Ranking genes based on topics

Nantia Iakovidou

Department of Informatics, Aristotle University of Thessaloniki
GR-54124 Thessaloniki, Greece
niakovid@csd.auth.gr

Abstract. The vast majority of existing methods for ranking gene datasets, do not include or take into account in their exported results other information that might accompany the genes, such as specie or terms. Given a higher order biological data set, we propose a methodology based on multilinear algebra for ranking genes across multiple dimensions. We apply PARAFAC decomposition on a Gene Ontology dataset (GO data) for multiple species and reveal interesting experimental results that provide to the user more information than other consisted methods.

Key words: Parafac decomposition, multi-dimensional biological data.

1 Introduction

Several methods for ranking genes have been proposed in recent years, such as GeneRank [8] and HITS [4] algorithm. All these methods perform ranking across one dimension, that of the totality of the genes and generate only one list of results. In other words they produce one list of scored genes that is consistent to a query gene.

The problem that arises here is how we can rank biological data in which more than one dimensions are involved. This is exactly the case that is investigated in this paper. In particular, we study a drosophila dataset which is a tensor that contains GO terms, species, genes and an integer number that represents the homolog genes within the genome. The question that emerges here is to apply a method so that genes can be ranked across terms and species at the same time. In this way the user would understand better and elicit more information from a higher-order representation of the genes.

The aim of such a task is to analyze genetic diversity and to use this knowledge for discerning the evolutionary relationships among species (i.e. phylogeny reconstruction), comparing different kinds of species and understanding complexities of biological processes (e.g. evolution of genetic regulation).

The solution that we suggest comes from the area of multilinear algebra. The idea that we develop in this paper is to decompose the initial tensor data set to its own distinct dimensions and assign a score to each element. In this way a ranked list of genes will result again, but this time will be followed by its relevant ranked list of terms and ranked list of species at the same time. By using PARAFAC-ALS (Alternating Least Squares) a number of factors are

derived and each one of them is associated with a scored list of "hosts" and a scored list of "terms". In this paper by the word hosts we mean genes and species, while the list of terms is produced from the GO-terms. The results from the aforementioned method are presented and discussed to the related section of this paper.

The remainder of the paper is organized as follows. Section 2 mentions some previous related work. Section 3 describes the method used in this paper. The next section introduces applications and examples that concern the proposed method. Section 5 provides an analytical description of the data set used in the particular study and apposes a discussion about the obtained experimental results. Lastly, we draw some conclusions derived from this work and mention future work that can extend the present paper.

2 Related Work

Authors in [1] describe the Singular Value Decomposition method (SVD) for transforming genome-wide expression data from genes × arrays space to reduced diagonalized "eigengenes" × "eigenarrays" space, where the eigengenes (or eigenarrays) are unique orthonormal superpositions of the genes (or arrays). They use this mathematical framework to prove that processing and modelling genome-wide expression data can lead to meaningful results for biology and medicine.

In linear algebra, the singular value decomposition (SVD) is an important factorization of a rectangular real or complex two-way matrix. Applications which employ the SVD include computing the pseudoinverse, least squares fitting of data, matrix approximation, and determining the rank, range and null space of a matrix. On the other hand, in multilinear algebra, there does not exist a multi-way svd that has all the properties of a matrix SVD. A matrix SVD simultaneously computes the orthonormal row/column matrices and a rank-R decomposition. This is generally not possible for multi-way arrays or "data-tensors". Instead, there exist two types of decompositions for multi-way arrays that capture different properties of the matrix svd. One decomposition represents a tensor as sum of rank-1 tensors (Parafac), while the second computes the othonormal spaces associated with the different axes or modes of a tensor. The latter decomposition has been known as HOSVD.

Authors in [9] describe the use of the higher-order singular value decomposition (HOSVD) method for transforming a data tensor of genes into a linear superposition of rank-1 subtensors. By using this framework for analysis of DNA microarray data from different studies, the authors revealed important results about the role of several genes on cell cycle progression.

Kolda et al. proposed a method called TOPHITS which analyzes a semantic graph that combines anchor text with the hyperlink structure of the web [5, 7]. The adjacency structure of the semantic graph is modelled by a three-way tensor containing both hyperlink and anchor text information. Then the authors apply the Parallel Factor (PARAFAC) decomposition, which is a higher-order

analogue of the two-way SVD, and produce triplets of vectors with authority and hub scores for the pages as well as topic scores for the terms. This algorithm is an extension of the Kleinberg's HITS algorithm [4] which uses the singular vectors of the hyperlink matrix (a two-way tensor) to produce multiple sets of hubs and authorities.

