comparative Biochemistry and Physiology - Part A: Molecular & Integrative Physiology*, 153(2, Supplement):–47, 2009. ISSN 1095-6433. doi: 10.1016/j.cbpa.2009.04.498. URL http://www.sciencedirect.com/science/article/pii/S1095643309001299.

Stefan Strobl, Klaus Maskos, Georg Wiegand, Robert Huber, F. Xavier Gomis-Rüth, and Rudi Glockshuber. A novel strategy for inhibition of α-amylases: yellow meal worm α-amylase in complex with the ragi bifunctional inhibitor at 2.5 ă resolution. *Structure*, 6(7):911–921, 1998.

Marius Sudol and Kieran F. Harvey. Modularity in the hippo signaling pathway. *Trends in biochemical sciences*, 35(11):627–633, 2010.

Mikita Suyama and Osamu Ohara. Domcut: prediction of inter-domain linker regions in amino acid sequences. *Bioinformatics*, 19(5):673–674, 2003. doi: 10.1093/bioinformatics/btg031.

B. Svensson. Protein engineering in the α-amylase family: catalytic mechanism, substrate specificity, and stability. *Plant molecular biology*, 25(2):141–157, 1994.

Birte Svensson and Štefan Janeček. Glycoside hydrolase family 13. http://www.cazypedia.org/index.php/Glycoside_Hydrolase_Family_13, June 2015. Accessed 20 June 2015.

Kazuhiro Takemoto and Suritalatu Borjigin. Metabolic network modularity in archaea depends on growth conditions. *PloS one*, 6(10):e25874, 2011.

Roman L. Tatusov, Michael Y. Galperin, Darren A. Natale, and Eugene V. Koonin. The cog database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic acids research*, 28(1):33–36, 2000. doi: 10.1093/nar/28.1.33.

Roman L. Tatusov, Darren A. Natale, Igor V. Garkavtsev, Tatiana A. Tatusova, Uma T. Shankavaram, Bachoti S. Rao, Boris Kiryutin, Michael Y. Galperin, Natalie D. Fedorova, and Eugene V. Koonin. The cog database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic acids research*, 29(1):22–28, 2001.

Ian W. Taylor and Jeffrey L. Wrana. Protein interaction networks in medicine and disease. *Proteomics*, 12(10):1706–1716, 2012.

Ian W. Taylor, Rune Linding, David Warde-Farley, Yongmei Liu, Catia Pesquita, Daniel Faria, Shelley Bull, Tony Pawson, Quaid Morris, and Jeffrey L. Wrana. Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nature biotechnology*, 27(2):199–204, 2009.

D. Thirumalai, Edward P. O'Brien, Greg Morrison, and Changbong Hyeon. Theoretical perspectives on protein folding. *Annu Rev Biophys*, 39:159–83, 2010. doi: 10.1146/annurev-biophys-051309-103835.

J. Thompson, D. Higgins, and T. Gibson. Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22:4673–4680, 1994. doi: 10.1093/nar/22.22.4673.

Robin Thompson. Estimation of quantitative genetic parameters. *Proceedings of the Royal Society B: Biological Sciences*, 275(1635):679–686, 2008.

Guido Tiana, Boris E. Shakhnovich, Nikolay V. Dokholyan, and Eugene I. Shakhnovich. Imprint of evolution on protein structures. *Proceedings of the National Academy of Sciences of the United States of America*, 101(9):2846–2851, 2004. doi: 10.1073/pnas.0306638101.

Alexandra Tonkova. Bacterial cyclodextrin glucanotransferase. *Enzyme and Microbial technology*, 22(8):678–686, 1998. doi: 10.1016/S0141-0229(97)00263-9.

Pamela V. Tran, Salil A. Lachke, and Rolf W. Stottmann. Toward a systems-level understanding of the hedgehog signaling pathway: defining the complex, robust, and fragile. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 5(1):83–100, 2013.

Edward N. Trifonov and Igor N. Berezovsky. Evolutionary aspects of protein structure and folding. *Current Opinion in Structural Biology*, 13(1):110–114, 2003. ISSN 0959-440X. doi: 10.1016/S0959-440X(03)00005-8. URL http://www.sciencedirect.com/science/article/pii/S0959440X03000058.

Johan P Turkenburg, A Marek Brzozowski, Allan Svendsen, Torben V Borchert, Gideon J Davies, and Keith S Wilson. Structure of a pullulanase from bacillus acidopullulyticus. *Proteins*, 76(2):516–9, 2009. doi: 10.1002/prot.22416.

Joshua R. Tyler, Dennis M. Wilkinson, and Bernardo A. Huberman. Email as spectroscopy: Automated discovery of community structure within organizations. Technical report, Hewlett Packard Labs, 2003.

