Global Description of Amino Acid & Nucleotide Sequences: Application to Predicting Protein Folds, Intron-Exon Discrimination, & RNA Gene Identification

Inna Dubchak[*], Stephen R. Holbrook[*], and Ilya Muchnik[**]

[*]E. O. Lawrence Berkeley National Laboratory, Berkeley, CA 94720,
[ildubchak, srholbrook]@lbl.gov
[**]DIMACS, Rutgers University, Piscataway, NJ 08855-1179. muchnik@dimacs.rutgers.edu

ABSTRACT

General global sequence descriptors have been developed which proved to be widely applicable to the prediction of properties of biological genes and gene products. These descriptors include composition, transition, and distribution of defined attributes in the amino acid or nucleotide sequence. We have tested this approach on three completely distinct biological problems: 1) prediction of protein three-dimensional folds; 2) discrimination between sequences of gene introns and exons; and 3) identification of putative RNA genes in genomic sequences.

Keywords: protein fold prediction, exon-intron discrimination, functional RNA genes, computer simulated neural networks

INTRODUCTION

Large-scale sequencing projects produce a massive amount of DNA, RNA, and protein sequence information, in contrast to our current ability to interpret it. We are interested in developing a general approach to predicting structural and functional properties of biological genes and gene products from their amino acid or nucleotide sequence. The approach should be easy to use and applicable to a variety of biological problems.

Successful empirical prediction of properties of biological molecules from sequence depends on: 1) a large database of known examples which contain information correlating the property to the sequence, 2) an appropriate parameterization of the input (sequence) data, and 3) the use of machine learning methods to extract the correlation and extrapolate to prediction of new examples.

Many properties of DNA, RNA, and proteins can not be assigned from examination of local sequences, but depend on global interactions within the entire sequence. Thus, a global parameterization of the entire sequence may be critical for the prediction of macromolecular properties. With this in mind, we have developed a general procedure that uses global parameters derived from the sequence together with computational neural networks (NN) to answer a variety of biological questions including the protein folding problem, recognition of DNA sequence signals, and gene annotation.

METHODS

Global Sequence Descriptors

Our approach uses global information computed from protein, DNA, or RNA sequences. The descriptors we developed are low-dimensionality parameter vectors. Fundamentally, the descriptors for proteins and nucleic acids are the same except that amino acids and their physical attributes are used for formulating the protein descriptors, while nucleotides and their physical attributes are used for nucleic acid descriptors as follows:

Protein sequences: A protein sequence can be represented by a set of parameter vectors based on various physico-chemical and structural properties of amino acids along the sequence. These parameter vectors are constructed in two steps. First, the sequence of the amino acids is transformed into sequences of certain physico-chemical or structural properties (attributes) of residues. The 20 amino acids are divided into three groups for each of six different amino acid attributes. Thus, for each attribute, every amino acid is replaced by the index 1, 2, or 3 according to one of the three groups to which it belongs. The attributes used are the predicted secondary structure, predicted solvent accessibility, hydrophobicity, normalized van der Waals volume, polarity, and polarizability. Second, the three descriptors, "Composition" (C), "Transition" (T), and "Distribution" (D), are calculated for a given attribute. These describe the global percent composition of each of the component attributes of the macromolecule (3 numbers), the percent frequencies with which the attribute changes its index along the entire length of the protein (3 numbers), and the distribution pattern of the attribute along the sequence (5 numbers for each group of the attribute), respectively (see details in [1]). Parameter vectors combine C,T,and D for all six attributes and thus contain 21 scalar components. Percent composition of the twenty amino acids was also used as a parameter vector for training.

DNA Sequences: Since nucleic acid sequences are composed of only 4 components (C,T/U,A,G) compared to 20 for proteins, we have used the nucleotide sequence itself rather than physico-chemical attributes as the basis for the calculated descriptors, i.e., C=% composition of the nucleotides, T= transition, defined as % composition of dinucleotides (AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, TT), and
D = distribution of A, G, C, T along the DNA sequence (5 numbers for each).

RNA Sequences: The basic RNA descriptors are the same as DNA, however for RNA we add "structural descriptors" representing the content of conserved structural elements in RNA sequences. They are: % UNCG tetraloops, % GNRA tetraloops, % CTAG sequence, % tetraloop receptors (AAPu), % uridine turns (CUGA), and calculated free energy of folding (related to base pairing).

Neural Networks

Three-layer feed-forward neural networks were used with the NN weights adjusted by conjugate gradient minimization as implemented in the computer program BIOPROP[2]. The descriptors were used as input to the NN with one node per scalar component of a descriptor. The number of hidden nodes was varied to optimize prediction. The output nodes correspond to the property being predicted. We primarily use two output nodes indicating whether a sequence has a certain property or not, for example, does the input protein sequence belong to a certain fold or not.

Databases

Protein Folds: We created a structural database that did not contain highly homologous protein sequences, yet adequately represented the most comprehensive Structural Classification of Proteins SCOP[1] at its fold level. The subset of SCOP we used contained 607 proteins (where no two proteins have more than 35% sequence identity for aligned subsequences longer than 80 residues) grouped into 128 folds.

