CoRAE: Concreate Relaxation Autoencoder for Differentiable Gene Selection and Pan-Cancer Classification

**Abdullah Al Mamun, Ananda Mondal\***

School of Computing and Information Sciences

Florida International University, Miami, US

**Supplementary**

* ***Supplementary-1.pdf*** contains a list of top-100 genes using all three methods, visualization of sample distributions, Figures of all five performance evaluation metrics, and annealing scheduling curve for lncRNA.
* ***Supplementary-2.xlsx*** contains data used to generate all the figures and tables in this paper.

1. **Selected Genes**

Table 1: Top-100 coding and non-coding genes selected using proposed and competing methods.

|  |  |  |
| --- | --- | --- |
| **Method** | **mRNA type** | **Gene name** |
| LASSO | mRNA | AMIGO2, ANXA8, ARHGAP6, ARHGEF6, ASB13, ATP8B2, B3GNT7, BAALC, BAMBI, BARX1, BMP7, CA12, CA9, CALML5, CCL25, CITED1, CPVL, CTSE, CYP2J2, DACT2, DAPP1, DHCR24, DHRS2, DNAJC12, EN1, FAM163A, FBXO41, GATA3, GATA4, GGT6, GINS1, GPSM2, HFE, HOXA9, HOXB13, HOXB7, HOXC10, HOXD8, IGFBP2, IRF5, IRX3, IRX5, ISL1, KAZALD1, KCNJ12, KIAA1161, KRT14, KRT23, KRT7, KRT80, KRT81, LIPE, LY6D, MGST1, MLC1, MLPH, MMP3, MT3, NDRG2, NEBL, NETO2, NKX2-1, NPM2, NR1H4, OLR1, PADI2, PAPSS2, PAQR5, PAX8, PCDH7, PCDHB9, PDLIM4, PITX1, PKNOX2, PRAME, PRKAR2B, PRLR, PTH2R, PVRL3, RAB31, RBMS3, RBPMS2, RXRG, S100P, SAMD5, SFTA2, SFTPB, SGCD, SH3BGRL2, SMC1B, SMPD3, SNCG, SOX17, SPNS2, STXBP6, TRNP1, TSHR, TTYH1, VAV3, ZNF280B |
| lncRNA | AC000123.4, AC005082.12, AC005083.1, AC005152.3, AC008268.1, AC016735.2, AC093850.2, AC108142.1, AC109642.1, AC114730.3, AC115522.3, ADIRF-AS1, ALDH1L1-AS2, AP000251.3, AP000439.3, AP001065.15, AP001626.1, AP003774.1, BMPR1B-AS1, C5orf66-AS1, CITF22-92A6.1, CKMT2-AS1, CTA-363E6.6, CTA-384D8.31, CTC-327F10.4, CTD-2015G9.2, CTD-2015H6.3, CTD-2089N3.1, CTD-2314G24.2, CTD-2377D24.4, CTD-3032H12.2, CTD-3094K11.1, DNMBP-AS1, DYNLL1-AS1, GATA3-AS1, H19, HAND2-AS1, HNF1A-AS1, HOXB-AS4, HOXC13-AS, HOXD-AS2, LA16c-316G12.2, LHFPL3-AS1, LINC00152, LINC00518, LINC00884, LINC00885, LINC00887, LINC00925, LINC01158, LINC01235, LINC01268, LINC01410, LINC01540, LOXL1-AS1, MIR202HG, MIR205HG, MIR4435-1HG, MIR503HG, PIK3CD-AS2, PTCSC2, RBMS3-AS3, RP1-232P20