Several other papers use drosophila datasets and compare GO terms across species as well [11, 10]. Since ontologies are identical, GO terms can be compared across species. It is worth to mention here that drosophila datasets [12] have been used in very important investigations that study experimental questions such as aging, DNA-damage response, immune response, resistance to DDT and embryonic development [10].

3 Methodology

As mentioned before, in this paper we use a Gene Ontology dataset that contains GO terms, species, genes and an integer number that represents the homolog genes within the genome, which is denoted as frequency. This particular dataset originates from a genus of small flies called drosophila and contains 1473 genes taken from 12 different species of it. It also contains 100 GO terms that are associated with the genes. Since we used a tensor (multidimensional array) to model the aforementioned data it seems inevitable to use in turn multilinear algebra methods that operate tensors, so that we can handle and process the data in a better way.

Here, we focus on PARAFAC (PARAllel FACtor analysis) decomposition method that is common in multilinear algebra. PARAFAC constitutes a generalization of the PCA method to higher orders [2]. In the following, scalars are indicated by lower-case letters, bold capitals are used for two-way matrices and italics capital letters are used for three-way arrays.

3.1 PARAFAC

Parafac is one of several decomposition methods for multi-way data. PARAFAC will decompose a tensor of order N, where $N \geq 3$ into the summation over the outer product of N vectors (a low-rank model). If the order of a tensor is 3 ($N = 3$) then the size of the tensor is for example I by J by K. For instance, given a third-order tensor $X \in R^{I \times J \times K}$ we wish to write it as in (1), where a_{if}, b_{jf}, c_{kf} are elements of the produced matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$. Number F represents the number of components of the PARAFAC decomposition [6].

$$X_{ijk} = \sum_{f=1}^{F} a_{if} b_{jf} c_{kf} \tag{1}$$

Fig. 1. The PARAFAC method provide a 3-way decomposition that yields terms, genes and species scores.

4 Applications and Examples

GeWare data warehouse system [3] is an application for microarray-based gene expression analysis, which offers high flexibility with multidimensional data models. In these models data is stored in several fact tables which are associated with multiple hierarchical dimensions holding describing annotations (e.g GO annotations) on genes, samples, experiments and processing methods.

GeWare can support algorithms for preprocessing and analyzing multidimensional gene datasets, e.g. to identify lists of interesting genes. The analysis methods are coupled in a simple and powerful way of exchanging experiment groups, gene groups and gene expression matrices.

The GeWare system has been employed in several research projects which study for example the role of the transcription factor IL-6 on the survival of myeloma cells and the factors influencing the binding behavior of sequences on microarrays.

5 Experimental Results

The data set used in this paper [1] is a sparse tensor which contains GO terms, species, genes and an integer number that represents the homolog genes within the genome (denoted as frequency). The size of the sparse tensor is 100 GO terms × 12 species × 1473 genes. In general, a GO term consists of a term name (e.g. cell) and a zero-padded seven-digit identifier (or accession number) prefixed by GO: (e.g. GO: 0005623), which is used as a unique identifier and database cross-reference. Species are numbered at Table 1, while genes are represented by a five-digit identifier prefixed by CG (e.g. CG31618).

1	2	3	4	5	6	7	8	9	10	11	12
dmel	*dsim*	*dsec*	*dyak*	*dere*	*dana*	*dper*	*dpse*	*dwil*	*dmoj*	*dvir*	*dgri*

Table 1. Drosophila species.

[1] Downloaded from http://insects.eugenes.org/species/

By applying the PARAFAC-ALS decomposition to the initial tensor dataset using the matlab tensor toolbox, three matrices are derived. In each one of them the number of the columns is equal to the number of factors (set by the user) and the number of lines is:

- equal to the number of GO terms for the first matrix
- equal to the number of species for the second matrix and
- equal to the number of genes for the third matrix.

All these matrices represent scored lists of topics and hosts, which are identified with the lists of factors. As mentioned before, in this paper by the word hosts we mean genes and species, while the list of topics is produced from the GO terms. The results from the aforementioned method are shown in Table 2.

Table 2 presents the parafac decomposition results for the first factor only. It is important to mention here that parafac can compute results, like those depicted in Table 2 for as many factors as the user needs. Here we examine the results of the first factor, since the results as well as the explanation of the rest of the factors is similar to the first.

