Detecting protein dissimilarities in multiple alignments using Bayesian variable selection

Sinae Kim\textsuperscript{1}, Jerry Tsai\textsuperscript{2}, Ioannis Kagiampakis\textsuperscript{3}, Patricia LiWang\textsuperscript{2}, and Marina Vannucci\textsuperscript{1}\textsuperscript{*}

\textsuperscript{1}Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA. \textsuperscript{2}Department of Biochemistry and Biophysics, Texas A&M University, College Station, TX, USA. \textsuperscript{3}Department of Statistics, Texas A&M University, College Station, TX, USA

Associate Editor: Martin Bishop

ABSTRACT

\textbf{Motivation:} We present an application of Bayesian variable selection to the novel detection of sequence elements that confer negative design to protein structure and function. As an illustration, we analyze the different dimer interfaces between the CXCL8 chemokine family with the CCL4 and CCL2 chemokine families to discover the changes that disfavor CXCL8 of quaternary structure.

\textbf{Results:} In comparison with known experimental results, our method identifies evolutionarily conserved sequence changes in the CC families that inhibit CXCL8 quaternary structure. Therefore, we find positive selection of negative design elements. Furthermore, our approach predicts that a two-residue deletion conserved in the CCL4 chemokine family disfavors CXCL8 dimerization.

\textbf{Availability:} The Matlab code for the Bayesian variable selection should be distributed for non-commercial purposes only and is freely available at http://stat.tamu.edu/~mvannucci/webpages/codes.html.

\textbf{Contact:} mvannucci@stat.tamu.edu

1 INTRODUCTION

The purpose of this short note is to show how Bayesian variable selection can detect sequence dissimilarities in multiple alignments of conserved sequences. This application is one of the first computational approaches to address an important and subtle biological concept of evolutionarily conserved negative design elements: sequences that disfavor a particular function and/or interaction. For illustration purposes, we investigate the different dimerization forms exhibited by chemokines with the same protein fold. We aim at detecting the conserved negative design elements in two CC chemokine families that inhibit them from forming the CXCL8 dimer. To identify these negative design elements, we frame this problem in terms of a classification model with variable selection and apply a Bayesian method, developed by Vannucci and collaborators (Sha et al., 2004). We provide a brief description on how to use the suite of Matlab functions available on the web. In our analysis, we compare our method’s results with experimental data and find that the positions identified are indicative of negative design. In addition, we predict a conserved two-residue deletion that disfavors CXCL8 dimer formation.

1.1 Chemokine Dimerization Interface

Chemokines are extracellular signaling proteins that play a general role in the innate and adaptive immune response, angiogenesis, cancer and wound healing. One family of chemokines possesses the same structural fold, a single \( \alpha \)-helix over a 3 stranded \( \beta \)-sheet, but exhibits different functions (Lodi et al., 1994). In mammals, these chemokines are characterized by four conserved cysteines and are categorized based on whether the first two cysteines are adjacent (CC) or separated by an amino acid (CXC) or 3 amino acids (CX3C). These chemokines’ quaternary structure is distinct for each group. For example, CXC chemokines dimerize using the first \( \beta \)-strand and \( \alpha \)-helix, whereas CC chemokines interact across their uncoiled, extended N-terminus (Lodi et al., 1994).

In this study, our aim is to identify the negative design sequences in the CC chemokine family that disfavors the formation of the CXCL8 dimer interaction. In order to do this we consider a multiple sequence alignment (MSA) of the CXCL8 protein separately with two CC chemokine families. While MSA is commonly used to discover evolutionary conservation, here we want to computationally detect residues that are conserved within the CC chemokine family but are not conserved in the sequences contributing to the CXCL8. We therefore limit the sequence comparison to only an evolutionarily related subset (clade) of the CC chemokine family and consider only those residues involved in the CXCL8 dimer interface. Based on the phylogenetic tree of the chemokine family (Husung et al., 2003), we selected the CCL4 and CCL2 clades consisting of 15 and 23 sequences, respectively, to compare with the CXCL8 sequence (39 total). A structural analysis (Fischer et al., 2006) of the molecular, quaternary contacts in the CXCL8 interface identified 58 positions (input variables) involved in the dimer interaction. The results we report were obtained by using as input data the profiles obtained by quantifying the amino acid information with their corresponding hydrophobicity scores. Similar results were obtained with log of the volume.

