Novel artificial intelligence-based identification of drug-gene-disease interaction through using protein-protein interaction

Y-h Taguchi[1*] and Turki Turki[2]

[1*]Department of Physics, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo, 112-8551, Japan.
[2]Department of Computer Science, King Abdulaziz University, Street, Jeddah, 21589, Saudi Arabia.

*Corresponding author(s). E-mail(s): tag@granular.com;
Contributing authors: tturki@kau.edu.sa;

Abstract

Identification of drug-gene-disease interaction is important to find drugs effective toward disease. Nevertheless, it is not easy since we have to identify drugs effective to genes which are also critical for diseases. When starting from diseases (disease centric approach), we need to identify genes critical in diseases and find drugs effective to the selected genes. When we start from drugs (drug centric approach), we need to find genes that drugs target and then find diseases in which identified genes are critical. These are complicated processes. If we can identify genes which are effective to the disease and can be targeted by drugs, i.e. if we can start from genes (gene centric approach), it is much easier. Nevertheless, none knows how we can identify such sets of genes without specifying either target diseases or drugs. Hence, our novelty is attributed to an artificial intelligence-based approach, employing unsupervised methods and identifying such genes without specifying either diseases or drugs. To inspect the feasibility, we have applied tensor decomposition (TD) based unsupervised feature extraction (FE) to perform drug repositioning from protein-protein interaction (PPI) without using any other information. Proteins selected by TD based unsupervised FE included many genes related to cancers as well as drugs that target selected proteins. Thus, we could identify drugs for cancers from only PPI. Since the selected proteins have more interaction, we replaced the selected proteins with hub proteins and found that hub proteins themselves can be used for drug repositioning. In contrast to hub proteins that can identify only cancer drugs, TD based unsupervised FE enables us drugs for other diseases. In addition to this, TD based unsupervised

FE can identify drugs effective in *in vivo* experiments, which is difficult by hub proteins. In conclusion, TD based unsupervised FE is a useful tool to perform drug repositioning only using PPI without other information.

Keywords: protein-protein interaction, drug repositioning, artificial intelligence, unsupervised learning, tensor decomposition

1 Introduction

Identification of drug-gene-disease information is a critical one to perform drug repositioning. Nevertheless, since it is not a simple identification of paired information, but that of triplet information, it is not easy to identify drug-gene-disease information. Identification of drug-gene-disease information was often performed by two steps. When starting from genes (drug-centric), identification of drug-gene relation was performed at first, then gene-disease information was extracted by finding disease related to the selected genes. When starting from diseases (disease-centric), identification of gene-disease relation was performed at first, then drug-gene information was extracted by finding drugs that target the selected genes. Since it is not guaranteed that there is a significant relation at the second stage (i.e. gene-disease information for drug-centric approach or drug-gene information for disease-centric approach), identification of drug-gene-disease information is much harder than identification of only gene-disease information or drug-gene information only.

To address this difficulty, linear algebra was often used to identify drug-gene-disease information, since linear algebra enables us to identify drug-gene-disease information with neither drug-centric nor disease centric-mannar. Wang et al [1] applied matrix factorization to gene expression matrices of drug treatment and diseases. Kim and Cho [2] employed tensor decomposition (TD) to extract drug-gene-disease information with starting from the product of ID embedding vectors for drug, gene, and disease. In these approaches, since identification of drug-gene-disease information was performed without passing through two stages approach, these approaches can often solve the difficulty of two stages approaches.

Another advantage of these linear algebra-based approaches is that it is fully unsupervised. We do not have to have any drug or disease specific information (e.g., differential expression between healthy controls and patients) and we can identify drug-gene-disease information in fully data-driven manner. Recently, we applied TD based unsupervised FE [3, 4] to integrate PPI with gene expression of cancers [5]. Then we have found that integrated analysis could enhance the coincidence with clinical labels. During this study, we have also found that PPI only has information related to cancers, although PPI itself is unlikely associated with cancers. In this paper, we try to find how much relationship with cancers PPI itself can have with using TD based unsupervised FE without integrating with other information, e.g., gene expression. Then we have found that applying TD based unsupervised FE to PPI enables us to find cancer related genes which are also used for drug repositioning. After identifying that the selected proteins are likely to be hub proteins in PPI, we also replaced proteins

selected by TD based unsupervised FE with hub proteins and found that hub proteins can be used for drug repositioning. The distinction between hub proteins and those selected by TD based unsupervised FE is the strength of correlation among interaction, i.e., proteins selected by TD based unsupervised FE have more number of shared interaction. Since only proteins selected by TD based unsupervised FE could have hits in *in vivo*-based drug database, DrugMatrix, the identification of proteins that have not only more interaction (i.e., hub proteins) but also more shared interaction might be the key to identify more promising proteins for drug repositioning.

2 Results

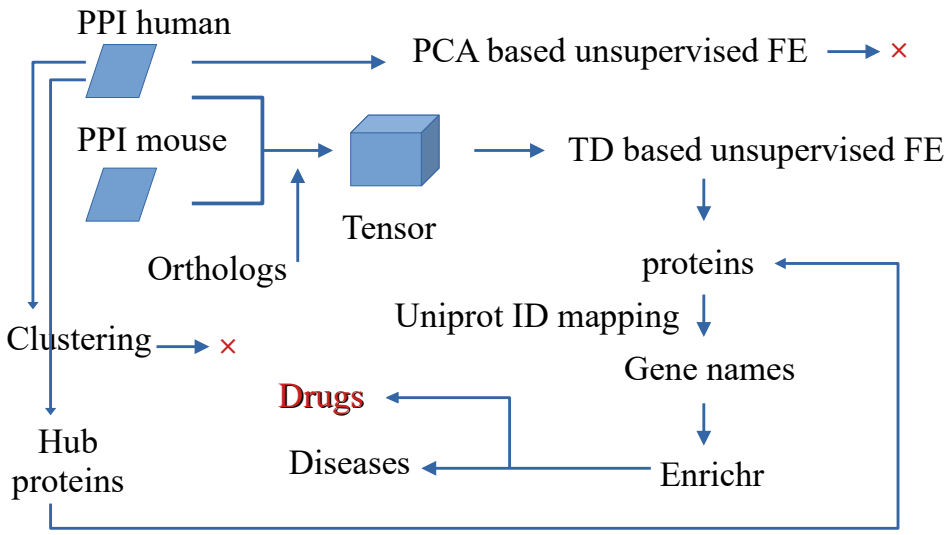

Fig. 1 Flowchart that describes the analyses in this study. PCA based unsupervised FE was applied to Human PPI, but failed. TD based unsupeervised FE applied to tensor generated from human and mouse PPI and gene names associated with the identified proteins were uploaded to Enrichr to identify associated diseases and drugs. Hub proteins and proteins selected by cluster analyses were tested and the former results in identification of diseases and drugs.

