Imaging genetics: bio-informatics and bio-statistics challenges

Jean-Baptiste Poline[1], Christophe Lalanne[1], Arthur Tenenhaus[2], Edouard Duchesnay[1], Bertrand Thirion[3], and Vincent Frouin[1]

[1] Neurospin, Institut d'Imagerie Biomédicale, CEA, 91191 Gif sur Yvette Cedex, France. jbpoline@cea.fr, ch.lalanne@gmail.com, edouard.duchesnay@cea.fr, vincent.frouin@cea.fr

[2] SUPELEC Sciences des Systèmes (E3S)-Department of Signal processing and Electronics systems, 91192 Gif-sur-Yvette Cedex. arthur.tenenhaus@supelec.fr

[3] Neurospin, INRIA-Parietal, 91191 Gif sur Yvette Cedex, France. Bertrand.Thirion@inria.fr

Abstract. The IMAGEN study—a very large European Research Project—seeks to identify and characterize biological and environmental factors that influence teenagers mental health. To this aim, the consortium plans to collect data for more than 2000 subjects at 8 neuroimaging centres. These data comprise neuroimaging data, behavioral tests (for up to 5 hours of testing), and also white blood samples which are collected and processed to obtain 650k single nucleotide polymorphisms (SNP) per subject. Data for more than 1000 subjects have already been collected. We describe the statistical aspects of these data and the challenges, such as the multiple comparison problem, created by such a large imaging genetics study (i.e., 650k for the SNP, 50k data per neuroimage). We also suggest possible strategies, and present some first investigations using uni or multi-variate methods in association with re-sampling techniques. Specifically, because the number of variables is very high, we first reduce the data size and then use multivariate (CCA, PLS) techniques in association with re-sampling techniques.

Keywords: Neuroimaging, Genome Wide Analyses, Partial Least Squares

1 Neuroimaging genetics and the IMAGEN project

Neuroimaging genetics studies search for links between biological parameters measured with brain imaging and genetic variability. These studies are based on the hypothesis that the brain endophenotype (e.g., size or activity of a brain region) is more linked to genetic variations than to behavioral or clinical phenotypes. There are several kind of neuroimaging genetics studies depending whether they address clinical or normal populations, which endophenotype is measured, or if family information is used. However, from both a statistical and a neuroscience point of view, an important classification is "how open are the genetic and imaging hypotheses?" Often the

neuroimaging genetics study considers a specific hypothesis about one polymorphism (e.g. the serotonin transporter) and involves few brain images of a small group of subjects.

The current somewhat low cost of full genome data acquisition makes possible to perform brain and genome wide analyses (BGWA). Genome wide analyses (GWA) are already statistically challenging and often require a very large cohort, but the challenge is even bigger with the large number of potential endophenotypes (see the description below) associated with a relatively small number of subjects, as it is time consuming and costly to acquire neuroimaging data in a large cohort. In fact, it is practically impossible for a single neuroimaging center to acquire data on thousands of subjects.

Despite these challenges, several studies are on the way such as the IMAGEN project which explores brain-genetic-behavior relations in a population of 2000 normal adolescents, with an emphasis on addiction disorders, including emotional, reward or impulsivity aspects. The consortium comprises eight Europeans neuroimaging centers, the data are centralized at Neurospin (CEA, I2BM) which deals with bioinformatics and biostatistics.

1.1 Genetic data: Single Nucleotide Polymorphisms (SNP)

GWAS focuses on the relationships between the genetic sequence information (i.e., the "genotype") and a trait or phenotype (e.g., cholesterol level) measured in vivo or in vitro in unrelated individuals. Single base pair changes occurring in at least 1% of the population are used as a proxy to reflect spatial loci of variability on the whole genome. In this data one must take into account the spatial correlation between markers on DNA strands; *linkage disequilibrium* (LD), which reflects the association between alleles present at each of two sites on a genome, because a set of SNPs may not directly explain the variations observed in the trait under consideration but may be correlated with a true disease creating variants of a known biomarker instead (for reviews see Cordell & Clayton, 2005, and Ioannidis et al., 2009).

However, GWAS are considered semi-exploratory and other techniques—relying on haplotypes, genes, and gene regulation pathways—are necessary to understand relations b etween genetic polymorphisms and a given phenotype.

