SEDGE

THE SOLUTIONS EDGE

SEDGE PLATFORM FOR LIFE SCIENCES

Validation Report

www.sedge.ai

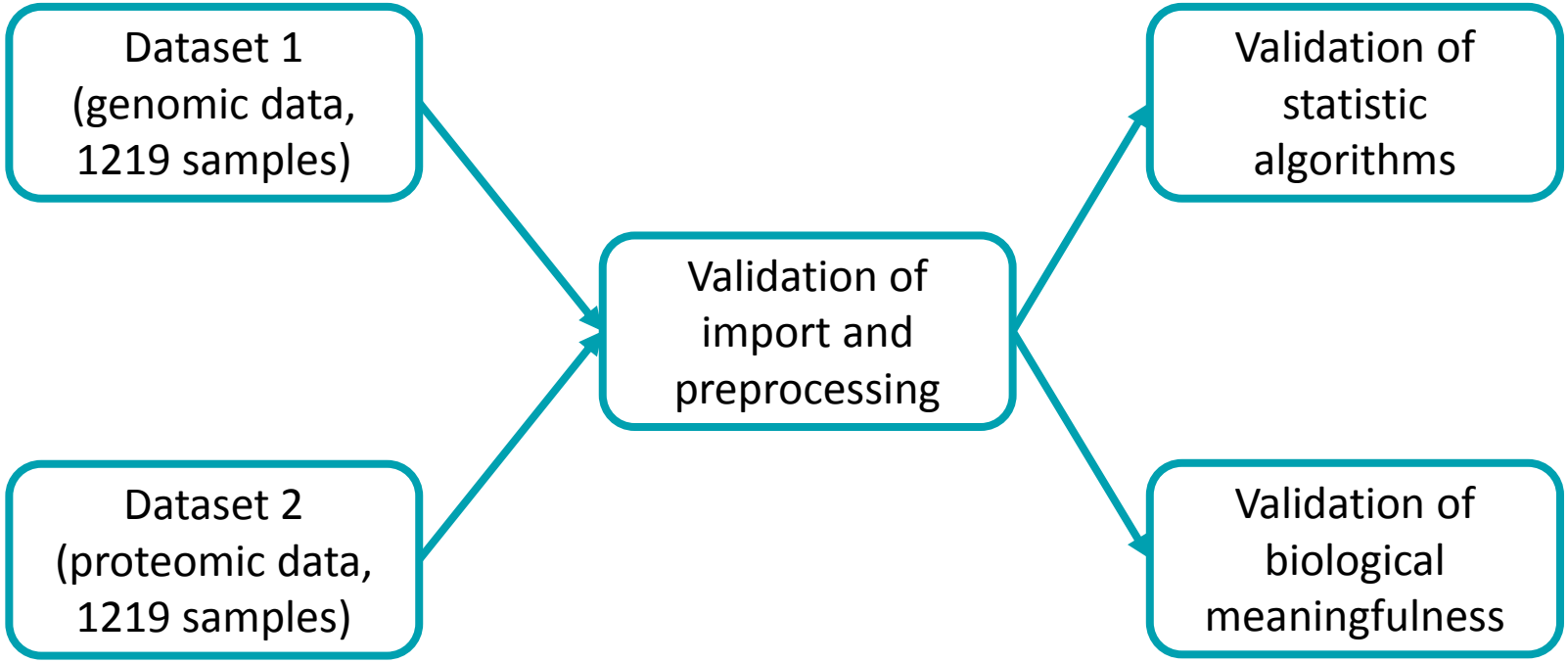
BioSistemika

RESEARCHERS DESERVE **CLEVER SOFTWARE**

SEDGE Platform for Life Sciences

Validation Report

Overview


Dataset 1 (genomic data, 1219 samples)
Dataset 2 (proteomic data, 1219 samples)
Validation of import and preprocessing
Validation of statistic algorithms
Validation of biological meaningfulness

Dataset 1

Brautbar et al 2012 – a pharmacogenomics study with patient SNP data and treatment response levels quantified by reduction of plasma APOC3 levels

Source: publicly available dataset, downloaded from: NCBI GEO Accession # GSE34945

