SciCell

Nova Biotechnol Chim (2020) 19(1): 52-60
DOI 10.36547/nbc.v19i1.577

Nova Biotechnologica *et* Chimica

A sequence-dependent classification algorithm for Crohn's Disease – causing NOD2 protein mutations

Jose Isagani B. Janairo[1,✉] and Marianne Linley L. Sy-Janairo[2]

[1]*Biology Department, De La Salle University, 2401 Taft Avenue, Manila 0922, Philippines*
[2]*Institute of Digestive and Liver Diseases, St. Luke's Medical Center – Global City, Rizal Drive, Taguig 1634, Philippines*

Article info

Article history:
Received: 10th December 2019
Accepted: 3rd March 2020

Keywords:
Artificial neural networks
Inflammatory bowel disease
Machine learning
Personalized medicine

Abbreviations:

ANN	artificial neural network
CD	Crohn's disease
DCM	disease-causing mutation
NOD2	Nucleotide-binding oligomerization domain-containing protein 2
NPV	negative predictive value
PPV	positive predictive value
RF	random forest
SVM	support vector machine

Abstract

Certain NOD2 protein mutations have been associated with the onset of the inflammatory bowel disease, Crohn's Disease (CD). NOD2 is involved in the inflammatory response of the gut to the microbial community, wherein its functional impairment through mutations may lead to CD progression. Considering the significant role that NOD2 plays in CD pathogenesis, predicting whether a specific type of NOD2 mutation is the cause of CD can greatly aid the accuracy of the disease diagnosis. Hence, a novel sequence-based classification algorithm built on artificial neural network (ANN) is herein presented that can predict whether a specific NOD2 mutation can cause CD or not. The NOD2 mutant types and their association with CD were taken from literature, and the calculated sequence-order coupling numbers were used as the classification predictors. The formulated ANN classifier exhibited satisfactory predictive ability, with 82.4 % accuracy, 62.5 % sensitivity, 100 % specificity, 100 % positive predictive value, and 75 % negative predictive value. The presented ANN classifier provides a proof-of-concept that predicting the onset of CD from NOD2 protein variant is possible.

Introduction

Crohn's Disease (CD) is characterized by chronic transmural inflammation of the gastrointestinal tract. Pathogenesis of CD is multifactorial, wherein one of the key drivers of the disease involves mutations in the NOD2 protein (Yamamoto and Ma 2009). NOD2 is encoded by the CARD15 gene in the human chromosome 16 (Strober and Watanabe 2011). This protein plays a critical role in microbe / pathogen sensing, wherein the leucine-rich region of NOD2 binds to the muramyl dipeptide (MDP) of the bacterial cell wall. Once activated by MDP, NOD2 initiates downstream signaling events relevant to the host immune response. Thus, NOD2 mutations may lead to the impaired regulation and response of the host to bacterial interactions, which increases the risk to unusual ileal inflammation (Sidiq *et al.* 2016).

Various NOD2 mutations have been associated with CD susceptibility, wherein the missense mutations and frameshift mutation appear to be the most common type of mutations associated with CD progression (Cuthbert *et al.* 2002; Hampe *et al.* 2002; Economou *et al.* 2004). Other less frequent mutations are also linked with CD pathogenesis, while other NOD2 mutations do not lead to CD (Lesage *et al.* 2002). Clearly, possible connections between the properties of NOD2 protein mutants and CD susceptibility exist, but remain to be uncovered. Thus, this study aims to use machine learning to formulate a predictive model that can classify NOD2 mutations as disease-causing or non-disease-causing based on

✉ *Corresponding author:* jose.isagani.janairo@dlsu.edu.ph

protein numerical representations. Artificial Neural Network (ANN) is a powerful machine learning technique that can uncover non-obvious patterns or associations from datasets of various characteristics. ANN has been widely used in medical diagnosis, particularly in cancer classification and prediction (Khan *et al.* 2001), tuberculosis (Er *et al.* 2010), among others. Having the ability to predict whether a specific NOD2 mutation maybe associated with CD can greatly improve disease detection and therapy. In addition, this ability becomes even more valuable after considering that NOD2 mutation type influences the response of the patient towards a particular treatment (Niess *et al.* 2012). Early detection of the disease is one of the main challenges in inflammatory bowel disease, such as CD (Flamant and Roblin 2018). Thus, such predictive model can potentially help improve disease diagnosis, as well as lay the groundwork for the greater adoption of personalized medicine for the management of CD.

