SEDGE

THE
SOLUTIONS
EDGE

SEDGE PREDICTION MODEL

for patients with Crohn's Disease treated with Adalimumab and Correlation Analysis of Genetic Data SNP Models

www.sedge.ai

PREDICTION MODEL OF SEDGE FOR PATIENTS WITH CROHN'S DISEASE, TREATED WITH ADALIMUMAB AND CORRELATION ANALYSIS OF GENETIC DATA SNP MODELS

SEDGE VALIDATION REPORT 1

Present report includes testing SEDGE software on a medium scale genetic database. The database consisted of 97 individuals with Crohn's disease who were treated with adalimumab drug and their genetic data for 100 co-dominant, dominant and recessive single nucleotide polymorphisms (SNP) models. Adalimumab is a biologic drug, an antibody against TNF-α. In autoimmune diseases, TNF-α binds to TNF-α receptors, leading to the inflammatory response and by binding of drug antibodies to TNF-α, adalimumab reduces this inflammatory response. Additionally, each individual in the dataset had assigned a response to the adalimumab after 4 weeks, 12 weeks, 20 weeks and 30 weeks of treatment. The response to adalimumab was based on Inflammatory Bowel Disease Questionnaire (IBDQ), which is a widely used questionnaire for assessment of health-related quality of life in patients with inflammatory bowel diseases.

The validation report consists of:

- Installation of SEDGE;
- Preparation of raw data;
- Database upload;
- Data analysis and visualization;
- Interpretation and comparison with existing data;
- SEDGE prediction evaluation;
- Recommendations.

Installation of SEDGE

SEDGE was installed with remote assistance from Solverminds. During installation we didn't encounter any problems and SEDGE was successfully installed on a server computer.

Preparation of raw data

Our database was prepared in Microsoft Excel and was carefully checked for any entry mistakes. All characters representing missing values were deleted, leaving empty cells. All variables in the present data are nominal with numerical coding.

Database upload

The database was converted to .csv comma delimited file format.

Data analysis

Database with 97 individuals and 100 SNP models was successfully loaded into SEDGE software.

Before further analysis, SEDGE asked us to delete the "Null/NA" values. After the deletion of missing values, the database consisted of 49 individuals, since SEDGE deleted entire rows that contained missing values.

Preprocessing

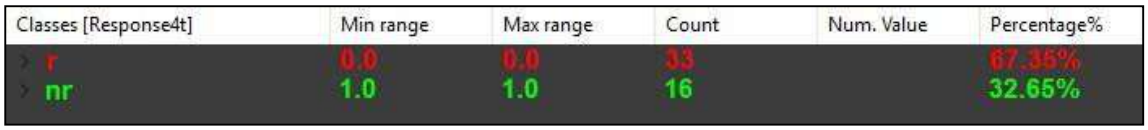
The "Preview" and "Statistics" tabs offer nice data overview and handling. Graphical analysis provides a quick overview of the data and is simple to use. The correlation type of graphical analysis is suitable for genetic data.

Under the "Pre process" tab, the response to adalimumab after 4 weeks, 12 weeks, 20 weeks and 30 weeks of treatment, was set as target attribute for further analyses. Each time point of measured response was independently set as a target attribute.

Visualization

Visualization was done with SEDGE version 1.5.0. The cluster of interest was the Level 1 Cluster 0, since we had responders and non-responders to the treatment with adalimumab. Results were ranked according to dCorrelation coefficient and first 15 with highest dCorr value were displayed and compared to existing data.

Results - target attribute: Response at 4 weeks



Classes [Response4t]	Min range	Max range	Count	Num. Value	Percentage%
r	0.0	0.0	33		67.35%
nr	1.0	1.0	16		32.65%