Figure 1 shows the analyses performed in this study.

2.1 TD based unsupervised FE

In this analysis, the first matrix to be integrated is human PPI, $n^1_{ii'} \in \mathbb{R}^{4901\times4901}$, and the second one is mouse PPI, $n^2_{ii'} \in \mathbb{R}^{2342\times2342}$. Since the number of common protein, N_{common}, is 1097, and the number of orthogonal protein, N_{ortho}, is 432, the resulting integrated tensor is $n_{ii'k} \in \mathbb{R}^{N\times N\times 2}$ where $N = 4901 + 2342 - 1097 - 432 = 5714$. After applying HOSVD to $n_{ii'k}$, we get eq. (7). P_is are attributed to is using u_{2i} or

u_{3i} by eq. (6) with replacing $u_{\ell i}$ with $u_{\ell_1 i}$ and are corrected by BH criterion [3, 4]. As a result, 196 and 59 proteins are associated with adjusted P-values less than 0.01, respectively. Uniprot accession numbers associated with these proteins are converted to 193 and 57 gene names by uniprot ID mapping, respectively (see supplementary information for the list of proteins and gene names). 193 and 57 gene names are uploaded to Enrichr [6] for the evaluation.

At first, we need to find what kinds of diseases are associated with two sets of gene names. At first, we considered "Jensen DISEASES" category. For 193 gene

Table 1 Top ranked 10 diseases in "Jensen DISEASES" category of Enrichr for 193 gene names.

Term	Overlap	P-value	Adjusted P-value
Lymphoid leukemia	13/152	2.51×10^{-9}	8.33×10^{-7}
Immune system cancer	11/137	8.40×10^{-8}	1.39×10^{-5}
Cancer	15/300	2.07×10^{-7}	2.29×10^{-5}
Familial adenomatous polyposis	5/25	3.43×10^{-6}	2.85×10^{-4}
Upper respiratory tract disease	4/13	5.39×10^{-6}	3.58×10^{-4}
Biliary tract cancer	4/14	7.49×10^{-6}	4.14×10^{-4}
Neurodegenerative disease	11/293	1.24×10^{-4}	5.15×10^{-3}
Ovarian cancer	5/51	1.24×10^{-4}	5.15×10^{-3}
Systemic lupus erythematosus	6/83	1.44×10^{-4}	5.31×10^{-3}
Ataxia telangiectasia	4/30	1.82×10^{-4}	6.04×10^{-3}

names (Table 1), not only there are highly significant diseases, but also more than half of diseases are cancer or tumors ("Lymphoid leukemia", "Immune system cancer", "Cancer", "Biliary tract cancer", and "Ovarian cancer"). In addition to this, "Familial adenomatous polyposis" is known to develope to cancers [7]. For 57 gene names

Table 2 Top ranked 10 diseases in "Jensen DISEASES" category of Enrichr for 57 gene names.

Term	Overlap	P-value	Adjusted P-value
Li-Fraumeni syndrome	5/12	9.26×10^{-11}	1.38×10^{-8}
Ovarian cancer	6/51	4.77×10^{-9}	3.56×10^{-7}
Immune system cancer	6/137	1.82×10^{-6}	8.19×10^{-5}
Breast cancer	9/434	2.20×10^{-6}	8.19×10^{-5}
Lymphoid leukemia	6/152	3.34×10^{-6}	9.94×10^{-5}
Biliary tract cancer	3/14	6.63×10^{-6}	1.65×10^{-4}
Ductal carcinoma in situ	3/18	1.48×10^{-5}	3.14×10^{-4}
Nijmegen breakage syndrome	3/22	2.76×10^{-5}	5.15×10^{-4}
ENSP00000256078	3/28	5.81×10^{-5}	8.91×10^{-4}
Hereditary breast ovarian cancer	2/5	7.12×10^{-5}	8.91×10^{-4}

(Table 2), not only there are also highly significant diseases, but also more than half of diseases are cancer or tumors ("Ovarian cancer", "Immune system cancer", "Breast cancer", "Biliary tract cancer", "Ductal carcinoma in situ", and "Hereditary breast

ovarian cancer"). In addition to this, "Li-Fraumeni syndrome" is known to be the cause of cancers [8].

Next we consider "OMIM Disease" of Enrichr. For 193 gene names (Table 3),

Table 3 Top ranked 10 diseases in "OMIM Disease" category of Enrichr for 193 gene names.

Term	Overlap	P-value	Adjusted P-value
Colorectal cancer	7/40	9.27×10^{-8}	2.97×10^{-6}
Parkinson disease	5/22	1.74×10^{-6}	2.78×10^{-5}
Ectodermal dysplasia	3/11	1.34×10^{-4}	1.43×10^{-3}
Breast cancer	3/28	2.36×10^{-3}	1.51×10^{-2}
Immunodeficiency	3/28	2.36×10^{-3}	1.51×10^{-2}
Fanconi anemia	2/13	6.60×10^{-3}	3.03×10^{-2}
Leukemia	4/78	6.64×10^{-3}	3.03×10^{-2}
Myocardial infarction	2/16	9.97×10^{-3}	3.99×10^{-2}
Orofacial cleft	2/19	1.39×10^{-2}	4.96×10^{-2}
Lymphoma	2/22	1.85×10^{-2}	5.92×10^{-2}

not only there are also highly significant diseases, but also four diseases are cancer or tumors ("Colorectal cancer", "Breast cancer", "Leukemia", and "Lymphoma"). In addition to this, "Fanconi anemia" is known to be often associated with cancers [9]. For 57 gene names (Table 4), not only there are also highly significant diseases, but

Table 4 Top ranked 10 diseases in "OMIM Disease" category of Enrichr for 57 gene names.

