Discovering Gene Functional Relationships Using a Literature-based NMF Model

Elina Tjioe

Dissertation Defense

Dec 23rd, 2008

1

OUTLINE

1. Introduction
2. Methods
3. FAUN Capabilities and Usability
4. Results
5. Summary and Future Work

2

1. Introduction

3

1.1 Research Problems

- Rapid growth of the biomedical literature
  - MEDLINE 2008 database contains over 17 million records in life sciences
  - The database is growing at an exponential rate

  → *Major challenge to keep track of all new discoveries.*

- Abundance of genomic information
  - Gene sequence analysis does not necessarily imply function
  - Interpretation of high throughput genomic data can be a challenging and daunting process

  → *Major challenge for determining functional relationship among genes.*

- Need a tool to facilitate both the discovery and classification of functional relationships among genes.

  → *Develop a Web-based bioinformatics tool:* ***FAUN (Feature Annotation Using Nonnegative matrix factorization).***

4

1.2 Overview of Previous Work

- Tools that utilize functional gene annotations:
  - Gene Ontology (GO)
  - Medical Subject Heading (MeSH)
  - Kyoto Encyclopedia of Genes and Genomes (KEGG)
- Tools that utilize MEDLINE database:
  - CoPub Mapper → co-occurrence of terms and gene descriptions
  - PubGene → co-occurrence of gene symbols
- Tools that use vector space models:
  - Semantic Gene Organizer (SGO) → based on Latent Semantic Indexing (LSI)

Main limitation of LSI: while it is robust in identifying *what* genes are related, it has difficulty in answering *why* they are related.

5

→ propose using nonnegative matrix factorization (NMF)

1.3 Brief Introduction of NMF

- Lee and Seung (1999) demonstrated the use of NMF in image analysis to both identify and classify image *features.*
- Xu et al.(2003) demonstrated how NMF-based indexing could outperform SVD-based LSI for some information retrieval tasks.
- NMF has been used in many areas including *protein fold recognition, analysis of NMR spectra, speech recognition, video summarization, and internet research.*
- Application of NMF in bioinformatics including *analysis of gene expression data, sequence analysis, gene tree labeling, and functional characterization of gene lists.*
- Chagoyen et al. (2006) demonstrated the use of NMF in extracting the semantic features in biomedical literatures
- Pascual-Montano et al. (2006) developed *bio-NMF* for simultaneous clustering of genes and samples.

6

2. Methods

7

2.1 FAUN Software Architecture

- Computational Core
  - Construct gene document collection
  - Parse the collection
  - Build NMF model
  - Classify new documents based on the NMF model
- Web-based user-interface
  - Interactive components that allow biologists to analyze gene datasets using the NMF model

FAUN utilizes a combination of technologies:

PHP, Javascript, Flash, and C++.

8

2.2 Gene Document Collection

- Express a document collection as a *m x n* matrix *A*
  - *m* = number of terms
  - *n* = *number of documents*

$$A = [w_{ij}] ,$$

$$w_{ij} = l_{ij} \times g_i .$$

- Apply log-entropy term weighting scheme
  → to give distinguishing terms more weight

$$l_{ij} = \log_2(1 + f_{ij}) ,$$

$$g_i = 1 + \left(\frac{\sum_j [p_{ij} \log_2(p_{ij})]}{\log_2 n} \right) ,$$

$$p_{ij} = \frac{f_{ij}}{\sum_j f_{ij}} ,$$

9

2.3 NMF Definition

Given a nonnegative matrix *A* and factorization rank *k*, find *W* and *H* such that $A \approx WH$

that minimize the cost function:

$$f(W,H) = \frac{1}{2} \|A - WH\|_F^2 = \frac{1}{2} \sum_{ij} (A_{ij} - (WH)_{ij})^2$$

- $0 < k \leq \min(m, n)$
- $W, H \geq 0$
- *W* has dimensions *m x k*
- *H* has dimensions *k x n*

10

2.3 continued……

- Initialization Methods
  - W and H are not unique.
    i.e., WD, $D^{-1}H$ for any invertible nonnegative D
  - →To start from a fixed starting point,
    use Nonnegative Double SVD (**NNDSVD**):
    NNDSVDz, NNDSVDa, NNDSVDe, NNDSVDme
- NMF Algorithm: **Multiplicative Update Method**

$$H_{cj} \leftarrow H_{cj} \frac{(W^T A)_{cj}}{(W^T W H)_{cj} + 10^{-9}},$$

$$W_{ic} \leftarrow W_{ic} \frac{(AH^T)_{ic}}{(WHH^T)_{ic} + 10^{-9}}.$$

O(kmn)

