Comparison of MLP NN Approach with PCA and ICA for Extraction of Hidden Regulatory Signals in Biological Networks

Zomorrodi, Alireza; Nasernejad, Bahram*[+]

Department of Chemical Engineering, Amirkabir University of Technology, Tehran, I.R. IRAN

Kabudian, Jahanshah

Department of Computer Engineering & Information Technology, Amirkabir University of Technology, Tehran, I.R. IRAN

ABSTRACT*: The biologists now face with the masses of high dimensional datasets generated from various high-throughput technologies, which are outputs of complex inter-connected biological networks at different levels driven by a number of hidden regulatory signals. So far, many computational and statistical methods such as PCA and ICA have been employed for computing low-dimensional or hidden representations of these datasets, but in most cases the results are inconsistent with underlying real network. In this paper we have employed and compared three linear (PCA and ICA) and non-linear (MLP neural network) dimensionality reduction techniques to uncover these regulatory signals, from outputs of such networks. The three approaches were verified experimentally using the absorbance spectra of a network of seven hemoglobin solutions, and the results revealed the superiority of the MLP NN to PCA and ICA. This study shows the capability of the MLP NN approach to efficiently determine the regulatory components in biological networked systems.*

KEY WORDS: *Regulatory signal, Biological network, PCA, ICA, MLP NN.*

INTRODUCTION

The identification and characterization of system-level features of biological organizations is a key issue of post-genomic biology. It is now widely recognized that thousands of components of a living cell are dynamically interconnected, so that the cell's functional properties are ultimately encoded into a complex intracellular web of molecular interactions. The biologists now face with masses of high dimensional datasets generated from various high-throughput technologies which are outputs of complex inter-connected biological networks at different levels driven by a number of hidden regulatory components. Uncovering these regulatory components in

To whom correspondence should be addressed.
+E-mail: banana@aut.ac.ir
1021-9986/06/4/1 7/$/2.70

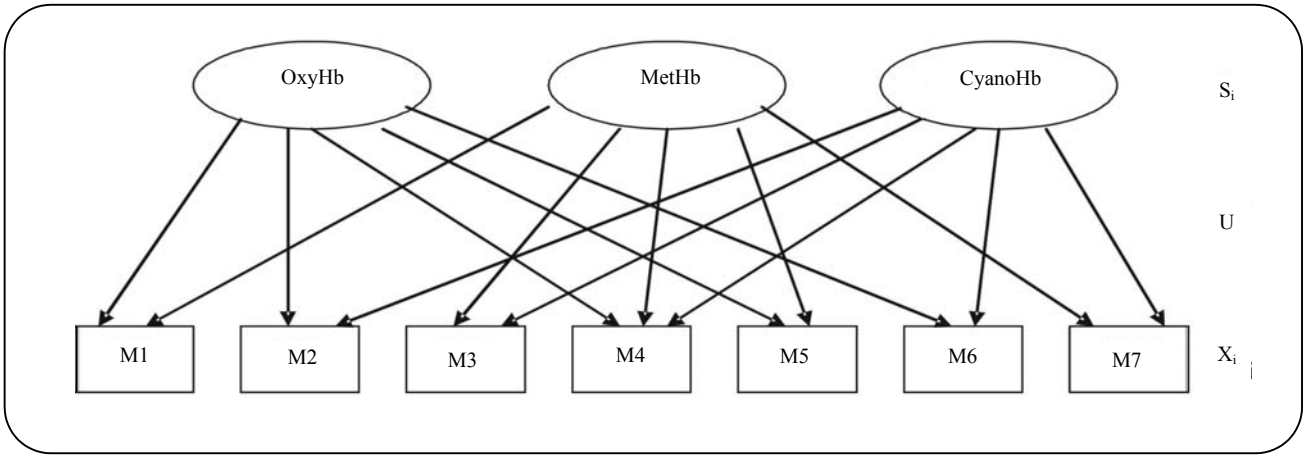

Fig. 1: The connectivity (mixing) diagram of the seven Hb solutions from three pure components: OxyHb (Oxyhemoglobin), MetHb (Methemoglobin), CyanoHb (Cyano-methemoglobin) that serve as the regulatory components.

a given biological network, to represent the high dimensional dataset in a lower-dimensional space has become a significant challenge for scientists in systems biology.