Term	Overlap	P-value	Adjusted P-value
Pancreatic cancer	3/11	3.02×10^{-6}	3.13×10^{-5}
Ovarian cancer	3/13	5.22×10^{-6}	3.13×10^{-5}
Colorectal cancer	3/40	1.71×10^{-4}	6.85×10^{-4}
Orofacial cleft	2/19	1.19×10^{-3}	3.56×10^{-3}
Breast cancer	2/28	2.59×10^{-3}	5.93×10^{-3}
Prostate cancer	2/30	2.97×10^{-3}	5.93×10^{-3}
Ectodermal dysplasia	1/11	2.93×10^{-2}	4.61×10^{-2}
Fanconi anemia	1/13	3.45×10^{-2}	4.61×10^{-2}
Melanoma	1/13	3.45×10^{-2}	4.61×10^{-2}
Immunodeficiency	1/28	7.30×10^{-2}	8.75×10^{-2}

also more than half of diseases are cancer or tumors ("Pancreatic cancer", "Ovarian cancer", "Colorectal cancer", "Breast cancer", "Prostate cancer", and "Melanoma"). In addition to this, "Fanconi anemia" is also listed. In conclusion, these two sets of proteins are highly related to cancers and tumors.

Next, we consider potential drug repositioning using these sets of genes. There are multiple categories that can be used for drug repositioning in Enrichr: "LINCS L1000 Chem Pert Consensus Sigs", "DSigDB", "DrugMatrix", "Drug Perturbations from GEO down", and "Drug Perturbations from GEO up". All of these categories return the list of compounds that are supposed to significantly target the set of uploaded

genes. Since 57 and 193 gene names are supposed to be deeply related to cancers and tumors, if there are drugs that are supposed to target these genes, these drugs can be drugs that can treat tumors and cancers. Table 5 lists the top ranked 10 compounds

Table 5 Top ranked 10 drugs in "LINCS L1000 Chem Pert Consensus Sigs" category of Enrichr for 193 gene names.

Term	Overlap	P-value	Adjusted P-value
KU-0060648 Up	14/236	6.59×10^{-8}	3.41×10^{-4}
Okadaic-Acid Up	14/246	1.10×10^{-7}	3.41×10^{-4}
Palmatine-Chloride Down	14/247	1.16×10^{-7}	3.41×10^{-4}
YL-55 Up	14/249	1.28×10^{-7}	3.41×10^{-4}
CHEMBL-399379 Up	13/231	3.61×10^{-7}	6.87×10^{-4}
LY-2606368 Up	13/233	3.98×10^{-7}	6.87×10^{-4}
RS-102221 Down	13/242	6.13×10^{-7}	6.87×10^{-4}
4-HQN Down	13/242	6.13×10^{-7}	6.87×10^{-4}
Butalbital Down	13/244	6.73×10^{-7}	6.87×10^{-4}
TL-HRAS-61 Up	13/247	7.72×10^{-7}	6.87×10^{-4}

in "LINCS L1000 Chem Pert Consensus Sigs" category for 193 gene names (no significant hit for 57 gene names). Tables 6 and 7 list the top ranked 10 compounds

Table 6 Top ranked 10 drugs in "DSigDB" category of Enrichr for 193 gene names.

Term	Overlap	P-value	Adjusted P-value
curcumin CTD 00000663	34/528	8.71×10^{-19}	2.63×10^{-15}
doxorubicin CTD 00005874	38/750	2.02×10^{-17}	3.05×10^{-14}
N-Acetyl-L-cysteine CTD 00005305	25/315	4.17×10^{-16}	4.19×10^{-13}
Arsenenous acid CTD 00000922	47/1283	6.07×10^{-16}	4.58×10^{-13}
MG-132 CTD 00002789	19/173	4.61×10^{-15}	2.78×10^{-12}
wortmannin CTD 00000504	17/146	5.05×10^{-14}	2.54×10^{-11}
LY 294002 CTD 00003061	20/227	6.01×10^{-14}	2.59×10^{-11}
ZINC CTD 00007011	50/1642	8.50×10^{-14}	3.21×10^{-11}
5-Fluorouracil CTD 00005987	42/1202	1.48×10^{-13}	4.96×10^{-11}
etoposide CTD 00005948	25/461	2.38×10^{-12}	7.18×10^{-10}

in "DSigDB" category for 193 gene names and 57 gene names, respectively. Table 8 lists the top ranked 10 compounds in "DrugMatrix" category for 193 gene names (no significant hit for 57 gene names). Tables 9 and 10 list the top ranked 10 compounds in "Drug Perturbations from GEO down" category for 193 gene names and 57 gene names, respectively. Tables 11 and 12 list the top ranked 10 compounds in "Drug Perturbations from GEO up" category for 193 gene names and 57 gene names, respectively.

As can be seen above, although it is not true for all categories, 193 and 57 gene names are often associated with lists of drugs that significantly target these sets of gene names. Thus, we can conclude that our proposed strategy, gene-centric drug repositioning, can be performed if we can employ TD based unsupervised FE to select set of protein based upon PPI network.

Table 7 Top ranked 10 drugs in "DSigDB" category of Enrichr for 57 gene names.

Term	Overlap	P-value	Adjusted P-value
Vorinostat CTD 00003560	23/425	1.11×10^{-24}	2.69×10^{-21}
FENRETINIDE CTD 00007166	16/228	6.88×10^{-19}	8.36×10^{-16}
curcumin CTD 00000663	20/528	2.65×10^{-18}	2.15×10^{-15}
resveratrol CTD 00002483	28/1601	3.80×10^{-17}	2.31×10^{-14}
doxorubicin CTD 00005874	21/750	1.40×10^{-16}	6.79×10^{-14}
aspirin CTD 00005447	19/561	1.77×10^{-16}	7.17×10^{-14}
bay 11-7082 CTD 00003959	11/83	2.62×10^{-16}	9.09×10^{-14}
Arsenenous acid CTD 00000922	25/1283	3.34×10^{-16}	1.02×10^{-13}
tamoxifen CTD 00006827	20/802	8.12×10^{-15}	2.12×10^{-12}
Leptomycin B CTD 00001805	9/50	8.70×10^{-15}	2.12×10^{-12}

Table 8 Top ranked 10 drugs in "DrugMatrix" category of Enrichr for 193 gene names.

