Journal of Neuroscience Methods 237 (2014) 69–78

Contents lists available at ScienceDirect

Journal of Neuroscience Methods

journal homepage: www.elsevier.com/locate/jneumeth

Computational Neuroscience
Review

Sparse models for correlative and integrative analysis of imaging and genetic data

Dongdong Lin[a,b], Hongbao Cao[c], Vince D. Calhoun[d,e], Yu-Ping Wang[a,b,*]

[a] Department of Biomedical Engineering, Tulane University, New Orleans, LA, 70118, USA
[b] Center of Genomics and Bioinformatics, Tulane University, New Orleans, LA, 70112, USA
[c] Unit on Statistical Genomics, Intramural Program of Research, National Institute of Mental Health, NIH, Bethesda 20852, USA
[d] The Mind Research Network & LBERI, Albuquerque, NM 87106, USA
[e] Department of Electrical and Computer Engineering, University of New Mexico, Albuquerque, NM 87131, USA

HIGHLIGHTS

- We review integration methods for imaging and genomic data analysis.
- We focus on our efforts in developing sparse models for imaging and genomic data integration.
- We show real examples on applications of sparse models to detecting genes and diseases diagnosis.
- We give a perspective on future research directions in imaging genomics.

ARTICLE INFO

Article history:
Received 5 May 2014
Received in revised form 27 August 2014
Accepted 1 September 2014
Available online 9 September 2014

Keywords:
Imaging genetics
Sparse modeling
Correspondence analysis
Integration
Classification

ABSTRACT

The development of advanced medical imaging technologies and high-throughput genomic measurements has enhanced our ability to understand their interplay as well as their relationship with human behavior by integrating these two types of datasets. However, the high dimensionality and heterogeneity of these datasets presents a challenge to conventional statistical methods; there is a high demand for the development of both correlative and integrative analysis approaches. Here, we review our recent work on developing sparse representation based approaches to address this challenge. We show how sparse models are applied to the correlation and integration of imaging and genetic data for biomarker identification. We present examples on how these approaches are used for the detection of risk genes and classification of complex diseases such as schizophrenia. Finally, we discuss future directions on the integration of multiple imaging and genomic datasets including their interactions such as epistasis.

Contents

* Corresponding author at: Biomedical Engineering Department, Tulane University, New Orleans, LA, USA. Tel.: +1 504 865 5867.
E-mail addresses: dlin5@tulane.edu (D. Lin), hongbao.cao@nih.gov (H. Cao), vcalhoun@unm.edu (V.D. Calhoun), wyp@tulane.edu (Y.-P. Wang).

http://dx.doi.org/10.1016/j.jneumeth.2014.09.001
0165-0270/

1. Introduction

In the past decades, increasing development of medical imaging and genomic techniques provides new opportunities to study tissue structure, function, and genetic variation as well as their relationship with human behavioral components, e.g., cognitive phenotypes and psychiatric disorders. For example, medical imaging measurements such as structural magnetic resonance imaging (sMRI), functional MRI (fMRI), diffusion tensor imaging (DTI) and positron emission tomography (PET), provide quantitative measurements of structural information at brain tissue level, dynamic blood oxygenation-level dependent (BOLD) response of neural activity and brain structural and functional connectivity. In addition, genetic measurements such as single polymorphism (SNP), gene expression, copy number variation (CNV) and proteomics can reveal structural and functional variations at molecular level. The goal of imaging genetics is to identify genetic factors that influence the intermediate quantitative measurements from anatomical or functional images, and the cognition and psychiatric disorders in humans. Rasetti and Weinberger (2011) described a cascade of imaging genetic studies, in which mutations start from genetic level to cellular processes, to the system level, e.g., brain structure, function and integrity, and eventually to human behaviors. Numerous examples like this demonstrate that the fusion of imaging and genetics will facilitate the understanding of the pathophysiology, the diagnosis of complex and heritable psychiatric disorders and the optimization of treatments in a personalized manner.

In recent imaging genetic studies, neural imaging endophenotypes (or intermediate phenotype) derived from diverse medical images, are commonly used for genetic analysis. By the definition of 'endophenotype' by psychiatric geneticists (Shen et al., 2013; Gottesman and Gould, 2003), an endophenotype should: (1) be associated with the illness/disorder of interest; (2) be heritable; (3) be state-independent; (4) exist temporally before the onset of the clinical illness in the pathophysiological pathway to the emergence of the clinical syndrome; (5) be found with higher frequency in healthy relatives of illness/disorder than in the general population. Each of these criteria is based on the hypothesis that the effects of susceptible genes will be more penetrated to the endophenotyes. The endophenotypes are considered to be closer to the biology of genetic function than the diagnostic results from self-reported and questionnaire-based clinical assessments (Gottesman and Gould, 2003; Meyer-Lindenberg and Weinberger, 2006), which can boost the causal variants detection power. Some quantitative endophenotypes derived from brain imaging are reproducible and reliable with high heritability, and can accommodate highly heterogeneous symptoms from patients in the same group. For these reasons, obtaining reliable and heritable endophenotypes is critical for imaging genetics studies.

