FEATURE RANKING BASED ON SYNERGY NETWORKS TO IDENTIFY PROGNOSTIC MARKERS IN DPT-1

Amin Ahmadi Adl¹, Xiaoning Qian¹, Ping Xu², Kendra Vehik², Jeffrey P. Krischer²

¹Dept. of Computer Science & Engineering, University of South Florida, Tampa, FL 33620
²Dept. of Pediatrics, College of Medicine, University of South Florida, Tampa, FL 33620

ABSTRACT

Traditional epidemiologic methods test hypotheses focusing on individual risk factors for studying disease of interest. However, complex diseases are triggered and progress due to complicated interactions among both genetic and environmental risk factors. In this paper, we propose a network-based approach by integration of pairwise synergistic interactions to identify potential risk factors and their interactions in disease development. Specifically, we study immunologic and metabolic indices that may provide prognostic and diagnostic information regarding the development of Type-1 Diabetes (T1D) by analyzing measurements from oral glucose tolerance tests (OGTTs) and intravenous glucose tolerance tests (IVGTTs) in subjects with high risk from the Diabetes Prevention Trial-Type 1 (DPT-1) study. Performance comparison of our network-based method with individual factor based analysis demonstrates that the systematic analysis of all potential factors by considering their synergistic relationships help predict the development of clinical T1D better.

Index Terms— Type-1 Diabetes, Immunologic markers, Metabolic markers, Synergy, Spectral methods

1. INTRODUCTION

Type-1 Diabetes (T1D) is an autoimmune disorder, one of common pediatric diseases with a diverse pathogenesis, clinical phenotype and outcome [1]. Despite the emergence of T1D as a global issue with a steady increase in incidence worldwide over the past decade [2], the etiology of T1D is still not fully understood. Recent studies, including the Diabetes Prevention Trial-Type 1 (DPT-1) [3], have suggested this complex disease has multiple risk factors, including genetic predisposition, diet, viruses and geography, in addition to autoimmunity [1, 4–7].

The previous epidemiology studies mostly focus on studying hypotheses regarding individual risk factors, which have obtained important initial understandings, including the predisposing roles from genetic markers such as Human Leukocyte Antigens (HLA) [5]. However, traditional hypothesis-driven approaches focusing on “essential” factors may not be sufficient for fully understanding T1D [6]. With large-scale perspective studies including DPT-1, we believe that data-driven investigation considering all candidate factors with their interactions can serve as a critical complement for previous hypothesis-driven research.

Data-driven methods may help identify risk factors and their interactions, which will not only provide valuable insights into probable mechanisms involved in the disease, but also identify potential markers that produce high predictive accuracy for early disease prediction [8, 9]. We propose a network-based mathematical model to systematically analyze all the collected immunologic and metabolic indices for the identification of potential markers for prognosis of the development of T1D. In this work, we consider that interactions among risk factors are manifested as joint statistical association with disease outcome, by which synergy networks are constructed. We compare our network-based method with the traditional forward feature selection method by ranking individual factors based on their statistical association with disease outcome. Our preliminary results show that our network-based method by the integration of pairwise synergy between factors can help identify better groups of markers that give higher predictive accuracy.

2. METHODS

To identify biomarkers from high dimensional biomedical measurements, different feature selection methods have been investigated [10–12]. Due to the high computational complexity of exhaustive feature enumeration, filtering features based on their individual predictive power has been the common practice in biomedical research. These greedy feature filters have been successful on identifying markers with strong individual effects on disease outcome. However, they may miss features with weak individual effects whose interactions produce high predicative accuracy [10, 12]. To avoid missing these critical factors with high synergistic interactions, we propose to first construct a synergy network integrating both individual effects and pairwise synergistic interactions among all features. Based on the constructed synergy network, we solve the problem to identify risk factors and their synergistic interactions with high predictive power by developing an efficient graph spectral algorithm.