11

2.3 continued……

- Additional application-dependent constraints:

$$f(W,H) = \frac{1}{2}\|A - WH\|_F^2 + \alpha J_1(W) + \beta J_2(H)$$

  - **Smoothness constraint**

$$J_1(W) = \|W\|_F^2 \qquad W_{ic} \leftarrow W_{ic} \frac{(AH^T)_{ic} - \alpha W_{ic}}{(WHH^T)_{ic} + 10^{-9}}$$

  - **Sparsity constraint**

$$J_2(H) = (\omega\|vec(H)\|_2 - \|vec(H)\|_1)^2 \qquad \omega = \sqrt{kn} - (\sqrt{kn} - 1)\gamma$$

$$H_{cj} = H_{cj} \frac{(W^T A)_{cj} - \beta(c_1 H_{cj} + c_2)}{(W^T W H)_{cj} + 10^{-9}}$$

$$c_1 = \omega^2 - \omega \frac{\|vec(H)\|_1}{2\|vec(H)\|_2} \qquad c_2 = \|vec(H)\|_1 - \omega\|vec(H)\|_2$$

12

2.3 continued.....

- Alternative NMF algorithm:

  sparse nonnegative matrix factorization (SNMF), which solves the following optimization problem:

$$\min_{W,H} \frac{1}{2}\left\{\|A - WH\|_F^2 + \eta\|W\|_F^2 + \beta\sum_{j=1}^{n}\|H(:,j)\|_1^2\right\}, \quad s.t.\ W,H \geq 0$$

  Each iteration involves solving two nonnegativity constrained least squares problems

$$\min_H \left\|\begin{pmatrix} W \\ \sqrt{\beta}\, e_{1\times k}\end{pmatrix} H - \begin{pmatrix} A \\ 0_{1\times n}\end{pmatrix}\right\|_F^2, \quad s.t.\ H \geq 0,$$

$$\min_W \left\|\begin{pmatrix} H^T \\ \sqrt{\eta} I_k\end{pmatrix} W^T - \begin{pmatrix} A^T \\ 0_{k\times m}\end{pmatrix}\right\|_F^2, \quad s.t.\ W \geq 0,$$

$O(k^4(m+n))$

13

2.4 FAUN Workflow

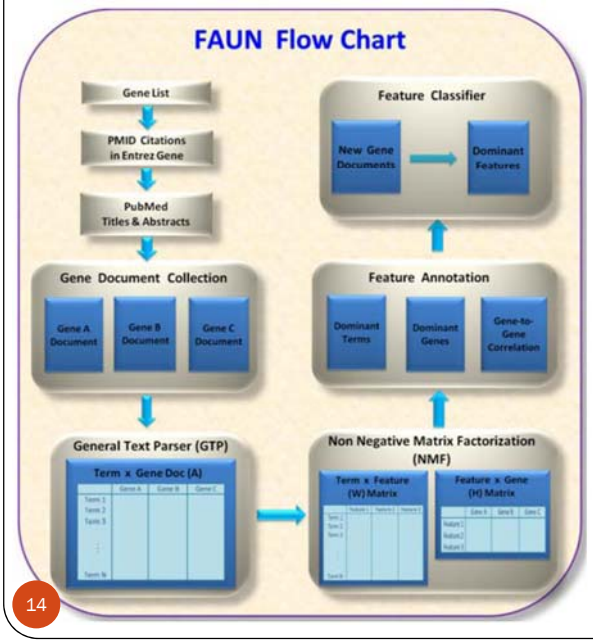


Table A. 2
Term-document matrix for the sample collection in Table A1.

	d1	d2	d3	d4	d5	d6	d7	d8	d9
Alcoholism	—	0.4338	—	—	—	0.2737	—	0.2737	0.4338
Anxiety	0.4745	—	—	—	0.4745	—	—	—	—
Attack	—	—	—	—	0.6931	—	—	—	—
Autism	—	—	—	—	—	—	0.752	—	0.752
Birth	—	—	—	—	—	0.4745	—	—	0.4745
Blood	—	—	—	0.3466	0.3466	0.3466	—	—	—
Bone	—	—	0.752	0.752	—	—	—	—	—
Cancer	—	0.4745	0.4745	—	—	—	—	—	—
Cells	—	—	—	0.6931	—	—	—	—	—
Children	—	—	—	—	—	—	0.4745	—	0.4745
Cirrhosis	—	0.752	—	—	—	—	—	0.752	—

