Supporting Information for

SIGNAL-3L 3.0: IMPROVING SIGNAL PEPTIDE PREDICTION THROUGH COMBINING ATTENTION DEEP LEARNING WITH WINDOW-BASED SCORING

Wei-Xun Zhang, Xiaoyong Pan, and Hong-Bin Shen

Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, and Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai, 200240, China

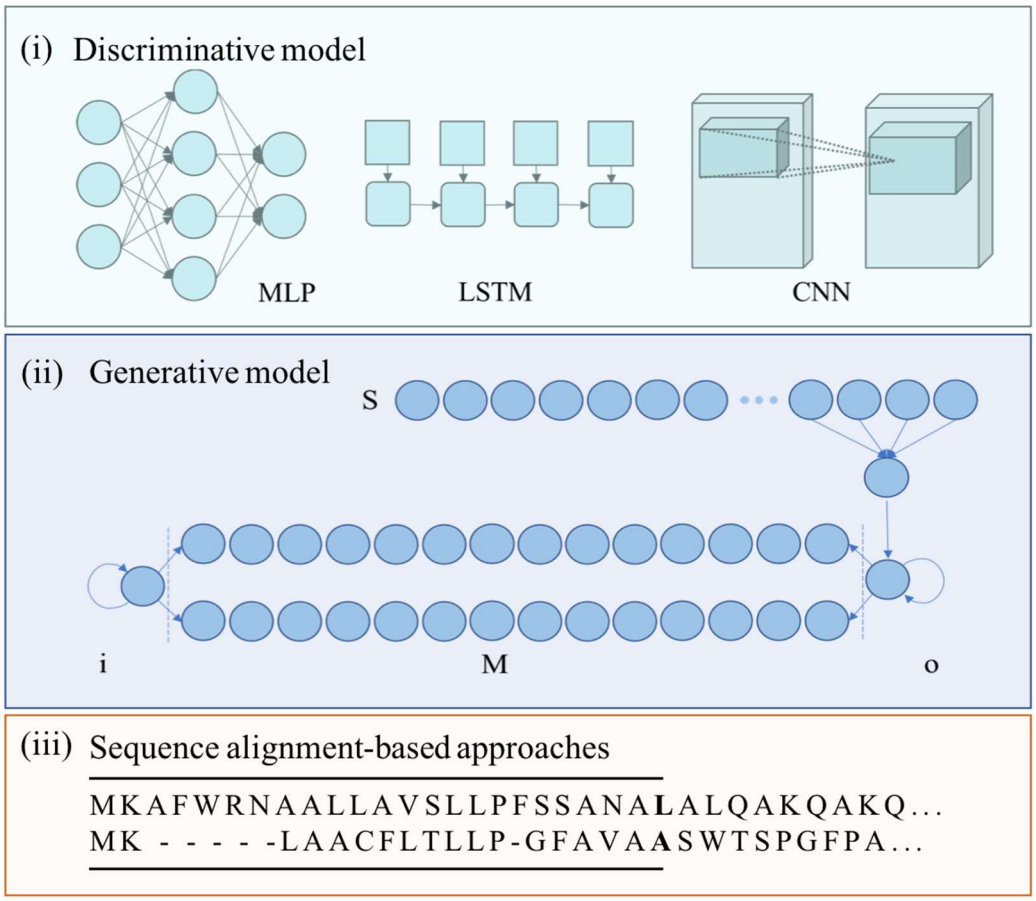

Figure S1. Existing prediction protocols of three different types of signal peptide predictors: (i) methods based on discriminative models, which inputs each residue feature (represented by a numeric feature vector) to generate a probability score; (ii) methods adopted Hidden Markov model to represent three regions (N−, H−, and C− regions) of signal peptide and gives a region-based estimation. (iii) sequence alignment based methods try to transfer the annotation from the annotated database according to evolutionary conservation.