Experimental

Data Mining

NOD2 protein disease-causing mutants (DCM), and non-disease-causing mutants (NDCM) were taken from (Lesage *et al.* 2002). The list contains 30 DCMs, and 13 non-DCMs, both which are mostly point mutations. Additional NOD2 mutant variants were taken from ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/), a database that shows the relationship between genetic variants and phenotypes (Landrum *et al.* 2014). The archive was searched using the search string [NOD2 AND (("Crohn Disease") AND (BENIGN OR PATHOGENIC))]. The search yielded 76 results, but after removing duplications, silent mutations, and inconclusive medical assignment for each variant, 16 NOD2 mutants were added to the dataset. The resulting 59 NOD2 protein variants (Table 1) were then numerically represented using the 30 sequence-order coupling number (SOCN) based from Schneider-Wrede descriptors (Schneider and Wrede 1994), calculated using ProtrWeb (Xiao *et al.* 2015). This web server for calculating protein descriptors require the protein sequence as the input. The canonical sequence of human NOD2 was taken from www.uniprot.org (Uniprot ID: Q9HC29), which also served as the basis of the mutant sequences. The functional impact of the mutations on the NOD2 protein was also assessed using the PROVEAN Protein tools as implemented in the PROVEAN web server version 1.1.3 (Choi *et al.* 2012) (http://provean.jcvi.org/seq_submit.php). The input in the PROVEAN web server is the wild type protein sequence, the position of the mutation, and the amino acid substitution. The utilized classification threshold was the default value of -2.5. From the provided information, the web server will then determine if the submitted mutation for analysis is either deleterious or neutral. The full dataset, which contains the 59 NOD variants and the corresponding 30 SOCNs is available in the supporting information.

Statistical Analysis

All statistical analyses were carried out using Tibco Statistica version 13.4.0.14. Statistical difference was probed between DCM and NDCM NOD2 protein variants through ANOVA using the 30 SOCN. The same descriptor set was also utilized to segregate the 43 NOD2 protein variants through a two-cluster solution using K-means clustering. After the exploratory data analysis, the dataset was then used to create various machine learning classification models. DCM / NDCM served as the categorical response variable, and the calculated sequence-order coupling numbers served as the continuous descriptors. For the artificial neural network (ANN) based - classification model, a feed-forward multilayer perceptron architecture was adopted, wherein sum of squares was the error function, the hidden unit activation function used was tanh, and identity was the output unit. Bootstrap subsampling was employed, wherein 10 subsamples were gathered in which 50 % was dedicated for training, 30 % for testing, and 20 % for validation. For support vector machine (SVM) classification, the radial basis function (RBF) kernel was utilized, leading to the automatic selection of the best Gamma and C parameters. The Gamma and C parameters are involved in the definition of the hyperplanes which leads to the separation and classification of the cases. 75 %

Table 1. Association of NOD2 mutations with CD progression as reported in Lesage *et al.* (2002) and information from the ClinVar dat. DCM refers to disease-causing mutation, while NDCM means non-disease-causing mutation.