Table A. 6
Top 5 weighted terms for each feature from the sample collection

f1	f2	f3	f4
Bone	Cirrhosis	Stress	Autism
Marrow	Alcoholism	Pressure	Children
Leukemia	Liver	Attack	Speech
Damage	Kidney	Anxiety	Defects
Cancer	Failure	Blood	Birth

Table A. 7
Rearranged term-document matrix for the sample collection.

	d3	d4	d2	d6	d8	d1	d5	d7	d9
Bone	0.752	0.752	—	—	—	—	—	—	—
Cancer	0.4745	—	0.4745	—	—	—	—	—	—
Cells	—	0.6931	—	—	—	—	—	—	—
Damage	0.6931	—	—	—	—	—	—	—	—
Leukemia	1.0986	—	—	—	—	—	—	—	—
Marrow	0.752	0.752	—	—	—	—	—	—	—
Tuberculosis	—	0.6931	—	—	—	—	—	—	—
Alcoholism	—	—	0.4338	0.2737	0.2737	—	—	—	0.4338
Cirrhosis	—	—	0.752	—	0.752	—	—	—	—
Failure	—	—	0.4745	0.4745	—	—	—	—	0.4745
Hypertension	—	—	—	0.6931	—	—	—	—	—
Kidney	—	—	0.4745	0.4745	—	—	—	—	0.4745
Liver	—	—	0.4745	—	0.4745	—	—	—	—
Scarring	—	—	—	—	0.6931	—	—	—	—
Anxiety	—	—	—	—	—	0.4745	0.4745	—	—
Attack	—	—	—	—	—	—	—	0.6931	—
Blood	—	0.3466	—	0.3466	—	—	0.3466	—	—
Pressure	—	—	—	0.4923	—	—	0.7804	—	—
Stress	—	—	—	—	—	0.4923	0.7804	—	—
Autism	—	—	—	—	—	—	—	0.752	0.752
Birth	—	—	—	0.4745	—	—	—	—	0.4745
Children	—	—	—	—	—	—	—	0.4745	0.4745
Defects	—	—	—	0.3466	—	—	—	0.3466	0.3466
Speech	—	—	—	—	—	—	—	0.6931	—

14

2.5 FAUN Classifier

- Classify new gene documents based on annotated NMF model
  - Inputs: a new document, term entropy weights, W matrix factor, stop words, entropy weight threshold, term frequency
  - Outputs: features sorted by weight

A. Pseudocode for FAUN Classifier:

```
1. Read terms T with their entropies from the key file: entropy (T)
2. Read term weights for each feature F: term_weight (T, F)
3. Read stopwords
4. Compute the frequencies of all terms in the document: frequency(T)
5. Total_terms = total number of terms in the document
6. Compute feature weights:
      for each feature F:
            weight(F) = 0
            for each term T in F:
            if entropy(T) >= entropy_thres and frequency(T) > doc_freq_thres:
               w = term_weight(T, F) * entropy(T) * log(1 + frequency(T)/total_terms)
               weight(F) = weight(F) + w
7. Sort all features by weight in decreasing order
8. The document is then classified based on its top feature (i.e., with the largest weight)
```

15

2.6 Automated FAUN Annotation

- Annotate features in the NMF models
  - Inputs: H matrix, known classification, NMF rank (k), a feature weight threshold
  - Output: feature label file

B. Pseudocode for automated FAUN Annotation:

```
1  Let weight (f, g) be the weight of feature f with respect to gene g
2  Let class (g) be the class assigned to gene g
3  For each feature F:
        Let G = the set of genes for which F is the top feature (the feature with max weight)
        Let L = { class(g) for all g in G }
        For each label l in L:
              Let max_weight (F, l) = max { weight (F, g) for each gene g in G
                                            such that class (g) = l }
              Let sum_weight (F, l) = sum { weight (F, g) for each gene g in G
                                            such that class (g) = l }
        Let max_sum (F) = max { sum_weight (F, l) for l in L }
        Let accepted_classes (F) = ( l for each l in L
              such that sum_weight (F, l) ≥ max_sum(F) × threshold
              sorted by max_weight (F, l) in descending order )
```