2.1 Synergy network

We first construct a synergy network for risk factor identi-
fication: \(G(V,E)\), in which \(V\) is the set of vertices \(v_i \in V\) corresponding to all the measured factors or features with corresponding weights \(f(v_i)\) measuring their individual predictive power for disease outcome. The edge set \(E\) is also weighted to measure the synergistic interactions \(s(v_i, v_j)\) between features. To measure individual and synergistic predictive power with respect to the outcome, one of potential definitions of \(f(v_i)\) and \(s(v_i, v_j)\) can be computed based on their statistical association with outcome. Specifically, we define \(f(v_i) = -\log p_i\), in which \(p_i\) is the coefficient p-value for \(v_i\) by fitting the logistic regression model: \(\log(g/(1-g)) = \beta_0 + \beta_1 v_i\) with the corresponding feature \(v_i\) and outcome \(y\) denoting the development of T1D in this paper, in which \(y\) denotes the posterior probability of \(y\) given measured features. Similarly, to define \(s(v_i, v_j)\), we fit the logistic regression model with the interaction term: \(\log(g/(1-g)) = \beta_0 + \beta_1 v_i + \beta_2 v_j + \beta_3 v_i v_j\). We let \(s(v_i, v_j) = -\log p_{ij}\), where \(p_{ij}\) is the coefficient p-value for \(v_i, v_j\) and gives the statistical significance of the predictive power of the interaction between features \(v_i\) and \(v_j\).

2.2 Finding subnetworks for risk factor identification

With previously defined metrics for synergy network construction, we identify potential prognostic markers by solving the following graph-based optimization problem:

\[
\max_{C \subseteq G} \sum_{v_i \in C} f(v_i) + \lambda \sum_{v_i, v_j \in C} s(v_i, v_j),
\]

where \(\lambda\) is a weighting coefficient between individual and synergistic effects, and \(C\) is a subnetwork with the fixed size \(\|C\| = K\) in \(G\). Without loss of generality, \(K\) can be an arbitrary number. This formulation (1) is in fact the problem of searching for a group of features with the largest objective function value, which combines the individual effects of the features in the subnetwork and the interactions among them. When \(C\) is constrained to be a clique with arbitrary size, we have a maximum weighted clique problem (MWCP), a generalized problem of the classical maximum clique problem (MCP) [13]. Furthermore, our new network-based risk factor identification problem considers both node weights and edge weights to overcome the problem in traditional feature selection of ignoring features with weak individual effects but high synergistic effects. As MCP is NP-hard [14], our network-based risk factor identification also is an NP-hard problem. Instead of deriving the exact optimal solution to this problem [13], we develop a new fast spectral approximate algorithm to solve the problem, which gives a ranked list of features with respect to their synergistic predictive power.

2.3 Feature ranking by a graph spectral algorithm

We can rewrite the original subnetwork finding problem (1) as a quadratic integer programming problem. We first introduce a set of integer variables \(x_i\), one for each feature node \(v_i\) in \(G\). When \(v_i\) is selected in \(C\) \((v_i \in C)\), \(x_i = 1\); otherwise, \(x_i = 0\). We then can rewrite (1) as

\[
\max_{x_i, 1 \leq i \leq n} \sum_{i=1}^{n} f(v_i)x_i^2 + \lambda \sum_{j=1}^{n} s(v_i, v_j)x_i x_j, \quad \text{where } n \text{ is the number of feature nodes in } G.
\]

Define the matrix \(M(v_i, v_j) = f(v_i)1(v_i, v_j) + \lambda s(v_i, v_j)\), in which \(1(v_i, v_j)\) is an indicator function which is equal to 1 only when \(i = j\). We have the matrix form for the final optimization problem:

\[
\max x^T M x, \quad \text{s.t. } x_i = 0, 1 \forall i \in [1, n],
\]

in which \(x = [x_1, \ldots, x_n]^T\) is an integer vector.