Attribute name	Attribute ID	Attribute type	dCor
CCNY_rs12777960_AA_ACvs_CC	76	Numeric	0.3392
CCNY_rs12777960	74	Numeric	0.32816
IL4R_Q576R_rs1801275	14	Numeric	0.325523
IL4R_Q576R_rs1801275_AAvs_AG_GG	15	Numeric	0.324663
DLG5_G113A_rs1248696	44	Numeric	0.260102
DLG5_G113A_rs1248696_CCvs_CT_TT	45	Numeric	0.260102
CTLA4_CT60_rs3087243_AAvs_AG_GG	27	Numeric	0.260102
10p11_2_rs4934697_CCvs_CT_TT	84	Numeric	0.259768
10p11_2_rs4934697	83	Numeric	0.253727
ORMDL3_rs2872507_AA_AGvs_GG	37	Numeric	0.249727
ORMDL3_rs2872507	35	Numeric	0.232039
ECM1S41G_AA_AGvs_GG	70	Numeric	0.213201
IL13_rs1295686	86	Numeric	0.212914
IL10_rs3024505_CC_CTvs_TT	91	Numeric	0.2076
IL13_rs1295686_AA_AGvs_GG	88	Numeric	0.198488

Results - target attribute: Response at 12 weeks

Classes [Response12t]	Min range	Max range	Count	Num. Value	Percentage%
r	0.0	0.0	35		71.43%
nr	1.0	1.0	14		28.57%

Attribute name	Attribute ID	Attribute type	dCor
TIMP_1_372T_C_rs4898_CC_CTvs_TT	64	Numeric	0.294514
PTGER4_5p13A_G_rs10512734	32	Numeric	0.281458
NOD2_L17fsn_rs2066847	20	Numeric	0.279372
NOD2_L17fsn_rs2066847_C_C_C_wtvs_wt_wt	22	Numeric	0.279372
IL4R_Q576R_rs1801275	14	Numeric	0.272692
IL4R_Q576R_rs1801275_AAvs_AG_GG	15	Numeric	0.271163
IL13_rs1295686_AA_AGvs_GG	88	Numeric	0.266053
ECM1S41G_AA_AGvs_GG	70	Numeric	0.258199
PTGER4_5p13A_G_rs10512734_AAvs_AG_GG	33	Numeric	0.258199
IL13_rs1295686	86	Numeric	0.257094
TIMP_1_372T_C_rs4898	62	Numeric	0.254668
CASP9_rs4645983_CC_CTvs_TT	58	Numeric	0.23625
PTGER4_5p13A_G_rs10512734_AA_AGvs_GG	34	Numeric	0.23625
DLG5_G113A_rs1248696_CCvs_CT_TT	45	Numeric	0.23625
DLG5_G113A_rs1248696	44	Numeric	0.23625

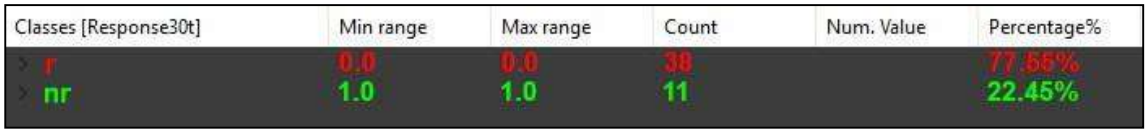
Results - target attribute: Response at 20 weeks

Classes [Response20t]	Min range	Max range	Count	Num. Value	Percentage%
r	0.0	0.0	34		69.39%
nr	1.0	1.0	15		30.61%

Attribute name	Attribute ID	Attribute type	dCor
IL4R_Q576R_rs1801275_AAvs_AG_GG	15	Numeric	0.298068
IL4R_Q576R_rs1801275	14	Numeric	0.296073
CASP9_rs4645983	56	Numeric	0.253805
CASP9_rs4645983_CC_CTvs_TT	58	Numeric	0.248112
10q21_rs10509115_AA_AGvs_GG	82	Numeric	0.235008
CASP9_rs4645983_CCvs_CT_TT	57	Numeric	0.224045
10p11_2_rs4934697_CC_CTvs_TT	85	Numeric	0.223906
CCNY_rs12777960_AAvs_AC_CC	75	Numeric	0.223906
PTPN22_R620W_rs2476601_AA_AGvs_GG	31	Numeric	0.214949
PTPN22_R62W_rs2476601	29	Numeric	0.21307

IL12RB1_Q214R_rs11575934	53	Numeric	0.211256
PTGER4_5p13A_G_rs10512734_AAvs_AG_GG	33	Numeric	0.207892
PTGER4_5p13A_G_rs10512734	32	Numeric	0.206305
IL12RB1_Q214R_rs11575934_AAvs_AG_GG	54	Numeric	0.198455
NOD2_L17fsn_rs2066847_C_C_C_wtvs_wt_wt	22	Numeric	0.173595

