NeuroImage 83 (2013) 384–396

Contents lists available at ScienceDirect

NeuroImage

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

Guided exploration of genomic risk for gray matter abnormalities in schizophrenia using parallel independent component analysis with reference

Jiayu Chen [a,b], Vince D. Calhoun [a,b,c,d,e], Godfrey D. Pearlson [d,e], Nora Perrone-Bizzozero [c], Jing Sui [b], Jessica A. Turner [b], Juan R. Bustillo [c,f], Stefan Ehrlich [g,h,i], Scott R. Sponheim [j,k,l], José M. Cañive [f,m], Beng-Choon Ho [n], Jingyu Liu [a,b,*]

[a] Department of Electrical and Computer Engineering, University of New Mexico, Albuquerque, NM 87131, USA
[b] The Mind Research Network, Albuquerque, NM 87106, USA
[c] Department of Neurosciences, University of New Mexico School of Medicine, Albuquerque, NM 87131, USA
[d] Olin Neuropsychiatry Research Center, Institute of Living, Hartford, CT 06106, USA
[e] Department of Psychiatry and Neurobiology, Yale University, New Haven, CT 06511, USA
[f] Department of Psychiatry, University of New Mexico School of Medicine, Albuquerque, NM 87131, USA
[g] Massachusetts General Hospital/Massachusetts Institute of Technology/Harvard Medical School, Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Charlestown, MA 02129, USA
[h] Department of Psychiatry, Harvard Medical School, Massachusetts General Hospital, Boston, MA 02114, USA
[i] Department of Child and Adolescent Psychiatry, University Hospital Carl Gustav Carus, Dresden University of Technology, Dresden, 01307, Germany
[j] Minneapolis Veterans Affairs Health Care System, One Veterans Drive, Minneapolis, MN 55417, USA
[k] Department of Psychiatry, University of Minnesota, Minneapolis, MN 55454, USA
[l] Department of Psychology, University of Minnesota, Minneapolis, MN 55454, USA
[m] Psychiatry Research Program, New Mexico VA Health Care System, Albuquerque, NM 87108, USA
[n] Department of Psychiatry, University of Iowa Carver College of Medicine, Iowa City, IA 52242, USA

ARTICLE INFO

Article history:
Accepted 11 May 2013
Available online 28 May 2013

Keywords:
Parallel ICA with reference
sMRI
SNP
Schizophrenia
Multivariate
Semi-blind

ABSTRACT

One application of imaging genomics is to explore genetic variants associated with brain structure and function, presenting a new means of mapping genetic influences on mental disorders. While there is growing interest in performing genome-wide searches for determinants, it remains challenging to identify genetic factors of small effect size, especially in limited sample sizes. In an attempt to address this issue, we propose to take advantage of a priori knowledge, specifically to extend parallel independent component analysis (pICA) to incorporate a reference (pICA-R), aiming to better reveal relationships between hidden factors of a particular attribute. The new approach was first evaluated on simulated data for its performance under different configurations of effect size and dimensionality. Then pICA-R was applied to a 300-participant (140 schizophrenia (SZ) patients versus 160 healthy controls) dataset consisting of structural magnetic resonance imaging (sMRI) and single nucleotide polymorphism (SNP) data. Guided by a reference SNP set derived from ANK3, a gene implicated by the Psychiatric Genomic Consortium SZ study, pICA-R identified one pair of SNP and sMRI components with a significant loading correlation of 0.27 ($p = 1.64 \times 10^{-6}$). The sMRI component showed a significant group difference in loading parameters between patients and controls ($p = 1.33 \times 10^{-15}$), indicating SZ-related reduction in gray matter concentration in prefrontal and temporal regions. The linked SNP component also showed a group difference ($p = 0.04$) and was predominantly contributed to by 1030 SNPs. The effect of these top contributing SNPs was verified using association test results of the Psychiatric Genomic Consortium SZ study, where the 1030 SNPs exhibited significant SZ enrichment compared to the whole genome. In addition, pathway analyses indicated the genetic component majorly relating to neurotransmitter and nervous system signaling pathways. Given the simulation and experiment results, pICA-R may prove a promising multivariate approach for use in imaging genomics to discover reliable genetic risk factors under a scenario of relatively high dimensionality and small effect size.