16

3. FAUN Capabilities and Usability

17

3 FAUN CAPABILITIES and Usability

3.1 Extracting concept-based features

3.2 Identifying genes in a feature

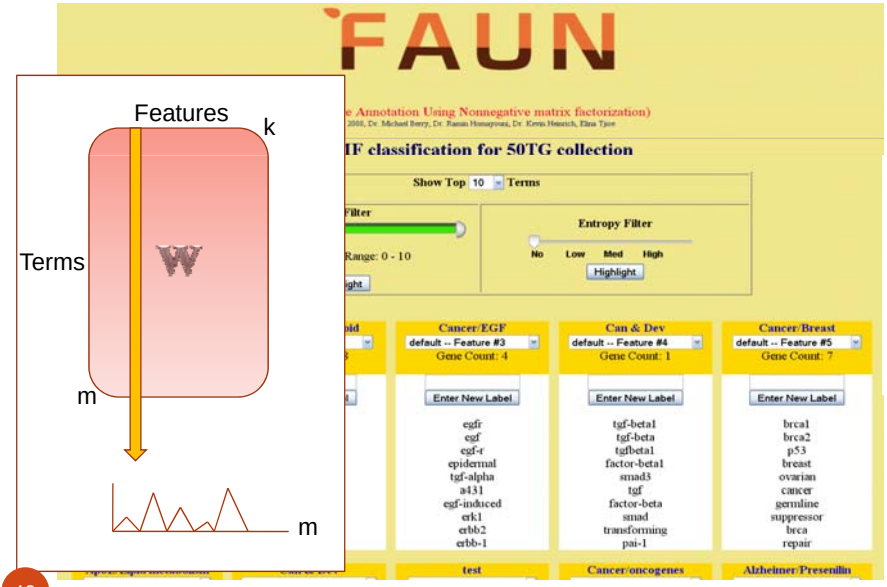
3.3 Exploring gene relationships

3.4 Classifying new gene documents

3.5 Discovering novel gene functional relationships

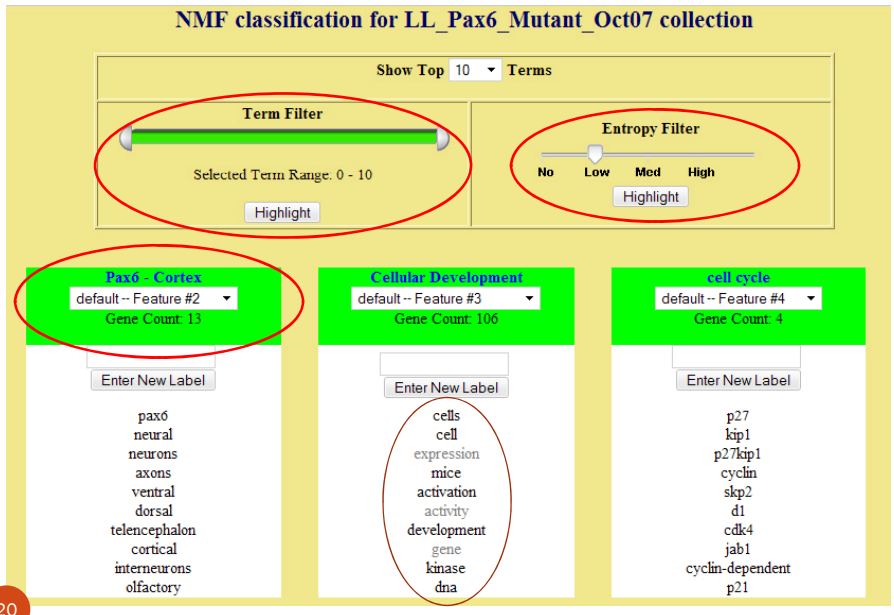
18

3.1 Extracting concept-based features

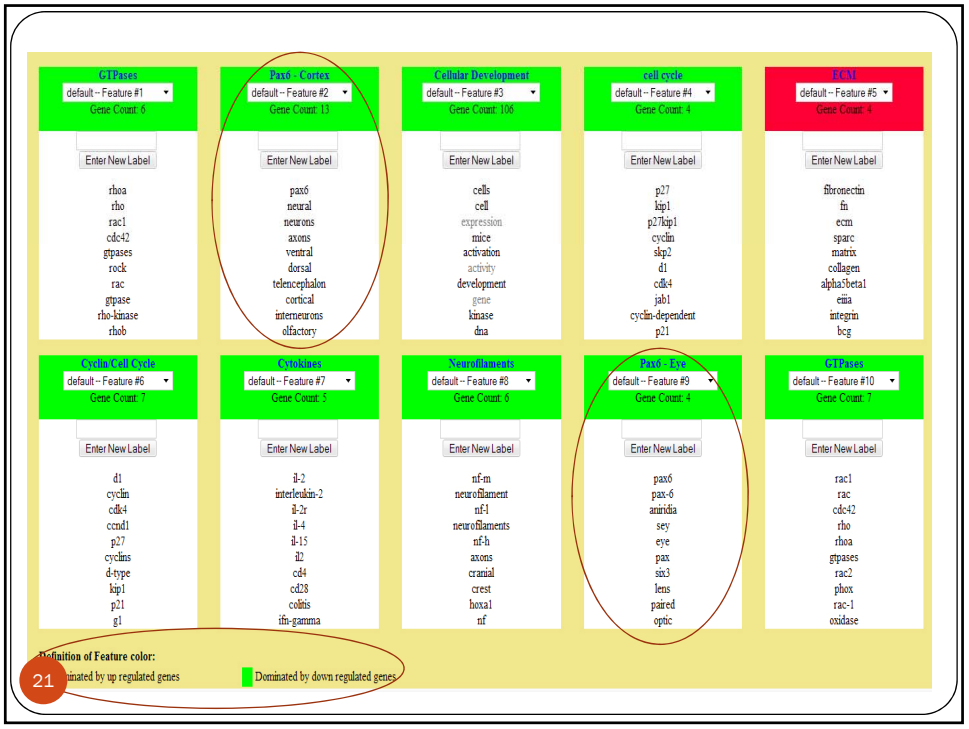


19

Tjioe E. Proceedings of the First Workshop on Data Mining in Functional Genomics, IEEE International Conference on Bioinformatics and Biomedicine, Philadelphia, PA, Nov. 3-5, 2008, pp.185-192.