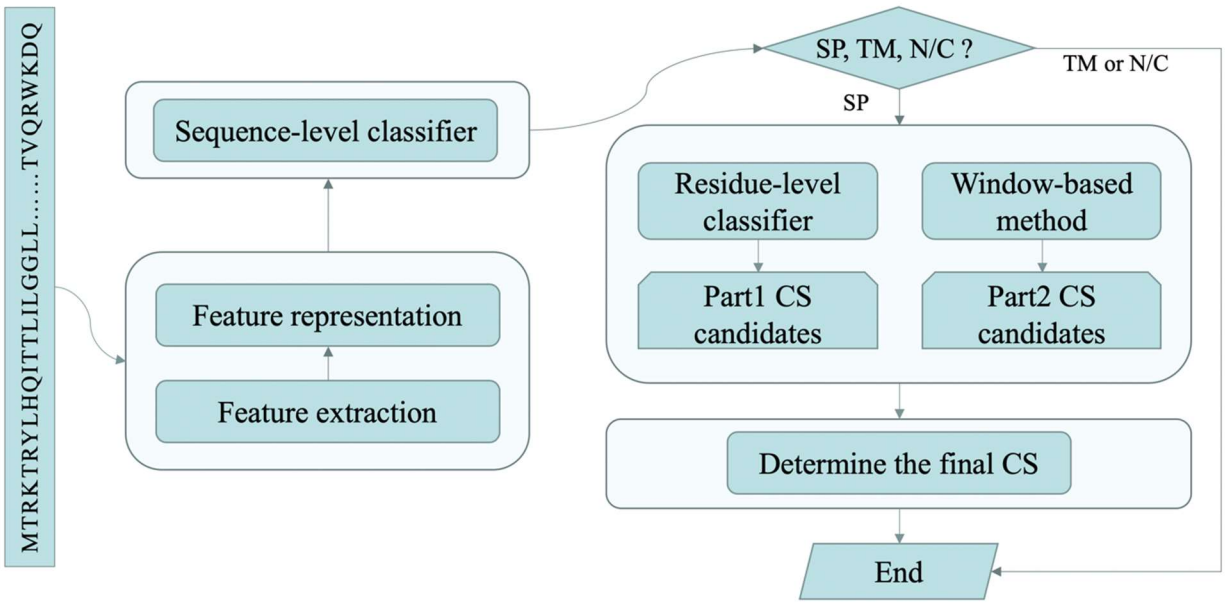

Figure S2. The overall prediction protocol of Signal-3L 3.0. First, we use programs (BLAST and HHblits) to encode each residue into a 50-dimensional vector. Then, a sequence-level classifier is used to judge whether the protein sequence contains signal peptide, transmembrane helix or is a globular protein. If the sequence is predicted to contain no signal peptide, the procedure is terminated. Otherwise, a residual-level classifier and a window-based method further generate the candidate cleavage sites, and the score of these two parts of candidate cleavage sites is weighted and fused to determine the final unique cleavage site. SP refers to signal peptide; TM refers to transmembrane helix; N/C refers to nuclear or cytosolic globular proteins; CS refers to cleavage site.

Window-based cleavage site scoring

The window-based scoring method uses a sliding window $-\xi_1, \ldots, -3, -2, -1, +1, +2, \ldots, +\xi_2$ along the protein sequence. Some amino acids occur more frequently on the three positions $(-3, -1, +1)$ when the cleavage site is located between -1 and +1. The probability of the true cleavage site between -1 and +1 within the kth sliding window with $\xi_1 = 13$ *and* $\xi_2 = 2$ is calculated as below:

$$\theta_k = p^+(R_{-13}) \ldots p^+(R_{-3})p^+(R_{-2})p^+(R_{-1}|R_{-3})p^+(R_{+1}|R_{-1})p^+(R_{+2}) - p^-(R_{-13}) \ldots p^-(R_{-3})p^-(R_{-2})p^-(R_{-1}|R_{-3})p^-(R_{+1}|R_{-1})p^-(R_{+2}) \quad \text{(S1)}$$

where $p^+(R_i)$and $p^-(R_i)$respectively correspond to the probability of residue R appearing at the i th position in the positive and negative datasets, respectively, and $p(R_i|R_i)$ is the conditional probability.