Term	Overlap	P-value	Adjusted P-value
Ceramide C2-44 uM in DMSO-Rat-Primary rat hepatocytes-0.67d-up	16/311	5.40×10^{-8}	4.14×10^{-4}
Ceramide C2-44 uM in DMSO-Rat-Primary rat hepatocytes-1d-up	13/280	3.12×10^{-6}	1.19×10^{-2}
Cerivastatin-7 mg/kg in Corn Oil-Rat-Skeletal muscle-5d-up	12/274	1.35×10^{-5}	3.46×10^{-2}
Methyl Salicylate-444 mg/kg in Corn Oil-Rat-Kidney-3d-up	12/290	2.37×10^{-5}	4.37×10^{-2}
Lithocholic Acid-74 uM in DMSO-Rat-Primary rat hepatocytes-0.67d-up	12/300	3.31×10^{-5}	4.37×10^{-2}
Stavudine-1120 uM in DMSO-Rat-Primary rat hepatocytes-0.67d-up	12/301	3.42×10^{-5}	4.37×10^{-2}
Carbon Tetrachloride-400 mg/kg in Corn Oil-Rat-Liver-0.25d-up	11/280	8.30×10^{-5}	9.09×10^{-2}
Microcystin-LR-0.07 uM in DMSO-Rat-Primary rat hepatocytes-0.67d-up	11/309	1.97×10^{-4}	1.66×10^{-1}
Pantoprazole-650 uM in DMSO-Rat-Primary rat hepatocytes-0.67d-up	11/310	2.03×10^{-4}	1.66×10^{-1}
Daunorubicin-1.5 uM in DMSO-Rat-Primary rat hepatocytes-1d-up	11/324	2.96×10^{-4}	1.66×10^{-1}

2.2 PCA based unsupervised FE

One might wonder why we needed to integrate human PPI and mouse PPI; simply using only human PPI might result in similar or even better performance. To address this problem, we applied PCA based unsupervised FE [3, 4] to human PPI and tried to get set of proteins associated with adjusted P-values less than 0.01. Nevertheless, when we consider the u_{2i}, we can find only one protein associated with adjusted P-values less than 0.01. Thus, since it is obvious that TD based unsupervised FE is superior to PCA based unsupervised FE, TD based unsupervised FE is worth while testing.

Table 9 Top ranked 10 drugs in "Drug Perturbations from GEO down" category of Enrichr for 193 gene names.

Term	Overlap	P-value	Adjusted P-value
dactinomycin DB00970 mouse GSE5324 sample 2701	22/522	6.08×10^{-9}	5.20×10^{-6}
dactinomycin DB00970 mouse GSE5324 sample 2700	22/544	1.28×10^{-8}	5.48×10^{-6}
cisplatin DB00515 human GSE6410 sample 2532	15/315	3.86×10^{-7}	1.10×10^{-4}
SODIUM BUTYRATE 5222465 mouse GSE4410 sample 3582	14/289	7.78×10^{-7}	1.33×10^{-4}
Etanercept DB00005 human GSE7524 sample 3295	16/379	7.80×10^{-7}	1.33×10^{-4}
Curcumin CID 969516 human GSE10896 sample 3259	17/458	2.04×10^{-6}	2.91×10^{-4}
actinomycin D 2019 human GSE6400 sample 3098	14/323	2.88×10^{-6}	3.52×10^{-4}
SODIUM BUTYRATE 5222465 mouse GSE4410 sample 3584	13/343	2.72×10^{-5}	2.91×10^{-3}
lapatinib DB01259 human GSE38376 sample 2585	12/301	3.42×10^{-5}	3.25×10^{-3}
HYPOCHLOROUS ACID 24341 human GSE11630 sample 3202	11/289	1.10×10^{-4}	9.07×10^{-3}

Table 10 Top ranked 10 drugs in "Drug Perturbations from GEO down" category of Enrichr for 57 gene names.

Term	Overlap	P-value	Adjusted P-value
trastuzumab DB00072 human GSE31432 sample 3123	7/280	9.87×10^{-6}	2.69×10^{-3}
ZINC ACETATE 11192 human GSE2964 sample 3589	6/189	1.17×10^{-5}	2.69×10^{-3}
lapatinib DB01259 human GSE38376 sample 2585	7/301	1.58×10^{-5}	2.69×10^{-3}
doxycycline DB00254 human GSE2624 sample 3074	8/425	1.72×10^{-5}	2.69×10^{-3}
genistein 5280961 mouse GSE2889 sample 3606	6/247	5.25×10^{-5}	6.57×10^{-3}
doxycycline DB00254 human GSE2624 sample 3077	7/391	8.30×10^{-5}	8.37×10^{-3}
2,3,7,8-tetrachlorodibenzo-p-dioxin 10852289 mouse GSE2812 sample 3602	7/400	9.57×10^{-5}	8.37×10^{-3}
Bisphenol A 6623 human GSE17624 sample 2654	6/281	1.07×10^{-4}	8.37×10^{-3}
lapatinib DB01259 human GSE38376 sample 2586	6/291	1.29×10^{-4}	9.00×10^{-3}
estradiol 5757 mouse GSE2889 sample 3609	7/433	1.56×10^{-4}	9.78×10^{-3}

2.3 Cluster analyses

Yet another concern is if TD based unsupervised FE is required when selecting set of proteins, since there are substantially many methods to select sub-clustrs within large network; such methods can replace TD based unsupervised FE to select set of proteins. Following Zhao et al [10], we employed three methods, multi-level [11], label-progration [12], and edge-betweenness [13], that enable us to identify sub-clusters. Table 13 lists the performances of identification of diseases and drugs significantly associated with selected proteins in individual identified sub-clusters (label-progration

Table 11 Top ranked 10 drugs in "Drug Perturbations from GEO up" category of Enrichr for 193 gene names.

Term	Overlap	P-value	Adjusted P-value
bortezomib DB00188 human GSE30931 sample 2686	18/364	1.41×10^{-8}	1.23×10^{-5}
Triiodothyronine-[13C6] hydrochloride (T3 thyronine) 53442265 human GSE29159 sample 3045	15/332	7.50×10^{-7}	2.53×10^{-4}
phenol 996 human GSE6907 sample 3528	15/336	8.72×10^{-7}	2.53×10^{-4}
resveratrol DB02709 human GSE36930 sample 3498	15/349	1.40×10^{-6}	2.96×10^{-4}
Promegestone 36709 human GSE67561 sample 3694	13/265	1.70×10^{-6}	2.96×10^{-4}
estradiol 5757 mouse GSE2195 sample 3619	13/271	2.17×10^{-6}	3.16×10^{-4}
Bisphenol A 6623 mouse GSE4650 sample 3574	11/202	4.02×10^{-6}	5.01×10^{-4}
pioglitazone DB01132 rat GSE20219 sample 2794	13/292	4.92×10^{-6}	5.37×10^{-4}
Triiodothyronine-[13C6] hydrochloride (T3 thyronine) 53442265 human GSE29159 sample 3044	13/300	6.60×10^{-6}	6.39×10^{-4}
NVP-TNKS656 (tankyrase inhibitor) 71227201 human GSE55624 sample 3247	13/307	8.45×10^{-6}	7.37×10^{-4}

Table 12 Top ranked 10 drugs in "Drug Perturbations from GEO up" category of Enrichr for 57 gene names.