.1, RP1-288H2.2, RP11-1017G21.5, RP11-1055B8.3, RP11-10A14.5, RP11-10C24.3, RP11-1149O23.2, RP11-1149O23.3, RP11-119F7.5, RP11-11N9.4, RP11-12M5.3, RP11-157J24.2, RP11-166D19.1, RP11-19E11.1, RP11-206M11.7, RP11-20F24.2, RP11-218E20.3, RP11-264I13.2, RP11-277P12.20, RP11-290H9.4, RP11-304L19.3, RP11-320N7.2, RP11-3P17.5, RP11-473M20.16, RP4-610C12.3, RP4-740C4.5, SATB2-AS1, SLCO4A1-AS1, SNHG14, SOX9-AS1, ST3GAL6-AS1, TINCR, TP73-AS1, VPS9D1-AS1, WDR86-AS1, ZFPM2-AS1, ZIM2-AS1, ZNF528-AS1 |
| SVM-RFE | mRNA | ACTG2, AZGP1, BARX1, C1orf186, CA12, CALML3, CAMK2N1, CDCA7, CDH1, CDH16, CDKN2A, CEACAM5, CLDN3, CLDN4, CLDN6, CP, DDX3Y, DES, DLK1, DNER, DSG2, EEF1A2, EIF3CL, EMX2, EPCAM, ESRP1, FGFR4, FOXA1, FOXA2, FOXE1, GATA3, GATA4, GFAP, GJB1, GNL3L, GPX2, GRHL2, GRIK5, HNF1B, HNF4A, HOXA9, HOXB7, HOXC10, IFFO1, IRX2, KIF1A, KRT19, KRT5, KRT7, KRT8, LGALS4, LYPLAL1, MAL, MALAT1, MFAP2, MGST1, MLANA, MLPH, MMP12, MSLN, NACA2, NFIX, NKX2-1, NME2P1, NPM3, NUDT16P1, PABPC3, PAX8, PITX1, PNMAL1, POU3F3, PRAME, PTPRH, PTPRN2, RAB25, REC8, RNF128, RPL39L, RPL41, RPS4Y1, S100A1, S100A14, SALL1, SERPINA5, SFN, SFRP2, SFTPB, SLC34A2, SOX15, SOX17, SOX2, SYTL1, TBX5, TM4SF4, TSPAN1, UCHL1, USH1C, WDR72, WNK2, ZBTB7A |
| lncRNA | AC005082.12, AC006042.6, AC007405.6, AC009299.3, AC016747.3, AC093850.2, AC133528.2, AFAP1-AS1, AL450992.2, AP000251.3, BBOX1-AS1, CASC9, CECR7, CRNDE, CTA-384D8.31, CTD-2015H6.3, CTD-2231H16.1, DNM3OS, EMX2OS, FAM83H-AS1, FENDRR, GATA2-AS1, H19, HNF1A-AS1, HOXA10-AS, HOXA11-AS, HOXD-AS2, LA16c-329F2.1, LINC00086, LINC00261, LINC00511, LINC00857, LINC00958, LINC01116, LINC01133, LINC01139, LINC01158, MAGI2-AS3, MALAT1, MEG3, MIR205HG, MIR99AHG, MNX1-AS1, NKX2-1-AS1, PIK3CD-AS2, PTCSC2, RP1-288H2.2, RP1-60O19.1, RP11-1149O23.3, RP11-11N9.4, RP11-132A1.4, RP11-13J10.1, RP11-164P12.4, RP11-166D19.1, RP11-223I10.1, RP11-264B14.2, RP11-276H19.2, RP11-284F21.7, RP11-304L19.1, RP11-304L19.3, RP11-329L6.2, RP11-350J20.12, RP11-357H14.17, RP11-373D23.2, RP11-392P7.6, RP11-395G23.3, RP11-3P17.5, RP11-449J21.5, RP11-44F21.5, RP11-465B22.8, RP11-465N4.4, RP11-47A8.5, RP11-530C5.1, RP11-532F12.5, RP11-567G11.1, RP11-680F8.1, RP11-739N20.2, RP11-760H22.2, RP11-977G19.5, RP3-404F18.5, RP3-406A7.7, RP3-416H24.1, RP4-639F20.1, SFTA1P, SLC38A3, SLCO4A1-AS1, SNHG18, SOX21-AS1, TBX5-AS1, TINCR, TRPM2-AS, TTTY14, TTTY15, U47924.27, UCA1, VPS9D1-AS1, XIST, ZFPM2-AS1, ZNF582-AS1, ZNF667-AS1 |
