2021 6th International Conference on Computational Intelligence and Applications (ICCIA) | 978-1-6654-3933-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICCIA52886.2021.00062

Multiscale network regression for associations between brain connectivity and cognitive and behavioural indices

1st Izaro Fernandez-Iriondo
Computer Science and Artificial Intelligence
University of the Basque Country (UPV/EHU)
San Sebastián, Spain
izaro.fernandez@ehu.eus

2nd Basilio Sierra
Computer Science and Artificial Intelligence
University of the Basque Country (UPV/EHU)
San Sebastián, Spain
b.sierra@ehu.eus

3rd José María Martínez-Otzeta
Computer Science and Artificial Intelligence
University of the Basque Country (UPV/EHU)
San Sebastián, Spain
josemaria.martinezo@ehu.eus

4th Antonio Jimenez-Marín
Computational Neuroimaging Lab
BioCruces-Bizkaia Health Research Institute
Barakaldo, Spain
antonio93jm@gmail.com

5th Paolo Bonifazi
Computational Neuroimaging Lab
BioCruces-Bizkaia Health Research Institute
Barakaldo, Spain
paol.bonifazi@gmail.com

6th Yosu Yurramendi*
Computer Science and Artificial Intelligence
University of the Basque Country (UPV/EHU)
San Sebastián, Spain
yosu.yurramendi@ehu.eus
* *Equal last-author contribution*

7th Jesus M. Cortés*
Computational Neuroimaging Lab
BioCruces-Bizkaia Health Research Institute
Barakaldo, Spain
jesus.m.cortes@gmail.com
* *Equal last-author contribution*

***Abstract*—The study of the relationship between brain connectivity and cognitive abilities is of great importance for a better understanding of the human mind. We propose to tackle this issue with the help of several connectivity measures, assessing the strength of their association with the subjects' performance in some standard neuropsychological tests. The novelty of this work is the use of these connectivity metrics outside the field were they originated, which is their association with the subjects' chronological age. The results obtained from an experiment with data from the Human Connectome Project show positive correlations between the proposed connectivity measures and the aforementioned neuropsychological indices. The outcome of this approach yields correlation values which compare favourably with state-of-the-art research and show that those measures could be relevant across several fields of study.**

***Keywords*—MRI, DTI, multi-scale, multi-modal, machine learning, regression**

I. Introduction

It is well-known that behavioural variability could be explained more accurately by general variability in the structure and/or function of the brain than just by taking into account isolated traits [1].

In [2] the authors explore this association by means of the Canonical Correlation Analysis (CCA), after a previous dimensionality reduction with Principal Component Analysis (PCA). On the other hand, in [3] the analysis of the latent relationship between the two sets of variables (connectivity and cognitive performance) is performed by Linked Independent Component Analysis (Linked ICA). This was the first study to discover a strong evidence of relationship between structural inter-individual variants and a wide spectrum of demographic and behavioral variants.

Those two works point to a strong evidence for a relationship between brain connectivity (functional [2] or structural [3]) and behavioral and cognitive measures. Nevertheless, when working with magnetic resonance imagery, widely used nowadays, the scale at which represent the data turns to be a crucial experimental design decision [4], [5]. In the finest detail or micro-scale the network nodes are neurons, while

978-1-6654-3933-6/21/$31.00 ©2021 IEEE
DOI 10.1109/ICCIA52886.2021.00062

in the meso-scale and the macro-scale the nodes would be neuron populations of different sizes. In this work we have focused on the macro-scale, but with a multi-scale approach inside this layer.

The measures of connectivity applied in this research are those defined in [4]. Our main contribution is the extension of their work beyond their original area (association with chronological age) to the association with neuropsychological indices. Our work shows that those measures provide results that outperform the research described in [2].

Further work could point to the study of pathological populations, which could be of great interest due to the changes in structural and functional networks in a resting state [6], [7].

II. Material and Methods

A. Participants

In this paper we have used open access data from the Human Connectome Project (HCP) database[1] [8], [9]. We have studied data from N=1,000 healthy adult subjects (ages 22-37 years with a mean of 28.68 years and standard deviation of 3.69 years) of which 536 are women and 464 men. Each of the participants has undergone an MRI scan and responded to different neuropsychological tests to obtain brain data (structural and functional) and scores of mental health and cognition.