Term	Overlap	P-value	Adjusted P-value
cisplatin DB00515 mouse GSE6206 sample 3412	8/279	7.74×10^{-7}	3.96×10^{-4}
Bisphenol A 6623 mouse GSE4650 sample 3574	7/202	1.14×10^{-6}	3.96×10^{-4}
lipopolysaccharide (LPS) 11970143 mouse GSE23639 sample 3443	7/261	6.24×10^{-6}	1.44×10^{-3}
Mehp 20393 rat GSE4514 sample 3578	7/293	1.32×10^{-5}	1.84×10^{-3}
Hydrogen Peroxide 784 mouse GSE3078 sample 3599	6/199	1.56×10^{-5}	1.84×10^{-3}
levetiracetam 5284583 rat GSE2880 sample 2669	6/201	1.65×10^{-5}	1.84×10^{-3}
2,3,7,8-tetrachlorodibenzo-p-dioxin 10852289 human GSE24193 sample 3189	7/309	1.87×10^{-5}	1.84×10^{-3}
Triiodothyronine-[13C6] hydrochloride (T3 thyronine) 53442265 human GSE29159 sample 3043	7/320	2.34×10^{-5}	2.02×10^{-3}
Bisphenol A 6623 mouse GSE4650 sample 3575	6/225	3.12×10^{-5}	2.40×10^{-3}
Phosgene 6371 mouse GSE2565 sample 3611	6/243	4.80×10^{-5}	3.32×10^{-3}

could not identify large enough clusters to be used for enrichment analyses). It is obvious that tensor outperforms clustering methods.

In conclusion, PPI can be a useful information source for drug repositioning if it is used with TD.

3 Discussion

One might wonder what kind of set of genes were selected in this analysis; although the practical procedure is clear, since it is a fully unsupervised method. To understand this point, we have done the following. At first, we computed the mean number of

Table 13 Performances of various protein selection methods. If there are more than or equal to 10 drugs/associated with adjusted P-values less than 0.05, ◯s are shown. Otherwise, the practical numbers of them are preseted.

Methods	Tesnor		multi-level		edge-betweenness	
Order	2nd	3rd	1st	2nd	1st	2nd
# of proteins	195	55	144	138	39	82
# of genes	193	57	128	164	30	111
Jensen DISEASES	◯	◯	◯	◯	7	5
OMIM Disease	9	9	8	0	0	0
LINCS L1000 Chem Pert Consensus Sigs	◯	0	4	1	0	◯
DSigDB	◯	◯	◯	◯	0	◯
DrugMatrix	6	0	0	0	0	0
Drug Perturbations from GEO down	◯	◯	0	0	0	2
Drug Perturbations from GEO up	◯	◯	6	0	1	0

bindings of individual proteins defined as

$$\langle n_{ii'1} \rangle_i = \frac{1}{N} \sum_{i'=1}^{N} n_{ii'1} \tag{1}$$

and compare $\langle n_{ii'1} \rangle_i$ between selected and not-selected proteins for human PPI. As

Table 14 Comparison of $\langle n_{ii'1} \rangle_i$ between the selected and not selected proteins when using various $u_{\ell_1 i}$ to select proteins. P-values were computed by t test with the alternative hypothesis that mean $\langle n_{ii'1} \rangle_i$ of the selected proteins is larger that that of not selected proteins. P-values were computed by Fisher's exact test.

$u_{\ell_1 i}$s used	u_{2i}		u_{3i}		u_{6i}	
	Selected	Not selected	Selected	Not selected	Selected	Not selected
Mean $\langle n_{ii'1} \rangle_i$	1.12×10^{-3}	4.17×10^{-4}	3.12×10^{-3}	4.13×10^{-4}	2.14×10^{-3}	4.26×10^{-4}
P-value	2.4×10^{-4}		5.3×10^{-5}		2.04×10^{-13}	

can be seen in Table 14, $\langle n_{ii'1} \rangle_i$ for 195 proteins selected by u_{2i} and that for 55 proteins selected by u_{3i} are always significantly larger than that of other (not selected) proteins. Thus, TD based unsupervised FE selected proteins that have more interactions in fully data driven and unsupervised manner.

One might wonder if we simply select proteins with more interaction (in the following, we call them as hub proteins with following convention) they can be used for drug repositioning, since proteins selected by TD based unsupervised FE have more interactions. To check this point, we selected top 200 and 50 proteins with more interaction (these numbers are selected so as to be closed to the number of proteins selected using u_{2i} and u_{3i}. See supplementary information for the list of proteins and gene names) using human PPI. Table 15 shows the confusion matrices of selected proteins. Although they are significantly overlapped, since majorities are not shared, they are distinct sets of proteins.

Table 15 Confusion matrix of selected proteins between TD based unsupervised FE and hub proteins.

TD based unsupervised FE		Hub proteins Top 200	
		Not selected	Selected
Using u_{2i}	Not selected	5353	168
	Selected	165	28
	Odds ratio	5.4	
	P-value	4.55×10^{-11}	
		Top 50	
		Not selected	Selected
Using u_{3i}	Not selected	5625	42
	Selected	30	17
	Odds ratio	75.4	
	P-value	3.04×10^{-23}	

To evaluate top 200 and 50 hub proteins, we uploaded associated gene names to Enrichr. Table 16 lists the performances achieved by hub proteins (For detailed 10 top ranked diseases and drugs in individual categories, see Tables S1 to S7 for top 200 hub proteins and Tables S8 to S14 for top 50 hub proteins). Their performances are essentially same as those by TD based unsupervised FE excluding DrugMatrix.

Table 16 Performances achieved by hub proteins. If there are more than or equal to 10 drugs/associated with adjusted P-values less than 0.05, ◯s are shown. Otherwise, the practical numbers of them are presented.