| CoRAE | mRNA | ACYP2, ADAM23, AKAP8L, AKR1B10, ALDH1A3, ANO9, ANXA3, APOB, ASB16, B3GAT1, BAZ2B, BCL11B, BEGAIN, C12orf10, CBS, CCDC77, CCDC85B, CD109, CEP55, CHRNA4, CHST13, CHTF18, CLEC2D, CMTM1, CNTFR, COL8A2, COX10, CWC25, CXADR, CYFIP2, CYP4F3, DCDC2, DCLK2, DFFB, DLL3, DUSP14, ELP3, EPHB3, EPS8L1, FAM182B, FAM83B, FBXO43, FCHO1, FGF2, FLI1, FLT4, GJB3, GPR35, GPSM2, HABP4, HBEGF, HOXA11, IGJ, IL17RD, INMT, INPP5J, IRX6, ISG20, ITGA9, KIAA1549, KLK3, KLK5, KREMEN2, LHFPL2, MAPT, MED9, MGAT5B, MSX2, MYEF2, NCF1B, NME5, OLR1, PDK1, PHOSPHO2, PHYHIPL, PPP1R3E, PRRX2, RGMA, RGS11, RPS6KC1, S100PBP, SLC17A5, SLC34A2, SLC39A5, SPAG1, TAF2, TAGAP, TAGLN, TFF1, TLN2, TLR7, TMEM229B, TTBK2, TTLL3, ZBTB25, ZNF43, ZNF486, ZNF561, ZNF665, ZNF770 |
| lncRNA | ABHD11-AS1, AC012360.4, AC016831.7, AC079630.4, AC106786.1, AC139100.3, CTA-212D2.2, CTA-217C2.2, CTA-331P3.1, CTC-444N24.6, CTC-487M23.5, CTD-2014E2.6, CTD-2020K17.4, CTD-2135J3.3, CTD-2331H12.7, CTD-2527I21.14, CTD-2554C21.3, CTD-2555C10.3, CTD-2561B21.4, EIF3J-AS1, HS1BP3-IT1, IGBP1-AS1, IGFBP7-AS1, ITGB2-AS1, KB-1410C5.5, KCNMB2-AS1, LINC00471, LINC00543, LINC00592, LINC00630, LINC00668, LINC00958, LINC01207, LINC01484, LINC01507, MIAT, MIR210HG, MIR99AHG, NBAT1, PWAR6, RP1-102K2.8, RP1-269M15.3, RP1-288H2.2, RP11-1017G21.5, RP11-1055B8.3, RP11-108M12.3, RP11-110I1.11, RP11-110I1.12, RP11-111K18.2, RP11-111M22.5, RP11-146F11.1, RP11-158M2.3, RP11-1D12.2, RP11-20F24.2, RP11-21M24.2, RP11-227F19.5, RP11-234B24.2, RP11-273G15.2, RP11-276H7.2, RP11-298D21.3, RP11-35G9.3, RP11-381N20.2, RP11-397A16.1, RP11-402G3.5, RP11-403I13.5, RP11-406H21.2, RP11-429J17.7, RP11-452H21.2, RP11-505E24.3, RP11-507K2.3, RP11-526F3.1, RP11-537H15.3, RP11-547D24.1, RP11-554D14.6, RP11-554D15.1, RP11-627G23.1, RP11-655C2.3, RP11-731J8.2, RP11-736N17.8, RP11-750H9.7, RP11-767N6.7, RP11-806O11.1, RP11-807H17.1, RP11-867G23.1, RP11-8L8.2, RP13-726E6.2, RP3-395M20.9, RP3-507I15.2, RP4-555L14.4, RP4-564M11.2, RP4-740C4.5, RP5-1085F17.3, RP5-1184F4.5, SATB2-AS1, SNHG20, THUMPD3-AS1, TUSC8, U91324.1, VPS9D1-AS1, YEATS2-AS1 |