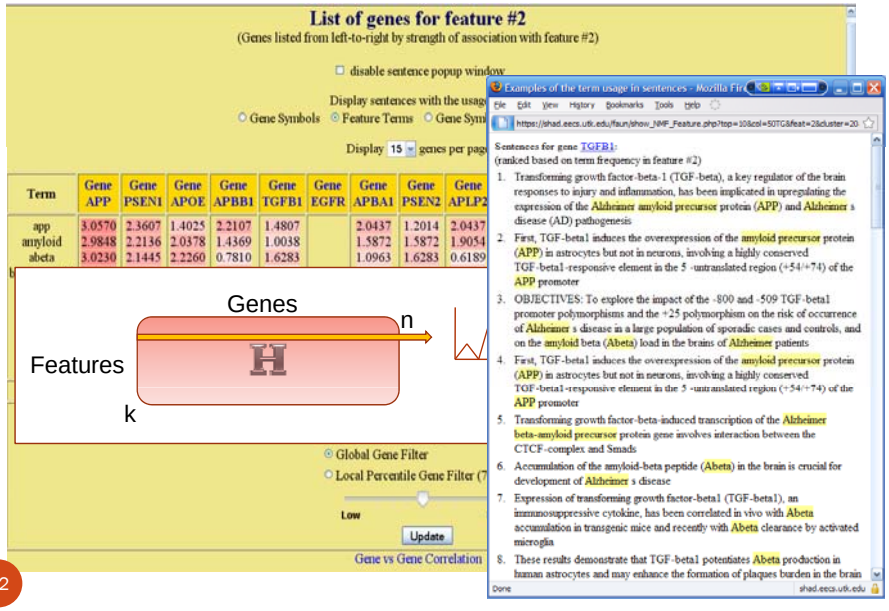


20



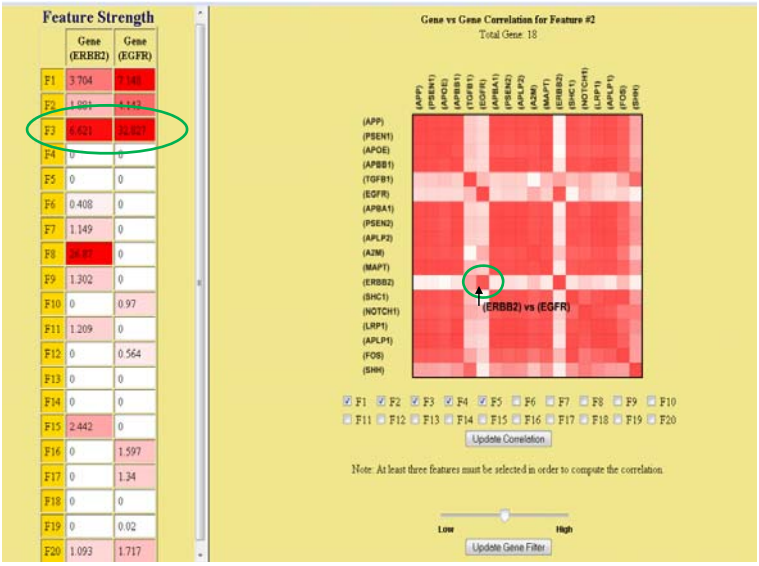
21

3.2 Identifying genes in a feature



22

3.3 Exploring gene relationships



23

3.4 Classifying new gene documents

- Built NMF models using 40 genes selected randomly from the 50 gene dataset
- Train FAUN classifier using the *W* matrix factor in newly built NMF models
- Test classification accuracy using the remainder 10 genes
  → Classifier accuracy ~80%

24

3.5 Discovering novel gene functional relationships

- 50TG dataset
  - Discover two cancer genes, ERBB2 and EGFR, involve in Alzheimer disease
- BGM dataset
  - Discover gene REN, involved in nephroblastoma, also involve in telomere maintenance
- Cerebellum dataset
  - Discover dataset contains a large component of transcription factors

25

4. Results

26

4.1 Gene Datasets

Table 1.
List of categories for each dataset used
to evaluate FAUN classification performance.

	Dataset 1 (50TG)		References
	Categories	**# of genes**	
1	Cancer	15	Homayouni et al. Bioinformatics 2005, 21(1):104-115
2	Alzheimer	11	
3	Development	5	
4	Cancer & Development	16	
5	Alzheimer & Development	3	
	Dataset 2 (BGM)		**References**
	Categories	**# of genes**	
1	Biocarta: Caspase cascade in apoptosis	21	Burkart MF et al. Bioinformatics 2007, 23(15):1995-2003