B. Data acquisition

Neuropsychological Tests: The chosen neuropsychological indices[2] that we will predict will be: Picture Vocabulary Test (age-adjusted and unadjusted), Penn Matrix Analysis Test (number of correct answers and number of skipped items) and the Delay Discounting Task (area under the ROC curve).

Magnetic Resonance Imaging: The acquisition parameters of the resting-state functional neuroimaging and diffusion neuroimaging have been acquired from the standard protocols[3] of the HCP database [10].

C. Image Preprocessing

The MRI data has been preprocessed according to the standard practice[4] of the Human Connectome Project [11].

D. Brain network nodes definition

The nodes of the structural and functional brain networks have been calculated using an unsupervised clustering of the voxels of the functional images as described in [12]. This has produced M = 2,308 regions of interest (ROIs), which, along with their mutual connections, will constitute the brain networks.

[1]Database available at: http://www.humanconnectomeproject.org/data

[2]A detailed description of each is available at https://wiki.humanconnectome.org/display/PublicData/HCP+Data+Dictionary+Public-+Updated+for+the+1200+Subject+Release.

[3]Data acquisition parameters: https://protocols.humanconnectome.org/HCP/3T/imaging-protocols.html.

[4]Detailed description available at:https://humanconnectome.org/storage/app/media/documentation/s1200/HCP_S1200_Release_Reference_Manual.pdf

E. Definition of the levels in the modular-hierarchical structure

From the partition obtained with the clustering procedure explained in the previous section we build a hierarchical tree with an agglomerative clustering. Therefore, we obtain a tree with the union of all the clusters in the root and M = 2,308 leaves in the bottom, which corresponds to all the detected clusters. The difference between one level in the tree and the next higher level is the fusion of two clusters on one, as determined by the agglomerative clustering algorithm. Therefore, there are M levels, each one corresponding to a different partition, and the partition sizes range from 1 (root) to M (leaves).

Depending on the size of the partition we choose, we could analyze the brain connectivity in a different scale. In this research we have focused in the partitions with sizes ranging from 20 to 1,000. The choice of this range is justified by, on the upper limit, the high computational cost of analysing lower levels of the hierarchical tree, with partition sizes greater than 1,000; and, on the lower limit, the insufficient spatial resolution to detect modules in small partitions (size < 20).

F. Calculation of multi-level brain features

After preprocessing the data and calculating the network nodes, we computed N=1,000 structural (SC) and functional (FC) connectivity matrices, one of both for each subject in the study, following [4]. On the one hand, the structural connectivity matrices for each subject will be constructed by counting the number of fibres between pairs of regions; and on the other hand, the functional connectivity matrices will be calculated by computing the Pearson correlation between pairs of time series of regions. Thus, the resulting two-dimensional matrices representing structural (SC) and functional (FC) connectivity between regions will be weighted and undirected.

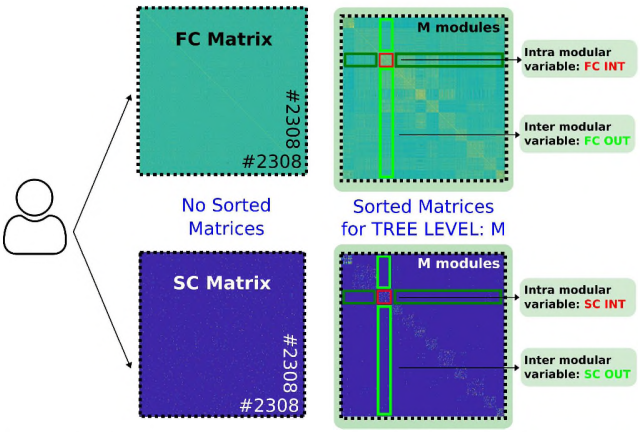

Fig. 1: **Extraction of SC and FC variables taking into account the hierarchical modular structure.** It shows how the 4 types of variables (SCINT, SCOUT, FCINT and FCOUT) have been extracted for each single module and for each participant from the two-dimensional connectivity matrices.