Methods	Tesnor			Hub proteins	
Order	2nd	3rd	6th	Top 200	Top 50
# of proteins	195	55	38	179	45
# of genes	193	57	41	190	59
Jensen DISEASES	◯	◯	◯	◯	◯
OMIM Disease	9	9	◯	◯	6
LINCS L1000 Chem Pert Consensus Sigs	◯	0	6	◯	0
DSigDB	◯	◯	◯	◯	◯
DrugMatrix	6	0	0	0	1
Drug Perturbations from GEO down	◯	◯	◯	◯	◯
Drug Perturbations from GEO up	◯	◯	◯	◯	◯

One might wonder that TD based unsupervised FE is useless if hub proteins can achieve similar performance. At first, to our knowledge, there are no other studies that use hub proteins derived from only PPI for drug repositioning without using other information since none thinks that hub proteins derived from PPI themselves without using other information can be used for drug repositioning for cancers. The reason why we could identify this fact is because proteins selected by TD based unsupervised FE are hub proteins. Without TD based unsupervised FE, we could not find that hub proteins derived only from PPI can be used for drug repositioning toward cancers. Second, hub proteins failed to identify hits for DrugMatrix. DrugMatrix is

in vivo specific database. This means that hub proteins do not have ability to find drug that might be useful in *in vivo* experiments. Thus, even if hub proteins can achieve similar performance toward *in vitro* experiments, TD based unsupervised FE is still useful. Third, using hub proteins we cannot target diseases other than cancers since hub proteins are enriched mainly in cancers. For top 200 hub proteins, there are as many as six cancers ("Cancer", "Lymphoid leukemia ", "Intestinal benign neoplasm", "Immune system cancer", "Biliary tract cancer", and "Stomach cancer ") within top ranked diseases in "Jensen DISEASES" category and as many as seven cancers ("colorectal cancer", "ovarian cancer", "breast cancer", "leukemia", "lung cancer", "melanoma", and "pancreatic cancer") within top ranked diseases in "OMIM Diseases" category (Tables S1 and S2). For top 50 hub proteins, there are as many as five cancers("Cancer", "DOID:9917"(Pleural cancer), "Thyroid cancer", "Esophageal carcinoma", and "Intestinal benign neoplasm") within top ranked diseases in "Jensen DISEASES" category and as many as five cancers ("colorectal cancer", "ovarian cancer", "breast cancer", "lung cancer", and "pancreatic cancer") within top ranked diseases in "OMIM Diseases" category (Tables S8 and S9). Nevertheless, since TD based unsupervised FE can have freedom to select $u_{\ell_1 i}$, it can provide us the set of proteins enriched in diseases other than cancers. Tables 17 and 18 list the diseases enriched in proteins selected by u_{6i} (See supplementary information for the list of proteins and gene names); they are mainly distinct from cancers (Only one cancer ("Cancer") within top ranked 10 diseases in "Jensen DISEASES" category (Table 17) and three cancers ("pancreatic cancer", "melanoma", and "lung cancer") within top ranked 10 diseases in "OMIM Diseases" category (Table 18)). On the other hand, there are enough number of significant drugs (For detailed drug names, see Tables S15 to S19). Thus considering TD based unsupervised FE is still useful even if hub proteins can identify drugs for cancers.

Table 17 Top ranked 10 diseases in "Jensen DISEASES" category of Enrichr for gene names selected by u_{6i}.

Term	Overlap	P-value	Adjusted P-value
Histoplasmosis	3/17	4.57×10^{-6}	6.86×10^{-4}
Hypohidrotic ectodermal dysplasia	2/12	2.42×10^{-4}	1.81×10^{-2}
Arthritis	4/186	4.62×10^{-4}	2.31×10^{-2}
Progressive supranuclear palsy	2/24	9.95×10^{-4}	3.73×10^{-2}
Lung disease	3/119	1.61×10^{-3}	3.80×10^{-2}
Wolf-Hirschhorn syndrome	2/31	1.66×10^{-3}	3.80×10^{-2}
Chagas disease	2/32	1.77×10^{-3}	3.80×10^{-2}
Exanthem	2/36	2.24×10^{-3}	4.20×10^{-2}
Cancer	4/300	2.70×10^{-3}	4.50×10^{-2}
Gastritis	2/43	3.18×10^{-3}	4.54×10^{-2}

One might also wonder what the difference between hub proteins and proteins selected by TD based unsupervised FE if they are not identical. To see this, the mean

Table 18 Top ranked 10 diseases in "OMIM Diseases" category of Enrichr for gene names selected by u_{6i}.

Term	Overlap	P-value	Adjusted P-value
ectodermal dysplasia	2/11	2.02×10^{-4}	2.22×10^{-3}
immunodeficiency	2/28	1.36×10^{-3}	7.46×10^{-3}
pancreatic cancer	1/11	2.12×10^{-2}	4.23×10^{-2}
dementia	1/12	2.32×10^{-2}	4.23×10^{-2}
melanoma	1/13	2.51×10^{-2}	4.23×10^{-2}
adenoma	1/14	2.70×10^{-2}	4.23×10^{-2}
malaria	1/14	2.70×10^{-2}	4.23×10^{-2}
migraine	1/16	3.08×10^{-2}	4.23×10^{-2}
orofacial cleft	1/19	3.64×10^{-2}	4.39×10^{-2}
lung cancer	1/21	4.02×10^{-2}	4.39×10^{-2}

correlation coefficient among proteins

$$\langle r \rangle = \frac{2}{N(N-1)} \sum_{i'' \neq i'} r(n_{ii'1}, n_{ii''1}) \tag{2}$$

where $r(n_{ii'1}, n_{ii''1})$ is Pearson's correlation coefficient between $\{n_{ii'1} | 1 \leq i \leq N\}$ and $\{n_{ii''1} | 1 \leq i \leq N\}$, were computed with human PPI. Table 19 list the $\langle r \rangle$s. Since

Table 19 $\langle r \rangle$s and P-values computed by one way t test

	$\langle r \rangle$	P-values
Proteins selected using u_{2i}	3.29×10^{-1}	0
Proteins selected using u_{3i}	9.64×10^{-2}	7.70×10^{-211}
Proteins selected using u_{6i}	8.14×10^{-2}	5.84×10^{-70}
Top 200 proteins	7.49×10^{-3}	3.97×10^{-128}
Top 50 proteins	1.19×10^{-3}	6.84×10^{-18}

proteins selected by TD based unsupervised FE have substantially larger $\langle r \rangle$, the one of distinctions between proteins selected by TD based unsupervised FE and hub proteins is definitely correlations. Since the larger $\langle r \rangle$ means that proteins share the proteins to be interacted, it might mean that proteins that share interations might be the key to find drugs effective in *in vivo* experiments.