Results - target attribute: Response at 30 weeks



Classes [Response30t]	Min range	Max range	Count	Num. Value	Percentage%
r	0.0	0.0	38		77.55%
nr	1.0	1.0	11		22.45%

Attribute name	Attribute ID	Attribute type	dCor
FCGR3A_158_V_F_rs396991_GG_GTvs_TT	73	Numeric	0.413876
FCGR3A_158_V_F_rs396991	71	Numeric	0.343097
IL13_rs1295686_AA_AGvs_GG	88	Numeric	0.279394
IL13_rs1295686	86	Numeric	0.278294
IL12RB1_Q214R_rs11575934	53	Numeric	0.277427
CCNY_rs12777960	74	Numeric	0.277153
IL12RB1_Q214R_rs11575934_AAvs_AG_GG	54	Numeric	0.274523
NOD2_G98R_rs2066845_CCvs_CG_GG	18	Numeric	0.268272
CCNY_rs12777960_AA_ACvs_CC	76	Numeric	0.268272
ATG16L1_promC_T_rs10210302	23	Numeric	0.259154
ATG16L1_promC_T_rs10210302_CC_CTvs_TT	25	Numeric	0.251175
NOD2_L17fsn_rs2066847_C_C_C_wtvs_wt_wt	22	Numeric	0.237661
NOD2_L17fsn_rs2066847	20	Numeric	0.237661
PTGER4_5p13A_G_rs10512734_AAvs_AG_GG	33	Numeric	0.233626
10p11_2_rs4934697	83	Numeric	0.225431

Interpretation and comparison with existing data

The top 15 selected SNPs were identical in all four analyses. Discrepancy was observed between SNPs CTLA4_CT6_rs3087243, IL10_rs3024505, 10q21_rs10509115, NOD2_G908R_rs2066845 and ORMDL3_rs2872507.

In the first analysis, where our target attribute was response to adalimumab at 4 weeks of treatment, the SEDGE visualization version 1.5.0., correctly classified CCNY_rs12777960 polymorphism with the highest dCor value. This particular polymorphism was also identified in our published study (Koder et al., 2015), which evaluated genetic polymorphisms in regard to response to adalimumab in Crohn's disease patients.

In the second analysis, where the target attribute was set to the response to the treatment at week 12, SEDGE also identified PTGER4_5p13A_G_rs10512734 polymorphism as the second highest ranked attribute after TIMP_1_372T_C_rs4898. rs10512734 was also identified as a suggestive SNP for response to adalimumab treatment (Koder et al., 2015). The highest ranked attribute in this analysis was TIMP_1_372T_C_rs4898. Levels of TIMP, which is an inhibitor of the matrix metalloproteinases, are elevated in autoimmune lymphoproliferative states as shown in the study performed by Boggio and colleagues (Boggio et al., 2010), and hence could play a significant role in the treatment with adalimumab and should be further evaluated.

When the target attribute was set to the response to the treatment at 20 weeks, SEDGE correctly identified the CASP9_rs4645983 polymorphism as the second ranked attribute after IL4R_Q576R_rs1801275. CASP9_rs4645983 was also identified in our published study (Koder et al., 2015) IL4R is also identified as an associated SNP with asthma, which is also an autoimmune disease (Li et al., 2016). Furthermore, rs1801275 was significantly associated with blood erythrocytes levels, blood thrombocytes levels, serum bilirubin levels, serum sodium levels in Slovenian IBD patients treated with adalimumab, and additionally it was included as a significant response variable into linear regression with quantitative measurement of response after 20 weeks of treatment (Rebernak, 2015).

In the last analysis, the target attribute was set to response to adalimumab at 30 weeks of treatment. Both, the first and second ranked attributes, FCGR3A_158_V_F_rs396991 and IL13_rs1295686, were correctly classified using SEDGE. Both SNPs are known to be associated with response to the treatment with adalimumab (Koder et al., 2015; Rebernak, 2015).