To avoid spurious associations between the trait of interest and genetic data, population substructure are assessed and SNPs with low minor allele frequency, not in Hardy-Weinberg equilibrium, or with low genotyping rate are discarded.

1.2 Magnetic Resonance Imaging (MRI) data

We describe below some endophenotypes acquired with MRI.

T1 images: brain macroscopic structure and the issue of anatomical structures variability. Studies of sulco-gyral anatomical variability across subjects (Riviere et al., 2002) established that variability is important even for normal subjects, and that the distance between two identical structures (e.g., sulci) can be as large as a centimeter after spatial normalization (i.e., "morphing") to a common template. These studies also showed that small structures may or may not be present in the brain of different subjects. However, characteristics of reproducible sulci can be heritable attributes and relevant endophenotypes in association studies (Rogers et al., 2010).

A popular alternative to studying individually identified structures is to use Voxel-Based Morphometry (VBM). VBM uses a the spatial normalization of the subjects brains (e.g., the MNI brain template) and then estimates from the number of voxels quantities such as grey matter density or regional volume (Ashburner and Friston, 2001). This convenient method is, however, sensitive to the values of the parameters of spatial normalisation procedures.

Diffusion Weighted Images (DWI) measures, for a voxel, the amount of water molecule diffusion in several directions. The spatial resolution is often of the order of 8mm^3 and the angular resolution (number of directions) varies from 6 to hundreds, but is often around 60 in standard settings. DWI is then used to reconstruct fiber tracks connecting brain regions (Assaf and Pasternak, 2008). As with T1 images, measurements can be made at the level of the voxel (mean diffusion, fractional anisotropy) or at the level of the fiber tracks reconstructed per subjects, or with hybrid strategies. The usefulness of the endophenotypes derived from DWI is still being assessed. Depending on the strategy, the number of features per subject ranges from a few (e.g., length of fiber tracks) to thousands (e.g., voxels).

fMRI processings. The Blood Oxygen Level Dependent (BOLD) signal measures the amount of brain regional blood flow and blood volume which correlates with neuronal activity at a spatial resolution of a few mms. In general, for a subject, an fMRI dataset is composed of several runs, each consisting of a few hundreds of three-dimensional scans acquired every few seconds. Prior to statistical analysis proper, a few essential pre-processing steps are necessary, such as intra-subject motion correction. Subject activation maps are then estimated (in general with the use of a linear model of the form $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, with several variables included in $\mathbf{X}$ to model the expected time variation of $\mathbf{Y}$). This step is crucial because the results depend strongly on the model. The model usually includes time courses designed to account for variation due to experimental conditions and confounding factors (see Poline et al, 2008). To compares subjects, a spatial normalization procedure is applied on each subject data. The group inference results are then obtained at the voxel level after a spatial smoothing of the individual data using mixed effect (or simple random effects) models (see, e.g., Mériaux et al., 2006).

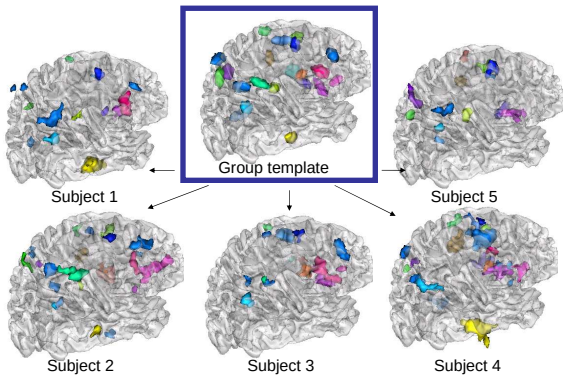

Fig. 1. How to build a model of the brain activity at the group level with a subject per subject representation (Thirion et al, 2007). This may provide more relevant endophenotypes than region of interest defined solely on the template space.

Parcellisation and functional landmarks techniques A parcellisation divides the brain into entities which are thought to correspond to well-defined anatomical or functional regions. In the context of group inference for neuroimaging, basing the analysis on parcels amounts to reducing spatial resolution to obtain a more reliable as well as interpretable matching of functional regions across subjects. Although atlas-based divisions are frequently used, their regions do not adapt to the individual functional anatomy.