4 Methods

4.1 Protein protein interaction (PPI)

PPI data was downloaded from DIP [14]. The data sets used are species-specific sets for *Homo Sapiens*, Hsapi20170205, and *Mus musculus*, Mmusc20170205. "full [.gz]" for tab limited files were downloaded and loaded into R by `read_csv` command as $n^k_{ii'} \in \mathbb{R}^{N_k \times N_k}$ ($k = 1$: human and $k = 2$:mouse) that takes 1 when the ith and i'th proteins are interacting with each other otherwise 0. Since these matrices are sparse, they are stored as sparse matrix format using Matrix [15] package.

4.2 Principal component analysis (PCA) based unsupervised feature extraction (FE)

Singular value decomposition was applied to using `irlba` function in irlba package [16] to human PPI, $n^1_{ii'}$, and we get

$$n^1_{ii'} = \sum_{\ell=1}^{L} \lambda_\ell u_{\ell i} u_{\ell i'} \tag{3}$$

where λ_ℓ is a singular value and $u_{\ell i} \in \mathbb{R}^{N \times N}$ is singular value vector and orthogonal vector.

P-values are attributed to ith protein as

$$\langle u_{\ell i} \rangle = \frac{1}{N} \sum_{i=1}^{N} u_{\ell i} \tag{4}$$

$$\sigma_\ell = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(u_{\ell i} - \langle u_{\ell i} \rangle \right)^2} \tag{5}$$

$$P_i = P_{\chi^2} \left[> \left(\frac{u_{\ell i} - \langle u_{\ell i} \rangle}{\sigma_\ell} \right)^2 \right] \tag{6}$$

where $P_{\chi^2}[> x]$ is the cumulative χ^2 distribution where the argument is larger than x and $\langle u_{\ell i} \rangle$ and σ_ℓ are mean and standard deviation, respectively. This means that we assume the null hypothesis that $u_{\ll i}$ obeys Gaussian distribution. P-values are corrected by BH criterion and proteins associated with adjusted P-values less than 0.01 are selected.

4.3 Tensor decomposition (TD) based unsupervised FE

To apply TD based unsupervised FE to PPI, we need to construct integrated tensor that stores multiple PPIs. Suppose we have two PPI matrices, $n^1_{ii'} \in \mathbb{R}^{N_1 \times N_1}$ and $n^2_{ii'} \in \mathbb{R}^{N_2 \times N_2}$. It is also supposed that there are N_{common} common proteins, N_{ortho} orthogonal proteins between two data sets. Then we merge these data set into one tensor, $n_{ii'2} \in \mathbb{R}^{N \times N \times 2}$ where $N = N_1 + N_2 - N_{\text{common}} - N_{\text{otrho}}$ as follows (Fig. 2). $n^k_{ii'}, (1 \leq i, i' \leq N_{\text{common}} + N_{\text{ortho}})$ are placed in $n_{ii'k}$ as it is (Yellow and red regions in Fig. 2). $n^1_{ii'}, (N_{\text{common}} + N_{\text{ortho}} < i, i' \leq N_1)$ are placed in $n_{ii'1}$ as it is (Blue region in Fig. 2). $n^2_{ii'}, (N_{\text{common}} + N_{\text{ortho}} < i, i' \leq N_2)$ are placed in $n_{ii'2}$ in a fragmented manner (Green region in Fig. 2). As a result, blue region of $n_{ii'2}$ and green region of $n_{ii'1}$ are blank. Shaded region is blank in $n_{ii'k}$.

HOSVD was applied to $n_{ii'k}$ and we got

$$n_{ii'k} = \sum_{\ell_1=1}^{N} \sum_{\ell_2=1}^{N} \sum_{\ell_3=1}^{2} G(\ell_1 \ell_2 \ell_3) u_{\ell_1 i} u_{\ell_2 i'} u_{\ell_3 k} \tag{7}$$

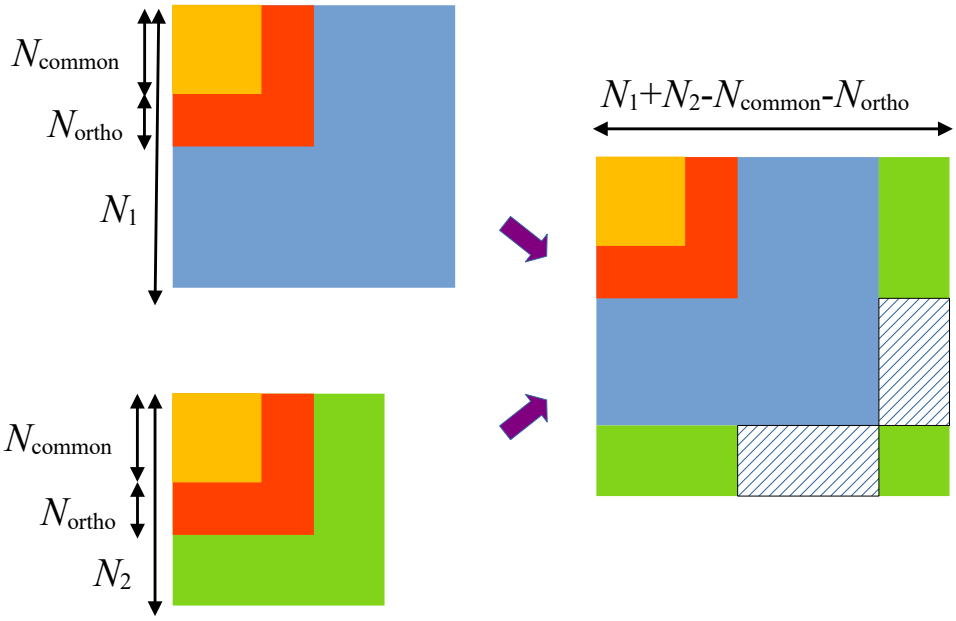

Fig. 2 How to merge two PPI matrices into one tensor.

where G is a core tensor that represents the weight(contribution) of $u_{\ell_1 i} u_{\ell_2 i'} u_{\ell_3 k}$ to $n_{ii'2}$ and $u_{\ell_1 i} = u_{\ell_2 i'} \in \mathbb{R}^{N \times N}$ and $u_{\ell_3 k} \in \mathbb{R}^{2 \times 2}$ are singular value matrices and orthogonal matrices. Attribution of P_is to is and the selection of proteins were performed by replacing $u_{\ell i}$ with $u_{\ell_1 i}$ in eq. (6).