NOD2 Variant	Association with CD	Reference	PROVEAN Score	Functional impact of mutation based on PROVEAN prediction
R138Q	DCM	Lesage *et al.* 2002	-2.120	Neutral
A140T	DCM	Lesage *et al.* 2002	-0.464	Neutral
W157R	DCM	Lesage *et al.* 2002	1.142	Neutral
T189M	DCM	Lesage *et al.* 2002	-0.494	Neutral
R235C	DCM	Lesage *et al.* 2002	-2.779	Deleterious
L248R	DCM	Lesage *et al.* 2002	-4.286	Deleterious
P268S	NDCM	Lesage *et al.* 2002	-0.614	Neutral
N289S	DCM	Lesage *et al.* 2002	-2.417	Neutral
D291N	DCM	Lesage *et al.* 2002	-2.097	Neutral
T294S	NDCM	Lesage *et al.* 2002	-3.033	Deleterious
A301V	NDCM	Lesage *et al.* 2002	-3.631	Deleterious
R311W	DCM	Lesage *et al.* 2002	-4.602	Deleterious
L348V	NDCM	Lesage *et al.* 2002	-2.291	Neutral
H352R	NDCM	Lesage *et al.* 2002	-4.166	Deleterious
R373C	DCM	Lesage *et al.* 2002	-2.984	Deleterious
N414S	DCM	Lesage *et al.* 2002	-1.796	Neutral
S431L	DCM	Lesage *et al.* 2002	-2.000	Neutral
A432V	NDCM	Lesage *et al.* 2002	-1.104	Neutral
E441K	DCM	Lesage *et al.* 2002	0.090	Neutral
558delLG	DCM	Lesage *et al.* 2002	-11.499	Deleterious
A612T	DCM	Lesage *et al.* 2002	-3.608	Deleterious
A612V	NDCM	Lesage *et al.* 2002	-3.645	Deleterious
R684W	DCM	Lesage *et al.* 2002	-3.092	Deleterious
R702W	DCM	Lesage *et al.* 2002	-3.285	Deleterious
R703C	DCM	Lesage *et al.* 2002	-3.313	Deleterious
R713C	DCM	Lesage *et al.* 2002	-2.838	Deleterious
A725G	NDCM	Lesage *et al.* 2002	-1.275	Neutral
A755V	NDCM	Lesage *et al.* 2002	-3.070	Deleterious
A758V	NDCM	Lesage *et al.* 2002	-0.953	Neutral
E778K	DCM	Lesage *et al.* 2002	-2.579	Deleterious
V793M	DCM	Lesage *et al.* 2002	-0.804	Neutral
E843K	DCM	Lesage *et al.* 2002	0.482	Neutral
N853S	DCM	Lesage *et al.* 2002	-4.637	Deleterious
M863V	DCM	Lesage *et al.* 2002	-0.070	Neutral
A885T	DCM	Lesage *et al.* 2002	-1.407	Neutral
G908R	DCM	Lesage *et al.* 2002	-5.822	Deleterious
A918D	DCM	Lesage *et al.* 2002	-4.932	Deleterious
G924D	DCM	Lesage *et al.* 2002	0.149	Neutral
V955I	NDCM	Lesage *et al.* 2002	-0.435	Neutral
V972I	NDCM	Lesage *et al.* 2002	-0.633	Neutral
G978E	DCM	Lesage *et al.* 2002	-1.646	Neutral
1007fs	DCM	Lesage *et al.* 2002	n.d.	n.d.
A292V	NDCM	ClinVar ID 319441	-2.391	Neutral
A612S	NDCM	ClinVar ID 319452	-2.850	Deleterious
A849V	NDCM	ClinVar ID 97855	-3.122	Deleterious
D154N	NDCM	ClinVar ID 319426	-0.914	Neutral
G1032S	NDCM	ClinVar ID 319475	0.737	Neutral
L682F	NDCM	ClinVar ID 319457	-3.467	Deleterious
Q902K	NDCM	ClinVar ID 319471	-1.023	Neutral
R391H	NDCM	ClinVar ID 319442	-0.414	Neutral
R471C	NDCM	ClinVar ID 319446	-2.087	Neutral
R708H	NDCM	ClinVar ID 319459	-1.413	Neutral
R716H	NDCM	ClinVar ID 319460	-2.092	Neutral
R791Q	NDCM	ClinVar ID 97850	-0.251	Neutral
T245M	NDCM	ClinVar ID 319434	-0.886	Neutral
V92I	NDCM	ClinVar ID 319425	-0.275	Neutral
V162I	NDCM	ClinVar ID 319427	0.486	Neutral
V955I	NDCM	ClinVar ID 97869	-0.435	Neutral

of the data set was used for training, while the remaining 25 % served as the test set, and the results were validated by applying a 10-fold cross validation. For random forest classification, the random test data proportion was set to 0.30, and 0.5 for the subsample proportion. The stopping parameters were set as follows: minimum *n* cases =5, maximum *n* cases = 10, minimum *n* child in node = 5, maximum *n* of nodes = 100. For the boosted trees regression, the learning rate was set to 0.1, with the following conditions: number of additive terms = 200, random test data proportion = 0.30, subsample proportion = 0.4. The stopping parameters were set as follows: minimum *n* of cases = 5, maximum *n* of levels = 10, minimum *n* in child node = 1, maximum *n* of nodes = 3.

