GWAS of CMR-derived LV morphology represented via graph-convolutional autoencoders in 30k UK Biobank subjects

Rodrigo Bonazzola[1], Andres Diaz-Pinto[1], Rahman Attar[1], Nishant Ravikumar[1], Eylem Levelt[2], Enzo Ferrante[3], Tanveer Syeda-Mahmood[4], and Alejandro F Frangi[1,2]

[1] Centre for Computational Imaging and Simulation Technologies in Biomedicine (CISTIB), School of Computing, University of Leeds, Leeds, UK r.bonazzola1@leeds.ac.uk
[2] Center for Computational Imaging and Simulation Technologies in Biomedicine, School of Medicine, University of Leeds, Leeds, UK
[3] Research Institute for Signals, Systems and Computational Intelligence, sinc(i), FICH-UNL / CONICET, Santa Fe, Argentina
[4] IBM Almaden Research Center, San Jose, USA

Abstract. The genetic basis of cardiac phenotypes is not yet well understood. Cardiac magnetic resonance (CMR) images, a rich source of structural phenotypes, as well as linked genotype information, were recently made available within the UK Biobank project. Subsequently, a genome-wide association study (GWAS) for left-ventricular (LV) function indices was published using this data, reporting novel genetic loci significantly associated with these phenotypes. However, LV morphological features are yet to be studied in this context. In this work, we performed dimensionality reduction on 3D LV meshes generated from CMR images, which allows us to learn non-local morphological features that constitute a latent basis for the shape. To this end, we used two approaches i) principal component analysis (PCA) and ii) graph-convolutional autoencoders. GWAS was performed on the components of the latent space obtained through each of the methods. Novel genetic loci were discovered which have not been reported in literature, in addition to other loci already known to be associated with cardiac phenotypes. Some of these features were found to be related to cardiac diseases, showing the potential of our approach to pinpoint genetic loci of clinical relevance.

Keywords: Cardiac MR · GWAS · graph-convolutional autoencoders.

1 Introduction

Highlight the importance of cardiovascular diseases in worldwide mortality.

The recent advances in cardiovascular imaging has allowed for a more accurate phenotyping of the heart, due to the increasing availability of high resolution images from different techniques. Such techniques include echocardiography, nuclear imaging, computerized tomography (CT) and cardiac magnetic resonance (CMR), and are used both in research and in clinical practice. Some of these image-derived phenotypes are known to have underlying genetic factors, and also to be associated with cardiovascular risk.

The UK Biobank is a prospective cohort study that between 2006 and 2010 recruited around half a million volunteers aged 40-69 years old, across the United Kingdom []. This sample aims to be representative of the whole UK population. The project collected a huge amount of phenotypic information about its participants, and also linked them to their electronic health records (EHR). The collected data include, among others, genetic data from SNP microarrays for all the individuals, and also CMR data for a subset of the participants.

On the other hand, genome-wide association studies (GWAS) have been very successful in identifying genetic variants associated with a broad range of phenotypes. To the best of our knowledge, only two studies have been published so far in the field of cardiac imaging genetics using CMR data. However, the breadth of image-derived cardiovascular phenotypes studied in this context has been hitherto limited to those with known clinical relevance. These include the volumes of the different cardiac chambers, parameteres related to the function of the heart (such as stroke volume, cardiac output and ejection fraction of the chambers) and myocardial thickness and mass.

The first one [**?**] studies left-ventricular wall thickness at end-diastole, performing and association test with a set of genetic variants in a vertex-by-vertex fashion.

The second one uses UK Biobank data and studies 6 left-ventricular phenotypes: 1) end-diastolic volume (LVEDV), 2) end-systolic volume (LVESV), 3) stroke volume (LVSV), 4) ejection fraction (LVEF), 5) mass (LVM) and 6) mass-to-EDV ratio (LVMVR).

The unprecedented amount of linked genetic and cardiac imaging data available within the UK Biobank allows for a different kind of approach: instead of extracting handcrafted features from the images (such as the ones mentioned above), techniques of unsupervised machine learning can be employed in order to automatically learn a set of features that best describe the morphology of the heart. The hypothesis is that these learnt features, by virtue of its greater variability across the population, will be good candidate phenotypes on which to perform genetic association analysis.