2	Biocarta: Sonic hedgehog pathway	8	
3	Biocarta: Adhesion and diapedesis of lymphocytes	10	
4	GO: Biological process: telomere maintenance	10	
5	GO: Cellular constituent: cornified cell envelope	7	
6	GO: Molecular function: DNA helicase	20	
7	MeSH: Disease: retinitis pigmentosa	8	
8	MeSH: Disease: chronic pancreatitis	8	
9	MeSH: Disease: nephroblastoma (Wilm's tumor)	10	
	Dataset 3 (NatRev)		**References**
	Categories	**# of genes**	
1	Autism	26	Abrahams et al. Nat Rev Genet 2008, 9(5):341-355
2	Diabetes	10	Frayling TM. Nat Rev Genet 2007, 8(9):657-662
3	Translation	25	Scheper GC. Nat Rev Genet 2007, 8(9):711-723
4	Mammary Gland Development	37	Robinson GW . Nat Rev Genet 2007, 8(12):963-972
5	Fanconi Anemia	12	Wnag, W. Nat Rev Genet 2007, 8(10):735-748

27

4.2 Input Parameters

- Initialization Methods:
  - Random
  - NNDSVD: NNDSVDz, NNDSVDa, NNDSVDe, NNDSVDme
- NMF ranks
  - $k = 10, 20, 30, 40, 50$
- Stopping criteria $\|W_{old} - W_{new}\| < \tau$ and $\|H_{old} - H_{new}\| < \tau$
  - 1000 maximum iterations with tolerance $\tau = 0.01$
  - 2000 maximum iterations with tolearance $\tau = 0.001$
- Smoothness and sparsity constraints
  - Smoothness parameters: 0.001, 0.01, 0.1
  - Sparsity parameters: 0.1, 0.5, 0.9
- NMF algorithm
  - Multiplicative update
  - Sparse NMF