Different linear methods such as Principal Component Analysis (PCA), or Independent Component Analysis (ICA) have been previously employed to extract the regulatory components in biological systems [1- 5], but in most cases a lack of compatibility with the real network has been an important drawback of these approaches. The main goal of this paper is to make a comparison between linear (PCA and ICA) and nonlinear (MLP neural network) dimensionality reduction techniques for reconstruction of hidden regulatory signals. To experimentally verify these approaches we used the data for a network of seven hemoglobin solutions [6], made up three regulatory components, as a test case. Having the absorbance spectra of seven hemoglobin solutions and three regulatory components at hand, we were able to verity and compare these three approaches. The comparison of the implantation results revealed the superiority of the proposed MLP NN to PCA and ICA. This study shows the capability of the MLP NN approach to successfully address the problem of reconstructing hidden regulatory signals in biological networks.

NETWORK OF HEMOGLOBIN SOLUTIONS

To verify experimentally the three linear and non-linear techniques, we used the data (absorbance spectra) for a network of seven hemoglobin (Hb) solutions [6] as a model system. Each solution contains a specific combination of three pure components: oxyhemoglobin (OxyHb), methemoglobin (MetHb), and cyano-methemoglobin (CyanoHb), that serve as the regulatory components for the network of hemoglobin solutions (see Fig. 1). In this experiment (which was conducted by Liao et al [6]), the absorbance spectra of various Hb solutions and the three regulatory components have been measured by using a UV/visible spectrophotometer (Beckman DU640) at wavelength from 380 to 700 nm. Spectral data were collected for a wavelength increment of 1 nm [6]. We aimed to determine the absorbance spectra of the three regulatory components, as the regulatory signals by using the three mentioned computational methods, having the absorbance spectra of seven Hb solutions at hand. Measuring the absorbance spectra of three regulatory components experimentally is to provide a criterion for comparison of the estimated regulatory signals from PCA, ICA and MLP NN with the true ones.

PRINCIPAL COMPONENT ANALYSIS (PCA)

Principal component analysis (PCA) is possibly the most widely used dimensional reduction technique in practice, perhaps due to its conceptual simplicity and to the fact that relatively efficient algorithms exist for its computation. In signal processing it is known as the Karhunen-Loeve transform.

Assume the set of N, D-dimensional column vectors, which are already in zero-mean form (If this is not the case, it is easily possible to make them zero-mean by subtracting the mean vector). If each of these vectors is represented by x_i (i=1,…,N), the elements of this vector

are correlated and dependent. The goal of principal component analysis is to find an orthogonal D×D transformation matrix U=[u_1,...,u_D] (u_i unit vectors), that determines a change of variable:

$$x_i=Us_i \quad (1)$$

with the property that the new variables $s_1,\ldots,s_D$ are uncorrelated and are arranged in order of decreasing variance. Notice that s_i can be obtained as

$$s_i=U^{-1}x_i=U^Tx_i \ , \ (i=1.....N) \quad (2)$$

where superscript T represent the transpose of matrix.

It is not difficult to verify that for any orthogonal U, the covariance matrix of $s_1,...,s_N$ is $D=U^TCU$ where C is covariance matrix of $x_1,...,x_N$. So the desired diagonal matrix is the one that makes D diagonal. Therefore if D is a diagonal matrix with eigenvalues $\lambda_1,\ldots \lambda_D$ of C on the diagonal, arranged so that $\lambda_1>\ldots>\lambda_D>0$, then U would be an orthogonal matrix whose columns are the corresponding unit eigenvectors of C,, since in this case CU=UD or $D=U^TCU$ [7].

The unit eigenvectors $u_1,...,u_D$ of the covariance matrix C are called the principal components of the data in the matrix of observation. The first principal component is the eigenvectors corresponding to the largest eigenvalue of C, the second principal component is the eigenvector corresponding to the second largest eigenvalue and so on.

Principal component analysis is potentially valuable for applications in which most of the variance or dynamic range in the data is due to the variance only in a few of the new variables $s,\ldots,s_d$ (d<D) [7]. Therefore for dimensionality reduction one can easily discard the variables with small variance, i.e. project on the subspace spanned by the first d (d<D) principal components. After transformation, the vectors in new space can be considered as the regulatory signals.