Model optimization

The nested 5-fold cross-validation is used to train and evaluate Signal-3L 3.0. In SignalP 5.0 dataset, the benchmark (Bench) dataset is a subset of the training data. When performing the nested 5-fold cross-validation procedure, the SignalP 5.0 dataset is randomly divided into five subsets. During each fold of nested 5-fold cross-validation, one is selected as the test set (contains one fifth of the Bench dataset), and the remaining 4 subsets are used to train four models. After completing

the 4-fold cross-validation of the internal loop, four models are obtained, and the average predicted results of these 4 models on this test set are obtained. Since the Bench dataset is a subset of the SignalP 5.0 dataset, some samples in this test set also belong to the Bench dataset. We pick out the prediction results of these shared samples in the Bench set. The above process is repeated 5 times, then the prediction results of Signal-3L 3.0 for the whole Bench dataset are obtained. In total, we obtain 20 models, which are used for webserver and predictions for SP19 dataset. The whole process is illustrated in Figure S3.

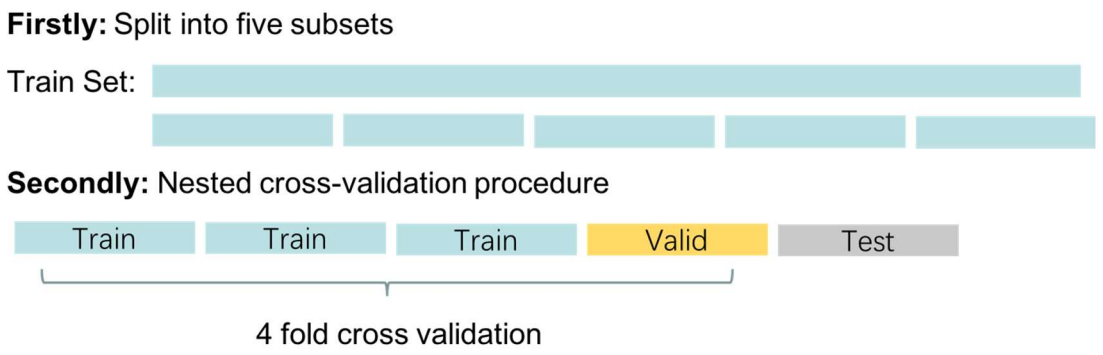

Figure S3. The process of nested cross-validation.

The hyperparameters, e.g. learning rate (0.1, 0.01, 0.001, and 0.0001), the number of Bi-LSTM layers (1, 2, 3) and hidden neurons (64, 128, 256), the number of 1D-CNN layers (1, 2, 3) and convolutional kernel size (2, 3, 4, 5), the number of neurons in the fully connected layer (64, 128 ,256), are optimized using cross-validation. The loss functions of the sequence-level classifier model and the residue-level classifier model are both cross entropy. Early-stopping is applied for model training. The early stop condition of the sequence-level classifier model is that the MCC score on the validation set for continuous multiple rounds does not increase. The early stop condition of the residue-level classifier model is that the signal peptide cleavage site recognition rate on the validation set for continuous multiple rounds does not increase.

In the nested 5-fold cross-validation, corresponding to the final 20 models, 20 scaling factor α are calculated. In Signal-3L 3.0, all the Bi-LSTM layers have 256 hidden neurons. In the sequence-level classifier, both two linear projection layers have a hidden size of 512, the three fully connected layers have a hidden size of 512, 512 and 3, respectively. In the residue-level classifier, the first-layer 1-D CNN has 64 feature maps with a kernel size of 3, the second-layer 1D-CNN has 32 feature maps with a convolution kernel size of 5. The two fully connected layers have a hidden size of 160 and 64, respectively. All codes were implemented using Pytorch version 1.1.0.