28

2.3 continued……

- Additional application-dependent constraints:

$$f(W,H) = \frac{1}{2}\|A - WH\|_F^2 + \alpha\, J_1(W) + \beta\, J_2(H)$$

- **Smoothness constraint**

$$J_1(W) = \|W\|_F^2 \qquad W_{ic} \leftarrow W_{ic}\frac{(AH^T)_{ic} - \alpha W_{ic}}{(WHH^T)_{ic} + 10^{-9}}$$

- **Sparsity constraint**

$$J_2(H) = (\omega\|vec(H)\|_2 - \|vec(H)\|_1)^2 \qquad \omega = \sqrt{kn} - (\sqrt{kn} - 1)\gamma$$

$$H_{cj} = H_{cj}\frac{(W^TA)_{cj} - \beta(c_1H_{cj} + c_2)}{(W^TWH)_{cj} + 10^{-9}}$$

$$c_1 = \omega^2 - \omega\frac{\|vec(H)\|_1}{2\|vec(H)\|_2} \qquad c_2 = \|vec(H)\|_1 - \omega\|vec(H)\|_2$$

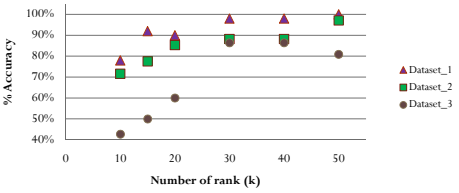
29

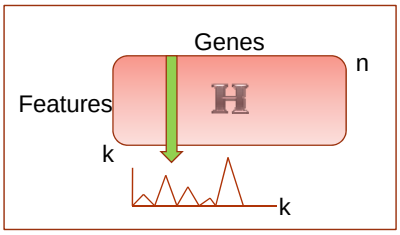
4.3 Evaluation approaches

List of categories for each dataset used to evaluate FAUN classification performance.

	Dataset 1	
	Categories	**# of genes**
1	Cancer	15
2	Alzheimer	11
3	Development	5
4	Cancer & Development	16
5	Alzheimer & Development	3
	Dataset 2	
	Categories	**# of genes**
1	Biocarta: Caspase cascade in apoptosis	21
2	Biocarta: Sonic hedgehog pathway	8
3	Biocarta: Adhesion and diapedesis of lymphocytes	10
4	GO: Biological process: telomere maintenance	10
5	GO: Cellular constituent: cornified cell envelope	7
6	GO: Molecular function: DNA helicase	20
7	MeSH: Disease: retinitis pigmentosa	8
8	MeSH: Disease: chronic pancreatitis	8
9	MeSH: Disease: nephroblastoma (Wilm's tumor)	10
	Dataset 3	
	Categories	**# of genes**
1	Autism	26
2	Diabetes	10
3	Translation	25
4	Mammary Gland Development	37
5	Fanconi Anemia	12

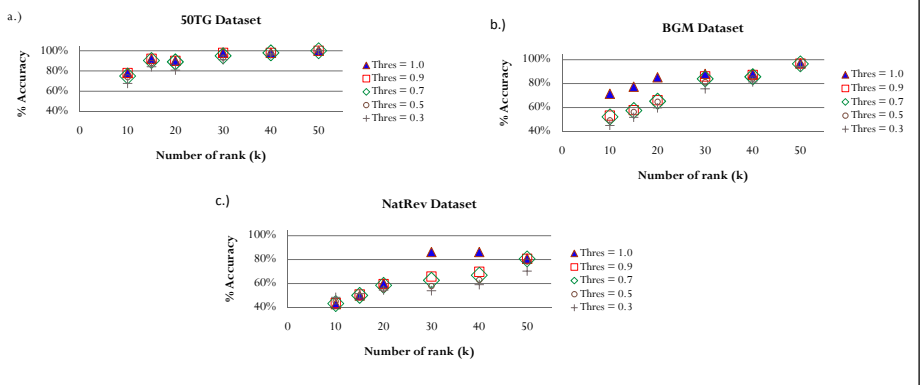
FAUN classification accuracy based on the strongest feature.