The predictive performances of the constructed models were evaluated using the diagnostic indices of accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). These indices are calculated as follows Eq. 1 – 5 (Trevethan 2017):

$$Accuracy = \frac{correct\ predictions}{total\ predictions}\ x\ 100 \quad (1)$$

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative}\ x\ 100 \quad (2)$$

$$Specificity = \frac{True\ Negative}{True\ Negative + False\ Positive}\ x\ 100 \quad (3)$$

$$PPV = \frac{True\ Positive}{True\ Positive + False\ Positive}\ x\ 100 \quad (4)$$

$$NPV = \frac{True\ Negative}{True\ Negative + False\ Negative}\ x\ 100 \quad (5)$$

The values used for the calculation were taken from the results of the validation sets of the ANN-based classification models, and test sets for the SVM and tree-based classification models.

Results

Variations between DCM and NDCM NOD2 mutants based on the location, and nature of the mutations were observed. Most of the DCMs were located at the leucine-rich region (LRR) of NOD2, while NDCMs occurred mostly at the non-domain part of the protein (Fig. 1). For the nature of the mutations, conservative mutations accounted for 71 % of the NDCMs and only 3 % for the DCMs. Most NDCMs involved mutations to an aliphatic residue, while the DCMs exhibited a scattered type of mutations (Fig. 2).

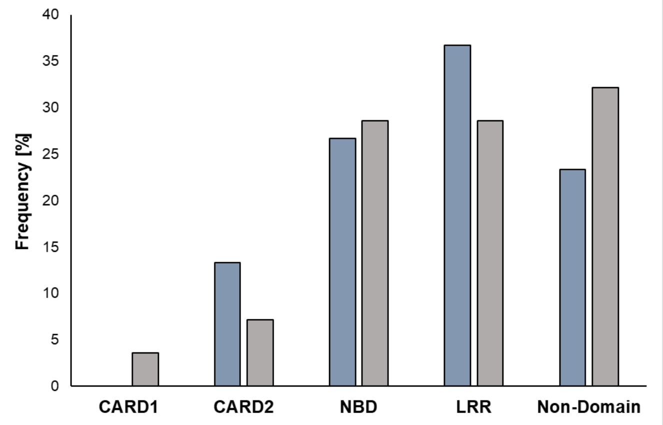

Fig. 1. Frequency of mutations based on the domain location within the NOD2 protein. Blue columns represent disease-causing mutation and gray columns represent non-disease-causing mutations. Designation of mutation type is based on Lesage *et al.* 2002, and ClinVar.

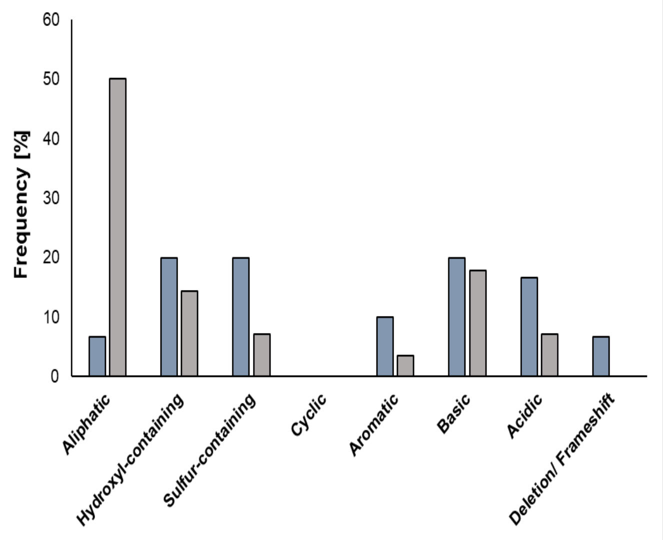

Fig. 2. Frequency of mutations based on the nature of the mutant amino acid. Blue columns represent disease-causing mutation and gray columns represent non-disease-causing mutations. Designation of mutation type is based on Lesage *et al.* 2002, and ClinVar.

Table 2. Confusion matrix for the comparison between the pathogenicity and functional impact of the NOD2 mutations.

	Deleterious	**Neutral**
DCM	14	15
NDCM	8	20

Table 3. Model architecture and performance of artificial neural network – based classification algorithms.