INDEPENDENT COMPONENT ANALYSIS (ICA)

Independent component analysis (ICA) was originally developed for blind source separation whose goal is to recover mutually independent but unknown source signals from their linear mixtures without knowing the mixing coefficients. Let x_i (i=1,…,N) denote the linear mixtures, which are formed from a linear combination (M) of source signals s_i (which can be considered as regulatory signals):

$$x_i=Ms_i \quad (3)$$

The goal of the ICA is to estimate s_i, having only the linear mixtures at hand, by

$$s_i=Ux_i \quad (4)$$

so that the estimated components of s_i ($s_1,\ldots,s_D$) are statistically independent. U in eq. (4) is called the un-mixing matrix.

The statistical independence implies that the joint probability density of the components of s_i is equal to the product of the marginal densities of the individual components. Thus, the higher order information of the original inputs is required for estimating s_i, rather than the second-order information of the sample covariance as used in PCA. For the identification of eq. (4), one fundamental requirement is that all the independent components of s_i, with the possible exception of one component, must be non-Gaussian [8].

A large amount of algorithms have been developed for performing ICA [8-11]. One of the best methods is the fixed-point-FastICA algorithm [8,9]. To estimate s_i, the FastICA algorithm finds a direction, i.e. a unit vector u_k (k=1,…,D) such that the projection $u_k^Tx_i$, maximizes non-Gaussianity. Maximizing non-Gaussianity leads to maximizing the negentropy (the most important parameter indicating the measure of non-Gaussianity), and this equally corresponds to minimizing the mutual information between the components [8,9]. However, estimation of negentropy is very difficult, since it requires an estimation (possibly non-parametric) of the probability density functions of components [8]. In the Fast ICA algorithm, the negentropy is approximated by using the contrast function which has the following form:

$$J(s_k)=\left[E\{G(u_k^Tx_i)-E\{G(\upsilon)\}\}\right]^2 \quad (5)$$

Where u_k is a D-dimensional vector, comprising one of the rows of the matrix U. υ is a standardized Gaussian variable (zero mean and unit variance). G is a non-quadratic function, and E represents the expected value. The point here is that by choosing G wisely, one can obtain good approximations of negentropy. The following choices of G have proved to be very useful [8,9]:

$$G_1(s_k)=\frac{1}{a_1}\log(\cosh(a_1 s_k)) \quad (6)$$

$$G_2(s_k)-\exp(-s_k^2/2)$$

where $1 \le a_1 \le 2$ is some suitable constant. Maximizing $j(s_k)$ in Fast ICA for one unit, is based on a fixed-point iteration scheme as follows [8]:

1- Choose an initial (e.g. random) vector u_k.

2- Let $E\{x_i g(u_k^T x_i)\}-E\{g^1(u_k^T x_i)\}u_k$

3- Let $u_k = u_k^+ / \|u_k^+\|$

4- If not converged, go back to 2.

where g and g^1 are, respectively, the first and second derivatives of G. Based on the maximal negentropy principal, the whole matrix U can be computed by maximizing the sum of one-unit contrast function and taking into account the constraint of decorrelation [8].

To simplify the Fast ICA algorithm, two preprocessing steps are applied to x_i. The most basic and necessary preprocessing is to center x_i, i.e. subtract the mean vector so as to make x_i a zero-mean variable. The second step is to whiten x_i, by transforming the observed vector x_i linearly so that the resulting new vector $\tilde{x}_i$ is white, i.e. its components are uncorrelated and their variances equal unity [8,9]. In other words the covariance matrix of $\tilde{x}_i$ equals the identity matrix. That is, $\tilde{x}_i = Vx_i$ and $E(\tilde{x}_i \tilde{x}_i^T)$=I. The transformation matrix V can be obtained by using eigenvalue decomposition such as PCA. The use of PCA whitening also has the property of reducing the dimension of $\tilde{x}_i = Vx_i$, eventually reducing the number of components of s_i.