HOSVD was performed using `irlba` function by applying SVD to unfolded matrices, since usual R function that can perform HOSVD does not accept sparse matrix as input.

4.4 Identification of orthogonal proteins

Orthologs between human and mouse were retrieved by querying "((organism_id:10090) OR ((organism_id:9606) AND (reviewed:true))) AND (database:orthodb)" in uniport search (https://www.uniprot.org/). Human uniprot accessions are converted to gene name as "XXX_HUMAN" using the above retrieved information. Then after seeking "XXX_MOUSE" in the above retrieved information (using the first hit if multiple hits are found), "XXX_MOUSE" is converted to the corresponding mouse uniprot accession numbers in the above retrieved information. This results in the corresponding table of uniprot ortholog corresponding table (The above retrieved information is provided as supplementary material).

4.5 Gene ID conversion

The obtained uniprot accession numbers attributed to proteins are converted to gene name by nuiprot ID mapping (https://www.uniprot.org/id-mapping).

5 Conclusion

In this study, we have demonstrated how useful for drug repositioning TD based unsupervised FE applied to PPI is. Although it is unlikely that only information retrieved from PPI is useful for disease-specific drug repositioning, it practically works. TD based unsupervised FE applied PPI is useful information for drug repositioning. More studies are expected to understand how much this strategy works.

Author Contributions

Y.-H.T. planned the research and performed the analyses. Y.-H.T. and T.T. evaluated the results, discussions, and outcomes and wrote and reviewed the manuscript. All authors have read and agreed to the published version of the manuscript

Funding

This work was supported by KAKENHI [grant numbers 24K15168] and Chuo Universiy Tokutei Kadai kenkyuu.

Data and Availability

All data used in this study can be downloaded from DIP [14].

Declarations

Not applicable.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no conflict of interest.

References

[1] Wang, L., Wang, Y., Hu, Q., Li, S.: Systematic analysis of new drug indications by drug-gene-disease coherent subnetworks. CPT: Pharmacometrics &

Systems Pharmacology **3**(11), 146 (2014) https://doi.org/10.1038/psp.2014.44 https://ascpt.onlinelibrary.wiley.com/doi/pdf/10.1038/psp.2014.44

[2] Kim, Y., Cho, Y.-R.: Predicting drug–gene–disease associations by tensor decomposition for network-based computational drug repositioning. Biomedicines **11**(7) (2023) https://doi.org/10.3390/biomedicines11071998

[3] Taguchi, Y.-h.: Unsupervised Feature Extraction Applied to Bioinformatics: A PCA Based and TD Based Approach, 1st edn. Springer, ??? (2020). https://doi.org/10.1007/978-3-030-22456-1 . http://dx.doi.org/10.1007/978-3-030-22456-1

[4] Taguchi, Y.-h.: Unsupervised Feature Extraction Applied to Bioinformatics: A PCA Based and TD Based Approach, 2nd edn. Springer, ??? (2024)

[5] Taguchi, Y.-H., Turki, T.: Integrated analysis of gene expression and protein–protein interaction with tensor decomposition. Mathematics **11**(17) (2023) https://doi.org/10.3390/math11173655

[6] Xie, Z., Bailey, A., Kuleshov, M.V., Clarke, D.J.B., Evangelista, J.E., Jenkins, S.L., Lachmann, A., Wojciechowicz, M.L., Kropiwnicki, E., Jagodnik, K.M., Jeon, M., Ma'ayan, A.: Gene set knowledge discovery with enrichr. Current Protocols **1**(3), 90 (2021) https://doi.org/10.1002/cpz1.90 https://currentprotocols.onlinelibrary.wiley.com/doi/pdf/10.1002/cpz1.90

[7] Carr, S., Kasi, A.: Familial Adenomatous Polyposis. StatPearls Publishing, Treasure Island (FL) (2024). Updated 2023 Feb 25. https://www.ncbi.nlm.nih.gov/books/NBK538233/

[8] Aedma, S.K., Kasi, A.: Li-Fraumeni Syndrome. StatPearls Publishing, Treasure Island (FL) (2024). PMID: 30335319. https://pubmed.ncbi.nlm.nih.gov/30335319/

[9] Bhandari, J., Thada, P.K., Puckett, Y.: Fanconi Anemia. StatPearls Publishing, Treasure Island (FL) (2024). Last Update: August 10, 2022. https://www.ncbi.nlm.nih.gov/books/NBK559133/

[10] Zhao, Q., Zhang, Y., Shao, S., Sun, Y., Lin, Z.: Identification of hub genes and biological pathways in hepatocellular carcinoma by integrated bioinformatics analysis. PeerJ **9**, 10594 (2021) https://doi.org/10.7717/peerj.10594

[11] Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment **2008**(10), 10008 (2008) https://doi.org/10.1088/1742-5468/2008/10/P10008

[12] Raghavan, U.N., Albert, R., Kumara, S.: Near linear time algorithm to detect community structures in large-scale networks. Phys. Rev. E **76**, 036106 (2007)

https://doi.org/10.1103/PhysRevE.76.036106

[13] Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E **69**, 026113 (2004) https://doi.org/10.1103/PhysRevE.69.026113

[14] Xenarios, I., Rice, D.W., Salwinski, L., Baron, M.K., Marcotte, E.M., Eisenberg, D.: DIP: the Database of Interacting Proteins. Nucleic Acids Research **28**(1), 289–291 (2000) https://doi.org/10.1093/nar/28.1.289 https://academic.oup.com/nar/article-pdf/28/1/289/9895160/280289.pdf

[15] Bates, D., Maechler, M., Jagan, M.: Matrix: Sparse and Dense Matrix Classes and Methods. (2024). R package version 1.7-0. https://CRAN.R-project.org/package=Matrix

[16] Baglama, J., Reichel, L., Lewis, B.W.: Irlba: Fast Truncated Singular Value Decomposition and Principal Components Analysis for Large Dense and Sparse Matrices. (2022). R package version 2.3.5.1. https://CRAN.R-project.org/package=irlba