LETTER TO THE EDITOR

On the Role of a Mutational Database in Building Artificial Intelligence Models for Understanding Gene Expression

T. Murlidharan Nair[*1,2] *and B. D. Kulkarni*[2]

[1]*Department of Medicinal Chemistry, University of Utah, Salt Lake City, UT; and* [2]*Chemical Engineering Division, National Chemical Laboratory, Pune, INDIA*

One of the most intriguing questions that biologists currently face is that concerning the mechanism that precisely controls the level of gene expression required foæ≤determining the fate of cell, cell proliferation, and, more importantly, the survival of the organism. Even more interesting is the mechanism underlying the switching on and off of a particular gene according to developmental programs; failure to follow such programs accurately can result in gross abnormalities. Most control mechanisms involved in the regulation of gene expression occur at the level of transcription, with a hierarchy of coarse and fine controls that together determine the transcriptional activity of each gene. Work toward understanding eukaryotic transcription regulation has been directed by studying transcriptional activation, negative regulatory elements, and other *cis*- and *trans*-acting elements. Understanding the process of transcription would mean a careful analysis of the different components involved in the control mechanism. Thus, understanding the regulatory regions of eukaryotic genomes that play a role in transcription control points toward the need for creating databases of regions harboring the signals involved in the control mechanism. The importance of

[*]Author to whom all correspondence and reprint requests should be addressed.

creating such databases is mainly owing to the need for integrating dispersed experimental data. Such an approach is also important in delineating the precise control regions and interacting factors.

One approach toward understanding such a process, theoretically, is to build first principle models; but building a phenomenological model for a complex system such as this is extremely difficult. An alternative approach is to build black-box models or artificial neural network models (ANN). ANNs are mathematical approximations of the biological synapse and were developed by researchers interested in modeling the brain mechanisms involved in perception. Recent technical advances in this area have made ANN a powerful tool that will help identify complex processes in the presence of noisy information, colinearity of data as also presence of transportation, and time delays. Thus resorting to such a powerful technique in understanding the complex problems involved in transcription control was the only other alternative. Moreover, ANNs are best suited in modeling systems for which only partial information is known. However, building even the black-box model becomes difficult owing to the nonavailability of data. In our recent study we have been able to successfully build a network model for predicting the rate of mRNA synthesis in the case of the β-globin gene *(1)*. The network model was built based on the mutation studies carried out by Myers et al. *(2)*. The system (gene) for which the black-box model was built was the β-globin gene.

The Globin gene expression is an ideal system for the study of differential gene expression, since it is regulated tissue specifically and developmentally. The globin gene encodes the α- and β-like polypeptide subunits of hemoglobin, an oxygen-binding protein. The globin genes were among the first eukaryotic genes to be cloned and are one of the best-characterized genes with respect to structural organization, expression, and evolution. Further, there are a large number of naturally occurring mutant globin genes that have been identified experimentally *(3)*. It is also important to note that the structures of the α- and β-like globin chains produced in the erythroid cells undergo changes at different stages during the development of the animal. These stage-specific globin genes are found in the α- and β-like globin gene cluster. The question as to what controls the activation and inactivation of the individual genes of the globin cluster at different developmental stages is as yet unanswered. While the problem the study addressed is being scrupulously attacked by experimentalists using mutagenesis techniques, there have been no attempts in predicting the rate of transcription theoretically. Such an approach would greatly help in the design of mutation experiments and in the understanding of as yet unknown genetic disorders based on simulation results.

A three-layered neural net to capture the internal representations associated with the transcription control signals harbored in the promoter region of the human β-globin gene has been developed. The three-layered network consisted of 484 neurons in the input layer corresponding to the sequence of the upstream region of the β-globin gene coded in binary, the output layer corresponded to the relative transcription level as reported experimentally. The network performed optimally with eight hidden neurons. The network was trained using the error-backpropagation algorithm *(4)*. In training the network, 117 mutations between positions –101 and +20 were used. The details of the network simulation are reported in our recent work *(1)*. The performance of the trained network was evaluated by testing its prediction capability on 12 mutations that were not part of the training data set. A correlation coefficient > 0.95 was obtained (*see* Table 1 in ref. *1*).

Further, the results of the simulation of mutations using the trained network helps in providing an *a priori* information of the outcome of a mutation. The simulation results delineated the presence of certain sequence elements within the conserved regions, which (when mutated) either did not affect or in some cases even caused a marginal increase in transcription levels. However, the validity of the simulation results could not be ascertained for lack of availability of experimental data. Such theoretical results can be taken as guidelines in designing mutation experiments. It is the importance of building a mutational database that we would like to emphasize in this letter, since there has been no attempt to do so in the past. The database in the form of sequences of regulatory region and their corresponding transcription levels would be a valuable data. This would assist theoretical analysts in building up models (blackbox/empirical). Pooling up of information will also help in validating simulation results. Further, applying highly developed models to important systems such as the globin gene would help in understanding unknown genetic disorders caused by the loss of transcription control signals. This approach would assist in determining the extent of functionality associated with a particular region. All this is possible only if there is a reservoir of mutation data available to build models.

"Data! Data! he cried impatiently, I can't make bricks without clay."

Conan Doyle

ACKNOWLEDGMENTS

This work is supported by the Council of Scientific and Industrial Research, India.

REFERENCES

1. Nair, T. M., Tambe, S. S., and Kulkarni, B. D. (1995) Analysis of transcription control signals using artificial neural network. *CABIOS* **3,** 293–300.
2. Myers, R. M., Tilly, K., and Maniatis, T. (1986) Fine structure genetic analysis of a β-globin promoter. *Science* **232,** 613–618.
3. Higgs, D. R., Vickers, M. A., Wilkie, A. O. M., et al. (1989) A review of molecular genetics of the human β-globin gene cluster. *Blood* **73,** 1081–1104.
4. Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986) Learning representations by back-propagating errors. *Nature* **323,** 533–536.