In conclusion, according to dCor values, SEDGE visualization was able to find important relations between nominal genetic data and rank the important attributes (polymorphisms), which were found in previous studies. Furthermore, the top 6 highest SEDGE ranked SNPs were the same SNPs as those indentified as most significant SNPs using univariate statistical tests in our studies.

Table of identical matches of SNPs in all 4 analyses and correlation to existing data.

CCNY_rs12777960	Associated with IBD in Slovenian patients. Repnik et al. Annual Research & Review in Biology, ISSN: 2347-565X,Vol.: 8, Issue.: 3. Suggestive association with response to adalimumab. Koder et al. Pharmacogenomics. 2015;16(3):191- 204. doi: 10.2217/pgs.14.172. Identified as statistically significant factor influencing IBDQ response using univariate and multivariate statistical approach. Master's thesis, J. Rebernak.
IL4R_Q576R_rs1801275	Suggestive association with asthma, which is also an immune disease. Li et al. Genet Mol Res. 2016; 15(4). doi: 10.4238/gmr15048873.
DLG5_G113A_rs1248696	Association with reduced risk for IBD in Europeans. Li et al. Sci Rep. 2016; 6:33550. doi: 10.1038/srep33550.
10p11_2_rs4934697	Associated with refractory Crohn's disease in Slovenian patients. Repnik et al. Annual Research & Review in Biology, ISSN: 2347-565X,Vol.: 8, Issue.: 3.
ECM1S41G	Strong association with hemoglobin levels in Slovenian patients with IBD treated with adalimumab. Master's thesis, J. Rebernak.
IL13_rs1295686	Suggestive association with response to adalimumab. Koder et al. Pharmacogenomics. 2015;16(3):191- 204. doi: 10.2217/pgs.14.172. Identified as statistically significant factor influencing IBDQ response using univariate statistical approach. Master's thesis, J. Rebernak.
TIMP_1_372T_C_rs4898	Levels of TIMPs (inhibitors of the matrix metalloproteinases) are elevated in autoimmune lymphoproliferative syndromes, and hence worsen the apoptotic defect in these diseases. Boggio et al. 2010; 95(11):1897-904. doi: 10.3324/haematol.2010.023085.
PTGER4_5p13A_G_rs10512734	Suggestive association with response to adalimumab. Koder et al. Pharmacogenomics. 2015;16(3):191- 204. doi: 10.2217/pgs.14.172.

	Identified as statistically significant factor influencing IBDQ response using univariate and multivariate statistical approach. Master's thesis, J. Rebernak.
NOD2_L17fsn_rs2066847	Identified in GWA study. Jostins et al. Nature. 2012 Nov 1;491(7422):119-24. doi: 10.1038/nature11582.
CASP9_rs4645983	Suggestive association with response to adalimumab. Koder et al. Pharmacogenomics. 2015;16(3):191- 204. doi: 10.2217/pgs.14.172. Identified as statistically significant factor influencing IBDQ response using univariate statistical approach. Master's thesis, J. Rebernak.
PTPN22_R620W_rs2476601	Identified autoimmune-associated variant in patients with rheumatoid arthritis. Bayley et al. Ann Rheum Dis. 2015 Aug;74(8):1588-95. doi: 10.1136/annrheumdis-2013-204796. Epub 2014 Mar 24.
IL12RB1_Q214R_rs11575934	Located in IL12RB1 gene, which codes for the receptor of IL-12 cytokine binding. IL-12 is secreted by immune cells in response to antigens.
FCGR3A_158_V_F_rs396991	Identified as statistically significant factor influencing IBDQ response using multivariate statistical approach. Master's thesis, J. Rebernak.
ATG16L1_promC_T_rs10210302	Identified as associated SNP for adalimumab treatment response, which is defined as decreased CRP levels in Slovenian IBD patients. Koder et al. Pharmacogenomics. 2015;16(3):191-204. doi: 10.2217/pgs.14.172.