Title of study: **LPL gene variants affect apoC-III response to combination therapy of statins and fenofibric acid in a randomized clinical trial of individuals with mixed dyslipidemia**

Disease: mixed dyslipidemia

Data Type: Single Nucleotide Polymorphism (SNP)

Groups: total 1219 samples (patients), treatment group 1: fenofibric acid, treatment group 2: fenofibric acid+statin, treatment group 3: statin alone

Dataset 2

Zupancic et al 2014 – a study of tumor plasma protein marker search using antibody arrays

Source: author's tab delimited file

Title of study: **Identification of plasma biomarker candidates in glioblastoma using an antibody-array-based proteomic approach**

Disease: glioblastoma tumor

Data type: protein abundance (analogous to gene expression data)

Groups: 17 healthy and 17 diseased individuals

Validation of import and preprocessing

Two datasets were supplied by Biosistemika Ltd:

VALIDATION REPORT: Both datasets were successfully preprocessed and imported to SEDGE.

Validation of Dataset 1

The authors of the study used a multivariate linear regression and two-way ANOVA for percent change in apoCIII level. They identified three SNPs associated with high treatment success of the drug combination of fenofibric acid and statin located in the lipase (LPL) gene region (Table 1).

Using SEDGE, the correlation between SNP alleles and apoCIII level reduction was calculated for the combination treatment. The correlation list was exported and sent to Biosistemika.

VALIDATION REPORT: The three LPL SNP-s were in the top 25 of the correlation list. SNP-s located in gene coding regions were annotated with Gene names and descriptions (Table 2). All three SNP-s from author's publication were ranked high in the SEDGE results top list: rs249 (1st), sr1801177 (14th) and rs7016529 (21st).

TABLE 1. Significant associations using multivariate regression analysis for percent change in apoC-III adjusted for age, sex, body mass index, smoking, baseline trait level, baseline triglyceride level, and diabetes by treatment group

				Combination therapy		Statin monotherapy		FA monotherapy	
Gene	SNP	MAF	Minor allele	P	Beta	P	Beta	P	Beta
LPL	rs1801177	2.0%	A	1.1×10^{-6}	37.1	0.59	2.2	0.67	−3.2
LPL	rs7016529	2.1%	G	3.0×10^{-6}	34.8	0.71	1.4	0.67	−3.2
LPL	rs249	7.5%	G	1.5×10^{-6}	15.6	0.3	2.2	0.08	−6.4

Based on NCBI Build 36.1; β, β coefficient; FA, fenofibric acid; MAF, minor allele frequency.

Validation of Dataset 1

Additionally, we can find another LPL SNP, rs7016529, in SEDGE results that was not reported in the author's report of the study. Also, SEDGE identified several SNP-s located in PPARA and CETP genes as well as on SPN in APOC2 as associated with the combination drug treatment success. PPARA gene encodes a major regulator of lipid metabolism in the liver and since fenofibrate is the agonist of this receptor SNP-s in this gene obviously can affect binding of the drug.

CETP and APOC2 both encode plasma proteins which facilitate the transport of triglycerides between the lipoproteins. The contribution of these SNP-s to combination treatment success makes sense because these genes are both connected to lipoprotein metabolism. Interestingly, another study reports that SNP-s in genes CETP and APOC2 are tentatively associated with fenofibrate response (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3485381/).

Validation of Dataset 1

Table 2: Top25 list of SNP-s correlated to apoCIII reduction using SEDGE platform. SNP-s are ordered by dCor value.