An alternative to parcellisation is functional brain landmarks. Here, one searches individual topographical features and estimates their frequencies in a population of subjects. By contrast with traditional approaches, this kind of inference follows bottom-up strategy, where objects are extracted individually and then compared. Typically, structural features or patterns relevant for descriptions are local maxima of activity, regions segmented by watershed methods or blob models. Whatever the pattern used, the most difficult questions are to 1) decide if these patterns represent true activity or noise, and 2) infer a pattern common to the population of subjects.

Which endophenotype? From the description of fMRI above, it is clear that a large number of endophenotypes can be chosen from the imaging data. These endophenotypes can be differentiated into 1) voxel based approaches which use spatial normalization prior to measuring the activity of brain structures, and 2) individual landmark/structure approaches that provide individual measures. Voxel-based approaches have the advantage of being easy to automatize, but are less precise, and depend on the normalization procedures. Individual structure detection have the advantage that the endophenotype defined are more relevant and therefore more sensitive, but they are difficult

to implement, rely on a model of the correspondence between subjects, and may not always define one endophenotype per subject.

The research to understand which endophenotypes are *heritable, sensitive, specific and reproducible* for association studies is only beginning and will certainly be a key aspect of imaging genetics in the near future.

Behavioral and clinical data. Clearly, imaging and genetic data have to be complemented by demographic, behavioral, and clinical data. Summarizing these data or constructing latent variables (e.g., with SEM or PLS) that can reveal association with genetic or imaging is also a challenge.

Indeed, using items as manifest variables to uncover the locations of a latent trait (e.g. extraversion, impulsivity) implies a measurement error whose magnitude depends on the reliability of the measurement scale. As a consequence, for example, correlations between latent constructs should be corrected for attenuation, group comparisons should account for possible differential item functioning (i.e., conditional on the true latent score, the probability of endorsing an item differs between the reference and a focal group, defined by external variables). As pointed by Ioannidis et al. (2009), these considerations apply when using latent variables in GWAS. However, higher-order latent variables should give a better account of the inter-subjects variance when integrated in a conceptual model, and so should constitute more sensitive indicators.

2 Biostatistics: challenges and methods

There are several challenges for the analyses[1] of these large datasets. The first challenge arises from the specificities of multiple complex types of data. To integrate these different types of data implies a good understanding of their acquisition, and pre-processing, as well as the neuroscience or clinical contexts. Second, the large number of variables requires appropriate statistical techniques (e.g., variable selections, use of sparse techniques). Third, there is an obvious multiple comparison issue. Fourth, it is not clear what should be the overall strategy of analysis. Figure 2 represents symbolically the data at hand and how they can be analysed.

2.1 Mappings one to many

Voxel based mappings: BWAS. The aim is to isolate brain regions or voxels associated with a genetic polymorphism or a trait/phenotype on the group of subjects. This corresponds to a simple standard statistical parametric mapping analysis in neuroimaging. The method consists in first computing

[1] The bioinformatics (database, computing) aspects of these large studies are not addressed here, but are vital.

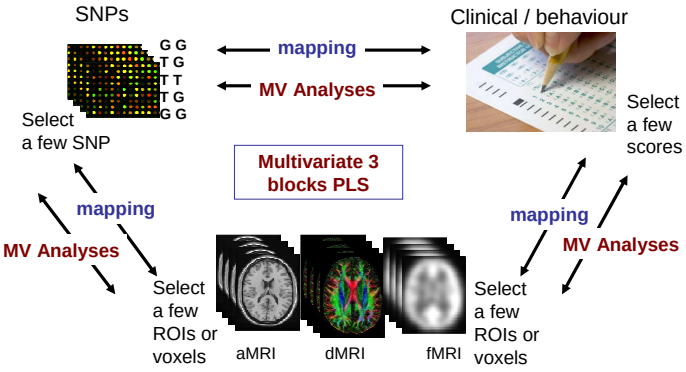

Fig. 2. The data available and how to combine them in mapping studies (one to many) or through multivariate PLS (many to many). Three PLS block can be used to integrate these data. Mappings and multivariate techniques often require variable selection or regularization because of the $N \ll P$ problem.

for each subject one brain volume summarizing the metabolic activity in one experimental condition (the so called contrast maps) and second to regressing for each voxel v the activity measured Y_v on one of the variables of interest that define the model X. The issue is then to select the brain regions with significant activity. This is done by first choosing a statistics (e.g., Fisher's F) and then estimating a threshold to correct for multiple comparisons involving 50k to 100k correlated voxels. The multiple comparison problem is often handled with random field techniques (Worsley, 2003) or permutation tests (Rorden et al., 2007). This approach is reasonable only if a limited number of candidate SNPs or scores are tested against few contrast maps.