MULTI-LAYER PERCEPTRON NEURAL NETWORK (MLP NN)

The PCA and ICA methodologies have been proposed for linear component analysis and separation. But in practice the sources and principal components are not commonly mixed linearly, and the mixing processes are non-linear in nature. One appropriate solution to do nonlinear principal component analysis is using Artificial Neural Networks (ANN), which have proved their successful applicability in many applications such as classification, control or regression. To perform non-linear component analysis, we used a Multi-Layer Perceptron (MLP) neural network in auto-associative mode, in which the desired output values are the same input values (see Fig. 2). Here the input to the MLP network is the available output signals from the regulatory network (e.g. network of Hb solutions) such that each node in the input layer corresponds to one output variable in regulatory network If the activation function of all layers are linear and we want to simulate a linear PCA, the network structure should be symmetric [12]. To do a non-linear PCA, we used the same symmetric architecture for a four-layer MLP neural network (which is a common strategy in most applications [12]). But here the activation functions of units were non-linear (Sigmoidal) for all layers except for the last one.

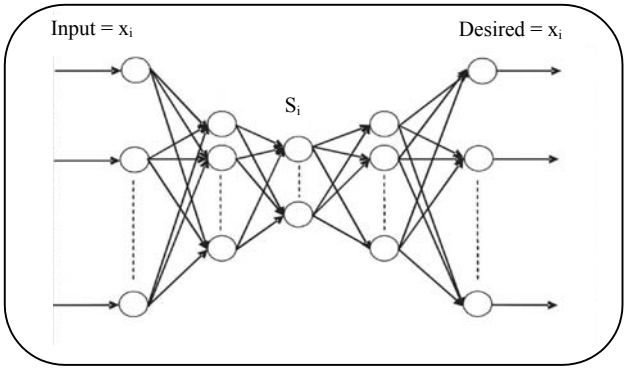

Fig. 2: A multi-layer perceptron neural network in auto-associative mode. The values of the middle layer s_i, are the same regulatory signals.

For dimensionality reduction, the number of nodes in the middle layer (d), should be smaller than those of the input layer (D). The values of the middle layer, s_i are in fact the same regulatory signals which can be obtained through suitable training of the neural network (such as a back-propagation training approach). The number of nodes in middle layer should be therefore equal to the number of regulatory signals.

IMPLEMENTATION AND EXPERIMENTAL VALIDATION

According to Beer–Lambert law, the absorbance spectra can be described as following:

$$[Abs]=[C][\varepsilon] \quad (7)$$

where the rows of [Abs] are the absorbance spectra of seven Hb solution at various wavelengths, the columns of

