Latent Semantic Analysis of the Languages of Life

Ryan Anthony Rossi*

Jet Propulsion Laboratory,
California Institute of Technology,
Pasadena, CA 91106, U.S.A.
ryan.a.rossi@jpl.nasa.gov

Abstract. We use Latent Semantic Analysis as a basis to study the languages of life. Using this approach we derive techniques to discover latent relationships between organisms such as significant motifs and evolutionary features. Doubly Singular Value Decomposition is defined and the significance of this adaptation is demonstrated by finding a phylogeny of twenty prokaryotes. Minimal Killer Words are used to define families of organisms from negative information. The application of these words makes it possible to automatically retrieve the coding frame of a sequence from any organism.

Keywords: Languages of Life, Motifs, Phylogeny, Minimal Killer Words, Doubly Singular Value Decomposition, Latent Semantic Analysis, Cross Language Information Retrieval, Knowledge Discovery, Data Mining.

1 Introduction

Latent Semantic Analysis (LSA) has been successfully used in applications such as natural language processing, speech recognition, cognitive modeling, document classification, search engines, and more recently security [1-8]. A significant application is in Cross Language Information Retrieval where direct matching of words is unlikely. A set of documents are used to create a reduced dimension space representation in which words that occur in similar contexts are near one another. This allows a query to retrieve relevant documents even if they have no words in common. LSA uses the Singular Value Decomposition (SVD) to model relationships between words that appear in similar contexts [2-4]. Cross Language LSA has been used to retrieve documents in different languages without having to translate the query [2]. From this method, queries in one language can retrieve documents in the same language and in different languages.

In this work we use techniques based on LSA to study the languages of life. In the second section we briefly describe the mathematical framework. In the third section we use cross language information retrieval of the languages of life to extract evolutionary features. The results validate the cross language information retrieval technique using unnatural languages. The algorithm proposed could be used for several applications such as studying the biological relationships between words and genes

* Work supported by NSF Grant ATM-0521002.

Z. Cai et al. (Eds.): ISICA 2009, CCIS 51, pp. 128–137, 2009.

from various organisms. It is shown how these techniques can be applied to model an arbitrary organism's language. In the fourth section, we validate the Cross Language of Life results using a method to extract the most significant motifs in an organism. We propose a technique called Doubly SVD to generate a phylogeny of twenty prokaryotes and show how this method can be extended to an arbitrary set of organisms. Finally we define Minimal Killer Words and describe how they can be used to automatically retrieve the frame of a coding sequence from an organism.

2 Mathematical Framework

The rows in our data set represent words and the columns represent gene sequences. If $M \in \Re^{nxm}$ then we decompose M into three matrices using the Singular Value Decomposition:

$$M \;=\; U \; S \; V^{T} \tag{1}$$

where $U \in \Re^{nxm}$, $S \in \Re^{mxm}$ and $V^{T} \in \Re^{mxm}$. The matrix S contains the singular values located in $[i, i]_{1,\ldots n}$ cells in descending order of magnitude and all other cells contain zero. The eigenvectors of MM^{T} make up the columns of U and the eigenvectors of $M^{T}M$ make up the columns of V. As a consequence, the singular values are the square roots of the eigenvalues of MM^{T} and $M^{T}M$. The matrices U and V are orthogonal, unitary and span vector spaces of dimension n and m, respectively. The inverses of U and V are their transposes.

$$\underset{U}{\begin{bmatrix} | & | & & | \\ d_1^g & d_2^g & \cdots & d_k^g \\ | & | & & | \end{bmatrix}} \underset{S}{\begin{bmatrix} s_1 & 0 & 0 & 0 \\ 0 & s_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & s_k \end{bmatrix}} \underset{V^T}{\begin{bmatrix} \text{---} & d_1^w & \text{---} \\ \text{---} & d_2^w & \text{---} \\ & \vdots & \\ \text{---} & d_k^w & \text{---} \end{bmatrix}}$$

The columns of U are the *principal directions of the genes* and the rows of V^{T} are the *principal directions of the words*. In a reduced space the rows of U are the coordinates of the words and the columns of V^{T} are the coordinates of the genes. The principal directions are ranked according to the singular values and therefore according to importance.