Genetic (SNP) based mappings. GWAS analysis seeks to isolate genetic markers that explain a significant part of the variance of a given trait for unrelated individuals. Usually, such associations are studied by analyzing SNPs with a GLM model in which the frequency of the minor allele predicts the trait under study. However, this amounts to run as many tests as there are SNPs and creates an obvious problem of multiple comparisons. To control for inflation of Type I error rate, FWER corrections (e.g., Bonferroni) will only retain SNP with a p-value as low as 5.10^{-8}. Such a drastically conservative approach is likely to mask functionally interesting variants with small effect size. Moreover, tests are not independent because adjacent loci are spatially correlated. Several authors (for a review, Dudoit and van der Laan, 2008) discussed alternative strategies to enhance signal to noise ratio and increase the likelihood of tagging reliable markers.

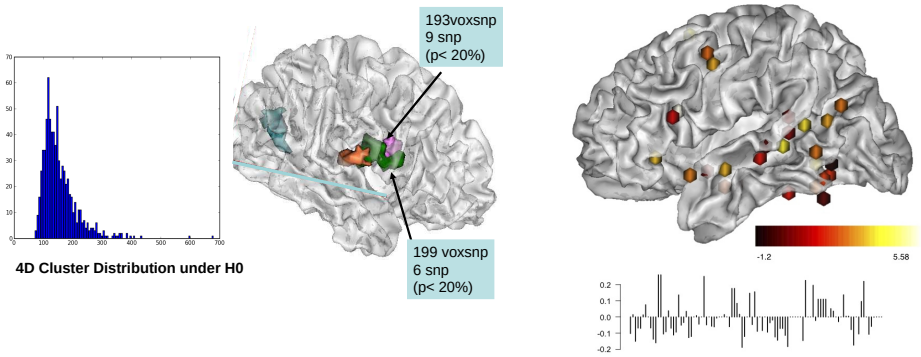

Fig. 3. Left: Constructing 4D clusters in a voxel × SNP space and permutation test. Right: Two blocks PLS. Loadings for the best 100 SNPs associated with 34 brain ROIs: positive (resp. negative) loadings in yellow (resp. red).

The voxels × SNPs challenge. Here we consider the endophenotype of an individual as the constructed 3D contrast map described above and study the association between all these voxels (approx. 50k voxel) and with each SNP within the set of more than 500k polymorphisms. For example the association of voxels with the allelic dosage (genetic additive model) for each SNP will generate around *25 billions* comparisons per contrast map.

In the QTL association study with SNP data, several techniques have been designed based on the idea that combining p-values of adjacent SNPs is more significant and more biologically relevant than considering SNPs independently. (e.g., Tippett's, Fisher's and Stoufers' methods). Recent contributions use a set of tests based on p-values aggregation (sliding window along the sequence or scan statistics). The multiple comparison issue is dealt with the usual techniques (e.g., Bonferroni, FDR, permutation tests). Theses ideas may be applied to imaging genetic data (voxel × SNPs) in order to detect contiguous brain regions linked to neighboring SNPs. The method detects clusters defined by a threshold in the product (4D) dataset, and calibrates the null hypothesis using permutations. While computationally intensive, this technique is conceptually simple, corrects for multiple comparisons in both imaging and genetic dimensions, and accounts for the spatial structure of the data. Preliminary results show that this method—illustrated in Figure 3—is efficient compared to other procedures.

2.2 Two-blocks methods

The main questions raised by two-blocks datasets with $N \ll P + Q$ are: 1) how to select the predictors of interest, 2) which multivariate model to choose, 3) how to evaluate its performance and 4) how to compare models.