Table of other ranked SNPs and correlation to existing data:

CTLA4_CT60_rs3087243	Associated with IBD in Slovenian patients. Repnik and Potocnik. DNA Cell Biol. 2010 Oct;29(10):603-10. doi: 10.1089/dna.2010.1021.
IL10_rs3024505	Identified in GWA study. Jostins et al. Nature. 2012 Nov 1;491(7422):119-24. doi: 10.1038/nature11582.
10q21_rs10509115	Identified as statistically significant factor influencing IBDQ response using multivariate statistical approach. Master's thesis, J. Rebernak.
NOD2_G908R_rs2066845	Crohn's disease associated variant. Bonen et al. Gastroenterology. 2003 Jan;124(1):140-6.
ORMDL3_rs2872507	Identified as statistically significant factor influencing IBDQ response using

	multivariate statistical approach. Master's thesis, J. Rebernak.

Table of the direct comparison of SEDGE results to our results.

Our rang 1 SNPs				SEDGE rang 1 SNPs		
Week	**SNP**	**P Value**	**Rang No. in SEDGE Top 15**	**Week**	**SNP**	**dCor**
4	CCNY rs12777960	0.0156	1	4	CCNY_rs12777960	0.3392
12	PTGER4 rs10512734	0.0276	2	12	TIMP_1 372T_C_rs4898	0.2945
20	Other variable	/	/	20	IL4R_Q576R_rs1801275	0.2981
30	CCNY rs12777960	0.037	6	30	FCGR3A_158_V_F_rs396991	0.4139

SEDGE prediction evaluation

Prediction models were built on one half of the database, of which individuals/cases were randomly selected amongst responders and nonresponders. Prediction was done using SEDGE fitting.

Prediction model according to 4 weeks of treatment:

Training set: 25 individuals

Case set: 24 individuals

SEDGE report set: 24 individuals

		IBDQ	
		Nonresponder	**Responder**
SEDGE	**Nonresponder**	**0**	**5**
	Responder	**8**	**11**

Sensitivity: 0

Specificity: 0.69

Accuracy: 0.46

Prediction model according to 12 weeks of treatment:

Training set: 25 individuals

Case set: 24 individuals

SEDGE report set: 24 individuals

		IBDQ	
		Nonresponder	**Responder**
SEDGE	**Nonresponder**	**2**	**6**
	Responder	**5**	**11**

Sensitivity: 0.29

Specificity: 0.65

Accuracy: 0.54

Prediction model according to 20 weeks of treatment:

Training set: 23 individuals

Case set: 26 individuals

SEDGE report set: 24 individuals (2 individuals from case set are missing in the report)

		IBDQ	
		Nonresponder	**Responder**
SEDGE	**Nonresponder**	**3**	**7**
	Responder	**5**	**9**

Sensitivity: 0.38

Specificity: 0.56

Accuracy: 0.5

Prediction model according to 30 weeks of treatment:

Training set: 27 individuals

Case set: 22 individuals

SEDGE report set: 20 individuals (2 individuals from case set are missing in the report)

		IBDQ	
		Nonresponder	**Responder**
SEDGE	**Nonresponder**	**1**	**6**
	Responder	**5**	**8**

Sensitivity: 0.17

Specificity: 0.57

Accuracy: 0.45

Our training and case databases consisted of relatively small samples and SEDGE fitting prediction accuracy remained approximately 50%.

Recommendations for additional features

- Implementation of a function, which allows the user to decide how to exclude rows/cases with missing values (listwise or pairwise) - reported before.
- Addition of correlations algorithms for nominal by nominal measures and ordinal by ordinal measures - reported before.
- Implementation of Generalized Linear Models analysis, maybe through a user friendly interface.
- Value labels should be optional for categorical variables and visualization should be able to distinguish between nominal, ordinal or continuous data.
- In the genetics, the P values are reported to demonstrate the significance of the findings. The SEDGE visualization tool reports several coefficients, but only one p value without any indication to which test (dCor, F-Test, KS-Test,…) does it belong.

Moreover, the association thresholds (strong - weak correlation) of the correlation coefficients should also be stated.

- Improving the PDF report of the visualization; the visualization tool crashes when generating the report.
- The points/end-nodes in graphical cluster view should be named according to a

name/sample ID column, which is not captured by the analysis itself, but serves only as ID's.

Conclusions

The real value of the SEDGE tool for genetic and pharmacogenomic data is SEDGE visualization and prediction. Further evaluation is warranted to test the SEDGE on large scale databases and to find appropriate ideas for graphical visualization of the complex genetic data. Furthermore, the behavior of SEDGE, when dealing with quantitative