PARAFAC RESULTS						
1st factor						
TOPICS			GENES		SPECIES	
$Score$	$Term$	$GOIDs$	$Score$	$Term$	$Score$	$Term$
0.7123	$nucleosome$	$GO:0000786$	0.5770	$CG31618$	0.5295	1
0.7008	$nucleosomeassembly$	$GO:0006334$	0.5451	$CG31613$	0.3960	5
0.0185	$molecularfunction$	$GO:0005554$	0.4836	$CG31617$	0.3832	3
0.0180	$cellularcomponent$	$GO:0008372$	0.3130	$CG31611$	0.3296	12
0.0132	$biologicalprocess$	$GO:0000004$	0.1133	$CG3379$	0.2898	11
0.0105	$monooxygenaseactivity$	$GO:0004497$	0.0963	$CG13329$	0.2552	6
0.0101	$microsome$	$GO:0005792$	0.0605	$CG5499$	0.2276	10
0.0096	$proteinubiquitination$	$GO:0016567$	0.0383	$CG5825$	0.1986	7
0.0091	$ubiquitin-proteinligaseactivity$	$GO:0004842$	0.0355	$CG7793$	0.1692	2
0.0089	$ubiquitinligasecomplex$	$GO:0000151$	0.0352	$CG3281$	0.1583	9
0.0085	$steroidmetabolism$	$GO:0008202$	0.0352	$CG11290$	0.1133	4
0.0076	$electrontransport$	$GO:0006118$	0.0342	$CG3509$	0.0223	8
0.0017	$peripheralnervoussystemdevelop.$	$GO:0007422$	0.0270	$CG32346$		
0.0017	$SCFubiquitinligasecomplex$	$GO:0019005$	0.0211	$CG8625$		
0.0017	$smoothenedsignalingpathway$	$GO:0007224$	0.0210	$CG5017$		
0.0014	$polysaccharidemetabolism$	$GO:0005976$	0.0210	$CG4236$		
0.0010	$proteinmodification$	$GO:0006464$	0.0206	$CG5330$		
0.0009	$proteinmetabolism$	$GO:0019538$	0.0206	$CG12109$		
0.0009	$transferaseactivity$	$GO:0016740$	0.0187	$CG9383$		
0.0008	$spermatogenesis$	$GO:0007283$	0.0184	$CG3708$		
etc	etc	etc	etc	etc		

Table 2. Ranked list of topics and hosts using the PARAFAC decomposition method.

From Table 2 it is quite obvious that the terms "nucleosome" and "nucleosome assembly" are ranked first in the list of topics and all the rest follow with great divergence. This means that the majority of genes scored on the next column should be mainly or only associated with these two GO terms. If someone takes a look to the initial dataset will see that indeed all the listed genes are associated with either both the first two topics or one of them. Exceptions to this rule constitute the genes CG15440 and CG4299 which, apart from the first topics, are also associated with the terms "protein metabolism" and "spermatogenesis". The same applies for the list of species, in relation with the other columns. Table 3 reveals the accuracy of the species ranking in Table 2, since it presents the correspondent part from the initial dataset that contains the first two genes and GO terms that are discussed in Table 2. In other words, Table 2 shows that the genes that are highly ranked on the second column are associated with the terms that are also highly ranked on the first column. The same applies with the third column, as well (species column) in relation with the other two columns. In this way, ranking across all these three dimensions is achieved, which was the target of this method. In Table 3 GO terms are indicated by their GO IDs.

6 Conclusions

In cases of multidimensional datasets, existing methods for ranking genes in biological databases cannot help the user analyze and extract useful information from them. We presented a multilinear algebra based methodology which uses the parafac decomposition to rank genes across multiple dimensions of the initial dataset. The proposed scheme can rank GO terms, genes and species at the same time, by providing accurate results that will help a researcher elicit more information from a higher order representation of the data and handle them in a better way.

7 Future Work

This paper can be extended with the insertion of additional similar methodologies to the one presented, as for example the Tucker decomposition method or other. Results from the aforementioned methods can be compared and discussed across species or across multiple datasets.

Acknowledgments. The following people were kind enough to take a look at earlier versions of this work and offer helpful advice on many fronts: Alexandros Nanopoulos and Yannis Manolopoulos.

References

1. Alter, O., Brown, P.O., Botstein, D.: Singular Value Decomposition for Genome-Wide Expression Data Processing and Modeling. In: Proceedings of the National Academy of Sciences 97 (18), pp. 10101–10106 (2000)

Table 3. Part from the initial dataset, before the application of decomposition.