Partial least squares (PLS) regression belongs to the type of methods used for modeling the association of original variables with latent variables. PLS builds successive (orthogonal) linear combinations of the variables belonging to each block, say $\mathbf{X}$ and $\mathbf{Y}$, with $\mathbf{u}_h$ and $\mathbf{v}_h$ denoting their associated canonical variates, such that their covariance is maximal:

$$\max_{|\mathbf{u}_h|=1,|\mathbf{v}_h|=1} \mathrm{cov}(\mathbf{X}_{h-1}\mathbf{u}_h, \mathbf{Y}\mathbf{v}_h)$$

where $\mathbf{X}_{h-1}$ denotes the residuals of $\mathbf{X}$ after deflation of component h. In other words, PLS seeks latent variables that account for the maximum of linear information contained in the $\mathbf{X}$ block while best predicting the $\mathbf{Y}$ block. For applications to genomics see Parkhomenko et al. (2007), to transcriptomics, see Lê Cao et al. (2009), and to SNPs X VBM, see Hardoon et al. (2009).

When predicting brain activation from SNPs, we face two issues created by the high-dimensionality of the problem. First we need to reduce the number of predictors and to design cross-validation procedures which avoid overfitting and facilitate interpretation of the resulting set of variables (Parkhomenko et al., 2007, Lê Cao et al., 2008). Second, we need to evaluate the significance of the X–Y links (e.g., with appropriate permutation schemes).

Figure 3 (right) illustrates the results obtained when maximizing PLS criterion across training samples and estimating correlation between factor scores in test samples. The significance of this test statistic was assessed using a permutation procedure embracing the whole statistical framework (cross-validation including feature selection). These preliminary results indicate that it is possible to spot significant relationships between genetic and MRI data.

2.3 Multi block analyses: RGCCA

To estimate conjointly relationships between 3 or more blocks of variables, we use Regularized Generalized Canonical Correlation Analysis (RGCCA, see Tenenhaus & Tenenhaus, submitted). With the following notations:

- J blocks $\{\mathbf{X}_1, \ldots, \mathbf{X}_J\}$ of centered variables measured on N observations,
- a design matrix $\mathbf{C} = (c_{jk})$ describing a network of connections between blocks ($c_{jk} = 1$ for two connected blocks, and 0 otherwise),
- a function g equal to the identity, absolute value or square function,
- shrinkage constants $\tau_1, ..., \tau_J$,

RGCCA is the solution of the following optimization problem:

$$\begin{cases} \underset{\mathbf{w}_1,\mathbf{w}_2,\ldots,\mathbf{w}_J}{\mathrm{argmax}} \sum_{1\le j<k\le J} c_{jk} g(\mathrm{cov}(\mathbf{X}_j\mathbf{w}_j, \mathbf{X}_k\mathbf{w}_k)) \\ \\ \text{with the constraints } (1-\tau_j)\mathrm{var}(\mathbf{X}_j\mathbf{w}_j) + \tau_j\|\mathbf{w}_j\|^2 = 1, j = 1, \ldots, J \end{cases}$$

RGCCA builds block components (i.e., latent variables) $\mathbf{y}_j = \mathbf{X}_j\mathbf{w}_j$, $j = 1, ..., J$ which explain their own block and are well correlated to their connected components. The RGCCA algorithm requires to invert—for each block—the shrunk estimation of the covariance matrices. This is computationally intractable for large blocks. To overcome this problem, we split the SNP block in blocks corresponding to chromosomes and add one block for neuroimaging. The method gives, for each block, the value of the the highest correlations is then associated to SNPs of interest. Our preliminary results with RGCCA show good sensitivity and interpretable results.

2.4 Biostatistics challenges and strategies for data analysis

The analysis of a large database such as IMAGEN, is also challenging at the level of the overall strategy as well as the computational methods and tools. Specific difficulties methodological, or even sociological are:

- The data are acquired continuously (this is necessarily the case for large imaging data studies) or by batch (genotyping). What intermediary steps should be taken, what is the likelihood that those will be confirmed with the full dataset analyses, how those should influence or not the remaining cohort recruitment are generally open questions.
- There are several approaches to study a particular neuroscience question, and controlling for the overall risk of error is difficult.
- While multivariate links may be better investigated first, this approach is technically challenged by the large number of variables (SNP, voxels) available; as multivariate variable selection is NP-hard and entails a combinatorial explosion, univariate procedures are often used in practice as initial screening.