An important theorem by Eckart and Young [9] demonstrates that the error in approximating M by M_k is given by:

$$\left\| M - M_k \right\|_F = \min_{rank(B) \le k} \left\| M - B \right\|_F = \sqrt{\sigma_{k+1}^2 + .. + \sigma_{rM}^2} \tag{2}$$

Where,

$$M_k = U_k \, S_k \, V_k^T \tag{3}$$

and is the closest rank-k least squares approximation of M. This theorem can be used in two ways. To reduce noise by setting insignificant singular values to zero or by setting the majority of the singular values to zero and keeping only the few influential singular values. The latter approach allows us to create a reduced space from which words used in the same context with one another are represented close together.

3 Cross Language of Life

As our training corpus or dual language semantic space we have selected 4000 gene sequences from Escherichia coli K12. From the training corpus, we construct a word by gene sequence matrix for DNA and Protein languages. The protein language has an alphabet of 20 amino acids,

$$\alpha_p = \{A, R, N, D, C, E, Q, G, H, I, L, K, M, F, P, S, T, W, Y, V\}$$

and the DNA language has an alphabet of 4 nucleotides,

$$\alpha_d = \{A, C, G, T\}$$

We select an arbitrary gene in both amino acids and nucleotides. We have to consider that one trinucleotide codes an amino acid. Therefore, we construct words of length three in amino acids and length nine in nucleotides. We use overlapping words. We check to see if the word w exists in our lexicon. If it does not, we add this word to our lexicon and add a 1 to the cell corresponding to the word and gene. We do this for 4000 Escherichia coli K12 genes which we use as our training corpus.

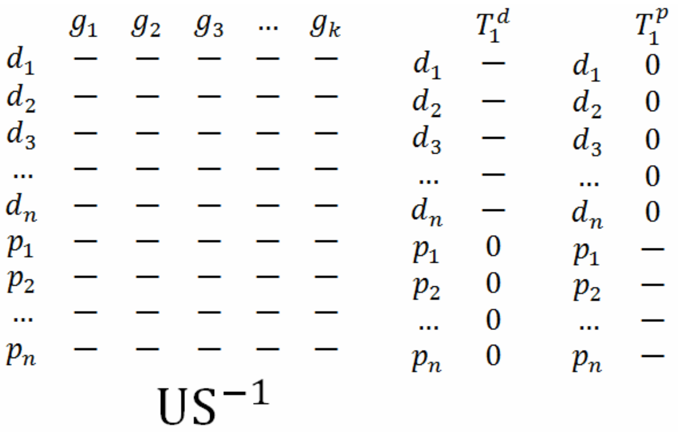
Let D be a word by gene matrix of m DNA sequences and n^d nucleotide words and P be a word by gene matrix of m protein sequences and n^p amino acid words.

$$M = \begin{Bmatrix} D \\ P \end{Bmatrix} \tag{4}$$

Such that M is of size *$(n^d + n^p)$ x m* where column i is a vector representing amino acid and nucleotide words appearing in the union of sequence i expressed in the two languages. After we have constructed our word by gene training corpus, we apply the Singular Value Decomposition:

$$M_k = \begin{bmatrix} U_k^d \\ U_k^p \end{bmatrix} S_k \ V_k \tag{5}$$

Where $U_k^d \mathit{and} \ U_k^p$ are k-dimensional vector lexicons for DNA and Protein sequences from Escherichia coli K12. In this example, we chose k to be 210. In this space, similar DNA and Protein words are given similar definitions, so this vector lexicon can be used for cross-language information retrieval.



$$\begin{array}{cccccc} & g_1 & g_2 & g_3 & \cdots & g_k \\ d_1 & - & - & - & - & - \\ d_2 & - & - & - & - & - \\ d_3 & - & - & - & - & - \\ \cdots & - & - & - & - & - \\ d_n & - & - & - & - & - \\ p_1 & - & - & - & - & - \\ p_2 & - & - & - & - & - \\ \cdots & - & - & - & - & - \\ p_n & - & - & - & - & - \end{array} \qquad \begin{array}{cc} & T_1^d \\ d_1 & - \\ d_2 & - \\ d_3 & - \\ \cdots & - \\ d_n & - \\ p_1 & 0 \\ p_2 & 0 \\ \cdots & 0 \\ p_n & 0 \end{array} \qquad \begin{array}{cc} & T_1^p \\ d_1 & 0 \\ d_2 & 0 \\ d_3 & 0 \\ \cdots & 0 \\ d_n & 0 \\ p_1 & - \\ p_2 & - \\ \cdots & - \\ p_n & - \end{array}$$