With \(\|C\| = K\), the subnetwork finding problem (2) has the constraint \(x^T x = K\). We can now develop the spectral approximate algorithm to solve this constrained quadratic integer programming problem. First, we relax the integer variable \(x_i \in \{0, 1\}\) to \(x_i \in \mathbb{R}\). The problem can be transformed to a quadratic programming by Lagrangian relaxation:

\[
\max x^T M x + \alpha(K - x^T x),
\]

where \(\alpha\) is the Lagrangian multiplier. This relaxed quadratic problem has the maximum with the necessary condition by Karush-Kuhn-Tucker condition [15] that the first derivative of the relaxed objective function in (4) is equal to 0:

\[
\frac{\partial}{\partial x} \left[ x^T M x + \alpha(K - x^T x) \right] = 0.
\]

By straightforward algebraic manipulations, we have the KKT condition for the potential solution \(x^*\) satisfying \(M x^* = \alpha x^*\). Hence, the solution to the relaxed MWCP (2) \(x^*\) is an eigenvector of matrix \(M\). To maximize the original objective function \(x^T M x^* = \alpha x^* x^* = \alpha K\), we would like \(\alpha\) to be as large as possible. Hence, we will have the relaxed solution \(x^*\) as the eigenvector of \(M\) corresponding to its largest eigenvalue. With that, if we are going to select \(K\) feature nodes in \(G\), the approximate solution is to take the top \(K\) nodes with the largest magnitudes of the corresponding eigenvector of \(M\). This naturally gives a ranked list of features by the integration of synergistic interactions among them in addition to their individual effects that are the only considered criterion in traditional feature selection methods.

3. EXPERIMENTS AND DISCUSSIONS

To illustrate the advantages of our network-based risk factor identification over traditional feature selection simply based on individual effects, we have applied our method to identify potential markers from measurements of oral glucose tolerance tests (OGTTs) and intravenous glucose tolerance tests (IVGTTs) considering their prognostic power for the development of T1D in DPT-1.

3.1 DPT-1

DPT-1 was a study designed to determine if T1D can be prevented or delayed by preclinical intervention of insulin supplement. It focuses on first and second degree non-diabetic relatives of patients with T1D before the age of 45, since they have more than ten-fold the risk of T1D in the general population [3]. DPT-1 screened 103,391 subjects altogether and categorized them into four risk groups based on genetic susceptibility, age, the presence of autoantibodies (including islet cell autoantibodies (ICA), insulin autoantibodies (IAA), glutamic acid decarboxylase (GAD), and insulinoma-associated protein 2 (ICA512)), and the change of metabolic markers during OGTT and IVGTT. The 3,483 subjects positive for ICA were staged to quantify the projected 5-year risk
of diabetes [7] and our analysis focuses on the study for the “high risk” group [7–9], which contains 339 subjects respectively, and there are 160 subjects diagnosed T1D at the end of the study. For these subjects, we here study immunologic markers by their titer values, and metabolic indices, including C-peptide, insulin, and glucose levels, to identify potential markers with desired prognostic accuracy. For DPT-1, previous analyses focus on understanding the prognostic power for different individual measurements [7–9]. To the best of our knowledge, this is the first time that multivariate analysis considering interactions is investigated in DPT-1.

3.2 Performance comparison

We study baseline characteristics in DPT-1, focusing on immunologic and metabolic markers. Specifically, we have taken the available titer values for different autoantibodies, including ICA, IAA, GAD, ICA512, and MIAA (micro-insulin autoantibodies). For metabolic indices, we have fasting glucose, glycated hemoglobin (HbA1c), fasting insulin, and first-phase insulin response (FPIR) from IVGTTs. Homeostasis model assessment of insulin resistance (HOMA-IR) and FPIR-to-HOMA-IR ratio are also computed as in [9]. From OGTTs, in addition to 2-hour glucose and fasting glucose, we have collected blood samples for C-peptide measurements in the fasting state and then 30, 60, 90, and 120 minutes after oral glucose. We also have computed peak C-peptide as the maximum point of all measurements and AUC (area under curve) C-peptide using the trapezoid rule. Furthermore, as age and Body Mass Index (BMI) have been conjectured to be important confounding factors, we also include them in our data-driven network-based analysis.