Joost CM Uitdehaag, Renée Mosi, Kor H. Kalk, Bart A. van der Veen, Lubbert Dijkhuizen, Stephen G. Withers, and Bauke W. Dijkstra. X-ray structures along the reaction pathway of cyclodextrin glycosyltransferase elucidate catalysis in the α-amylase family. *Nature Structural & Molecular Biology*, 6(5):432–436, 1999. doi: 10.1038/8235.

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. URL http://www.stats.ox.ac.uk/pub/MASS4. ISBN 0-387-95457-0.

Stella Veretnik, Philip E Bourne, Nickolai N Alexandrov, and Ilya N Shindyalov. Toward consistent assignment of structural domains in proteins. *J. Mol. Biol.*, 339 (3):647–78, 2004. doi: 10.1016/j.jmb.2004.03.053.

Malene Bech Vester-Christensen, Maher Abou Hachem, Birte Svensson, and Anette Henriksen. Crystal structure of an essential enzyme in seed starch degradation: barley limit dextrinase in complex with cyclodextrins. *J. Mol. Biol.*, 403(5):739–50, 2010. doi: 10.1016/j.jmb.2010.09.031.

Saraswathi Vishveshwara, K. V. Brinda, and N. Kannan. Protein structure: insights from graph theory. *Journal of Theoretical and Computational Chemistry*, 1(01): 187–211, 2002. doi: 10.1142/S0219633602000117.

Saraswathi Vishveshwara, Amit Ghosh, and Priti Hansia. Intra and inter-molecular communications through protein structure network. *Current Protein and Peptide Science*, 10(2):146–160, 2009. doi: 10.2174/138920309787847590.

K Visuri and A M Klibanov. Enzymatic production of high fructose corn syrup (hfcs) containing 55% fructose in aqueous ethanol. *Biotechnol. Bioeng.*, 30(7): 917–20, 1987. doi: 10.1002/bit.260300715.

Christopher A. Voigt, Carlos Martinez, Zhen-Gang Wang, Stephen L. Mayo, and Frances H. Arnold. Protein building blocks preserved by recombination. *Nature Structural & Molecular Biology*, 9(7):553–558, 2002.

Noreen von Cramon-Taubadel, Brenda C. Frazier, and Marta Mirazon Lahr. The problem of assessing landmark error in geometric morphometrics: theory, methods, and modifications. *American journal of physical anthropology*, 134(1):24–35, 2007. doi: 10.1002/ajpa.20616.

A. Wagner. Robustness and evolvability: a paradox resolved. *Proceedings of the Royal Society B: Biological Sciences*, 275(1630):91–100, 2008.

Ivo Van Walle, Ignace Lasters, and Lode Wyns. Sabmark–-a benchmark for sequence alignment that covers the entire known fold space. *Bioinformatics*, 21(7):1267–1268, 2005. doi: 10.1093/bioinformatics/bth493.

Bruce Walsh and Mark W. Blows. Abundant genetic variation+ strong selection= multivariate genetic constraints: a geometric view of adaptation. *Annual Review of Ecology, Evolution, and Systematics*, 40:41–59, 2009. doi: 10.1146/annurev.ecolsys.110308.120232.

Qin Wang, Jinli Yan, and Xiaoqin Li. Protein fold recognition based on functional domain composition. *Comput Biol Chem*, 48:71–6, 2014. doi: 10.1016/j.compbiolchem.2013.12.001.

Hidemichi Watari, E. Joan Blanchette-Mackie, Nancy K. Dwyer, Jane M. Glick, Shutish Patel, Edward B. Neufeld, Roscoe O. Brady, Peter G. Pentchev, and Jerome F. Strauss. Niemann-pick c1 protein: obligatory roles for n-terminal domains and lysosomal targeting in cholesterol mobilization. *Proceedings of the National Academy of Sciences*, 96(3):805–810, 1999.

R. K. Wierenga. The tim-barrel fold: a versatile framework for efficient enzymes. *FEBS letters*, 492(3):193–198, 2001. doi: 10.1016/S0014-5793(01)02236-0.

Angela Wilkins, Serkan Erdin, Rhonald Lua, and Olivier Lichtarge. Evolutionary trace for prediction and redesign of protein functional sites. In Riccardo Baron, editor, *Computational Drug Discovery and Design*, pages 29–42. Springer, 2012.

Alastair J. Wilson, Denis Reale, Michelle N. Clements, Michael M. Morrissey, Erik Postma, Craig A. Walling, Loeske EB Kruuk, and Daniel H. Nussey. An ecologist's guide to the animal model. *Journal of Animal Ecology*, 79(1):13–26, 2010.

Inken Wohlers, Noël Malod-Dognin, Rumen Andonov, and Gunnar W. Klau. Csa: comprehensive comparison of pairwise protein structure alignments. *Nucleic Acids Research*, 40(W1):–303, 2012. doi: 10.1093/nar/gks362. URL http://nar.oxfordjournals.org/content/40/W1/W303.abstract.