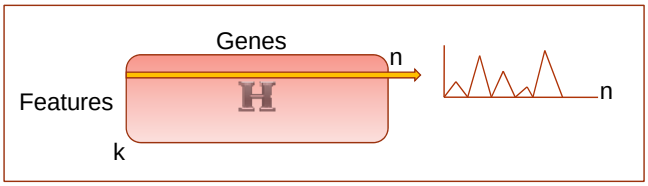




30

FAUN classification accuracy based on the total gene recall





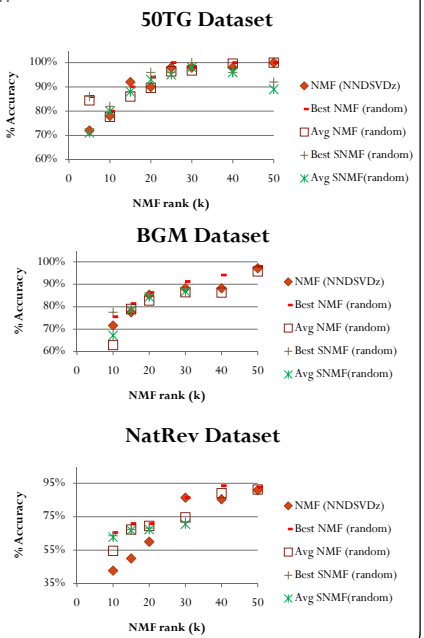
31

Comparison with the sparse NMF (SNMF) algorithm

SNMF (Matlab)									
NMF Rank *k*	Number of Iterations			CPU Time (s)			Number of operations per iteration: $O(k^2(m+n))$		
	50TG	BGM	NatRev	50TG	BGM	NatRev	50TG	BGM	NatRev
10	237	202	281	145	251	335	8.8E+07	1.3E+08	1.3E+08
20	282	270	234	829	2,175	2,952	1.4E+09	2.0E+09	2.1E+09
30	217	451	235	3,408	13,533	7,341	7.1E+09	1.0E+10	1.1E+10
40	30	325	231	1,544	22,202	12,318	2.3E+10	3.2E+10	3.4E+10
50	26	330	37	1,763	24,174	10,750	5.5E+10	7.9E+10	8.2E+10

Default NMF (C++)									
NMF Rank *k*	Number of Iterations			CPU Time (s)			Number of operations per iteration: $O(kmn)$		
	50TG	BGM	NatRev	50TG	BGM	NatRev	50TG	BGM	NatRev
10	86	130	92	3.57	14.4	8.2	4.4E+06	1.3E+07	1.4E+07
20	114	133	106	11.55	28.96	15.88	8.8E+06	2.6E+07	2.9E+07
30	162	154	119	28.34	46.57	23.91	1.3E+07	3.9E+07	4.3E+07
40	166	165	147	36.27	70.4	40.59	1.8E+07	5.1E+07	5.7E+07
50	634	171	180	197.96	94.7	61.33	2.2E+07	6.4E+07	7.2E+07

	50TG	BGM	NatRev
Number of terms (m)	8,750	12,590	13,038
Number of gene docs (n)	50	102	110



32

Effect on classification accuracy using different NMF parameters

- NMF Rank effect
  - Classification accuracy in general increases with the increase of NMF rank
- Initialization effect
  - All initializations in general show very similar accuracy trends
- Stopping criteria effect
  - Increasing the maximum number of iterations beyond 2000 and the tolearance 0.01 does not appear to increase the accuracy
- Smoothing effect
  - Smoothing on *W* matrices has little or no effect
  - Smoothing on *H* matrices could increase or decrease the accuracy to ~6%
- Sparsity effect
  - Sparsity constraints on *W* or *H* matrices show little or no effect on the accuracy

33

5. Summary and Future Work

Summary

- FAUN classifies genes with promising accuracy.
- FAUN assists in understanding why genes are related.
- FAUN allows researchers to reveal hidden but published knowledge of functional relationships among genes.
- FAUN provides utilities for knowledge discovery.
  - A FAUN-based analysis of a new cerebellum gene set has revealed new knowledge – the gene set contains a large component of transcription factors.

Future Work

- Enhancing FAUN utilities such as dragging and selecting multiple cells on gene-to-gene correlation matrix
- Implement gene query system

34

ACKNOWLEDGEMENTS

- Dr. Michael Berry
- Dr. Ramin Homayouni
- Dr. Michael Langston
- Dr. Igor Jouline
- Dr. Robert Ward
- Dr. Kevin Heinrich
- Cerebellum Group
- GST Program

35

Thank you....!!

36

FAUN site: http://grits.eecs.utk.edu/faun