Protein-Protein Interaction Network and Novel Biomarkers of Celiac Disease

Selia Chowdhury, Mehedi Hasan Bappy, Md. Shahraj Chowdhury, Nurjahan Shilda Chowdhury, Athanasios Patousis, Anna Gerothanasi, Panagiotis Patousis, Samia Chowdhury, and Shajeda Chowdhury

ABSTRACT

**Background:** Celiac disease is a small intestine enteropathy. Gluten exposure in the diet of those who are vulnerable causes it. Susceptibility is determined by genetics. Despite advancements in technology, small intestinal biopsy remains the gold standard for diagnosing Celiac disease.

**Methods:** Differentially expressed proteins related to celiac disease were collected from published research articles and considered for this study. Gene IDs of all proteins contained in the data set were identified using the NCBI Gene Database. The list of entrez gene IDs was then entered into the bio profiling software to look for protein-protein interactions between the gene IDs.

**Results:** Using the Protein-Protein Interaction (PPI) spider model, numerous important signaling pathways implicated in the proteome of Celiac disease have been discovered. The p values of model D1 and D2 were <0.03 and <0.05, respectively. Model D1 depicts the pathway involved, which includes reverse cholesterol transport and model D2 depicts the pathway involved, which include anti-apoptosis, response to virus, and positive regulation of I-kappaB cascade.

**Conclusions:** The deregulation of several pathways and protein interaction networks is suggested by bioinformatic analysis of the differentially expressed proteins. Following validation, the unique biomarkers discovered can be utilized to better understand this illness and identify potential pharmaceutical therapies.

**Keywords:** Celiac disease, bioinformatics, gene expression, intestine biopsy, proteomics, protein-protein interactions.

I. INTRODUCTION

In most populations, celiac disease affects roughly 1% of the population [1]. The rise in diagnosis rates appears to be related to an increase in incidence rather than greater awareness and detection. Celiac disease occurs in genetically vulnerable people who build an immunological response in reaction to unknown environmental causes, which is then activated by gluten consumption [1]. The condition manifests itself in a variety of ways, from severe malabsorption to nonsymptomatic or less symptomatic presentations. The presence of duodenal villous atrophy and circulating antibodies against tissue transglutaminase are required for diagnosis; however, in children, European guidelines allow a diagnosis without a duodenal biopsy if rigorous clinical and serological criteria are satisfied [1]. Although a gluten-free diet is successful for the majority of people, a significant minority have persistent or recurring problems. Due to difficulties adhering to a gluten-free diet, non-dietary treatments have been developed, with some of them now undergoing human trials.

II. MATERIALS AND METHODS

Bioinformatic evaluations of proteins relevant to celiac disease that were differentially expressed (whether increased or reduced) were obtained from a variety of research papers. The information was gathered from public databases (proteomics and microarray-based techniques documents). Full text publications published till January 2022 on the relevant keywords "proteomic" and "celiac disease" were searched in databases. In comparison to controls, these proteins were found in blood or peripheral tissue samples. After removing duplicates, six distinct investigations [1]-[6] discovered a total of 66 proteins that meet these characteristics. The DAVID tool [7], [8] was used to identify the genes responsible for encoding these proteins more quickly. Gene IDs of all proteins contained in the data set
were identified using the NCBI Gene Database. The list of entrez gene IDs was then entered into the bioprofiling software, a web-based tool which reports the statistical probability of interactions between proteins to look for protein-protein interactions between the gene IDs [9]. Using the PPI spider database, we identified protein-protein interactions. 34 of these proteins were identified as part of the PPI spider database. Three models of protein interaction were generated: the D1 model forecast direct interactions between candidates of the data set, the D2 model forecast interactions between candidates in the data set while allowing for up to one skipped gene, and the D3 model for up to two skipped genes. Following this, cytoscape was used to visualize the PPI spider models into two-dimensional interaction maps to visualize these interactions and pathways involved [10]. The models were rendered as organic protein interaction networks, and genes were color-coded based on function. Afterwards, Protein Analysis Through Evolutionary Relationships (PANTHER) was used to model the functional classifications in the form of pie and bar charts [11]. PANTHER provides comprehensive functional information about queried gene lists and presents them graphically. For the computed gene list from the PPI spider, PANTHER was used to classify the genes according to biological process, cellular component, protein class and pathway.

III. RESULTS

Bioinformatic analyses of the differentially expressed proteins compiled from various research articles highlighted the genes identified are involved in binding, catalytic activity, molecular function regulation, molecular transducer activity, structural molecule activity and transporter activity. Also included are charts showing the classification of genes according to biological process, cellular component, protein class and pathway.