Model	Descriptor / Selection basis	Network structure	Algorithm	Prediction accuracy	Diagnostic performance
A	1.30 Full model	Multilayer perceptron Input layer: 30 Hidden layer: 22 Output layer: 2	BFGS 6	Training = 70.6 % Testing = 52.9 % Validation = 27.3 %	Sensitivity = 40 % Specificity = 17 % PPV = 28.6 % NPV = 25 %
B	1.10 Segmented	Multilayer perceptron Input layer: 10 Hidden layer: 8 Output layer: 2	BFGS 4	Training = 47.1 % Testing = 58.8 % Validation = 63.6 %	Sensitivity = 60 % Specificity = 67 % PPV = 60 % NPV = 66.7 %
C	11.20 Segmented	Multilayer perceptron Input layer: 10 Hidden layer: 8 Output layer: 2	BFGS 12	Training = 77.8 % Testing = 52.9 % Validation = 54.5 %	Sensitivity = 50 % Specificity = 60 % PPV = 60 % NPV = 50 %
D	21-30 Segmented	Multilayer perceptron Input layer: 10 Hidden layer: 8 Output layer: 2	BFGS 16	Training = 80 % Testing = 82.4 % Validation = 45.5 %	Sensitivity = 62.5 % Specificity = 100 % PPV = 100 % NPV = 75 %

The functional impact of the mutations on NOD2 was also assessed using PROVEAN. As seen in Table 2, a deleterious mutation weakly associates with NOD2 pathogenicity since it only accounts for 48 % of disease-causing mutations. This is in contrast to mutations with neutral functional impact, which account for 71 % of non-disease-causing mutations.

The difference between DCM and NDCM NOD2 mutants based on the 30 SOCN was initially probed

Table 4. Model architecture and performance of SVM – based classification algorithms.

Model	Descriptor / Selection basis	Prediction accuracy	Diagnostic performance	Gamma	Capacity	Support vectors
E	1.30 Full model	Training = 52.3 % Test = 46.7 % Overall = 50.5 % Validation = 40.9 %	Sensitivity = 0 % Specificity = 0 % PPV = 0 % NPV = 0 %	0.033	1.000	44 (44 bounded)
F	1.10 Segmented	Training = 52.3 % Test = 46.7 % Overall = 50.5 % Validation = 40.9 %	Sensitivity = 0 % Specificity = 0 % PPV = 0 % NPV = 0 %	0.100	1.000	44 (44 bounded)
G	11.20 Segmented	Training = 52.3 % Test = 46.7 % Overall = 50.5 % Validation = 40.9 %	Sensitivity = 0 % Specificity = 0 % PPV = 0 % NPV = 0 %	0.100	1.000	44 (44 bounded)
H	Segmented	Test = 46.7 % Overall = 50.5 % Validation = 40.9 %	Specificity = 0 % PPV = 0 % NPV = 0 %			(44 bounded)

Table 5. Model architecture and performance of Random Forest – based classification algorithms.

Model	Descriptor / Selection basis	Prediction accuracy	Diagnostic performance

I	1.30
	Full model
		Training = 85 %
		Test = 58 %
		Overall = 76 %
			Sensitivity = 37.5 %
			Specificity = 73 %
			PPV = 50 %
			NPV = 61.5 %

J	1.10
	Segmented
		Training = 65.7 %
		Test = 62.5 %
		Overall = 64.4 %
			Sensitivity = 50 %
			Specificity = 71 %
			PPV = 55.6 %
			NPV = 66.7 %

K	11.20
	Segmented
		Training = 77.1 %
		Test = 41.7 %
		Overall = 62.7 %
			Sensitivity = 30 %
			Specificity = 50 %
			PPV = 30 %
			NPV = 50 %

L	21-30
	Segmented
		Training = 74.2 %
		Test = 41.7 %
		Overall = 62.7 %
			Sensitivity = 30 %
			Specificity = 50 %
			PPV = 30 %
			NPV = 50 %

through one way-ANOVA and *k*-means clustering. As anticipated, ANOVA yielded a non-significant difference ($p > 0.05$) between the two classes of NOD2 mutants, owing to the small variations in their respective SOCNs. A similar case was observed for the two-cluster solution created through k-means clustering. The first cluster only contained the truncated NOD2, the 1007fs, while the second cluster contained the remaining 42 NOD variants. Evidently, these two statistical methods were unable to discriminate DCM from NDCM NOD2 variants based on the 30 SOCNs.

Several binary classification models using the SOCNs as the predictors were then formulated utilizing various machine learning algorithms, which include ANN (Table 3), SVM (Table 4), Random Forest (Table 5), and Boosted Trees (Table 6).