Performance evaluation metrics

$$Recall = \frac{TP}{TP+FN} \tag{S2}$$

$$Precision = \frac{TP}{TP+FP} \tag{S3}$$

$$F1-score = 2 \times \frac{Precision \times Recall}{Precision+} \tag{S4}$$

$$MCC = \frac{TP \cdot TN - FN \cdot FP}{\sqrt{(TP+FN)(TP+FP)(TN+FN)(TN+FP)}} \tag{S5}$$

In SP prediction, we define proteins containing signal peptides as positive samples and transmembrane helices and other non-secretory proteins as negative samples. In cleavage site prediction, we define the protein sequence correctly predicted to contain signal peptide in SP prediction and correctly identified the cleavage site as TP. The exact match between the predicted and true positions of cleavage site is denoted as the distance 0 according to SignalP 5.0 definition. Similarly, the performance for $\pm 1, \pm 2, \pm 3$ nearby the true cleavage sites is also calculated, which considers the prediction is correct if the distance between the position of the predicted cleavage site and the position of the true cleavage site is less than or equal to 1, 2, and 3, respectively.

Table S1. Details of the SignalP 5.0 benchmark dataset used in this study [a].

	SP		TM		N/C		Total	
Organism	Train	Bench	Train	Bench	Train	Bench	Train	Bench
Euk	2614	210	1044	318	13612	6929	17270	7457
GP	509	90	220	50	202	103	931	243
GN	189	25	50	25	140	64	379	114

[a] Euk: Eukaryotes; GP: Gram-positive; GN: Gram-negative; SP: proteins with signal peptide; TM: transmembrane-helical proteins; N/C=Nuclear and/or Cytosolic proteins; Bench: benchmark set.

Table S2. Details of the SP19 independent dataset in this study [a].

Organism	Eukaryotes	Gram-positive	Gram-negative
Raw (Test)	67 (33)	9 (9)	5 (3)

[a] Raw represents the number of newly annotated SP proteins from Uniprot Knowledgebase released from 2018_04 to 2019_07. Test represents the remaining proteins in Raw after removing redundancy against SignalP 5.0 dataset and the internal redundancy.

Table S3. Cross-validation performance comparison of one vs. two or three LSTM layers for signal peptide recognition measured by MCC.

Number of	Eukaryotes	Gram-negative	Gram-positive
LSTM layers	MCC	MCC	MCC
One	0.984	0.985	0.979
Two	0.980	0.978	0.968
Three	0.978	0.976	0.960

Table S4. Confusion matrices for the different type of predictions that Signal-3L 3.0 makes on the benchmark dataset.

	Eukaryotes			Gram-negative			Gram-positive		
real/predicted	SP	TM	CN	SP	TM	CN	SP	TM	CN
SP	198	4	8	87	3	0	24	1	0
TM	7	290	21	0	49	1	0	24	1
CN	12	21	6896	1	3	99	0	2	62

Table S5. Signal peptide recognition measured by MCC, the MCCs of other methods are directly from SignalP 5.0.

Method	Eukaryotes	Gram-negative	Gram-positive
	MCC	MCC	MCC
Signal-3L 3.0	0.925	0.965	0.974
SignalP 5.0	0.883	0.860	0.922
SignalP4.1	0.808	0.851	0.949
DeepSig	0.819	0.792	0.870
Signal-3L 2.0	0.597	0.806	0.922
LipoP	0.363	0.787	0.922
Philius	0.421	0.802	0.843
Phobius	0.510	0.818	0.818
PolyPhobius	0.456	0.844	0.852
PrediSi	0.553	0.802	0.781
PRED-LIPO	0.234	0.775	0.896
PRED-SIGNAL	0.272	0.724	0.830
PRED-TAT	0.326	0.769	0.830
Signal-CF	0.326	0.561	0.558
SOSUIsignal	0.375	0.693	0.722
SPEPlip	0.655	0.746	0.646
SPOCTOPUS	0.492	0.860	0.922
TOPCONS2	0.477	0.860	0.897

Table S6. Signal peptide cleavage site prediction measured by recall on SignalP 5.0 dataset. The results of other methods are directly from SignalP 5.0.

