

Prediction of protein subcellular locations using Markov chain models

Zheng Yuan*

National Laboratory of Biomacromolecules, Institute of Biophysics, Academia Sinica, Beijing 100101, China

Received 8 February 1999; received in revised form 7 April 1999

Abstract A novel method was introduced to predict protein subcellular locations from sequences. Using sequence data, this method achieved a prediction accuracy higher than previous methods based on the amino acid composition. For three subcellular locations in a prokaryotic organism, the overall prediction accuracy reached 89.1%. For eukaryotic proteins, prediction accuracies of 73.0% and 78.7% were attained within four and three location categories, respectively. These results demonstrate the applicability of this relative simple method and possible improvement of prediction for the protein subcellular location.

© 1999 Federation of European Biochemical Societies.

Key words: Protein subcellular location; Markov chain; Protein sequence analysis; Maximum propensity; Residue pair probability; Jack-knife analysis

1. Introduction

The number of new genomes has dramatically increased over recent years and it has once again brought to the forefront the problem of protein function prediction. More and more modern bioinformatic methods have been utilized to provide structural characterization of proteins, for example, predicting the location of secondary structure elements [1,2] and transmembrane helices [3,4], assigning sequences to structure classes [5]. But the prediction of protein functions is primarily depending on the similarity searches against the protein sequence databank [6,7]. For the *Arabidopsis* genome database, the putative cellular roles of 54% of the predicted proteins were established by sequence similarity to proteins from other plants and other organisms [8]. It is evident that predicting the cellular location of a new protein sequence would be very fruitful for determining its function. As yet, only two automatic methods for assignment of the subcellular location are publicly available. One method would be to predict the location only based on the amino acid composition. This approach was suggested by the results of Nakashima and Nishikawa [9]. They found the discrimination between intracellular and extracellular proteins by amino acid composition and residue pair frequencies. Cendano et al. [10] adopted Mahalanobis distance measures to predict the subcellular locations. Furthermore, Reinhardt and Hubbard [11] used neural network models to predict subcellular locations. As for proteins from eukaryotic and prokaryotic cells, the prediction rates are 66% and 81%, respectively, corresponding to four and three classes of subcellular locations. Recent discovery drawn from protein structures suggested that the locations

were in a better correlation with the surface composition than the interior and whole composition [12]. This may lead to an improvement of the prediction accuracy by decreasing the noise, but it must firstly rely on reliable methods to predict the protein surface accessibility. The other method is an expert system based not only on composition but also strongly on the existence of signal peptides [13,14]. In many cases, the signal peptide cannot be found or partly assigned, thereby leading to some problems depending on it. It is known that when protein sequences are decomposed into the amino acid composition, they lose much information for prediction. Hence, it is expected that a higher accuracy should be gained when predicting the subcellular locations directly from sequences.

This study has used Markov chains to predict protein subcellular locations. Markov models are well-known tools for analyzing biological sequence data and they have been used in finding new genes from the open reading frames [15,16], database searching and multiple sequence alignment of protein families and protein domains [17], predicting protein α -turn types [18]. Here, we applied a first-order Markov chain and extended the residue pair probability to higher-order models. This approximation was properly tested during the predicting process. With appropriate Jack-knife tests, this method achieved a prediction accuracy that was 8% higher than the neural networks method, based on the amino acid composition.

2. Materials and methods

2.1. Sequence data

The non-redundant dataset was previously used by Reinhardt and Hubbard [11]. It was taken from SWISSPROT release 33.0 [19] and only included globular proteins, because the transmembrane proteins were predicted with a much higher accuracy [4,20]. No special step was taken to identify or exclude the signal peptides. As shown in Table 1, there are 2427 protein sequences from eukaryotic species classified into four location groups, cytoplasmic, extracellular, nuclear and mitochondrial. 997 prokaryotic sequences were assigned to three location categories, cytoplasmic, extracellular and periplasmic.

2.2. Markov models

The Markov models [21] were used to predict the location. Let S be a protein sequence of length n , which can be generally expressed by

$$r_1 r_2 \dots r_i r_{i+1} \dots r_n$$

where r_i is the amino acid residue at sequence position i . As in the Markov model, this sequence can be considered as being generated from r_1 to r_n . The probabilities of peptides are calculated, taking into account a correlation between residue frequencies in different positions. For a first-order Markov model, the frequencies in position i depend on the residue in position $i-1$. The second-order model takes into account the frequencies of position $i-1$ and $i-2$. A k -order model assumes that the frequencies depend on k former consecutive positions and so backward. Since the correlations between residues are different in every location categories, the Markov models are also different from each other. The following formalism described for the

*Fax: (86) (10) 64872206.