As we can see in the left side of figure 1, we will have one SC and one FC matrix of dimensions 2308 x 2308 regions for each subject. This matrix has been reordered according to the clustering resulting from each level of the hierarchical tree. Thus, when we are at level M of the tree our SC and FC matrices will have M sub-matrices (red square in figure 1) on the main diagonal where the connectivity values inside each of them will be higher than outside (green rectangle in figure 1).

For each of the levels of the hierarchical tree ($M = [20, 1000]$) and each of the modules present in the connectivity matrix corresponding to that level, we will extract four types of variables: internal (SCINT, red rectangle in figure 1) and external (SCOUT, green rectangle in figure 1) structural connectivity variables; and internal (FCINT, red rectangle in figure 1) and external (FCINT, green rectangle in figure 1) functional connectivity variables.

From one level to the next only two modules are merged to create a new one, while the rest of the modules remain unchanged. Therefore a lot of modules show up many times along the tree. We only keep the variables belonging to different modules, so only one out of the many possible multiple appearances of a module is taken into account. As there are 981 levels, in the first level there are 1,000 clusters, and only one new cluster is created between levels, there are a total of 1,000 + 980 = 1,980 different modules. Therefore, we are going to explore 1,980 * 4 = 7,920 variables.

G. Data processing

1) Missing values, exceptional variables and normalization: Once we have defined the variables we are going to study, we have seen different scales of measuring these variables, missing values and exceptional situations where some values cannot be computed. Regarding the different units of measurement our variables have been scaled and centred prior to any further analysis. In some exceptional situations the internal structural and functional connectivity could not been computed because those modules correspond to very small brain regions (1 ROI). This limit case amounts to 645 modules, and affect to 645 * 2 = 1,290 variables which invalid value. Therefore we are left with 7,920 - 1,290 = 6,630 variables Half of them (3,315) are related to functional connectivity, and the other half (3,315) to structural connectivity. As for the missing values, those related to brain connectivity have been ignored (38 for structural connectivity) and those related to neuropsychological indices have been imputed using the K-nearest neighbours technique.

2) Splitting training and testing data: Our aim has been to explore the predictive value of the SC variables, the FC variables and the union of both. Therefore, we have three different scenarios and in all of them we split the data into training and test with the proportion 80/20. A stratified split has been applied over each discretized target variable, which could be different for each neuropsychological measure.

H. Dimensionality reduction

We have used the minimum redundancy maximum relevance feature selection (mRMR) technique, which is particularly attractive for finding a set of relevant and low-redundant features, from which accurate prediction models can be built [13].

The mathematical formula to rank our variables has been $q_j = I(x_j, y) - \frac{1}{|S|} \sum_{x_k \in S} I(x_j, x_k)$ being y the variable to predict, $X = x_1, ...x_N$ the set of N input variables and S the set of selected variables; the method makes a ranking of X by maximizing I with y (maximum relevance) and minimizing the average of I with all previously selected variables (minimum redundancy).

I. Regression model

The regression model we have used for training has been the General Linear Model. We will train this model with the training data set and validate it with the test set.

III. Results

A. Filter Method Feature Selection

This filter has as input the final number of variables to which we want to reduce our dataset, let's call it k. By entering the parameter k and the dataset the filter returns the first k variables of the ranking it has made based on the redundancy and relevance of the variables.

To determine the optimal k, we have separated the data in training and testing according to the corresponding neuropyschological variable and then, with the training data set, we have made a search of the values of k. This search has been carried out for all the scores for the range $k = [2, 400]$ taking into account the computational cost involved and the reduction we want to apply, however, for some scores a larger range $k = [2, 800]$ has been necessary to see the trend of the curve.

For each value of k we trained the general linear regression model and then predicted the neuropsychological scores on the test data. Thus, we obtained a curve with the number of variables (k) on the x-axis and with the coefficient of determination R^2 we evaluated the prediction on the y-axis. The value of k that maximises the curve will be the optimum.

B. Regression Model Performance

Once we have determined the number of variables to which we will reduce the different data modalities in each neuropsychological score, we are going to keep only the best result, that is, the optimal k. Thus, for each modality and score we will have a Training Correlation, which will determine the relationship between the predicted and real score with the training data; and a Testing Correlation, which will determine the relationship between the predicted and real score with the testing data. The results are summarized for each data modality: structural (Table I), functional (Table II) and both (Table III).