Method	Eukaryotes				Gram-negative				Gram-positive			
	0	±1	±2	±3	0	±1	±2	±3	0	±1	±2	±3
Signal-3L 3.0	0.676	0.752	0.800	0.829	0.667	0.722	0.789	0.822	0.760	0.800	0.880	0.880
SignalP 5.0	0.729	0.762	0.795	0.833	0.733	0.767	0.800	0.800	0.840	0.840	0.880	0.880
SignalP 4.1	0.695	0.729	0.762	0.786	0.644	0.711	0.733	0.744	0.840	0.840	0.840	0.840
DeepSig	0.624	0.652	0.690	0.724	0.600	0.656	0.667	0.678	0.760	0.760	0.840	0.840
Signal-3L 2.0	0.648	0.686	0.733	0.762	0.644	0.700	0.722	0.733	0.800	0.800	0.840	0.840
LipoP	0.343	0.386	0.419	0.448	0.733	0.767	0.789	0.789	0.600	0.600	0.640	0.640
Philius	0.619	0.686	0.743	0.781	0.700	0.744	0.789	0.811	0.600	0.600	0.600	0.600
Phobius	0.667	0.700	0.738	0.786	0.644	0.722	0.789	0.811	0.600	0.600	0.600	0.600
PolyPhobius	0.681	0.733	0.776	0.833	0.644	0.733	0.811	0.822	0.680	0.680	0.720	0.720
PrediSi	0.652	0.695	0.719	0.767	0.722	0.789	0.811	0.822	0.640	0.640	0.760	0.800
PRED-LIPO	0.095	0.114	0.152	0.181	0.467	0.522	0.567	0.600	0.760	0.760	0.760	0.760
PRED-SIGNAL	0.224	0.290	0.329	0.362	0.444	0.522	0.622	0.644	0.680	0.680	0.720	0.720
PRED-TAT	0.410	0.510	0.571	0.614	0.711	0.767	0.800	0.822	0.720	0.720	0.760	0.760
Signal-CF	0.652	0.676	0.724	0.762	0.689	0.711	0.744	0.778	0.720	0.720	0.800	0.800
SOSUIsignal	0.176	0.329	0.467	0.576	0.267	0.367	0.567	0.622	0.200	0.240	0.280	0.440

SPEPlip	0.710	0.733	0.771	0.810	0.611	0.678	0.722	0.733	0.680	0.680	0.720	0.720
SPOCTOPUS	0.390	0.533	0.686	0.757	0.467	0.689	0.833	0.867	0.640	0.760	0.800	0.880
TOPCONS2	0.371	0.505	0.638	0.729	0.544	0.622	0.733	0.767	0.240	0.320	0.400	0.440

Table S7. Signal peptide cleavage site prediction measured by precision on SignalP 5.0 dataset. The results of other methods are directly from SignalP 5.0.

Method	**Eukaryotes**			
	0	±1	±2	±3
Signal-3L 3.0	0.654	0.728	0.774	0.802
SignalP 5.0	0.671	0.702	0.732	0.732
SignalP4.1	0.613	0.643	0.672	0.693
DeepSig	0.604	0.631	0.668	0.700
Signal-3L 2.0	0.322	0.341	0.365	0.379
LipoP	0.159	0.178	0.194	0.207
Philius	0.151	0.168	0.182	0.191
Phobius	0.226	0.237	0.250	0.267
PolyPhobius	0.176	0.190	0.201	0.216
PrediSi	0.273	0.291	0.301	0.321
PRED-LIPO	0.069	0.083	0.110	0.131
PRED-SIGNAL	0.066	0.085	0.096	0.106
PRED-TAT	0.08	0.099	0.111	0.119
Signal-CF	0.105	0.109	0.117	0.123
SOSUIsignal	0.037	0.069	0.098	0.121
SPEPlip	0.366	0.378	0.398	0.418
SPOCTOPUS	0.120	0.164	0.211	0.233
TOPCONS2	0.107	0.146	0.184	0.210