Many endophenotypes have been used in imaging genetic studies such as voxel-, vertex-, surface- or connection-based measures from structural, functional or diffusion images, respectively. For example, on structural MRI, the volumetric measures of total cerebral and gray and white matters (Nymberg et al., 2013; Baaré et al., 2001), cortical thickness and cortical area (Winkler et al., 2010) have been studied as quantitative traits. On rest- or task-functional MRI images, there are functional endophenotypes utilized such as the extent of activation or deactivation for each voxel responding to the task-related stimuli, *t*-test contrast map, and functional brain connectivity (van den Heuvel et al., 2013). On DTI images, brain integrity (e.g., fractional anisotropy and mean diffusivity), measures of coherent direction of axons (e.g., radial diffusivity and axial diffusivity) (Jahanshad et al., 2013; Kochunov et al., 2010) and the anatomical connectivity such as the measurement of fiber density or integrity have also been explored. Moreover, due to the high resolution of brain imaging (e.g., structural MRI), we can analyze the genetic influence on diverse imaging endophenotypes across the entire brain, which facilitates the understanding of underlying neurobiological mechanism of psychiatric disorders.

Fig. 1 shows an integrative approach for combining imaging and genetics techniques for biomarker detection, from which multiple psychiatric disorders or subtypes can be better classified. We will elaborate on this approach in the following three aspects: (1) Between modality analysis to explore the correspondence or association between imaging and genetic data. One type of data is taken as an endophenotype (e.g., brain structure (Winkler et al., 2010), fiber integrity (Jahanshad et al., 2013), functional connectivity and network (Thompson et al., 2013)) to find the correlated or associated variables in the other data (e.g., genotype (Filippini et al., 2009)). The imaging endophenotypes are usually used as quantitative traits to explore the potential genetic risk factors. As reviewed in (Liu and Calhoun, 2014; Ge et al., 2013), the research in this area has evolved from candidate approaches to the whole genome and whole brain investigation, and many complicated models are proposed to account for the increasing number of variables. (2) Integration of imaging and genetic data for biomarker identification. A variety of medical imaging modalities (e.g., sMRI, fMRI and DTI) provide different insights on the change of brain or neuron activity at tissue-level, while genetic data (e.g., SNP, mRNA expression, DNA methylation and proteomics) measure different layers of genetic information at the molecular level. These different types of data are complementary, so combing these multiple modalities is likely to facilitate a better identification of biomarkers and a more comprehensive diagnosis of complex diseases. (3) Subtyping or classification of diseases by using biomarkers extracted from multimodal data. The identified biomarkers from imaging and genetic data contain complementary information about multiple psychiatric disorders (e.g., schizophrenia, bipolar and unipolar disorders). By using these biomarkers as features and input into a linear or non-linear classifier, we can achieve better disease classification, translating into more accurate diagnosis and ideally a clinical impact.

Many processing strategies and analysis approaches have been proposed to combine imaging and genetic information. For example, as reviewed in (Hibar et al., 2011a), between modality analysis methods can be categorized into univariate and multivariate imaging genetic analysis. Voxel-wise genome-wide association study (vGWAS; Stein et al., 2010) and voxel-wise gene-wide study (vGeneWAS; Hibar et al., 2011b) have been used to screen each pair of SNP/gene and voxel in maps of regional brain volume under the control of multiple comparisons. Canonical correlation analysis (CCA; Correa et al., 2008; Sui et al., 2010), partial least square (PLS; Wold et al., 1983; Krishnan et al., 2011) and parallel ICA (Liu et al., 2009) have been applied to extract a pair of correlated latent variables from imaging and genetic datasets. Kernel machine based (Ge et al., 2012) and Bayes methods (Stingo et al., 2013) have also been proposed for imaging genetics analysis. For biomarker identification, joint ICA (Sui et al., 2011), multi-set CCA (Correa et al., 2010), multi-table PLS (Caplan et al., 2007) and multi-task learning methods (Zhou et al., 2011) have been used in modeling multimodal imaging, genetic and human behavioral data. Based on the identified biomarkers, a variety of classifiers such as support vector machine (Mourão-Miranda et al., 2005; Yang et al., 2010a) and multiple kernel learning (Castro et al., 2014; Ji et al., 2008) have been applied to the classification of complex diseases.

Despite the success of these methods in the analysis of imaging and genetic data, there are still challenges due to the high dimensionality and heterogeneity of these datasets. For example, many conventional statistical methods such as CCA, PLS and ICA perform poorly for data with smaller sample size but with larger number of features/variables (e.g., voxels and SNPs). A dimensional reduction