C. Location of variables in the hierarchical tree

One of the contributions of this research is the proposal of a multiscale analysis, i.e., to extract brain connectivity variables not only from one level of the hierarchical tree but from a range (M=[20, 1000]) that we have defined following criteria of spatial resolution and computational cost.

Therefore, once we have fixed the modality (structure, function or a union of both) that has given us the best results in each neuropsychological score, we have represented from which level of the tree the selected variables have been extracted.

We have not found in any neuropsychological score a level of the tree where a large part of the selected variables are concentrated. Therefore, we have been able to see that spatially larger brain regions (near the M=20 level) as well as smaller regions (near M=1000) are equally necessary for our prediction.

TABLE I: Results of feature selection with the mRMR filter and prediction of neuropsychological scores using structural connectivity descriptors.

Score Name	Number of Predictors	Training Correlation	Testing Correlation
Picture Vocabulary Test (Age Adjusted)	50	0.5094	0.3506
Picture Vocabulary Test (Unadjusted)	98	0.5900	0.2761
PMAT (Correct Responses)	382	0.7160	0.1870
PMAT (Skipped Items)	250	0.6221	0.2330
Delay Discounting Task	62	0.5336	0.1242

TABLE II: Results of feature selection with the mRMR filter and prediction of neuropsychological scores using functional connectivity descriptors.

Score Name	Number of Predictors	Training Correlation	Testing Correlation
Picture Vocabulary Test (Age Adjusted)	490	0.8229	0.2155
Picture Vocabulary Test (Unadjusted)	438	0.8055	0.1367
PMAT (Correct Responses)	30	0.4315	0.2526
PMAT (Skipped Items)	54	0.4467	0.2204
Delay Discounting Task	782	0.9930	0.1728

TABLE III: Results of feature selection with the mRMR filter and prediction of neuropsychological scores using structural and functional connectivity descriptors.

Score Name	Number of Predictors	Training Correlation	Testing Correlation
Picture Vocabulary Test (Age Adjusted)	94	0.5861	0.3400
Picture Vocabulary Test (Unadjusted)	290	0.7675	0.2467
PMAT (Correct Responses)	62	0.5458	0.2106
PMAT (Skipped Items)	22	0.4585	0.2492
Delay Discounting Task	126	0.6639	0.1585

D. Neuroanatomical representation of results

The methodology we have used allows us to identify the anatomical areas that constitute the brain connectivity variables selected in each case. Therefore, we are going to see test by test the percentage of overlap (in decreasing order and greater than 5%) that our selected variables have with the anatomical areas of the Automated Anatomical Labeling atlas, AAL [14] and on the other hand, we will also show with which networks our selected variables have greater overlap so that, knowing which networks and which test we are dealing with, we can look for a neuroscientific interpretation of the results.

For example, the underlying brain circuits associated with the age-adjusted picture vocabulary test score have a percentage overlap with the Cerebellum of 14.327%, Temporal Mid 10.357%, Precuneus 10.062%, Parietal Up 6.0568% and Temporal Pole Up 5.9672%. On the other hand, these circuits have a higher percentage of overlap with the Default Mode Network (DMN) (linked to the comprehension and recall of narratives [15]) and the Limbic Network (related to people's memory and attention).

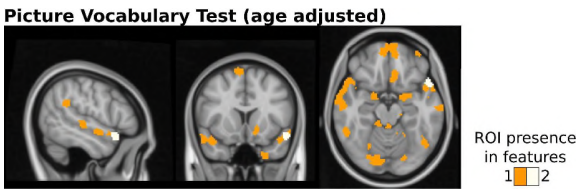

Fig. 2: Anatomical localization of brain regions whose connectivity profiles are linked to the association with the results of the age-adjusted picture vocabulary test.

In conclusion, the neuropsychological tests that the participants have completed are very heterogeneous where attention, hearing, comprehension, mental speed, vision, memory and other cognitive functions have to be alert. Therefore, as it might be expected, figure 2 does not show large anatomical areas but are scattered throughout the brain overlapping with

different anatomical areas that are related to different networks. However, we can say that attention networks are more predominant in the results of the analysis than networks of primary cognitive functions such as hearing, vision, smell, ... This is because in order to respond to the tests, subjects have to make use of several primary cognitive functions (represented by modules segregated throughout the brain) to perform high-level cognition that allows them to attend, reason and act coherently.