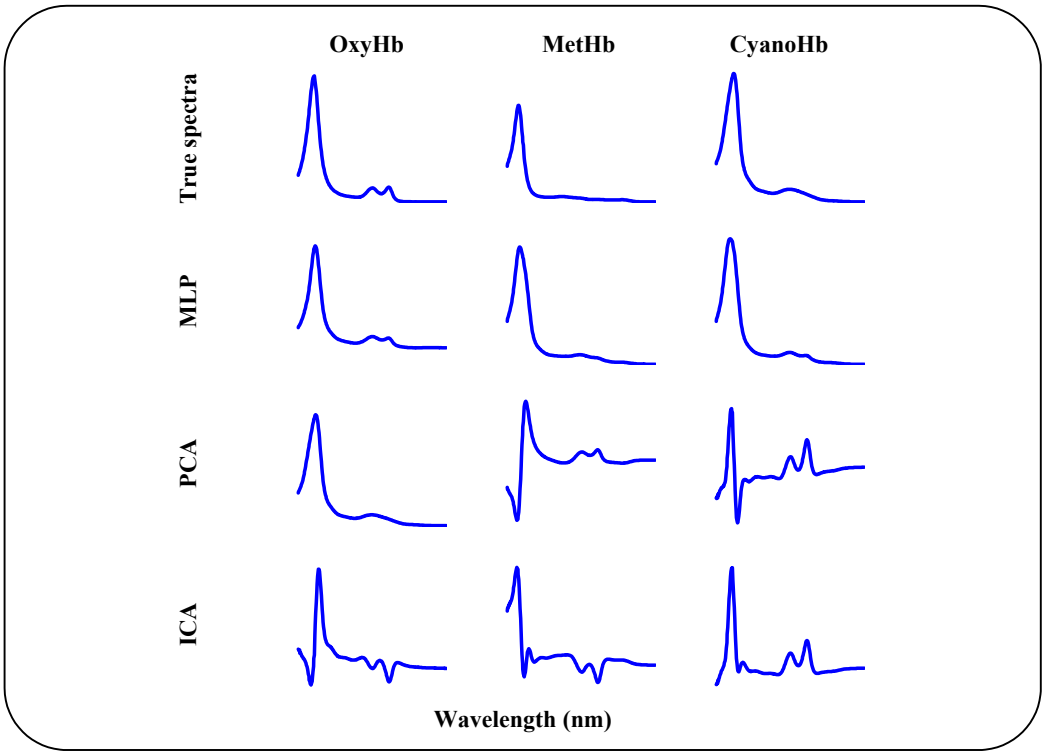

Fig. 3: Experimental validation of the MLP neural network using absorbance spectra of Hb. solutions. The regulatory signals (pure component spectra) derived from MLP approach agree well with the true values, whereas those derived from PCA or ICA do not.

the connectivity matrix [C] are the compositions of three components in seven Hb solutions, encoding the connectivity strength between the regulatory layer and the output signals, and the rows of [ε] are the spectra of pure components. The connectivity diagram of this solution network is shown in Fig. 1. The pure-components spectra [ε] are assumed to be unknown and will be estimated using the three computational methods described above. Determining regulatory signals using the four-layer MLP neural network described above and training with a back-propagation approach, showed that the resulted pure components spectra, [ε], agreed well with the true spectra obtained from independent measurements for pure components, (see Fig. 3). As depicted in Fig. 3, despite the similarity among the pure components spectra, the MLP approach was able to acceptably resolve the differences. In contrast PCA and ICA could not reconstruct the pure components spectra faithfully. It is important to note that here the similarity between the estimated regulatory signals and the true ones is of our special interest and the scale is not considered as an important factor. Therefore to gain a better insight, we have also used a normalized criterion to measure the distance between the estimated regulatory signals and the real signals as the following [13]:

$$d_i = \sqrt{1 - \left(R_i P_i^T / \left(\|R_i\| \cdot \|P_i\|\right)\right)^2} \qquad (8)$$

$$d_{ave} = \left(\sum d_i\right)/3$$

where R_i, denotes the i-th regulatory signal (spectra of the i-th regulatory component resulted from each method) and P_i is the i-th true regulatory signal (the true spectra of the i-th regulatory component determined by independent measurement).

As it is seen in table 1 the average distance between the true spectra of regulatory components and the spectra obtained from the MLP neural network is much less than those of PCA and ICA.

DISCUSSION OF RESULTS AND CONCLUSION

We used a MLP neural network with a symmetric structure and with non-linear activation functions for all layers except for the first layer and the network is then trained based on the back-propagation approach to achieve the minimum square error between the targets and inputs. As it was observed the results of MLP neural

network were significantly superior to those of PCA and ICA, and it was able to successfully reconstruct the regulatory signals for the test case of hemoglobin solutions. It is important to note that the traditional dimensionality reduction approaches such as PCA and ICA are not basically designed to address the problem of hidden dynamics reconstruction in our systems of interest, biological systems, and they usually ignore the underlying network structure. The main reason is that PCA and ICA provide decompositions based purely on a priori theoretical and statistical constraints on the computed regulatory signals. In PCA approach the resulted regulatory signals are constrained to be orthogonal and in ICA they are constrained to be statistically independent. However for cases other than biological systems the assumption of statistical independence or orthogonality may be a reasonable assumption which agrees well with the real system, but this is not usually the case for biological systems. The resulted decompositions thus provide only a phenomenological model for the observed data and do not necessarily contain physically or biologically meaningful signals. Another point is that PCA and ICA try to extract the principal components and regulatory signals based on a linear approach, whereas in real biological networks the regulatory signals are commonly mixed through a complex non-linear process to produce the output signals. Therefore in this research we mainly tried to use a non-linear approach in which no theoretical and statistical assumption is made on regulatory signals. In MLP approach we used, no statistical constrain is posed on the hidden regulatory signals and this naturally allows proper reconstruction of the regulatory signals which is more consistent with underlying real network.