2 METHODS

2.1 Bayesian Variable Selection

In order to identify structural dissimilarities in protein alignments we have applied the well-known variable selection method for classification with probit models proposed by Sha et al., 2004. Latent variables are introduced to transform the model into a normal linear model. The number of explanatory variables, \( p \), can greatly exceed the sample size. Bayesian variable selection is done via a binary vector with \( p \) entries that identifies the different sets of variables. The marginal posterior distribution of this binary vector is derived and Markov chain Monte Carlo algorithms are used to sample from its posterior distribution. The method allows the identification of sets of discriminating variables and the classification of future samples (via least squares or Bayesian model averaging).

2.2 Matlab Suite
For our example we have used a Matlab suite, written by M. Vannucci with contributions by N. Sha (UT El Paso), recently made available on the web. The software was originally written for an application of Bayesian variable selection to classification problems involving microarray data. The implementation of the probit model is quite general and allows nominal, binary or multinomial, and ordinal responses. The software also includes various parameter choices for the prior on the regression coefficients and on the binary selection vector. Inference is done efficiently by integrating out the model parameters. Posterior probabilities are calculated via fast QR updating. Users must pre-process the data by centering the training data and subtracting the training means from the future data (if any). Code and related documentation (bivprob.tar) can be freely downloaded from the webpage of the corresponding author.

The main function of bivprob.tar is bivsm Prob.m and requires as input: the data, the hyperparameter specifications (the variance parameters of the prior on the intercept, h, and on the regression coefficients, H, and the parameter, w, of the Beta-Bernoulli prior on the binary indicator) and the Metropolis parameters (initial number of included variables, total number of MCMC iterations and burn-in). Guidelines for the choices of these parameters, together with a discussion of sensitivity issues, are given in Sha et al., 2004. The main function bivsm_prob.m returns as output: the list of all visited variable subsets and their relative posterior probabilities, the normalized ordered relative probabilities of distinct visited models and the marginal probabilities of inclusion of the single variables. Distinct visited models and their posterior probabilities can serve as input to to miscler.m for least squares or Bayesian model averaging predictions, or to crosspred.m for cross-validation predictions.

3 RESULTS

We first looked at pair-wise comparisons of CXCL8 versus CCL2, CXCL8 versus CCL4 and also CCL4 versus CCL2. We set the hyperparameters to $H_2 = I, h = 10^6$ and $w = 5$, and ran MCMC chains with 50,000 iterations and 10,000 burn-in. Each chain took about 7 minutes on a desktop computer. The method identified 4 sites with high probability of being discriminatory of CXCL8 versus CCL2 and 7 sites for CXCL8 versus CCL4. These are reported in Table 1. Cross-validated prediction errors were 0 and 1 for CXCL8 versus CCL2 and CXCL8 versus CCL4, respectively. When comparing CCL4 versus CCL2, positions corresponding to residue numbers 27 and 66 were identified with a 0 misclassification error. Finally, sites 26, 27, 38, 62 and 66 were identified in the 3-group comparison of CXCL8, CCL2 and CCL4.

To assess the predictive ability of the Bayesian variable selection method in identifying negative design changes in the chemokine family, we have compared the identified mutational sites against experimental results. We concentrated on the CXCL8 sequence against the two sequence clades CCL4 and CCL2. Table 1 compares our results with those from the 4 mutational studies that have tried to inhibit the CXCL8 dimer, while retaining the chemokine fold. The data shows that there are more ways to push monomerization in vivo than to disrupt the CXCL8 dimer, whereas lines with black text are strong monomers.

Secondary structure is given by sheet (E), helix (H), and coil (C) as defined by DSSP (Kabsch and Sander, 1983). The ‘.‘ in residue number represents a break in the sequence. *BVS stands for Bayesian variable selection. The black dots (•) represent strongly predicted sites and grey dots (••) weakly predicted sites. Under “Experiments”, we show the amino acids to which residues are mutated. The L′ represents an N-methyl leucine, and the dots (•) indicate deleted residues. Lines with grey text indicate a weak dimer, whereas lines with black text are strong monomers.

Acknowledgements

M.V. is supported by NIH-R01HG003319 and NSF/DMS-0600416. J. T. has support from NIH/NCI-R25CA090801. P. L. and I. K. are supported by NIH-R01AI47832, NIH-R01AI070993 and a Welch Foundation Grant A1472.

REFERENCES