In this paper, we implement this approach. In particular, we use two different dimensionality reduction techniques:

MENTION PREVIOUS STUDIES WITH BOTH KINDS OF DATA (FRAMINGHAM, MESA?) AND COMPARE THE SAMPLE SIZES.

2 Methods

2.1 Description of the data

The data used for this work comes from the UK Biobank project, data accession number —.

Cardiovascular Magnetic Resonance (CMR) data. (There is no need to describe the temporal aspect of the data as I'm only using end-diastole) The CMR imaging protocol used to obtain the raw imaging data is described elsewhere **??**. The cardiac segmentation algorithm is described in detail in [1].

Each cardiac mesh, $\mathcal{G}_i^{(t)}$, consists of a set of vertices and edges $\mathcal{G}_i^{(t)} = \{\mathcal{V}_i^{(t)}, \mathcal{E}_i^{(t)}\}$ describing the LV shape for individual i at instant $t \in [0, 1)$ of the cardiac cycle, where the cardiac period is normalized to 1 for each individual and $t = 0$ is the end-diastolic phase (whereas end-systole varies across individuals but lies typically in the range ??? 18-19/50). Notice that in order to convert from these normalized time units to real time units, the pulse rate needs to be used.

Each element of $\mathcal{V}_i$ is a 3D point $\mathbf{x}_{ij}^{(t)} = (x_{ij}^{(t)}, y_{ij}^{(t)}, z_{ij}^{(t)})$, $j = 1, ..., M$, $i = 1, ..., N$, and $(\mathbf{x}_{ij_1}^{(t)}, \mathbf{x}_{ij_2}^{(t)}) \in \mathcal{V}_i^{(t)}$ means that there is an edge between vertices j_1 and j_2 in the cardiac mesh of subject i at time t.

In our particular case, this becomes significantly simpler since the connectivity of the meshes is inherited from the connectivity of a reference shape (atlas), and thus all the meshes have not only the same number of vertices but also the same connectivity, both across individuals and across phases; i.e. $M = |\mathcal{V}_k^{(t_1)}| = |\mathcal{V}_l^{(t_2)}|$ and $\mathcal{E} := \mathcal{E}_k^{(t_1)} = \mathcal{E}_l^{(t_2)}$, $\forall k, l, t_1, t_2$. [5]

An alternative way to express the connectivity of the mesh is through the adjacency matrix $A \in \{0, 1\}^{M \times M}$, where $A_{ij} = 1$ if vertex i is connected to vertex j, and is zero otherwise. This way of representing the connectivity is useful in the context of spectral graph theory, as will be explained in more detail in subsection ??.

Genotypic data SNP microarray data is available for all the individuals in the UK Biobank cohort. This microarray covers 801526genetic variants including SNPs and short indels. The design of this microarray has been described in detail in .

From these genotyped markers, an augmented set of 10000000variants was imputed. The GWAS was performed across this latter set, filtering by a minor allele frequency (MAF) threshold of 1% and a HWE p-value threshold of 1e-4.

2.2 Segmentation algorithm

2.3 Dimensionality reduction

In order to extract a reduced set of features that describe left-ventricular shape, two methods were utilised: principal component analysis (PCA) and mesh-convolutional autoencoders. In the first case only 3D point clouds are provided as input, whereas in the second case we also leverage information about connectivity between the vertices. However, both approaches can be thought of as special cases of a common paradigm.

In such paradigm, there is a pair of encoding and decoding functions, $E_{\theta_E} : \mathbb{R}^{3M} \to \mathbb{R}^\gamma$ and $D_{\theta_D} : \mathbb{R}^\gamma \to \mathbb{R}^{3M}$ that are parameterized by a set of parameters θ_E and θ_D, respectively. γ is usually chosen so that $\gamma \ll M$ (hence the dimensionality reduction).

The optimal parameters θ_E^* and θ_D^* for reconstruction can be learnt by making the composite function $D_{\theta_D} \circ E_{\theta_E}$ as close to the identity function I as possible for the data in the training set, using some reasonable loss function such as the squared L_2 norm of the element-wise difference; that is to say

$$L(\mathbf{s}|\theta_E, \theta_D) = \sum_{i=1}^{N} \left|\left|\big(D_{\theta_D} \circ E_{\theta_E} - I\big)(\mathbf{s}_i)\right|\right|_2^2 \quad (1)$$

$$(\theta_E^*, \theta_D^*) = \text{argmin}_{\theta_E, \theta_D} L(\mathbf{s}|\theta_E, \theta_D). \quad (2)$$

By calling $\hat{\mathbf{s}}_i = \big(D_{\theta_D^*} \circ E_{\theta_E^*}\big)(\mathbf{s}_i)$

$E(\mathbf{s}_i) \in \mathbb{R}^\gamma$ would then be a low-dimensional representation of the shape $\mathbf{s}_i$.