On the other hand a large amount of ever-increasing datasets from biological systems are now available by means of many high-throughput experimental technologies such as DNA microarrays, and developing appropriate and powerful approaches to analyze, and identify the regulatory components, hidden in their underlying networks, is now a significant challenge. This study shows the potential capability of the MLP neural network approach stated above, to efficiently uncover and characterize the hidden regulatory components in many types of biological or biomedical networked systems using a wide variety of large-scale data, such as DNA microarrays, neuronal signals, signal transduction data, metabolic fluxes or protein-protein interactions.

It is important to note however that the Network Component Analysis (NCA) is an approach which has been previously developed for this purpose [6], but it requires a priori knowledge from the underlying network topology. Although such knowledge is going to become available for some biological systems by means of many types of experiments [14-16], it is currently limited to a few number of microorganisms and there are still many organisms and biological systems for which such information from the network structure, are not available at present. Therefore our proposed MLP neural network approach can be applied to less-characterized organisms and biological systems for which there is still only little or no information available from their network structure, to obtain a primary inference from the regulatory components and their dynamic in the network.

Acknowledgments

We highly thank Prof. James C. Liao and Dr. Young-Lyeol Yang from Department of Chemical Engineering at UCLA, for providing us with the experimental data for Hemoglobin solutions.

Received : 1st March 2005 ; Accepted : 6th February 2006

REFERENCES

[1] Raychaudhuri, S., Stuart, J. M. and Altman, R. B., Principal components analysis to summarize microarray experiments: application to sporulation time series, *Pac. Symp. Biocomput.*, **5**, 455 (2000).

[2] Holter, N. S., Mitra, M., Maritan, A., Cieplak, M., Banavar, J. R. and Fedoroff, N. V., Fundamental patterns underlying gene expression profiles: simplicity from complexity, *Proc. Natl. Acad. Sci. USA*, **97**, 8409 (2000).

[3] Holter, N. S., Maritan, A., Cieplak, M., Fedoroff, N. V. and Banavar, J. R., Dynamic modeling of gene expression data, *Proc. Natl. Acad. Sci. USA*, **98**, 693 (2001).

[4] Yeung, M. K., Tegner, J. and Collins, J. J., Reverse engineering gene networks using singular value decomposition and robust regression, *Proc. Natl. Acad. Sci. USA*, **99**, 6163 (2002).

[5] Liebermeister, W., Linear models of gene

expression determined by independent component analysis, *Bioinformatics*, **18**, 51 (2002).

[6] Liao, James C., Riccardo Boscolo Young-Lyeol Yang, Linh My Tran, Chiara Sabatti, and Vwani P. Roychowdhury, Network component analysis: Reconstruction of regulatory signals in biological systems, *PNAS*, **100** (26), 15522, December 23 (2003).

[7] Lay, D. C., Linear Algebra And Its Applications, 2nd ed.; Addison-Wesley Longman Inc.: Reading, MA (1997).

[8] Hyvarinen, A., Oja, E., Independent component analysis: algorithms and applications, *Neural Networks*, **13**, 411 (2000).

[9] Hyvarinen, A., and Oja, E., A fast fixed-point algorithm for independent component analysis, *Neural Computation*, **9** (7), 1483 (1997).

[10] Yamaguchi, T., Itoh1, K., An algebraic solution to independent component analysis, *Optics Communications*, **178**, 59 (2000).

[11] Karhunen, J., Oja, E., Wang, L., Vigario, R. and Joutsensalo, J., A class of neural networks for independent component analysis, *IEEE Trans. Neural Networks*, **8**, 486 (1997).

[12] Haykin, S., Neural Networks: A Comprehensive Foundation, Prentice Hall (1999).

[13] Mansour, D., and Juang, B.H., A family of distortion measures based upon projection operation of robust speech recognition, *IEEE Trans. Acoustics, Signal and Speech Processing, ASSP*, **57(**11), 659 (1989).

[14] Lee, T. I., et al., Transcriptional regulatory networks in Saccharomyces cerevisiae, *Science*, **298,** 799 (2002).

[15] Gardner, T. S., di Bernardo, D., Lorenz, D. and Collins, J. J., Inferring genetic networks and identifying compound mode of action via expression profiling, *Science* **301,** 102 (2003).

[16] Iyer, V. R., Horak, C. E., Scafe, C. S., Botstein, D., Snyder, M., and Brown, P. O., Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF, *Nature*, **409,** 533 (2001).