$$\mathrm{US}^{-1}$$

If we have two sets of sequences in both languages denoted T^d for DNA sequences and T^p for Protein sequences, we can retrieve sequences in either languages. As an example, we have a query in DNA and want to retrieve similar protein sequences. We find the similarity between the DNA query T_1^d and the protein database T^p using:

$$sim\left\langle US^{-1}T_1^D,\ US^{-1}T^P\right\rangle \tag{6}$$

Where,

$$sim\left\langle d_1,\ \ p_1\right\rangle = \frac{d_1^T p_1}{\left\|d_1^T\right\|_2\ \left\|p_1\right\|_2} \tag{7}$$

This gives us a ranking of the relevant protein sequences.

3.1 Empirical Results

In the experiments below we use 4000 genes from Escherichia coli K12 as our dual-language semantic space. We use a technique called cross-language mate retrieval [2] to test the significance of our dual-language semantic space. We extract at random 1000 nucleotide sequences and their corresponding translated sequences in amino acids from Escherichia coli HS. These two strains of Escherichia coli are considered very similar. We use Escherichia coli HS as a starting point to see if our choice of k = 210 is at all reasonable for our dual-language semantic space.

The cross-language mate retrieval technique works by considering each of the 1000 nucleotide gene sequences as queries that have only one relevant ‘mate’ sequence in the corresponding amino acid language and conversely. We compute the accuracy by strictly testing if the cross language mate returned as the most relevant.

From the 1000 Escherichia coli HS gene sequences the cross-language mate is retrieved 94.8% of the time using the nucleotide sequences as queries against the amino acid sequences. Similarly, the cross-language mate is retrieved 94.4% of the time using the amino acid sequences as queries against the nucleotide sequences. We find

Table 1. Results using Escherichia coli K12 as our semantic space where we have genes from Escherichia coli HS as queries and sequences in both amino acids and nucleotides

Semantic Space	Query	Sequences	Accuracy
Escherichia coli K12	Escherichia coli HS Nucleotide	Escherichia coli HS Amino Acid	94.8%
	Escherichia coli HS Amino Acid	Escherichia coli HS Nucleotide	94.4%

that most of the sequences where the mate was not returned as the most significant are very short sequences (around 30-40 amino acids or 90-120 nucleotides). From this analysis we find that these two strains of Escherichia coli strongly share the same set of words of length three and are considered very similar.

In the experiment below we use the same Escherichia coli K12 dual-language semantic space. As our testing set we select 1000 genes from Shigella sonnei in both nucleotides and amino acids.

The same phenomenon is seen using Shigella sonnei and Yersinia pestis KIM.

We decided to do an experiment using genes from a wide range of organisms as our testing set. We picked up 2800 genes from 34 prokaryotes, 12 eukaryotes and 13 archaea. From each of these organisms we randomly selected on average 50 genes.

The cross-language mate is retrieved as the most relevant 59.57% of the time using the nucleotide sequences as queries against the amino acid sequences. Similarly, the cross-language mate is retrieved as the most relevant 57.78% of the time using the amino acid sequences as queries against the nucleotide sequences. Nevertheless, these results provide evidence that there exists a structure and similarity in the language that defines all organisms. A more systematic study is warranted to find this universal structure [11].

Table 2. Results using Escherichia coli K12 as our semantic space where we have genes from Shigella sonnei as queries and sequences in both amino acids and nucleotides

Semantic Space	Query	Sequences	Accuracy
Escherichia coli K12	Shigella sonnei Nucleotide	Shigella sonnei Amino Acid	92.00%
	Shigella sonnei Amino Acid	Shigella sonnei Nucleotide	93.30%

Table 3. Results using Escherichia coli K12 as our semantic space where we have genes from Yersinia pestis KIM as queries and sequences in both amino acids and nucleotides

Semantic Space	Query	Sequences	Accuracy
Escherichia coli K12	Yersinia pestis KIM Nucleotide	Yersinia pestis KIM Amino Acid	88.80%
	Yersinia pestis KIM Amino Acid	Yersinia pestis KIM Nucleotide	88.10%