The ANN-based classification algorithm and Boosted Trees yielded better predictive models compared with those produced using SVM and Random Forest. Optimization of the selection of the descriptors was done to improve the performance of the classification models, and was conducted systematically through the systematic segmentation of the 30 SOCNs. Comparing Tables 3-6, models D, M, and P yielded the reliable classification models as demonstrated by its satisfactory prediction accuracy for the training, test, and validation sets. However, model D was deemed as the best predictive model since it required only 10 predictors, and it exhibited a good balance in accuracy and diagnostic performance. The diagnostic indices of sensitivity, specificity, PPV, and NPV were used to further probe the performance of the constructed models. Thus, the overall performance of model D is satisfactory, based on the algorithm accuracy, which indicates how often the classifier is correct. Moreover, the sensitivity and specificity of the model are also satisfactory. Sensitivity demonstrates the ability of the model for positive classification, while specificity for negative identification (Wong and Lim 2011). PPV and NPV demonstrate how many were indeed true positives and true negatives from the ratings given by the classifier (Trevethan 2017).

Discussion

Understanding the relationship between NOD2 mutant type, and CD susceptibility is of paramount importance, considering the pivotal role of NOD2 in CD pathogenesis. However, such connection remains to be fully understood. As what has been previously demonstrated from a study of 612

Table 6. Model architecture and performance of Boosted Trees – based classification algorithms.

Model	Descriptor / Selection basis	Prediction accuracy	Diagnostic performance

M	1.30
	Full model
		Training = 90.2 %
		Test = 72.2%
		Overall = 84.7 %
			Sensitivity = 88.9 %
			Specificity = 56 %
			PPV = 66.7 %
			NPV = 83.3 %

N	1.10
	Segmented
		Training = 73.2 %
		Test = 66.7 %
		Overall = 71.2 %
			Sensitivity = 55.6 %
			Specificity = 78 %
			PPV = 71.4 %
			NPV = 63.6 %

O	11.20
	Segmented
		Training = 95.1 %
		Test = 55.6 %
		Overall = 83.1 %
			Sensitivity = 55.6 %
			Specificity = 56 %
			PPV = 55.6 %
			NPV = 55.6 %

P	21-30
	Segmented
		Training = 82.9 %
		Test = 77.8 %
		Overall = 81.4 %
			Sensitivity = 77.8 %
			Specificity = 78 %
			PPV = 77.8 %
			NPV = 77.8 %

European CD patients, NOD2 mutations can either be disease-causing, or non-disease-causing (Lesage *et al.* 2002). Several studies have identified that certain NOD2 mutations, such as the frameshift mutation 1007fs, as markers or indicators of CD susceptibility (Ogura *et al.* 2001).

It is believed that the frameshift mutation leads to a truncated NOD2 in the leucine-rich region (LRR), thereby impairing its function. Consequently, most known CD DCMs occur at the LRR (Fig. 1). Apart from impairing the ligand-binding function of the protein, mutations at the LRR may also lead to a destabilized protein structure resulting to a loss of function for NOD2 (Maekawa *et al.* 2016). However, analysis on the functional impact of the mutations revealed that less than half of analyzed DCMs have deleterious functional impact. This suggests that mutational pathogenicity observes multiple possible mechanisms apart from impairment brought by the mutation. Aside from the location of the mutation, a striking difference between DCM and NDCM is the nature of the mutation. It was observed that the known DCM are mostly non-conservative (3 % conservative mutations), as opposed to NDCM that are 71 % conservative. While several diseases are also caused by conservative mutations, it is possible that the non-conservative mutations may have greater impact on the NOD2 protein. The non-conservative mutations may significantly alter the microenvironment in which the mutation is located owing to change in property of the amino acid substitution. On the other hand, the conservative non-disease-causing mutations may have little effect on the NOD2 loss-of-function since most conservative NDCM involve aliphatic to aliphatic substitutions (Fig. 2).

These findings therefore present a viable opportunity to deploy statistical methodologies in order to uncover associations between NOD2 mutant type and CD progression. However, creating predictive models of protein mutants based on the primary structure is challenging due to the minute variations introduced by the point mutations. Sequence – order coupling numbers are therefore ideal descriptors to be used, since this numerical representation of proteins reflects the sequence-order effect. For example, the NOD2 mutants A432V and A612V have identical amino acid composition, but the substitution is located at different positions. This positional difference is adeptly captured by this class of descriptors since these two mutants have different values for the 30 SOCNs. In addition, SOCN can fully describe the observed differences between DCM and NDCM with respect to the location and nature of the mutation. This frequently used protein descriptor class is based on distance matrices derived from amino acids, their sequence-order, and physicochemical properties (Schneider and Wrede 1994; Chou 2000). The 30 different SOCN represents the rank of the SOCN. For example, the first SOCN describes the coupling of adjacent residues, the second SOCN describes the coupling of between all second most contiguous residues, so forth. Thus, this protein descriptor class can potentially reveal hidden association between NOD2 mutation effect and CD susceptibility.