The major signaling pathways in the panther model show that the genes are mostly involved in areas included Inflammation mediated by chemokine and cytokine signaling pathway (11 genes: RSG1, STAT1, CXCL10, PTGS2, CXCL8, REL, IL6, IL15, CCR3, CCR9, and NFKB1), CCKR signaling map (5 genes: CLU, SERPINE1, PTGS2, BAX, and CXCL8), Interleukin signaling pathway (5 genes: STAT1, CXCL8, IL6, IL11, and IL15), Apoptosis signaling pathway (4 genes: NFKB1, BAX, RIPK1, and REL), toll receptor signaling pathway (4 genes: PTGS2, NFKB1, TNFAIP3 and REL), EGF receptor signaling pathway (2 genes: STAT1, PEBP1), p53 pathway (2 Genes: SERPINE1, BAX), and PDGF signaling pathway (2 genes: STAT1, ELF5).
Network Models

Using the PPI spider model several major signaling pathways have been identified as involved in the proteomics of Celiac disease.

![Gene classification according to protein class](image1)

**Fig. 4.** Gene classification according to protein class.

![Gene classification according to pathway](image2)

**Fig. 5.** Gene classification according to pathway.

**Network Model D1**

Model D1 shows enriched networks with 0 missing genes allowed. This model covered 3 input genes. There was a P value of <0.05. The model D1 shows the pathway involved includes reverse cholesterol transport.

![Network model D1](image3)

**Fig. 6.** Network model D1.

**Network Model D2**

Model D2 shows enriched networks with 1 missing gene allowed. This model covered 9 input genes. There was a P value of <0.03. The model D2 shows the pathway involved include: antiapoptosis, response to virus, and positive regulation of I-kappaB kinase/NF-kappaB cascade.

![Network model D2](image4)

**Fig. 7.** Network model D2.

**Network Model D3**

Model D3 which shows enriched networks with 2 missing genes allowed. This model covered 24 input genes. There was a P value of <0.58. Model D3 shows the pathways involved include: antiapoptosis, response to drug, response to virus, mRNA processing, endocytosis, protein autophosphorylation, negative regulation of apoptosis, cell surface receptor linked signaling pathway, phospholipid efflux, negative regulation of endothelial cell proliferation, positive regulation of DNA-dependent transcription, and endocytosis. It was excluded for being statistically insignificant as the p value was >0.05.
The goal of this study is to use a bioinformatics technique to provide a visual overview of protein-protein interactions that are implicated in celiac disease. Proteomic research has proven to be quite effective in the creation of datasets for prospective markers that may be utilized in diagnostic, prognostic, and therapeutic techniques in the evaluation of human illnesses [2]. Celiac disease has a small number of protein expression investigations. Advanced proteomics technology provides novel biomolecular markers for early identification as well as new endpoints for evaluating treatment effectiveness, toxicity, and response to therapy [2]. Molecular profiling investigations of affected cells have the potential to uncover novel biomarkers that will help to improve disease diagnosis approaches as well as therapy choices. Celiac disease biomarkers are a quantifiable sign of the degree or existence of the disease that can aid in patient management. The P value of network model D1 was <0.05 and model D2 was <0.03 which are statistically significant. Model D1 showed the involvement of pathway reverse cholesterol transport. Model D2 showed the pathways involved include anti-apoptosis, response to virus, and positive regulation of I-kappaB kinase/NFkappaB cascade.

The purpose of bioinformatics in high-throughput operations is to limit the amount of data by filtering down a big list of genes to a few groups of genes known to be engaged in interactions [7]. The networks generated by these genes indicated additional possibilities for involvement in the cascade, necessitating more research. This enables a more focused approach to discovering which proteins and genes are involved in disease pathogenesis. Bioinformatics also eliminates the need to fish for problematic genes involved in interactions at random [7]. Identification of proteins and the networks to which they belong will increase our understanding of the molecular basis of disease pathology, as well as provide a better picture of the complex molecular cascade via analysis of these protein-protein interactions [2]. It is critical for the advancement of innovative medicines to target several pathways concurrently rather than a single element of a complicated disease [2]. PPI spider, for example, provides a rigorous statistical approach for commutated models that assess the chance of random gene lists existing in nature. Researchers can evaluate if novel genes provided in these models have a meaningful relevance to a disease process by manually analyzing computed networks in the lab. These proteins might therefore be employed as biomarkers or therapeutic targets in the future.

V. CONCLUSION

The deregulation of several pathways and protein interaction networks is suggested by bioinformatic analysis of the differentially expressed proteins in celiac disease. The novel biomarkers identified can be used after validation to better understand this condition and designing possible pharmacological interventions. It is critical for the advancement of innovative medicines to target several pathways concurrently rather than a single element of a complicated disease. Through a rigorous statistical approach for commutated models that assess the chance of random gene lists existing in nature, researchers can evaluate if novel genes provided in these models have a meaningful relevance to a disease process by manually analyzing computed networks in the lab. These proteins might therefore be employed as biomarkers or therapeutic targets in the future.

FUNDING

The authors received no support from any funding source for this project.

CONFLICT OF INTEREST

Authors declare that they do not have any conflict of interest.
REFERENCES