Table S8. Benchmarking of signal peptide cleavage site prediction measured by F1-score on SignalP 5.0 dataset. The results of other methods are directly from SignalP 5.0.

Method	**Eukaryotes**			
	0	±1	±2	±3
Signal-3L 3.0	0.665	0.740	0.787	0.815
SignalP 5.0	0.699	0.731	0.762	0.779
SignalP4.1	0.651	0.683	0.714	0.737
DeepSig	0.614	0.641	0.679	0.712
Signal-3L 2.0	0.430	0.456	0.487	0.506
LipoP	0.217	0.244	0.265	0.283
Philius	0.243	0.270	0.292	0.307
Phobius	0.338	0.354	0.373	0.399
PolyPhobius	0.280	0.302	0.319	0.343
PrediSi	0.385	0.410	0.424	0.453
PRED-LIPO	0.080	0.096	0.128	0.152

PRED-SIGNAL	0.102	0.131	0.149	0.164
PRED-TAT	0.134	0.166	0.186	0.199
Signal-CF	0.181	0.188	0.201	0.212
SOSUIsignal	0.061	0.114	0.162	0.200
SPEPlip	0.483	0.499	0.525	0.551
SPOCTOPUS	0.184	0.251	0.323	0.356
TOPCONS2	0.166	0.227	0.286	0.326

Table S9. Benchmarking of signal peptide recognition and its cleavage site prediction on SP19 independent dataset. (ID 1-33 are Eukaryotic sequences, 34-36 are Gram-negative sequences, 37-45 are Gram-positive sequences. Column *Name* records the name of the protein sequence. Column *GT* indicates the experimental label of the signal peptide cleavage site annotated in the database. P_Y represents the recognition result of the method on whether the sequence contains signal peptide, Y is inclusive, N is exclusive. P_{cs} represents the result of the method for predicting signal peptide cleavage site.