The presented ANN classifier provides a proof-of-concept that predicting the onset of CD from NOD2 protein variant is possible. The presented classification model (model D) is reliable after considering that the other models exhibited overfitting as characterized by the high training set accuracy but extremely low test accuracy. In addition, the other models were unable to classify NDCM NOD2 variants, as demonstrated by the low scores obtained for specificity and NPV. Out of the 16 classification models created, only model D demonstrated a satisfactory accuracy for the training and test sets, in addition to respectable scores for the diagnostic indices. Statistical endeavors that aimed to enhance CD detection involved the formulation of machine learning classification algorithms based on endoscopic data (Mossotto *et al.* 2017), multivariate analysis of magnetic resonance spectroscopic data of gastrointestinal tissues (Bezabeh *et al.* 2001), neuro-fuzzy classifier based on multitudes of clinical data (Ahmed *et al.* 2017), and serological, genetic, and inflammatory markers-dependent random forest classifier (Plevy *et al.* 2013). Recently, an SVM classification model that can categorize individuals into healthy or CD patients based on exome variations was reported (Wang *et al.* 2019). The SVM classifier used over 10,000 genes for the classification, including the NOD2 gene. In the present study, the focus of the classifier is to categorize whether mutations are disease-causing or not, based on variations in the NOD2 protein. Thus, the present study has therefore demonstrated a new way that is solely dependent on the sequence of the NOD2 protein which can potentially enhance detection and diagnostics. While the created algorithm is presently constrained by availability of data for training and validation, it is expected that the model will improve its predictive ability as more mutation types are incorporated in the system. It should also be taken into consideration the relationship between population predisposition, NOD2 mutations, and CD progression. For example, NOD2 mutations were absent in Japanese CD patients (Yamazaki *et al.* 2002). Thus, the present utility of the algorithm may be restricted to the population group from which the data was taken.

Conclusion

Differences between NOD2 Crohn's disease-causing mutations and non-disease-causing mutations were observed. The variations were related to the location and nature of the mutations. Based from these, a comprehensive statistical analyses were conducted which demonstrated the possibility of predicting the association of NOD2 mutations with CD susceptibility. The ANN model exhibited satisfactory capability to predict whether a specific NOD2 mutation is associated with the onset of CD, based on sequence-order coupling numbers. The presented classifier sets itself apart from previously reported algorithms by using the primary structure of NOD2 as the predictor. The formulated predictive model is potentially useful for the enhanced diagnosis and understanding of Crohn's Disease.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

Ahmed SS, Dey N, Ashour AS, Sifaki-Pistolla D, Balas-Timar D, Balas VE, Tavares JMRS (2017) Effect of fuzzy partitioning in Crohn's disease classification: a neuro-fuzzy-based approach. Med. Biol. Eng. Comput. 55: 101-115.

Bezabeh T, Somorjai RL, Smith IC, Nikulin AE, Dolenko B, Bernstein CN (2001) The use of 1H magnetic resonance spectroscopy in inflammatory bowel diseases: distinguishing ulcerative colitis from Crohn's disease. Am. J. Gastroenterol. 96: 442-448.

Choi Y, Sims GE, Murphy S, Miller JR, Chan AP (2012) Predicting the functional effect of amino acid substitutions and indels. PLoS One 7: e46688.

Chou KC (2000) Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. Biochem. Biophys. Res. Commun. 278: 477-483.

Cuthbert AP, Fisher SA, Mirza MM, King K, Hampe J, Croucher PJP, Mascheretti S, Sanderson J, Forbes A, Mansfield J, Schreiber S, Lewis CM, Mathew CG (2002) The contribution of NOD2 gene mutations to the risk and site of disease in inflammatory bowel disease. Gastroenterology 122: 867-874.

Economou M, Trikalinos TA, Loizou KT, Tsianos EV, Ioannidis JPA (2004) Differential effects of NOD2 variants on Crohn's disease risk and phenotype in diverse populations: A metaanalysis. Am. J. Gastroenterol. 99: 2393-2404.