Within the subjects of the high risk group, there are 169 subjects who received parenteral insulin supplement while the other 170 subjects were randomly assigned to the placebo arm of the study. We first construct the synergy networks for both placebo (Fig. 1A) and study subgroups (Fig. 1B). In the figure, we illustrate the features with the coefficient p-values lower than 0.05 by normal nodes with features of significant individual effects. We have plotted synergistic interactions in red if the corresponding coefficient p-values are lower than 0.05 and in green if p-values are lower than 0.1. For each subgroup, we respectively rank these features based on their individual predictive power and their eigenvectors with the integration of their synergistic interactions with $\lambda = 1.0$. We compare the classification performance of these features using quadratic discriminant analysis (QDA) in MATLAB [16] to make sure that interactions among features are considered in the classification model. Forward feature selection algorithm [11] based on different ranked lists has been applied to select a subset of features that give the best classification accuracy with QDA. Specifically, we sequentially add features into the subset based on the ranked lists if adding it improves the classification accuracy based on 10-fold cross validation.
validation. Table 1 gives the estimated test accuracy by 1,000 runs of 10-fold cross validation for two subgroups with individual and network-based rank lists. It is clear that the integration of interactions among features by our network-based method leads to higher classification accuracy in both subgroups, which is statistically significant with p-value < 0.01. We note that both methods choose similar markers, including age (or height), 2-hour glucose, and FPIR-to-HOMAIR ratio. Interestingly, our network-based method also takes other factors such as HbA1c, which in fact has been added as another marker for diagnosis of diabetes by American Diabetes Association (ADA) in 2010. It also is interesting that when we predict the progress to clinical T1D for placebo subgroups without intervention and study subgroups with intervention, different groups of features are selected as prognostic markers as shown in Table 2, which may suggest that different subgroups respond differently to insulin treatment and the selected features may characterize these subgroups as different subtypes of diabetes.

Table 1. The test classification accuracy for placebo and study groups for high risk subjects: “Ind” column includes the results based on forward feature selection using the ranked lists by individual effects; “Syn” column includes the results based on forward feature selection using the ranked lists by our network-based method.

|       | Placebo | Study |
|-------|---------|-------|
| Ind   | 0.653   | 0.675 |
| Syn   | 0.678   | 0.653 |

Table 2. The selected features (ranked by individual or interactive effects respectively) to obtain the classification accuracy in Table 1.

| Placebo-Ind | Age, 2-h glucose, Peak C-Peptide, FPIR-to-HOMAIR ratio |
| Placebo-Syn | Height, Fasting glucose (IVGTT), FPIR-to-HOMAIR ratio, IAA, HbA1c, 2-h glucose, ICA512 |
| Study-Ind   | Height, ICA512 |
| Study-Syn   | ICA, HbA1c, Age, Height, ICA512 |

4. CONCLUSIONS AND FUTURE RESEARCH

Our network-based risk factor identification considers interactions among candidate factors in addition to their individual predictive power with respect to the outcome. Our preliminary results have shown that our new method identifies prognostic markers with higher prediction accuracies of diagnosed T1D in the high risk group in DPT-1. For our future research, we will develop effective and efficient algorithms for risk factor identification by systematically analyzing all the available measurements in large-scale studies, develop new prognostic models, and carry out thorough performance evaluation for them. Ultimately, we would like to develop systematic data-driven methods to analyze available factors to understand the interactions of genetic and various environmental factors regarding the development of autoimmunity and T1D, and to identify better prognostic or diagnostic markers for effective screening of subjects for the risk of developing T1D, so that they can benefit from early disease prediction and prevention.