3 Conclusions

To conclude, we believe that neuroimaging genetics—a new field that emerges at the interaction of several domains such as neuroimaging, cognitive neuroscience, genetics, experimental psychology—is particularly challenging for computational statistics, because it requires to adapt, tailor, or even create statistical methods suitable for high dimensional and heterogeneous data but also to develop specific software and databasing tools.

Acknowledgements

We are very grateful to H. Abdi for his help in editing the manuscript. Support was provided by the IMAGEN project, which receives research funding from the European Community's Sixth Framework Programme (LSHM-CT-2007-037286). This manuscript reflects only the author's views and the Community is not liable for any use that may be made of the information contained therein.

References

ASHBURNER J, and FRISTON KJ. (2001). Why voxel-based morphometry should be used. *NeuroImage, 14*, 1238-1243.

ASSAF Y, and PASTERNAK O. (2008). Diffusion tensor imaging (DTI)-based white matter mapping in brain research: a review. *J Mol Neurosci, 34*, 51–61.

DUDOIT, S. and VAN DER LAAN, M. J. (2008). Multiple Testing Procedures with Applications to Genomics. *Springer, New York.*

CORDELL HJ, and CLAYTON DG. (2005). Genetic association studies. *Lancet, 366*, 1121–1131.

HARDOON DR, ETTINGER U, MOURO-MIRANDA J, ANTONOVA E, *et al.*, (2009). Correlation-based multivariate analysis of genetic influence on brain volume. *Neuroscience letters, 450*, 281–286.

IOANNIDIS JP, THOMAS G, and DALY MJ (2009). Validating, augmenting and refining genome-wide association signals. *Nature Reviews Genetics, 10*, 318–329.

LÊ CAO KA, ROSSOUW D, ROBERT-GRANIÉ C, and BESSE P. (2008). A sparse PLS for variable selection when integrating omics data. *Statistical Applications in Genetics and Molecular Biology, 7*, 35.

MÉRIAUX S, ROCHE A, DEHAENE-LAMBERTZ G, THIRION B, and POLINE JB. (2006). Combined permutation test and mixed-effect model for group average analysis in fMRI. *Hum Brain Mapp, 27*, 402-410.

PARKHOMENKO E, TRITCHLER D, and BEYENE J (2007). Genome-wide sparse canonical correlation of gene expression with genotypes. *BMC Proceedings, 1*, S119.

POLINE JB, ROCHE A, CIUCIU P, and THIRION B. (2008). Intra- and inter-subject aspects of fMRI data analysis. In Paragios N., Duncan J., Ayache N. (Eds.) *Handbook of Biomedical Imaging.*

ROGERS J, KOCHUNOV P, ZILLES K, SHELLEDY W, *et al.*, (in press). On the genetic architecture of cortical folding and brain volume in primates. *Neuroimage.*

RORDEN C, BONILHA L, and NICHOLS TE. (2007). Rank-order versus mean based statistics for neuroimaging. *NeuroImage, 35* 1531–1537.

SMITH SM., JENKINSON M, JOHANSEN-BERG H, RUECKERT D, *et al.*, (2006). Tract-based spatial statistics: Voxelwise analysis of multi-subject diffusion data. *NeuroImage, 31* 1487–1505.

TENENHAUS A, and TENENHAUS M. (in revision). Regularized generalized canonical correlation analysis, *Psychometrika.*

THIRION B, PINEL P, and POLINE, JB (2005): Finding landmarks in the functional brain: detection and use for group characterization. *Med Image Comput Comput Assist Interv Int Conf, 8*, 476–483.

THIRION B., P. PINEL, A. TUCHOLKA, A. ROCHE, P. CIUCIU, J.-F. MANGIN, and J.-B. POLINE. (2007). Structural analysis of fMRI data revisited: Improving the sensitivity and reliability of fMRI group studies. *IEEE Transactions on Medical Imaging, 26*, 1256–1269.

WORSLEY KJ. (2003). Detecting activation in fMRI data. *Stat Methods Med Res, 12*, 401–418.