Table 4. Results using Escherichia coli K12 as our semantic space where we have genes from a wide range of organisms as queries and sequences in both amino acids and nucleotides

Semantic Space	Query	Sequences	Accuracy
Escherichia coli K12	Variety of genes Nucleotide	Variety of genes Amino Acid	59.57%
	Variety of genes Amino Acid	Variety of genes Nucleotide	57.78%

4 Organism Motifs and Profiles

We start with a word by protein sequence matrix and compute the Singular Value Decomposition. The first few principal directions of U can be interpreted as containing the most significant characteristics or motifs of a particular organism [8]. As an example, the most significant value in the first principal direction provides an indication of the 'most important' motif of length three in that organism. The first few principal directions are viewed as a profile for an arbitrary organism. This can also be used to model a particular organism's language.

In the table below we select the first principal direction of several organisms including a dataset with 59 organisms from the three domains. We extract the ten most significant motifs for each organism/dataset.

Table 5. Ranking of motifs of length three in various organisms

Organism	Top 10 Motifs From Different Organisms
Escherichia coli K12	KLL, FFA, GAL, GLA, NAA, TAI, TRL, LAA, HLA, TAA
Shigella sonnei	KLL, FFA, GAL, GLA, NAA, TRL, VLL, LAA, PKE, HLA
Yersinia pestis KIM	VVG, IAL, QLA, CLA, MLL, LLA, CLL, SNL, QLE, MSL
59 Organisms	NVT, GEI, NTI, GRL, MAM, DHK, DIN, PGD, NKQ, HLH

Escherichia coli K12 and Shigella sonnei share most of the significant motifs with the exception of a few. Using the dataset with 59 organisms from the 3 domains we extract Universal Motifs. These are motifs that can be found in all organisms. A more systematic study of the Universal Motifs is needed to find biological meaning.

5 Phylogeny Using Doubly Singular Value Decomposition

We start with 20 prokaryotes. From each of these organisms we extract 1000 genes in amino acids and construct a word by gene matrix for each organism. We use overlapping words of length three.

We compute the Singular Value Decomposition for each word by gene matrix (corresponding to a prokaryote) and extract the first principal direction of the genes (of U) and the first principal direction of the words (of V^T) from each organism. From these principal directions, we derive two matrices:

$$\begin{array}{cccccc} & p_1^g & p_2^g & p_3^g & \cdots & p_{20}^g \\ c_1^w & - & - & - & - & - \\ c_2^w & - & - & - & - & - \\ c_3^w & - & - & - & - & - \\ c_4^w & - & - & - & - & - \\ c_5^w & - & - & - & - & - \\ c_6^w & - & - & - & - & - \\ c_7^w & - & - & - & - & - \\ \cdots & - & - & - & - & - \\ c_{8000}^w & - & - & - & - & - \end{array} \qquad \begin{array}{cccccc} & c_1^g & c_2^g & c_3^g & \cdots & c_{1000}^g \\ p_1^w & - & - & - & - & - \\ p_2^w & - & - & - & - & - \\ p_3^w & - & - & - & - & - \\ p_4^w & - & - & - & - & - \\ p_5^w & - & - & - & - & - \\ p_6^w & - & - & - & - & - \\ p_7^w & - & - & - & - & - \\ \cdots & - & - & - & - & - \\ p_{20}^w & - & - & - & - & - \end{array}$$

Computing the SVD of the matrix on the left; we find that the columns of U are the **principal directions of prokaryotes** and the highest absolute value of the principal direction of prokaryotes represents the **most influential word**. The rows of V^T are the **principal directions of words** for prokaryotes and the highest absolute value of the principal direction of words for prokaryotes represents the **most influential prokaryote**. Similarly, taking the SVD of the matrix on the right; we find that the columns of U are the **principal directions of genes** for prokaryotes and the highest absolute value of the principal direction of genes for prokaryotes represents the **most influential prokaryote**. The rows of V^T are the **principal directions of prokaryotes** and the highest absolute value of the principal direction of prokaryotes represents the **most influential gene**.