E. Validation of the results

In addition to validating the results by separating the data into 80% train and 20% test, we have performed a 5-fold cross-validation which shows that the mean is very similar to the results obtained with the previous setup, and that the standard deviation is quite low, as shown in Table IV.

TABLE IV: 5-fold cross validation results corresponding to the modality where the best test correlation has been obtained previously.

Score Name	Modality	Training Correlation ($\mu \pm \sigma$)	Testing Correlation ($\mu \pm \sigma$)
Picture Vocabulary Test (Age Adjusted)	Structural	0.5210 ± 0.0099	0.2602 ± 0.0713
Picture Vocabulary Test (Unadjusted)	Structural	0.5834 ± 0.0034	0.2225 ± 0.0819
PMAT (Correct Responses)	Functional	0.4489 ± 0.0138	0.2066 ± 0.0391
PMAT (Skipped Items)	Bimodal	0.4703 ± 0.0115	0.2683 ± 0.0643
Delay Discounting Task	Bimodal	0.6344 ± 0.0184	0.1197 ± 0.0655

F. Comparison of results with previous work

The most relevant previous work, and the one that has motivated us for this project, was that reported by Smith et al. in 2015 [2]. In 2015 the HCP database did not contain the 1000 subjects that we have today, so the analysis was performed on 461 subjects and they used unimodal magnetic resonance imaging, functional to be more precise. The analysis was carried out on the entire dataset where they used the principal components technique to reduce dimensionality and then, using canonical correlation analysis, found associations between the components and neuropsychological scores.

TABLE V: **Comparison of results**. On the one hand, we have the results obtained in this work using different modalities (SC, FC and both) and on the other hand, we have the results obtained in [2] using only functional data. In all of them, we can see the Pearson correlation coefficient that varies from -1 to 1.

Score Name	SC data results	FC data results	SC FC data results	Smith et al. results
Picture Vocabulary Test (Age Adjusted)	0.5094	0.8229	0.5861	0.39
Picture Vocabulary Test (Unadjusted)	0.5900	0.8055	0.7675	0.41
PMAT (Correct Responses)	0.7160	0.4315	0.5458	0.38
PMAT (Skipped Items)	0.6221	0.4467	0.4585	-0.36
Delay Discounting Task	0.5336	0.9930	0.6639	0.38

IV. Conclusiones

The multimodal and multiscale study of the HCP database of 1,000 subjects has resulted in an improvement of the state of the art. Brain variables extracted from neuroimages (functional and structural) with metrics already known in another context have been regressed to predict neuropsychological indices using machine learning techniques. The correlations between both groups of variables have been positive and we have concluded that none of the three modalities (structure, function and a union of both) has been better than the other, all levels of the hierarchical tree in which the brain is organized have participated in the prediction and most of the selected variables are related to attention networks.

Acknowledgment

Data were provided by the Human Connectome Project, MGH-USC Consortium (Principal Investigators: Bruce R. Rosen, Arthur W. Toga and Van Wedeen; U01MH093765) funded by the NIH Blueprint Initiative for Neuroscience Research grant; the National Institutes of Health grant P41EB015896; and the Instrumentation Grants S10RR023043, 1S10RR023401, 1S10RR019307.

References

[1] Cedric Huchuan Xia, Zongming Ma, Zaixu Cui, Danilo Bzdok, Bertrand Thirion, Danielle S. Bassett, Theodore D. Satterthwaite, Russell T. Shinohara, and Daniela M. Witten. Multi-scale network regression for brain-phenotype associations. *Human Brain Mapping*, 41(10):2553–2566, 2020.

[2] Stephen M. Smith, Thomas E. Nichols, Diego Vidaurre, Anderson M. Winkler, Timothy E. J. Behrens, Matthew F. Glasser, Kamil Ugurbil, Deanna M. Barch, David C. Van Essen, and Karla L. Miller. A positive-negative mode of population covariation links brain connectivity, demographics and behavior. *Nature Neuroscience*, 18(11):1565–1567, November 2015.

[3] Alberto Llera, Thomas Wolfers, Peter Mulders, and Christian F Beckmann. Inter-individual differences in human brain structure and morphology link to variation in demographics and behavior. *eLife*, 8:e44443, July 2019.