GO IDs	Sp.	Genes	F	GO IDs	Sp.	Genes	F
GO:0000786	1	CG31618	25	GO:0000786	1	CG31613	23
GO:0000786	2	CG31618	10	GO:0000786	2	CG31613	10
GO:0000786	3	CG31618	18	GO:0000786	3	CG31613	16
GO:0000786	4	CG31618	7	GO:0000786	4	CG31613	0
GO:0000786	5	CG31618	22	GO:0000786	5	CG31613	17
GO:0000786	6	CG31618	15	GO:0000786	6	CG31613	14
GO:0000786	7	CG31618	10	GO:0000786	7	CG31613	9
GO:0000786	8	CG31618	0	GO:0000786	8	CG31613	0
GO:0000786	9	CG31618	10	GO:0000786	9	CG31613	7
GO:0000786	10	CG31618	11	GO:0000786	10	CG31613	12
GO:0000786	11	CG31618	14	GO:0000786	11	CG31613	17
GO:0000786	12	CG31618	15	GO:0000786	12	CG31613	20
GO:0006334	1	CG31618	25	GO:0006334	1	CG31613	23
GO:0006334	2	CG31618	10	GO:0006334	2	CG31613	10
GO:0006334	3	CG31618	18	GO:0006334	3	CG31613	16
GO:0006334	4	CG31618	7	GO:0006334	4	CG31613	0
GO:0006334	5	CG31618	22	GO:0006334	5	CG31613	17
GO:0006334	6	CG31618	15	GO:0006334	6	CG31613	14
GO:0006334	7	CG31618	10	GO:0006334	7	CG31613	9
GO:0006334	8	CG31618	0	GO:0006334	8	CG31613	0
GO:0006334	9	CG31618	10	GO:0006334	9	CG31613	7
GO:0006334	10	CG31618	11	GO:0006334	10	CG31613	12
GO:0006334	11	CG31618	14	GO:0006334	11	CG31613	17
GO:0006334	12	CG31618	15	GO:0006334	12	CG31613	20

Sp. stands for Specie and F for Frequency.

2. Bro, R.: PARAFAC: Tutorial and applications. In: Chemom. Intelligent Lab. Systems., V. 38. pp. 149–171 (1997)
3. Kirsten, T., Do, H., Rahm, E.: A Data Warehouse for Multidimensional Gene Expression Analysis. Technical Report, Interdisciplinary Centre for Bioinformatics, University of Leipzig (2004)
4. Kleinberg, J.M.: Authoritative Sources in a Hyperlinked Environment. In: Journal of the ACM, Vol.46, No.5, pp.604–632 (1999)
5. Kolda, T.G., Bader, B.W.: The TOPHITS model for higher-order web link analysis. In: Workshop on Link Analysis, Counterterrorism and Security (2006)
6. Kolda, T.G., Bader, B.W.: Tensor Decompositions and Applications. Technical report Number SAND2007-6702, Sandia National Laboratories (2007)
7. Kolda, T.G., Bader, B.W., Kenny, J.P.: Higher-Order Web Link Analy-sis Using Multilinear Algebra. In: ICDM 2005: Proceedings of the 5th IEEE International Conference on Data Mining, pp. 242–249 (2005)
8. Morrison, J.L., Breitling, R., Higham, D.J., Gilbert, D.R.: GeneRank: Using search engine technology for the analysis of microarray experiments. In: BMC Bioinformatics, vol.6, pp.1–12 (2005)
9. Omberg, L., Golub, G.H., Alter, O.: A Tensor Higher-Order Singular Value Decomposition for Integrative Analysis of DNA Microarray Data from Different Studies. In: Proceedings of the National Academy of Sciences 104 (47), pp. 18371–18376 (2007)
10. Spellman, P.T., Rubin, G.M.: Evidence for large domains of similarly expressed genes in the Drosophila genome. In: Journal of Biology 1(1):5 (2002)
11. Ueda, H.R., Matsumoto, A., Kawamura, M., Iino, M., Tanimura, T., Hashimoto,S.: Genome-wide Transcriptional Orchestration of Circadian Rhythms in Drosophila. In: The Journal of Biological Chemistry Vol. 277, No. 16, Issue of April 19, pp. 14048–14052 (2002)
12. Affymetrix GeneChip Drosophila Genome Array, [http://www.affymetrix.com/products/arrays/specific/fly.affx]