Principal component analysis. PCA is a standard linear technique for dimensionality reduction. In terms of the framework detailed above, it can be obtained by requiring D and E to be linear transformations/mappings.

Given a set of 3D shapes $\mathbf{s}_i^{(t)} = (x_{i1}^{(t)}, y_{i1}^{(t)}, z_{i1}^{(t)}, ..., x_{iM}^{(t)}, y_{iM}^{(t)}, z_{iM}^{(t)}) \in \mathbb{R}^{3M}$, we derive the mean shape and the shape covariance matrix:

$$\bar{\mathbf{s}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{s}_i, \quad (3)$$

$$\mathbf{S} = \frac{1}{N-1} \sum_{i=1}^{N} (\mathbf{s}_i - \bar{\mathbf{s}})(\mathbf{s}_i - \bar{\mathbf{s}})^t. \quad (4)$$

The idea is to find a basis of vectors $\mathcal{B} = \{\}_{i=1}^K$ for a fixed $K < 3M$, such that the K-dimensional linear subspace generated by $\mathcal{B}$ captures as much of the shape variability in LV shape as possible. It can be shown that such basis corresponds to the K eigenvectors of the $\mathbf{S}$ matrix with the top largest eigenvalues; i.e. if $\mathbf{S} = \Omega^t \Lambda \Omega$ where $\Lambda = \delta_{ij}\lambda_i$ and $\lambda_i \geq \lambda_j$ if $i \leq j$, then $\mathcal{B} = \{\Omega \mathbf{e}_i\}_{i=1}^K$.

Convolutional mesh-AE. In order to incorporate/leverage information about the topology of the mesh, a convolutional approach was used. The encoder E consists of convolutional and pooling layers, whereas D consists of unpooling layers. Since the vertices are not in a rectangular grid, the usual convolution, pooling and unpooling operations defined for such geometry are not adequate for this task and need to be suitably adapted/generalized. There is a number of methods to do this, but they all can be classified into two large groups: spatial or spectral . In this work we'll apply a method belonging to the latter category, which relies on expressing the features in the Fourier basis of the graph, as will be explained below.

The Laplace-Beltrami operator of a graph with adjacency matrix A is defined as

$$\mathcal{L} = D - A, \quad (5)$$

where D is the degree matrix, i.e. a diagonal matrix where $D_{ii} = \sum_j A_{ij}$ is the number of edges that connect to vertex i. The Fourier basis of the graph can be obtained by diagonalizing the Laplace operator, $\mathcal{L} = U^t \Lambda U$. The columns of U constitute the Fourier basis, and the operation of convolution $\star$ for a graph can be defined in the following manner

$$x \star y = U(U^t x \odot U^t y) \quad (6)$$

where $\odot$ is the element-wise product (also known as Hadamard product) [6]

All spectral methods for convolution rely on this definition of convolution, and differ from one another in the form of the kernel/filter utilized. In this work, a parameterization proposed in will be used. The said method is based on the Chebyshev polynomials $\{T_i\}$. The kernel g_ξ is defined as

$$g_\xi = \sum_{i=1}^{K} \xi_i T_i. \quad (7)$$

[5] CLARIFICATION (This could go to supplementary material): in more rigor, the endocardial and epicardial meshes are derived each from a different atlas and then a merging procedure is applied in order to obtain a single mesh for both walls. Notice that since the whole shape does not come from one single atlas, there is no guarantee that the connectivity of the resulting mesh will be the same for different individuals (it depends on the nature of the merging procedure selected, and if it is based on distances between vertices, it could certainly not be the case since the neighboring vertices in different meshes). To overcome this inconvenient situation, the merging procedure is applied to a single mesh, and the resulting connectivity is applied to the rest of the meshes.

[6] Can $\mathbf{S}$ (the covariance matrix) be thought of as the adjacency matrix of a weighted graph and thus the principal components are just the elements of the Fourier basis of this weighted graph?