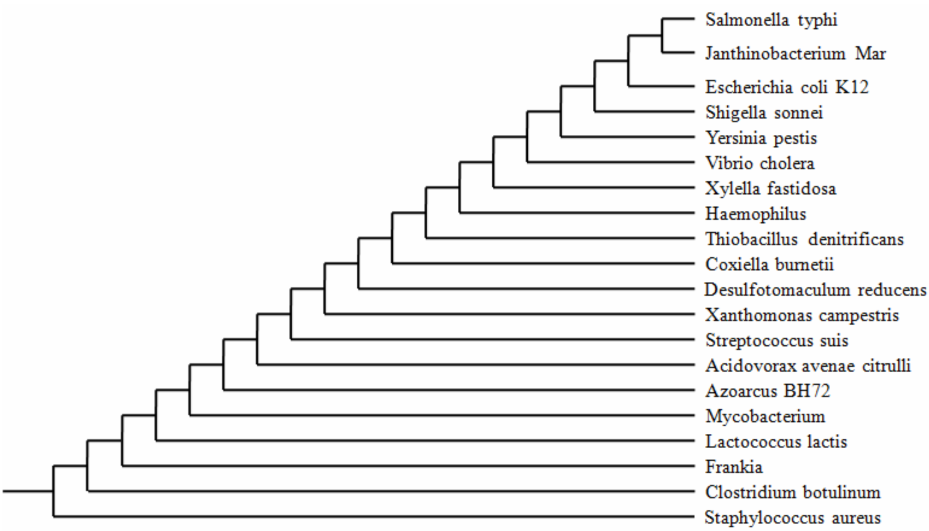


Fig. 1. Phylogeny of 20 prokaryotes using the principal direction of words

Our technique uses the principal directions of the words to build a comparatively accurate phylogenetic tree. We call this method doubly singular value decomposition (DSVD). These results are fairly strong. The influence values are close together with the exception of Staphyloccoccus aureus. The majority of the organisms are represented in the first principal direction, while Staphylococcus aureus is more accurately represented in the second direction which is orthogonal to the first. This could be due to major evolutionary differences between Staphylococcus aureus and the other prokaryotes.

6 Minimal Killer Words

Let A be a finite alphabet with the symbols {A, C, G, T} and A* be the set of words drawn from the alphabet A.

Let L $\subseteq$ A* be a language consisting of all factors of its words from a given organism. Let F_0, F_1 and F_2 be sets containing all factors of its words from the frames of an organism. A word w $\in$ A* is called a Minimal Killer Word for L if w $\notin$ F_0 and w $\in$ F_2 and all proper factors of w do not belong to M K (L). We denote by M K (L) the language of Minimal Killer Words for L.

Consider the subsequence AGGCTAGCT. We derive F_0, F_1 and F_2.

F_0 = {AGG, CTA, GCT, AGGCTA, AGGCTAGCT}
F_1 = ~~{GGC, TAG, GGCTAG}~~
F_2 = {GCT, AGC, GCTAGC}

Therefore, M K (L) = {AGC}. The word GCTAGC is not minimal because a proper factor is in the set M K (L).

The following circular codes $\mathbf{X_0}$, $\mathbf{X_1}$, and $\mathbf{X_2}$ have been found as subsets of the genetic code [11-12].

$\mathbf{X_0}$ = {AAC, AAT, ACC, ATC, ATT, CAG, CTC, CTG, GAA, GAC, GAG, GAT, GCC, GGC, GGT, GTA, GTC, GTT, TAC, TTC}

$\mathbf{X_1}$ = {ACA, ATA, CCA, TCA, TTA, AGC, TCC, TGC, AAG, ACG, AGG, ATG, CCG, GCG, GTG, TAG, TCG, TTG, ACT, TCT}

$\mathbf{X_2}$ = {CAA, TAA, CAC, CAT, TAT, GCA, CCT, GCT, AGA, CGA, GGA, TGA, CGC, CGG, TGG, AGT, CGT, TGT, CTA, CTT}

However, it is the codes

$$T_0 = X_0 \bigcup \{AAA, TTT\},\ \ T_1 = X_1 \bigcup \{CCC\}\ \ and\ \ T_2 = X_2 \bigcup \{GGG\}$$

that we will consider as their union forms the entire genetic code. These codes have remarkable properties and have been used to find a universal coding frame [11]. We associate three T-Representations to any coding sequence u:

The first representation, T, is obtained by replacing each codon by 0 if it belongs to T_0, 1 if it belongs to T_1 and 2 if it belongs to T_2. This representation corresponds to the coding frame, while the two others represent the shifted frames. The second representation T^+ is obtained by elimination of the first letter of u and applying the preceding construction. Finally, the third representation T^{++} is obtained by eliminating a second letter from u and again applying the same construction.