1. **Data distribution**

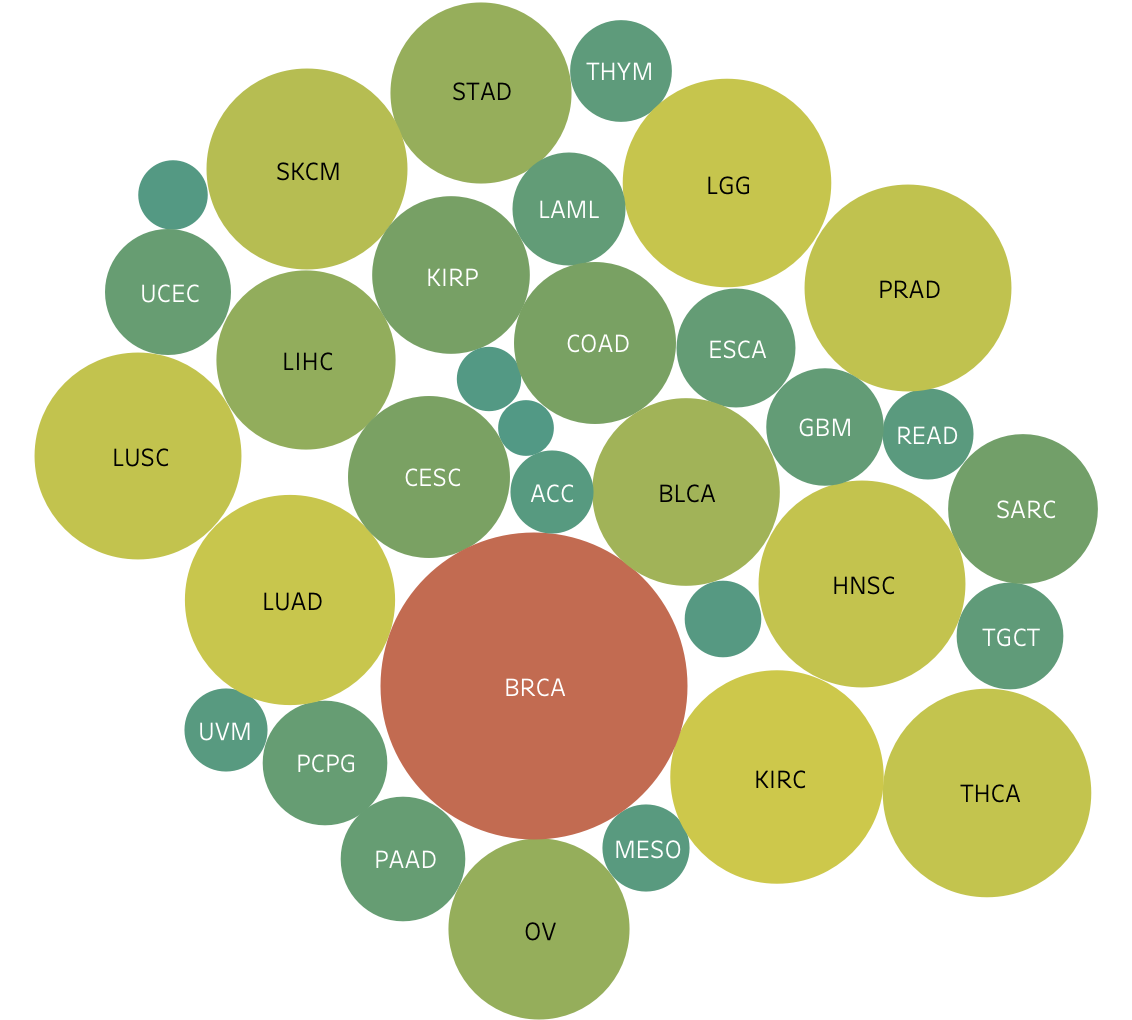


Fig 1: **Bubble chart of cancer samples of 33 cancer types.** Here, the size of bubbles refers the number of samples of a particular cancer type. For example, Breast cancer (BRCA) has highest number of sample whereas Adrenocortical Cancer (ACC) has lowest number of samples available.