[4] Paolo Bonifazi, Asier Erramuzpe, Ibai Diez, Iñigo Gabilondo, Matthieu P. Boisgontier, Lisa Pauwels, Sebastiano Stramaglia, Stephan P. Swinnen, and Jesus M. Cortes. Structure–function multi-scale connectomics reveals a major role of the fronto-striato-thalamic circuit in brain aging. *Human Brain Mapping*, 39(12):4663–4677, 2018.

[5] Javier Rasero, Mario Pellicoro, Leonardo Angelini, Jesus M. Cortes, Daniele Marinazzo, and Sebastiano Stramaglia. Consensus clustering approach to group brain connectivity matrices. *Network Neuroscience*, 1(3):242–253, October 2017.

[6] Javier Rasero, Carmen Alonso-Montes, Ibai Diez, Laiene Olabarrieta-Landa, Lakhdar Remaki, Iñaki Escudero, Beatriz Mateos, Paolo Bonifazi, Manuel Fernandez, Juan Carlos Arango-Lasprilla, et al. Group-level progressive alterations in brain connectivity patterns revealed by diffusion-tensor brain networks across severity stages in Alzheimer's disease. *Frontiers in aging neuroscience*, 9:215, 2017.

[7] Verónica Mäki-Marttunen, Ibai Diez, Jesus M. Cortes, Dante R. Chialvo, and Mirta Villarreal. Disruption of transfer entropy and inter-hemispheric brain functional connectivity in patients with disorder of consciousness. *Frontiers in Neuroinformatics*, 7:24, 2013.

[8] David C Van Essen, Kamil Ugurbil, Edward Auerbach, Deanna Barch, Timothy EJ Behrens, Richard Bucholz, Acer Chang, Liyong Chen, Maurizio Corbetta, Sandra W Curtiss, et al. The Human Connectome Project: a data acquisition perspective. *Neuroimage*, 62(4):2222–2231, 2012.

[9] Kawin Setsompop, R Kimmlingen, E Eberlein, Thomas Witzel, Julien Cohen-Adad, Jennifer A McNab, Boris Keil, M Dylan Tisdall, P Hoecht, P Dietz, et al. Pushing the limits of in vivo diffusion MRI for the Human Connectome Project. *Neuroimage*, 80:220–233, 2013.

[10] Boris Keil, James N. Blau, Stephan Biber, Philipp Hoecht, Veneta Tountcheva, Kawin Setsompop, Christina Triantafyllou, and Lawrence L. Wald. A 64-channel 3T array coil for accelerated brain MRI. *Magnetic Resonance in Medicine*, 70(1):248–258, July 2013.

[11] Qiuyun Fan, Aapo Nummenmaa, Thomas Witzel, Roberta Zanzonico, Boris Keil, Stephen Cauley, Jonathan R Polimeni, Dylan Tisdall, Koene RA Van Dijk, Randy L Buckner, et al. Investigating the capability to resolve complex white matter structures with high b-value diffusion magnetic resonance imaging on the MGH-USC Connectom scanner. *Brain connectivity*, 4(9):718–726, 2014.

[12] R Cameron Craddock, G Andrew James, Paul E Holtzheimer III, Xiaoping P Hu, and Helen S Mayberg. A whole brain fmri atlas generated via spatially constrained spectral clustering. *Human brain mapping*, 33(8):1914–1928, 2012.

[13] Nicolas De Jay, Simon Papillon-Cavanagh, Catharina Olsen, Nehme El-Hachem, Gianluca Bontempi, and Benjamin Haibe-Kains. mRMRe: an R package for parallelized mRMR ensemble feature selection. *Bioinformatics*, 29(18):2365–2368, September 2013.

[14] Edmund T. Rolls, Marc Joliot, and Nathalie Tzourio-Mazoyer. Implementation of a new parcellation of the orbitofrontal cortex in the automated anatomical labeling atlas. *NeuroImage*, 122:1–5, November 2015.

[15] Jessica R. Andrews-Hanna. The brain's default network and its adaptive role in internal mentation. *The Neuroscientist: A Review Journal Bringing Neurobiology, Neurology and Psychiatry*, 18(3):251–270, June 2012.