Er O, Temurtas F, Tanrıkulu AÇ (2010) Tuberculosis disease diagnosis using Artificial Neural Networks. J. Med. Syst. 34: 299-302.

Flamant M, Roblin X (2018) Inflammatory bowel disease: towards a personalized medicine. Therap. Adv. Gastroenterol. 11: 1-15.

Hampe J, Grebe J, Nikolaus S, Solberg C, Croucher PJP, Mascheretti S, Jahnsen J, Moum B, Klump B, Krawczak M, Mirza MM, Foelsch UR, Vatn M, Schreiber S (2002) Association of NOD2 (CARD 15) genotype with clinical course of Crohn's disease: A cohort study. Lancet 359: 1661-1665.

Khan J, Wei JS, Ringnér M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C, Meltzer PS (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat. Med. 7: 673-679.

Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR (2014) ClinVar: Public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 42: D980-5.

Lesage S, Zouali H, Cézard J-P, Colombel J-F, Belaiche J, Almer S, Tysk C, O'Morain C, Gassull M, Binder V, Finkel Y, Modigliani R, Gower-Rousseau C, Macry J, Merlin F, Chamaillard M, Jannot A-S, Thomas G, Hugot J-P (2002) CARD15/NOD2 mutational analysis and genotype-phenotype correlation in 612 patients with inflammatory bowel disease. Am. J. Hum. Genet. 70: 845-857.

Maekawa S, Ohto U, Shibata T, Miyake K, Shimizu T (2016) Crystal structure of NOD2 and its implications in human disease. Nat. Commun. 7: 11813.

Mossotto E, Ashton JJ, Coelho T, Beattie RM, MacArthur BD, Ennis S (2017) Classification of paediatric inflammatory bowel disease using machine learning. Sci. Rep. 7: 2427.

Niess JH, Klaus J, Stephani J, Pfluger C, Degenkolb N, Spaniol U, Mayer B, Lahr G, von Boyen GBT (2012) NOD2 polymorphism predicts response to treatment in Crohn's disease-first steps to a personalized therapy. Dig. Dis. Sci. 57: 879-886.

Ogura Y, Bonen DK, Inohara N, Nicolae DL, Chen FF, Ramos R, Britton H, Moran T, Karaliuskas R, Duerr RH, Achkar J-P, Brant SR, Bayless TM, Kirschner BS, Hanauer SB, Nunez G, Cho JH (2001) A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. Nature 411: 603-606.

Plevy S, Silverberg MS, Lockton S, Stockfisch T, Croner L, Stachelski J, Brown M, Triggs C, Chuang E, Princen F, Singh S (2013) Combined serological, genetic, and inflammatory markers differentiate Non-IBD, Crohn's disease, and ulcerative colitis patients. Inflamm. Bowel Dis. 19: 1139-1148.

Schneider G, Wrede P (1994) The rational design of amino acid sequences by artificial neural networks and simulated molecular evolution: de novo design of an idealized leader peptidase cleavage site. Biophys. J. 66: 335-344.

Sidiq T, Yoshihama S, Downs I, Kobayashi KS (2016) Nod2: A critical regulator of ileal microbiota and Crohn's disease. Front. Immunol. 7: 367.

Strober W, Watanabe T (2011) NOD2, an intracellular innate immune sensor involved in host defense and Crohn's disease. Mucosal Immunol. 4: 484-495.

Trevethan R (2017) Sensitivity, specificity, and predictive values: foundations, pliabilities, and pitfalls in research and practice. Front. Public Heal. 5: 307.

Wang Y, Miller M, Astrakhan Y, Petersen B-S, Schreiber S, Franke A, Bromberg Y (2019) Identifying Crohn's disease signal from variome analysis. Genome Med. 11: 59.

Wong HB, Lim GH (2011) Measures of diagnostic accuracy: Sensitivity, specificity, PPV and NPV. Proc. Singapore Healthc. 20: 316-318.

Xiao N, Cao DS, Zhu MF, Xu QS (2015) Protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. Bioinformatics 31: 1857-1859.

Yamamoto S, Ma X (2009) Role of Nod2 in the development of Crohn's disease. Microbes Infect. 11: 912-918.

Yamazaki K, Takazoe M, Tanaka T, Kazumori T, Nakamura Y (2002) Absence of mutation in the NOD2/CARD15 gene among 483 Japanese patients with Crohn's disease. J. Hum. Genet. 47: 469-472.