1. **Performance Evaluations**



1. Accuracy at different #feature selected: mRNA (left) and lncRNA (right)



1. Precision at different #feature selected: mRNA (left) and lncRNA (right)



1. Recall at different #feature selected: mRNA (left) and lncRNA (right)



1. f1 score at different #feature selected: mRNA (left) and lncRNA (right)

Fig 2: **Classification performance using selected RNA features.** Comparison of CoRAE with other feature selection methods. Throughout the all values of *k* tested on both mRNA (left) and lncRNA (right) expressions. For all the performance metrics (a) Accuracy (b) Precision (c) Recall and (d) f1 score.



Fig 3: Reconstruction mean squared error using selected mRNA (left) and lncRNA (right)



Fig 7: Annealing schedule of CoRAE for *k* = 100 using lncRNA

Table 2: Classification and Reconstruction performance for different number of selected mRNA and lncRNAs

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | **mRNA** | | | | | **lncRNA** | | | | |
| **#Feature** | **Method** | **Accuracy** | **Precision** | **Recall** | **F1** | **MSE** | **Accuracy** | **Precision** | **Recall** | **F1** | **MSE** |
| 500 | LASSO | 0.96 | 0.94 | 0.94 | 0.94 | 6.79 | 0.94 | 0.92 | 0.92 | 0.92 | 11.19 |
| CoRAE | 0.99 | 0.98 | 0.98 | 0.98 | 2.86 | 0.96 | 0.94 | 0.93 | 0.94 | 7.7 |
| SVM-RFE | 0.95 | 0.93 | 0.93 | 0.93 | 9.47 | 0.94 | 0.92 | 0.91 | 0.91 | 10.01 |
| 400 | LASSO | 0.93 | 0.92 | 0.92 | 0.92 | 7.22 | 0.94 | 0.92 | 0.92 | 0.93 | 10.63 |
| CoRAE | 0.99 | 0.96 | 0.97 | 0.96 | 3.7 | 0.95 | 0.94 | 0.93 | 0.93 | 7.95 |
| SVM-RFE | 0.94 | 0.91 | 0.92 | 0.92 | 9.2 | 0.94 | 0.91 | 0.92 | 0.92 | 11.88 |
| 300 | LASSO | 0.93 | 0.93 | 0.92 | 0.93 | 10.05 | 0.9 | 0.89 | 0.88 | 0.89 | 14.87 |
| CoRAE | 0.98 | 0.96 | 0.97 | 0.97 | 3.04 | 0.95 | 0.93 | 0.93 | 0.93 | 9.48 |
| SVM-RFE | 0.94 | 0.9 | 0.91 | 0.9 | 11.56 | 0.93 | 0.91 | 0.9 | 0.9 | 13.38 |
| 250 | LASSO | 0.94 | 0.92 | 0.92 | 0.91 | 9.77 | 0.94 | 0.91 | 0.91 | 0.91 | 11.57 |
| CoRAE | 0.98 | 0.97 | 0.97 | 0.97 | 4.03 | 0.95 | 0.94 | 0.92 | 0.93 | 9.41 |