DNA Intron-Exons: We used the multi-exon-GB.dat database[3] containing carefully selected genes from GenBank. This database includes 304 genes with 1798 exons of length ranging from 2-5,558 (average 164) nucleotides and introns of length between 23 and 71,717 (average 800) nucleotides. From this database, we selected exons in the range of 100-300 nucleotides and introns from 100-2000 bases for training of the neural networks.

RNA Genes and Non-Assigned Sequences: The *E. coli* chromosome consisting of two strands of 4,697,212 nucleotides (A,T,C,G) each contains 115 annotated RNA genes[4]. These known RNA genes together with the predicted protein coding regions were removed from the genomic sequence resulting in a database of unannotated sequence segments. From the set of RNA gene sequences and the unannotated sequences, we selected windows of 80 nucleotides each to use for training/testing neural networks for the prediction of RNA genes.

RESULTS AND DISCUSSION

We have applied the concept of global sequence descriptors as input to computer simulated neural networks for empirical prediction of protein fold, exon-intron discrimination, and RNA gene identification. The neural networks have been trained using databases of known protein folds, exons and introns, and RNA genes respectively. Prediction schemes can be tested in three different ways: 1) characterization of the performance of each individual attribute; 2) cross-validation on the training set; and 3) use of an independent data set.

Protein Fold Prediction

Predicting a protein fold and therefore an implied function from the amino acid sequence is a problem of great interest. The direct prediction of a protein's 3D structure from sequence remains elusive. However, considerable progress has been made in assigning a sequence to a fold. There have been two general approaches to this problem. One is to use threading algorithms that solve the inverse folding problem: given a group of structures and a sequence, identify the structure that is most compatible with this sequence. The other is a taxonometric approach which presumes that the number of folds is restricted and thus the focus is on structural predictions in the context of a particular classification of 3D folds. Proteins are defined as having a common fold if they have the same major secondary structures in the same arrangement and with the same topological connections. Proteins possessing the same fold in these classifications sometimes have little similarity at the sequence level. The availability of several representatives for each fold allows for extraction of the common features of its members. The level of generalization in this method is higher than in direct sequence-sequence and sequence-structure comparison approaches. Two sequences belonging to the same fold can differ significantly at the amino acid level but the vectors of their global descriptors will be located very close in parameter space. Thus, utilizing these aggregate properties for fold recognition has an advantage over using detailed sequence comparisons. The more classes there are in a classification scheme and the more similar they are, the more difficult it is to distinguish between them. The availability of fine-grained classifications of known structures, such as SCOP (describing more than 400 different protein folds), has encouraged us to choose a taxonometric approach for the development of the scheme for searching for one or a few protein folds which may be similar to that of a target sequence whose structure is unknown. We have developed a neural network based expert system which, given a classification of protein folds, can assign a protein to a fold using sequence data. The prediction procedure is simple, efficient, and incorporated into easy-to-use-software. An important point is that our general descriptors are length independent.

Performance of Attributes: In the case of protein fold prediction, the number of folds predicted at 60% and higher accuracy levels, that we considered satisfactory, totaled 24 (out of 128 folds in our SCOP subset) for all attributes. Among them, 17 folds were predicted by only one attribute, six folds by two attributes, and one fold by three attributes. Assignment to thirteen folds was possible by the predicted secondary structure attribute alone. Note that the accuracy of a random correct prediction to a particular fold in our database equaled N/607 and varied from 2/607 = 0.003 (0.3%) to 30/607 = 0.05 (5%) for the least and the most populated folds in the database. It is important to discover an individual set of descriptors, which work best for a particular fold of interest.

Cross-validation test: The reliability of assignment to a particular fold is directly related to the number of fold representatives used for training. That is why we selected the 27 most populated folds of our database (7 or more proteins) to use in the cross-validation test. Every protein of the fold was individually separated from the training set, the set was trained for the recognition of this particular fold and testing was performed on the separated protein. For every protein, the seven NNs trained on the parameter sets derived from the six amino acid attributes plus the amino acid composition made a fold assignment. The protein sequence was predicted as having the fold if more than one-half of the NNs positively predicted the fold. The percentage of accurately predicted fold representatives was calculated for each of 27 folds. Prediction performance varies widely for different folds – from 9% for the Flavodoxin-like fold to 82.7% for the (TIM)-barrel, and the best prediction accuracy (66.6 – 82.7%) was achieved for the most populated folds with 10-25 representatives. This proves of critical importance to the growth of protein databases for machine learning.

Performance of the method on an independent test set: To test our prediction scheme on independent data, we used the PDB40D set created by the authors of SCOP[1]. This set contains sequences with less than 40% identity to each other and to the proteins of our training set. For this test, the complete database (SCOP subset) was used for NN training and then fold assignments were made for the PDB40D proteins of 27 common folds previously selected. Of the 386 proteins tested there were a total of 161 (41.7%) correctly assigned proteins, and the accuracy of prediction varied in the range 18.5 - 100% for different folds. These predictions are excellent when taking into consideration the low probability of a random assignment (1 – 5%, see above).