Introduction

Imaging genomics is an emerging field dedicated to the study of genetic variants associated with brain structure and function. Structural or functional imaging markers are believed to be closer to the underlying

* Corresponding author at: The Mind Research Network, 1101 Yale Blvd. NE. Albuquerque, NM 87106-3834, USA. Fax: +1 505 272 8002.
E-mail address: jliu@mrn.org (J. Liu).

1053-8119/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.neuroimage.2013.05.073

biological mechanisms affected by genetic variants than behavioral or symptom-based measures (Rasch et al., 2010; Turner et al., 2006). A recent meta-analysis lent support for this notion, where schizophrenia (SZ) risk variants were found to show larger effects at the level of brain structure and function than behavior (Rose and Donohoe, 2013). Consequently, interest in studying imaging measures has increased. In the case of structural imaging, measurements can be obtained via different approaches, ranging from single region-of-interest (ROI) methods, to image-wide approaches such as voxel based morphometry (VBM) (Ashburner and Friston, 2005) and surface-based measures such as FreeSurfer (Fischl and Dale, 2000).

High-throughput genotyping employing genome-wide techniques has made it feasible to sample the entire genome of a substantial number of individuals (Oliphant et al., 2002; Shen et al., 2005). More targeted candidate gene strategies examining a limited number of points of genetic variations have been successfully applied to the study of illnesses such as Fragile X syndrome (Lightbody and Reiss, 2009). Yet, the candidate gene approach is less applicable when the genetic basis of a disease is complex and less understood. For instance, little success has been achieved in replicating evidence for causal genes in schizophrenia (SZ) (Duan et al., 2010) using traditional candidate gene approaches. In contrast, recent works (Derks et al., 2012; Purcell et al., 2009) lent support for a polygenic model (Gottesman and Shields, 1967) in many cases of SZ, where an aggregate of common genetic variants was shown to collectively account for a substantial proportion of variation in risk, despite concomitant evidence for rare mutations of large effect size (Xu et al., 2009). Given such evidence, an unbiased search of the entire genome may more effectively describe the genetic architecture underlying complex disorders in which a significant proportion of risk for the disorder is likely due to many genetic variants, each carrying a small proportion of disease risk and failing to reach genome-wide significance individually.

While there is growing interest in image-wide and genome-wide approaches which allow unbiased searches over a large range of variants, novel mathematical and computational methods are desired to optimally combine these two strategies. One of the most challenging problems is the correction for the huge number of statistical tests used in univariate models. The correction makes it highly difficult to identify a factor of small effect size with a practical sample size. In addition, univariate approaches are not well-suited to identify weak effects across multiple variables. For this reason, multivariate approaches show specific advantage for simultaneously assessing many variables for an aggregate effect. To better identify aggregate effects across many variables, a number of models have been derived, including principal component regression (PCReg) (Wang and Abbott, 2008), sparse reduced-rank regression (sRRR) (Vounou et al., 2010) and parallel independent component analysis (pICA) (Liu et al., 2009).

PCReg, sRRR, and pICA are designed to deal with datasets of high dimensionality and yield interpretable results. However these approaches are not able to take prior information into account. Such information can be useful to enable a guided yet flexible approach and can improve the robustness of the results compared to a fully blind approach. For instance, some genes known to participate in a biological pathway critical to a disease may help identify a set of genes contributing in a coordinated way to a larger network. The incorporation of prior information may be especially helpful in analyzing genomic data, where a component usually accounts for a small amount of variance in the data and is more difficult to identify (Liu et al., 2012). Thus, we propose parallel independent component analysis with reference (pICA-R), which extends pICA to incorporate prior information to provide a reference to guide analyses. While pICA is designed based on regular (blind) ICA to enhance correlation between two modalities, pICA-R further takes advantage of a priori knowledge to guide the analysis and pinpoint a particular component of interest embedded in a large complex dataset. In this work, we compare pICA-R with other multivariate models through simulated data and evaluate the models under several scenarios. In addition, we apply pICA-R to a real dataset consisting of whole-brain gray matter concentration images and genome-wide single nucleotide polymorphisms (SNPs) to test whether pICA-R is able to yield reliable and interpretable components given a sample size of 300.