E-mail: zxwang@sun5.ibp.ac.cn

first-order Markov model gives an idea of how the method works. The probability of a sequence in the X location category is given by the ordinary Markov chain formula:

$$P^X(S) = P_1^X(r_1)P_2^X(r_2|r_1)P_3^X(r_3|r_2)\dots P_n^X(r_n|r_{n-1}) \quad (1)$$

Here, P_1 is the probability of residue r occurring at the initial position and its value can be derived from a training set consisting of X location proteins only. $P_n^X(r_n|r_{n-1})$ is the conditional probability (also called transition probability) of observing residue r in position n , given that r is in $n-1$ position and it can also be estimated based on the statistics of duplet peptide frequencies occurring in the same data set. Because r_n and r_{n-1} can be any of the 20 amino acids, the statistics of consecutive pair-residues will generate a matrix with 20×20 elements. Each element represents the occurring frequency of amino acid pair $r_{n-1}r_n$, denoted by $F^X(r_{n-1}, r_n)$. When calculating the conditional probability $P_n^X(r_n|r_{n-1})$ for a given amino acid r_{n-1} , we use the following formula:

$$P_n^X(r_n|r_{n-1}) = \frac{F^X(r_{n-1}, r_n)}{\sum_{\{r_n\}} F^X(r_{n-1}, r_n)} \quad (2)$$

where $\{r_n\}$ means the sum over all 20 amino acids. Thus, for every r_{n-1} , 20 conditional probabilities can be obtained. As a result, a total of 20×20 conditional probabilities should be calculated.

The difference of higher-order models from first-order is reflected by the statistics on frequencies of longer peptides. For example, the transition probability of a residue in second-order Markov chains can be presented as $P(r_n|r_{n-1}, r_{n-2})$, which is dependent on its preceding two residues. Generalization of Eq. 1 to higher-order Markov models is straightforward. For higher-order Markov models, the calculation of the conditional probability becomes more complicated. For the current study, a sequence can be classified into four or three groups according to their subcellular locations. Therefore, X can represent any one of cytoplasmic, extracellular, nuclear, periplasmic and mitochondrial groups and four or three sequence probabilities should be calculated by Eq. 1 in correspondence to different location categories.

For a given sequence, its propensity to any one of the four or three possible states is determined by its probabilities. A maximum probability reflects the highest tendency to this type of location. We change Eq. 1 into an exponential form and define the propensity Ω as follows:

$$\Omega^X(S) = \sum_{i=2}^n \log P_i^X(r_i|r_{i-1}) + \log P_1^X(r_1) \quad (3)$$

Here, X can be any of the location categories. We can assign this sequence to the category of locations with a maximum value.

2.3. Residue pair probability approximation

In general, a higher-order Markov model for a protein sequence analysis is better than a lower-order model but it requires much more probabilities to be calculated. For example, a k -order Markov model demands 20^{k+1} probabilities, which cannot be well-estimated due to the relative small sample of protein sequences available. As to a second-order Markov model, 8000 transition probabilities have to be calculated. As shown in Table 1, when the smallest data set is taken into account, a triplet only has about six averaged frequencies. Although such a database fits for a first-order Markov model, it is not large enough to yield meaningful statistic results for higher-order models. However, we can also take into account other pair residue information besides the neighboring residues (that is the only consideration in a first-order Markov chain). For example, the next nearest residue pair or more can be reliably calculated. We assume the transition probability as:

$$P(r_n|r_{n-1}, \dots, r_{n-k}) = \prod_{i=1}^k P(r_n|r_{n-i}) \quad (4)$$

This is a probability with which all the preceding residues ($r_{n-k}, r_{n-k+1}, \dots, r_{n-1}$) transit to the same residue r_n , given that they are independent of each other. When k equals 1, it degenerates to a first-order Markov model transition probability. The conditional probability $P(r_n|r_{n-1})$ can be easily estimated by counting the residue duplet from different sites in the database. Eq. 3 is changed by superposing $P(r_i|r_{i-1})$ with $P(r_i|r_{i-1}, r_{i-2}, \dots)$. With the decision rule, any

sequence can be predicted to the location category with a maximum propensity.

2.4. Accuracy measures

For the assessment of prediction performance, a Jack-knife test is performed as the cross-validation examination. The learning step is performed with all sequences of a given database except the one for which the location is to be predicted. Matthew's correlation coefficients between the observed and predicted locations over a dataset were calculated after a Jack-knife test, as given by [22]:

$$C(s) = \frac{p(s)n(s) - u(s)o(s)}{\sqrt{(p(s) + u(s))(p(s) + o(s))(n(s) + u(s))(n(s) + o(s))}} \quad (5)$$

Here, $p(s)$ is the number of properly predicted proteins of location s , $n(s)$ is the number of correctly predicted proteins not of location s , $u(s)$ is the number of under-estimated and $o(s)$ is the number of over-estimated sequences.