	Signal proteins		**DeepSig**		**Phobius**		**PRED-TAT**		**PrediSi**		**SignalP 5.0**		**Signal-3L 3.0**	
ID	*Name*	*GT*	P_Y	P_{cs}	P_Y	P_{cs}	P_Y	P_{cs}	P_Y	P_{cs}	P_Y	P_{cs}	P_Y	P_{cs}
1	AMP1_PARCM	23	Y	23	Y	23	Y	23	N	-	Y	23	Y	23
2	AMP2_PARCM	23	Y	23	Y	23	Y	21	Y	23	Y	23	Y	23
3	AMP3_PARCM	23	Y	23	Y	23	Y	23	Y	23	Y	23	Y	23
4	AMPU1_AMPCP	26	Y	24	Y	24	Y	24	N	-	Y	24	Y	24
5	AMPU2_AMPCP	26	Y	24	Y	24	Y	24	N	-	Y	24	Y	24
6	AMPU3_AMPCP	17	N	-	N	-	N	-	N	-	Y	20	Y	17
7	ANIS5_ANISI	18	Y	18	Y	18	Y	18	Y	18	Y	18	Y	18
8	APIM_APIME	24	N	-	Y	24	Y	24	Y	24	Y	24	Y	24
9	CG2RA_CONMI	29	Y	22	Y	19	Y	19	Y	19	Y	19	Y	22
10	CMU1_USTMA	21	Y	21	Y	21	Y	22	Y	21	Y	21	Y	23
11	CROS1_CROSA	26	Y	25	Y	25	Y	25	Y	25	Y	25	Y	25
12	DEP23_DERPT	21	Y	21	Y	21	Y	21	Y	21	Y	21	Y	21
13	GH5FP_CHAOB	28	N	-	Y	28	Y	28	Y	28	Y	28	Y	26
14	IBPKA_TYPIS	20	N	-	Y	20	Y	20	Y	20	Y	20	Y	20
15	IVOB_EMENI	18	Y	19	Y	19	Y	19	Y	19	Y	19	Y	19
16	PNX36_SCOSD	25	Y	25	Y	30	Y	25	Y	25	Y	25	Y	25
17	SCX5B_MESEU	19	Y	19	Y	19	Y	20	Y	19	Y	19	Y	19
18	THAP1_THAPI	18	Y	16	Y	16	Y	16	Y	16	Y	16	Y	16
19	TX13A_MYRGU	24	N	-	Y	24	Y	24	Y	24	Y	24	Y	24
20	TX14A_MYRGU	25	Y	25	Y	19	Y	23	N	-	Y	25	Y	25
21	TX51A_SCOSD	24	Y	24	Y	24	Y	24	Y	24	Y	24	Y	24
22	TXA1A_SCOSD	23	Y	23	Y	23	Y	21	N	-	Y	23	Y	23
23	TXA2A_SCOSD	23	Y	23	Y	23	Y	20	N	-	Y	23	Y	50
24	TXF1A_SCOMU	23	Y	23	Y	22	Y	23	Y	23	Y	23	Y	23
25	TXF1A_SCOSD	20	Y	20	Y	20	Y	20	Y	20	Y	20	Y	20

26	TXF2A_SCOSD	20	Y	18	Y	19	N	-	N	-	Y	20	Y	20
27	TXF3A_SCOSD	23	Y	23	Y	22	Y	23	Y	23	Y	23	Y	23
28	TXF3B_SCOSD	23	Y	23	Y	22	Y	23	Y	23	Y	23	Y	23
29	TXG1A_SCOSD	23	Y	23	Y	26	Y	21	Y	23	Y	23	Y	23
30	TXX1A_ETHRU	24	Y	24	Y	24	Y	22	Y	24	Y	24	Y	24
31	TXX1B_ETHRU	24	Y	24	Y	24	Y	22	Y	24	Y	24	Y	24
32	XEG1_PHYSP	19	Y	19	Y	19	Y	19	Y	19	Y	19	Y	19
33	ZONAD_KOMPC	19	Y	19	Y	19	Y	19	Y	20	Y	19	Y	19
34	BPAC_BURP2	71	N	-	N	-	N	-	Y	58	N	-	Y	25
35	CDIA_ECONC	32	Y	22	N	-	Y	29	N	-	Y	31	Y	30
36	MTO_HYPSQ	24	Y	24	Y	24	Y	24	Y	24	Y	24	Y	24
37	CHPD_STRCO	23	Y	30	Y	23	Y	23	N	-	Y	23	Y	23
38	CHPE_STRCO	27	Y	34	Y	27	Y	27	Y	27	Y	27	Y	27
39	CHPF_STRCO	36	Y	42	Y	35	Y	35	Y	32	Y	35	Y	35
40	CHPG_STRCO	27	Y	34	Y	37	Y	27	N	-	Y	27	Y	31
41	CHPH_STRCO	25	Y	31	Y	25	Y	25	Y	25	Y	25	Y	25
42	INLC_LISMG	34	Y	34	Y	34	Y	34	Y	29	Y	34	Y	34
43	PSRP_STRPN	72	N	-	N	-	N	-	Y	34	N	-	Y	34
44	RDLA_STRCO	28	Y	28	Y	28	Y	22	Y	21	Y	28	Y	28
45	RDLB_STRCO	28	Y	28	Y	22	Y	22	Y	20	Y	28	Y	28