P G Wolynes. Recent successes of the energy landscape theory of protein folding and function. *Q. Rev. Biophys.*, 38(4):405–10, 2005. doi: 10.1017/S0033583505004075.

Peter E. Wright and H. Jane Dyson. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *Journal of molecular biology*, 293(2): 321–331, 1999. doi: 10.1006/jmbi.1999.3110.

Cathy H Wu, Anastasia Nikolskaya, Hongzhan Huang, Lai-Su L Yeh, Darren A Natale, C R Vinayaka, Zhang-Zhi Hu, Raja Mazumder, Sandeep Kumar, Panagiotis Kourtesis, Robert S Ledley, Baris E Suzek, Leslie Arminski, Yongxing Chen, Jian Zhang, Jorge Louie Cardenas, Sehee Chung, Jorge Castro-Alvear, Georgi Dinkov, and Winona C Barker. Pirsf: family classification system at the protein information resource. *Nucleic Acids Res.*, 32(Database issue):D112–4, 2004. doi: 10.1093/nar/gkh097.

Yu Xia and Michael Levitt. Simulating protein evolution in sequence and structure space. *Curr. Opin. Struct. Biol.*, 14(2):202–7, 2004. doi: 10.1016/j.sbi.2004.03.001.

Y. Xing and C. J. Lee. Protein modularity of alternatively spliced exons is associated with tissue-specific regulation of alternative splicing. *PLoS genetics*, 1(3):e34, 2005.

Ying Xu, Dong Xu, and Harold N. Gabow. Protein domain decomposition using a graph-theoretic approach. *Bioinformatics*, 16(12):1091–1104, 2000. doi: 10.1093/bioinformatics/16.12.1091.

Zhidong Xue, Dong Xu, Yan Wang, and Yang Zhang. Threadom: extracting protein domain boundary information from multiple threading alignments. *Bioinformatics*, 29(13):i247–56, 2013. doi: 10.1093/bioinformatics/btt209.

Hari Krishna Yalamanchili and Nita Parekh. Graph spectral approach for identifying protein domains. In *Bioinformatics and Computational Biology*, pages 437–448. Springer, 2009. doi: 10.1007/978-3-642-00727-9_40.

Takuji Yamada and Peer Bork. Evolution of biomolecular networks—lessons from metabolic and protein interactions. *Nature Reviews Molecular Cell Biology*, 10 (11):791–803, 2009.

Keizo Yamamoto, Hideo Miyake, Masami Kusunoki, and Shigeyoshi Osaki. Steric hindrance by 2 amino acid residues determines the substrate specificity of isomaltase from *Saccharomyces cerevisiae*. *Journal of bioscience and bioengineering*, 112 (6):545–550, 2011. doi: 10.1016/j.jbiosc.2011.08.016.

Kavestri Yegambaram, Esther M M Bulloch, and Richard L Kingston. Protein domain definition should allow for conditional disorder. *Protein Sci.*, 22(11):1502–18, 2013. doi: 10.1002/pro.2336.

Takehiro Yokota, Takashi Tonozuka, Yoichiro Shimura, Kazuhiro Ichikawa, Shigehiro Kamitori, and Yoshiyuki Sakano. Structures of thermoactinomyces vulgaris r-47 α-amylase ii complexed with substrate analogues. *Bioscience, biotechnology, and biochemistry*, 65(3):619–626, 2001. doi: 10.1271/bbb.65.619.

P D Yoo, A R Sikder, J Taheri, B B Zhou, and A Y Zomaya. Domnet: protein domain boundary prediction using enhanced general regression network and new profiles. *IEEE Trans Nanobioscience*, 7(2):172–81, 2008. doi: 10.1109/TNB.2008.2000747.

Chao Yuan, Hao Chen, and Daisuke Kihara. Effective inter-residue contact definitions for accurate protein fold recognition. *BMC Bioinformatics*, 13(1):292, 2012. doi: 10.1186/1471-2105-13-292.

M. L. Zelditch, D. L. Swiderski, H. D. Sheets, and W. L. Fink. *Geometric morphometrics for biologists: A primer*. Elsevier Academic Press, New York, 2004.

Daohai Zhang, Nan Li, Shee-Mei Lok, Lian-Hui Zhang, and Kunchithapadam Swaminathan. Isomaltulose synthase (pali) of klebsiella sp. lx3 crystal structure and implication of mechanism. *Journal of Biological Chemistry*, 278(37):35428–35434, 2003.

Wanding Zhou and Luay Nakhleh. Convergent evolution of modularity in metabolic networks through different community structures. *BMC Evolutionary Biology*, 12 (1):181, 2012.

Robert Zwanzig, Attila Szabo, and Biman Bagchi. Levinthal's paradox. *Proceedings of the National Academy of Sciences of the United States of America, Volume 89, Issue 1, pp. 20-22*, 89:20–22, jan 1992. doi: 10.1073/pnas.89.1.20.