SNP ID	Pearson	Spearman	dCor	GENE	GENE NAME
rs249	0.189566489	0.110857717	0.162077206	LPL	lipoprotein lipase
rs4253772	0.130964379	0.103465186	0.150405595	PPARA	peroxisome proliferator-activated receptor alpha
rs4823613	0.13957997	0.133226008	0.145370408	PPARA	peroxisome proliferator-activated receptor alpha
rs5767700	0.134470049	0.102942663	0.137434319	PPARA	peroxisome proliferator-activated receptor alpha
rs1368516	0.133653855	0.108371028	0.132311592		
rs12330015	0.136884988	0.117548365	0.131300366	PPARA	peroxisome proliferator-activated receptor alpha
rs1842746	0.126128506	0.132614434	0.128812247		
rs2288911	0.122493974	0.132156929	0.127288494	APOC2	cholesteryl ester transfer protein, plasma
rs948384	0.107811143	0.0659661	0.126320819		
rs1800774	0.087225157	0.117016399	0.12551818	CETP	cholesteryl ester transfer protein, plasma
rs4253755	0.117125892	0.10100828	0.125262013	PPARA	peroxisome proliferator-activated receptor alpha
rs11076176	0.083076674	0.112343681	0.123851458	CETP	cholesteryl ester transfer protein, plasma
rs5882	0.085097123	0.096274069	0.118814573	CETP	cholesteryl ester transfer protein, plasma
rs1801177	0.130918789	0.079409136	0.118741622	LPL	lipoprotein lipase
rs2075440	0.117511417	0.113920419	0.118001859	MYO1H	myosin IH
rs1801706	0.110988919	0.070843699	0.117962044	CETP	cholesteryl ester transfer protein, plasma
rs3818730	0.12436117	0.098073397	0.116899677	RXRA	retinoid X receptor, alpha
rs1920325	0.105846657	0.118293026	0.115244773		
seq-rs1800777	0.118508532	0.129059252	0.11502426	CETP	cholesteryl ester transfer protein, plasma
rs12708974	0.096196312	0.088560354	0.114929844	CETP	cholesteryl ester transfer protein, plasma
rs7016529	0.123588202	0.076395916	0.111711634	LPL	lipoprotein lipase
rs1160985	0.086180242	0.105414738	0.10974908	TOMM40	translocase of outer mitochondrial membrane 40 homolog
rs1046661	0.093221211	0.104960196	0.10952089	EYA2	EYA transcriptional coactivator and phosphatase 2
rs281	0.079870173	0.091432341	0.107286917	LPL	lipoprotein lipase
rs5880	0.107692136	0.103030261	0.106145992	CETP	cholesteryl ester transfer protein, plasma

Validation of Dataset 2

The authors identified 11 plasma proteins that are statistically most strongly associated with the presence of glioblastoma (Table 3). SEDGE was used to find proteins that differentiate Healty and Diseased protein abundance. Generated SEDGE rank list was exported and sent to Biosistemika.

VALIDATION REPORT: SEDGE ranked 8 of 11 dentified protein markers in the top25 results. This shows that SEDGE is giving reasonable results without having in-depth knowledge of the biological problem. 'H.Pylori antigen' was the protein that was ranked first in the SEDGE list. This protein was not reported in the publication because it is, as the author of the study explains, a biological false positive (it doesn't have to anything with glioblastoma but with patient's age).

Validation of Dataset 2

Table 3: Potential plasma protein biomarker candidates for glioblastoma, as identified by the antibody array screening approach in Zupancic et al 2014

Protein name	Gene name	Difference in protein abundance	Molecular class	RANK IN SEDGE
Increased in GBM patients				
Ferritin light chain	FTL	1.65	Storage protein	6
Guanine nucleotide binding protein, alpha	GNAO1	1.65	G protein	85
S100 calcium binding protein A9	S100A9	1.66	Calcium binding protein	5
Deacreased in GBM patients				
Cyclin dependent kinase inhibitor 1B	CDKN1B	0.62	Cell cycle protein	10
FAS-associated death domain protein	FADD	0.52	Adapter molecule	2
Intercellular adhesion molecule 1	ICAM1	0.66	Adhesion molecule	61
DNA mismatch repair protein Mlh1	MLH1	0.57	DNA repair protein	17
Matrix metalloproteinase 11	MMP11	0.62	Metalloprotease	49
DNA polymerase, gamma	POLG	0.54	DNA polymerase	3
S phase kinase associated protein 1A (p19A)	SKP1	0.60	Ubiquitin proteasome protein	4
Sialyltransferase 8	ST8SIA1	0.59	Sialyltransferase	13

Conclusion

The results of validations described above show that SEDGE platform is suitable for research in life sciences and bioinformatics. It does however lack features specifically needed for life science applications in order to be competitive.