We find the Minimal Killer Words of an organism using the representations T, T^{+} and T^{++}. We arbitrarily selected Escherichia coli K12 and Shigella sonnei to use as a starting point. The set of Minimal Killer Words of length nine for Escherichia coli K12 and Shigella sonnei are shown below.

M K₉ (E) = {122121222, 122222121, 211222122, 212221112, 212221221, 221221121, 221222121, 222121221, 222122211}

M K₉ (S) = {122222121, 211222122, 221222121, 222122211}

Interestingly, the Minimal Killer Words of length nine for Escherichia coli K12 and Shigella sonnei are formed from strictly T_1 and T_2. These words highlight evolutionary features in the organisms. As an example, M K_9(S) is a subset of M K_9 (E) indicating similarity. The Minimal Killer Words provide a way to define families of organisms using negative information.

The Minimal Killer Words allow us to automatically retrieve the coding frame of a sequence from an organism. This is a direct consequence of the definition and is a very powerful property. As an example, consider the set L of coding sequences where μ belongs to the set M K (L). Therefore, if μ appears in a coding sequence of L we can infer the coding frame of that sequence.

7 Conclusion

We describe techniques based on Latent Semantic Analysis to study the languages of life. Latent relationships between organisms such as motifs and evolutionary features are identified. Using DSVD we build a phylogeny of twenty prokaryotes. The definition of Minimal Killer Words allows for the retrieval of the coding frame from an organism. A more systematic study is warranted to find biological meaning of the motifs and Minimal Killer Words. A natural future direction will be to define more precisely this universal structure in the language of life.

Acknowledgments. We thank Jean-Louis Lassez for his help in clarifying our ideas as well as the many insightful discussions. This work was partly supported by a research fellowship through the National Aeronautics and Space Administration.

References

1. Landauer, T.K., Foltz, P.W., Laham, D.: Introduction to Latent Semantic Analysis. Discourse Processes 25, 259–284 (1998)
2. Landauer, T.K., Littman, M.L.: Fully automatic cross language document retrieval using latent semantic indexing. In: Cross Language Information Retrieval. Kluwer, Dordrecht (1998)
3. Deerwester, S., Dumais, S.T., Landauer, T.K., Furnas, G.W., Harshman, R.A.: Indexing by latent semantic analysis. JSIS 41(6), 391–407 (1990)
4. Lemaire, B.: Tutoring Systems Based on LSA. AIED, 527–534 (1999)

5. Alter, O., Brown, P.O., Botstein, D.: Singular value decomposition for genome-wide expression data processing and modeling. PNAS 97, 10101–10106 (2000)
6. Kintsch, W.: Comprehension: A Paradigm for Cognition. Cambridge University Press, Cambridge (1998)
7. Lassez, J.-L., Rossi, R., Jeev, K.: Ranking Links on the Web: Search and Surf Engines. In: Nguyen, N.T., Borzemski, L., Grzech, A., Ali, M. (eds.) IEA/AIE 2008. LNCS (LNAI), vol. 5027, pp. 199–208. Springer, Heidelberg (2008)
8. Lassez, J.-L., Rossi, R., Sheel, S., Mukkamala, S.: Signature Based Intrusion Detection System using Latent Semantic Analysis. IJCNN, 1068–1074 (2008)
9. Eckart, C., Young, G.: The approximation of one matrix by another of lower rank. Psychometrika 1, 211–218 (1936)
10. Berry, M., Browne, M.: Understanding Search Engines: Mathematical Modeling and Text Retrieval. SIAM, Philadelphia (1999)
11. Lassez, J.-L., Rossi, R.A., Bernal, A.E.: Crick's Hypothesis Revisited: The Existence of a Universal Coding Frame. AINA/BLSC, 745–751 (2007)
12. Lassez, J.-L.: Circular Codes and Synchronization. IJCIS 5, 201–208 (1976)