Material and methods

pICA-R

pICA-R is formulated by incorporating a reference constraint into pICA to guide the component extraction towards a priori knowledge. Typical pICA builds on regular infomax (Amari et al., 1996; Bell and Sejnowski, 1995) to extract independent components in parallel for each modality, followed by a conditional enhancement of the inter-modality correlations (Liu et al., 2009). In comparison, pICA-R imposes an additional constraint upon the infomax framework to minimize the distance between a certain component and the reference. The mathematical model is shown below, and Fig. 1 illustrates the flow of the approach.

$$X_m = A_m S_m \rightarrow S_m = W_m X_m, A_m = W_m^{-1}, m = 1,2 \tag{1}$$

$$\begin{aligned} &Y_m = \frac{1}{1+e^{-U_m}}, U_m = W_m X_m + W_{m0} \\ &F1 = \max\{H(Y_1)\} = \max\left\{-E\left[\ln f_{y_1}(Y_1)\right]\right\} \\ &F2 = \max\left\{\lambda H(Y_2) + (1-\lambda)\left[-\mathrm{dist}^2\left(\tilde{r}, |\tilde{S}_{2k}|\right)\right]\right\} \\ &\quad = \max\left\{\lambda\left(-E\left[\ln f_{y_2}(Y_2)\right]\right) + (1-\lambda)\left(-\left\||W_{2k}\tilde{X}_2| - \tilde{r}\right\|_2^2\right)\right\} \end{aligned} \tag{2}$$

$$F3 = \max\left\{\sum_{i,j} \mathrm{Corr}^2\left(A_{1i}, A_{2j}\right)\right\} = \max\left\{\sum_{i,j} \frac{\mathrm{Cov}^2\left(A_{1i}, A_{2j}\right)}{\mathrm{Var}(A_{1i})\mathrm{Var}\left(A_{2j}\right)}\right\} \tag{3}$$

Given a dataset **X** with dimension of sample (i.e., subjects) × feature (i.e., voxels [m = 1], SNPs [m = 2]), Eq. (1) illustrates the mathematical model of data decomposition, where the observed dataset **X** is decomposed into a linear combination of the underlying independent components, or sources. **S** is the component matrix, **A** is the loading or mixing matrix (estimated as the pseudo inverse of W), **W** is the unmixing matrix, and the subscript m runs from 1 to 2, denoting the data modality. Specifically, pICA-R iteratively solves the unmixing matrices $\mathbf{W}_1$ and $\mathbf{W}_2$ simultaneously for the two modalities, gradually maximizing the objective functions *F*1, *F*2 and *F*3 in the manner described in Fig. 1. In particular, *F*1 is the objective function of the regular infomax

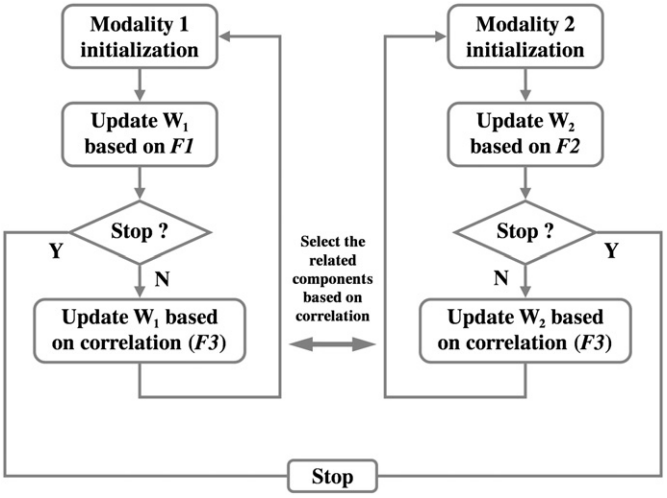

Fig. 1. Flow chart of pICA-R. $\mathbf{W}_1$ and $\mathbf{W}_2$ denote the unmixing matrices of the two modalities, respectively. *F*1, *F*2 and *F*3 represent the objective functions based on which unmixing matrices are updated.