Acknowledgement The project was supported in part by Award R21DK092845 from the National Institute Of Diabetes And Digestive And Kidney Diseases, National Institutes of Health.

References

[1] A Lernmark and J Ott, “Sometimes it’s hot, sometimes it’s not,” Nat Genet, vol. 19, no. 3, pp. 213–214, July 1998.
[2] DERI Study Group, “Secular trends in incidence of childhood IDDM in 10 countries. Diabetes Epidemiology Research International Group,” Diabetes, vol. 39, pp. 858–864, 1990.
[3] Diabetes Prevention Trial-Type 1 Diabetes Study Group, “Effects of insulin in relatives of patients with type 1 diabetes mellitus,” N Engl J Med, vol. 346, pp. 1685–1691, 2002.
[4] GF Bottazzo, A Florin-Christensen, and D Doniach, “Islet-cell antibodies in diabetes mellitus with autoimmmune polyendocrine deficiencies,” Lancet, vol. 2, no. 7892, pp. 1280–1283, 1974.
[5] J Nerup, P Platz, OO Andersen, M Christy, J Lyngsoe, JE Poulsen, LP Ryder, LS Nielsen, M Thomsen, and A Svejgaard, “HL-A antigens and diabetes mellitus,” Lancet, vol. 2, no. 7885, pp. 864–866, 1974.
[6] P Bougnères and AJ Valleron, “Causes of early-onset Type 1 Diabetes: Toward data-driven environmental approaches,” J Exp Med, vol. 205, pp. 2953–2957, 2009.
[7] JP Krischer, DD Cuthbertson, L Yu, T Orban, N Maclaren, R Jackson, WE Winter, DA Schatz DA, JP Palmer, and GS Eisenbarth GS, “Screening strategies for identification of multiple antibody-positive relatives of individuals with type 1 diabetes,” J Clin Endocrinol Metab, vol. 88, pp. 103–108, 2003.
[8] J Sosenko, JP Palmer, CJ Greenbaum, J Mahon, C Cowie, JP Krischer, HP Chase, NH White, B Buckingham, KC Herold, D Cuthbertson, JS Skyler, and the Diabetes Prevention Trial-Type 1 Study Group, “Increasing the accuracy of oral glucose tolerance testing and extending its application to individuals with normal glucose tolerance for the prediction of type 1 diabetes,” Diabetes Care, vol. 30, pp. 38–42, 2007.
[9] P Xu, Y Wu, Y Zhu, G Daghe, G Johnson, D Cuthbertson, JP Krischer, JM Sosenko, JS Skyler, and the DPT-1 Study Group, “Prognostic performance of metabolic indexes in predicting onset of Type 1 Diabetes,” Diabetes Care, vol. in press, no. doi: 10.2337/dc10-0802, 2010.
[10] HC Peng, F Long, and C Ding, “Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 8, pp. 1226–1238, 2005.
[11] Y Saeys, I Inza, and P Larrañaga, “A review of feature selection techniques in bioinformatics,” Bioinformatics, vol. 23, no. 19, pp. 2507–2517, 2007.
[12] J Watkinson, X Wang, T Zheng, and D Anastassiou, “Identification of gene interactions associated with disease from gene expression data using synergy networks,” BMC Systems Biology, vol. 2, pp. 10, 2008.
[13] SJ Sajjadi, AA Adl, B Zeng, and X Qian, “Finding the most discriminating sets of biomarkers by maximum weighted clique,” in Proceedings of the 6th INFORMS Workshop on Data Mining and Health Informatics, 2011, vol. DM-HI 2011.
[14] PM Pardalos and J Xue, “The maximum clique problem,” Journal of Global Optimization, vol. 4, no. 3, pp. 301–328, 1994.
[15] DP Bertsekas, Nonlinear Programming, Athena Scientific, Belmont,MA, 1995.
[16] WJ Krzanowski, Principles of Multivariate Analysis: A User’s Perspective, New York: Oxford University Press, 1988.