| SVM-RFE | 0.93 | 0.92 | 0.91 | 0.91 | 11.69 | 0.93 | 0.9 | 0.9 | 0.89 | 13.64 |
| 200 | LASSO | 0.94 | 0.91 | 0.91 | 0.91 | 10.73 | 0.9 | 0.9 | 0.89 | 0.89 | 15.2 |
| CoRAE | 0.97 | 0.96 | 0.96 | 0.96 | 4.41 | 0.95 | 0.93 | 0.92 | 0.92 | 9.99 |
| SVM-RFE | 0.93 | 0.89 | 0.9 | 0.89 | 13.56 | 0.92 | 0.89 | 0.89 | 0.89 | 14.01 |
| 150 | LASSO | 0.93 | 0.9 | 0.9 | 0.9 | 10.22 | 0.93 | 0.9 | 0.91 | 0.9 | 12.87 |
| CoRAE | 0.97 | 0.95 | 0.94 | 0.95 | 5.8 | 0.95 | 0.93 | 0.91 | 0.92 | 10.41 |
| SVM-RFE | 0.93 | 0.89 | 0.9 | 0.89 | 12 | 0.91 | 0.88 | 0.89 | 0.88 | 13.77 |
| 100 | LASSO | 0.93 | 0.89 | 0.89 | 0.89 | 10.4 | 0.92 | 0.9 | 0.9 | 0.89 | 12.83 |
| CoRAE | 0.96 | 0.95 | 0.95 | 0.95 | 6.4 | 0.94 | 0.93 | 0.91 | 0.92 | 11.37 |
| SVM-RFE | 0.92 | 0.88 | 0.88 | 0.88 | 15.04 | 0.91 | 0.87 | 0.87 | 0.87 | 14.36 |
| 50 | LASSO | 0.86 | 0.85 | 0.84 | 0.84 | 20.21 | 0.9 | 0.88 | 0.87 | 0.87 | 15.05 |
| CoRAE | 0.94 | 0.9 | 0.9 | 0.9 | 11.64 | 0.92 | 0.88 | 0.88 | 0.88 | 13.82 |
| SVM-RFE | 0.9 | 0.89 | 0.86 | 0.86 | 16.71 | 0.86 | 0.82 | 0.79 | 0.79 | 25.44 |
| 40 | LASSO | 0.79 | 0.78 | 0.78 | 0.78 | 32.06 | 0.85 | 0.83 | 0.83 | 0.84 | 28.04 |
| CoRAE | 0.94 | 0.88 | 0.9 | 0.89 | 12.54 | 0.92 | 0.87 | 0.88 | 0.87 | 14.95 |
| SVM-RFE | 0.86 | 0.82 | 0.83 | 0.82 | 24.99 | 0.84 | 0.82 | 0.78 | 0.79 | 29.42 |
| 30 | LASSO | 0.73 | 0.71 | 0.69 | 0.72 | 29.25 | 0.82 | 0.81 | 0.81 | 0.81 | 34.13 |
| CoRAE | 0.93 | 0.88 | 0.88 | 0.88 | 16.36 | 0.91 | 0.86 | 0.87 | 0.86 | 16.73 |
| SVM-RFE | 0.85 | 0.81 | 0.81 | 0.81 | 23.87 | 0.8 | 0.79 | 0.72 | 0.73 | 35.52 |
| 20 | LASSO | 0.6 | 0.59 | 0.58 | 0.59 | 51.01 | 0.76 | 0.73 | 0.72 | 0.73 | 45.99 |
| CoRAE | 0.91 | 0.86 | 0.86 | 0.85 | 18.06 | 0.84 | 0.83 | 0.8 | 0.81 | 29 |
| SVM-RFE | 0.77 | 0.74 | 0.72 | 0.72 | 48.78 | 0.74 | 0.74 | 0.68 | 0.69 | 47.64 |
| 10 | LASSO | 0.55 | 0.51 | 0.52 | 0.51 | 66.02 | 0.55 | 0.51 | 0.5 | 0.52 | 81.56 |
| CoRAE | 0.79 | 0.73 | 0.67 | 0.68 | 36.07 | 0.7 | 0.6 | 0.58 | 0.57 | 60.11 |
| SVM-RFE | 0.63 | 0.6 | 0.53 | 0.53 | 64.69 | 0.5 | 0.4 | 0.36 | 0.35 | 88.19 |