Exon-Intron Discrimination

Two approaches have been generally used for identifying specific regions of DNA: search by signal and search by content. We believe that our general global sequence description can contribute to the second of these. We tested this approach to DNA content identification by attempting to discriminate between human gene exon and intron sequences. Discrimination by composition had been previously used, but transition and distribution discriminators are novel as well as is their combination. We trained neural networks on 219 examples of intron and exon sequences and tested on a large set of 2000 examples with the results shown in Table 1.

Table 1. Prediction of DNA sequence as intron or exon*

Parameter set	# of inputs	% correctly predicted examples
Composition (C)	4	69.5
Transition (T)	16	83.6
Distribution (D)	20	82.1
C + T + D	40	93.1

*The neural networks utilized 2 hidden nodes

Identification of Putative RNA Genes

The location of functional RNA genes in genomic sequences is much more difficult than the assignment of ORFs as potential protein coding genes. This is because RNA genes do not utilize the genetic code, lack start and stop codons, and lack a ribosome-binding site (Shine-Dalgarno sequence). To date, the only method of identifying functional RNA genes is by homology to known RNA species,

Table 2. Identification of functional RNA genes*

Parameter	# of inputs	Predicted / Predicted correct / % correct			
		10% threshold		20% threshold	
		RNA	non-RNA	RNA	non-RNA
Composition (C)	4	286/260/90.9	281/231/82.2	286/260/90.9	281/231/82.2
Transition (T)	16	301/279/92.7	301/267/88.7	295/275/93.2	296/266/89.9
Distribution (D)	20	279/194/69.5	283/192/67.8	239/176/73.6	227/147/64.7
C + T + D	40	305/272/89.2	303/274/90.4	305/272/89.2	303/274/90.4
Structural	6	281/262/93.2	291/220/75.6	264/242/91.6	283/219/77.3

*All networks had 2 hidden nodes. Threshold = (pred1-pred2)*100.0/(pred1+pred2), where pred1 & pred2 are activities to two NN output nodes.

such as transfer and ribosomal RNAs, snoRNAs and RnaseP RNA. Elegant computational methods have been developed for locating conserved structural and sequence patterns characteristic of tRNA and snoRNA [5,6].

In an attempt to develop a general method for locating putative genes for novel, functional RNAs, we have applied the composition, transition and distribution attributes discussed above to sequence segments (windows) representing known RNA genes and non-assigned regions of the *E. coli* chromosome. A total of 610 examples of these sequence windows (305 from RNA genes, 305 from non-assigned regions) were used to calculate the descriptors and train neural networks. The results of training a variety of neural networks of different architecture, tested by the jackknife approach are shown in Table 2.

A more analytical description of RNA genes is to assume conservation of structural elements found in all known functional RNAs. These include tetraloops, uridine turns, tetraloop receptors, adenosine platforms, and a high percentage of double helical base pairing. The abundance of these structural elements can be calculated for the sequence windows and also used as input in training neural networks (see Table 2, Structural). A consensus among the various neural networks gives highly accurate prediction of sequences found in functional RNA genes and thus can be used to identify novel RNA genes in genomic sequences.

SUMMARY

Biological systems are based on genes and gene products that are macromolecules composed of strings of nucleotides or amino acids. We have tested a general knowledge-based method that makes predictions of properties based on the global characteristics of the building blocks of sequences. We have developed this computational approach for the assignment of a protein sequence to 3D folds. This method uses global descriptors of protein sequence in terms of the physical, chemical, and structural properties of the constituent amino acids. Neural networks are utilized to combine these descriptors in a way to discriminate members of a given fold from members of all other folds. The method is applicable to the fold assignment of any protein sequence with or without significant sequence homology to known proteins. Our method performed well in the protein fold prediction category of the 1996 Critical Assessment of Techniques for Protein Structure Prediction (CASP2)[7]. A WWW page for predicting protein folds is available at URL http://cbcg.lbl.gov/.

The application of this predictive approach to discrimination of intron and exon sequences and identification of novel RNA genes has been shown to accurately identify these targets in test cases. We believe the approach described here will be generally applicable to a variety of biological problems that are sequence dependent.

REFERENCES

[1] Hubbard TJP, Bart A, Brenner SE, Murzin AG, Chothia C. Nucleic Acids Res. 27:254-56., 1999

[2] Muskal SM & Kim S-H. J. Mol. Biol. 225: 713-27, 1992

[3] Kulp, D., Haussler, D., Reese, M.G., and Eeckman, F.H. Proc. Conf. on Intelligent Systems in Molecular Biology '96 edited by D.J.States, et al. St. Louis, Missouri, AAAI/MIT Press, 134-142, 1996

[4] Blattner FR, Plunkett G 3rd, Bloch CA, et al, Science, 277:1453-1474, 1997

[5] Lowe, T.M. Nucl. Acids Res. 25:955- 964, 1997

[6] Lowe, T.M. Science, 283:168-171, 1999

[7] Levitt M. Proteins, Suppl. 1:92-104, 1997