Chebyshev polynomials have the advantage that they can be computed recursively through the relation $T_i(x) = xT_{i-1}(x) - T_{i-2}(x)$ and the base cases $T_1(x) = 1$ and $T_2(x) = x$.

The pooling and unpooling operations used in this work were the ones proposed in cite**??**.

2.4 GWAS

According to the traditional GWAS scheme, we tested each genetic variant, $X_i \in \{0, 1, 2\}$, for association with each of the LV latent features Z_j through a univariate linear model:

$$Z_j = \beta_{ij} X_i + \epsilon_{ij} \tag{8}$$

where ϵ_{ij} is the component not explained by the genotype, which we assumed to be normally distributed. The null hypothesis tested is that $\beta_{ij} = 0$.

2.5 Implementation details

The architecture of the autoencoder is explained in table **??**.

All of the executions were performed using Amazon Web Services EC2 virtual machines.

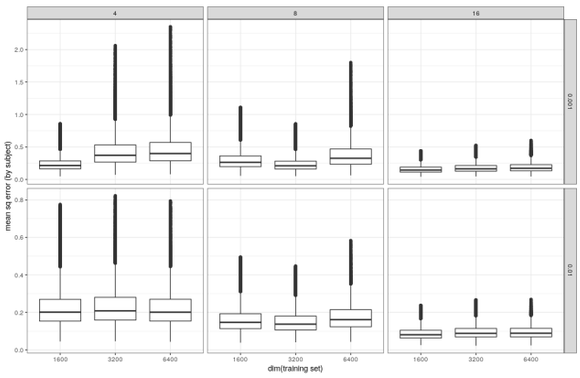
3 Results

Statements:

1. The convolutional mesh autoencoder learns a latent basis that allows to reconstruct the cardiac meshes with a lesser amount of components compared to PCA.
2. The components of the latent basis have easy visual interpretations.
3. Some of the components have genetic loci associated.
4.

3.1 Comparison PCA vs. mesh-AE

Plot reconstruction error.



3.2 Morphological interpretation of the latent basis

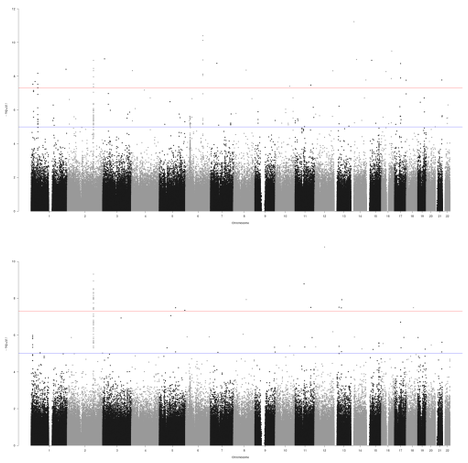
PCA Plots of the meshes obtained by varying 1 of the codes each time.

Mesh AE Plots of the meshes obtained by varying the codes.

3.3 GWAS

Manhattan plots.

THESE FIGURES ARE ONLY PLACEHOLDERS



4 Discussion

5 Conclusions

One possible future direction of this work is to consider spatio-temporal patterns by using all the phases of the cardiac cycle.

References

1. Rahman Attar, Marco Pereaez, Ali Gooya, Xnia Alb, Le Zhang, Milton Hoz de Vila, Aaron M Lee, Nay Aung, Elena Lukaschuk, Mihir M Sanghvi, Kenneth Fung, Jose Miguel Paiva, Stefan K Piechnik, Stefan Neubauer, Steffen E Petersen, Alejandro F Frangi. Quantitative CMR population imaging on 20,000 subjects of the UK Biobank imaging study: LV/RV quantification pipeline and its evaluation. Medical Image Analysis. Volume 56, August 2019, Pages 26-42.
2. Author, F.: Article title. Journal **2**(5), 99–110 (2016)
3. Author, F., Author, S.: Title of a proceedings paper. In: Editor, F., Editor, S. (eds.) CONFERENCE 2016, LNCS, vol. 9999, pp. 1–13. Springer, Heidelberg (2016). https://doi.org/10.10007/1234567890
4. Author, F., Author, S., Author, T.: Book title. 2nd edn. Publisher, Location (1999)
5. Author, A.-B.: Contribution title. In: 9th International Proceedings on Proceedings, pp. 1–2. Publisher, Location (2010)
6. LNCS Homepage, http://www.springer.com/lncs. Last accessed 4 Oct 2017