3. Results and discussion

In order to examine whether residue pair correlation was taken into account properly, from first to eighth order Markov models were tested for all eukaryotic and prokaryotic proteins, as shown in Fig. 1. All prediction accuracies were based on Jack-knife tests. Using first-order Markov models, 70.0% eukaryotic protein sequences and 87.6% prokaryotic sequences were correctly identified. Due to neighboring residue pair correlation, about 5% prediction improvement was achieved in contrast with the neural networks based only on the amino acid composition. As the order of the Markov model increased, the percentage of correctly predicted proteins improved. fourth-order Markov models for both eukaryotic and prokaryotic proteins attained the cumulative accuracies. It is interesting to note that the prediction accuracy was slightly decreased when more than four former consecutive sites were included. This result probably indicates that long distance residue pair correlation (e.g. site 1 and site 6) is comparatively weak, thereby generating noise. The correlations of other pair residues, for example, site 1 and site 5, 2 and 5, 3 and 5, 4 and 5, account for about 3% accuracy improvement. All the following results reported for the location predictions were based on fourth-order Markov models.

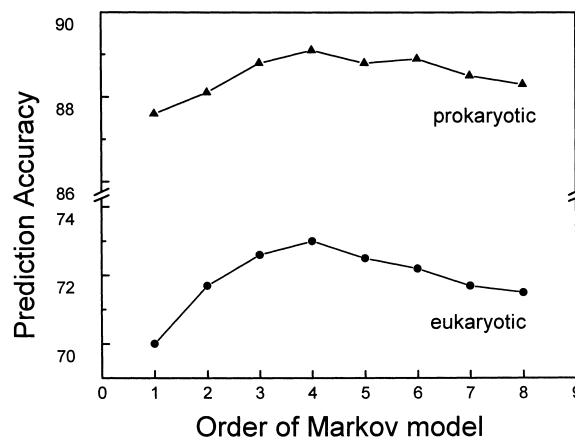


Fig. 1. The prediction accuracy versus the order of Markov model. Prokaryotic proteins were classified into three location categories and eukaryotic proteins were classified into four location categories. The prediction accuracy is the percentage of correctly predicted proteins over the total number of proteins in the dataset for eukaryotic or prokaryotic sequences. \blacktriangle corresponds to prokaryotic proteins and \bullet represents eukaryotic proteins.

Table 1
Observed and predicted sequences for each subcellular location group

Species	Location	Number of sequences	Number of residues
Eukaryotic	Cytoplasmic	684	319 618
	Extracellular	325	123 642
	Nuclear	1 097	551 145
	Mitochondrial	321	137 686
Prokaryotic	Cytoplasmic	688	270 348
	Extracellular	107	51 823
	Periplasmic	202	70 769

Table 2 shows the results for prokaryotic proteins. They were classified into three categories, cytoplasmic, extracellular and periplasmic. An overall prediction accuracy is 89%, which is 8% higher than neural networks based on the amino acid composition. The prediction for periplasmic proteins is worse than that of neural networks, but it is quite better for cytoplasmic proteins (93.6% accuracy and 0.83 correlation coefficient).

Eukaryotic proteins were classified into four categories, cytoplasmic, extracellular, nuclear and mitochondrial. Using this algorithm, 73% of this set were identified correctly (Table 3). In comparison to neural network methods, the prediction accuracy of cytoplasmic proteins was increased up to 23%, although the prediction of extracellular proteins was not, as well as that of neural networks. It was also observed that cytoplasmic and mitochondrial proteins were predicted with low coefficients. Even with the residue pair information, the proteins in overlapping regions were not easy to be identified. Further testing showed that they were liable to be assigned to each other. It was previously found that these two types of proteins could not be well-identified by their amino acid compositions [11]. Since the two kinds of proteins were found in the cell, outside the nuclear and close together, it is reasonable to consider them as one group. The prediction of the mixed proteins could reach 88%, meanwhile, it only changed marginally for the remaining two categories (Table 4). The overall accuracy for three categories of a subcellular location is 78.1%.

Although a simple assumption was given in Eq. 3, the prediction improvement is evident. The results prove that predicting from protein sequences is better than from their amino acid compositions. In this work, transition probabilities are the only estimated parameters, so this model seems simple. Markov models are also biologically meaningful. When proteins are generated from N-terminal to C-terminal, a residue is chosen correlated with its preceding residues and the location

signal is reflected in the selection procedure and in the transition probabilities as well.

Our results also suggest that it may be possible to improve the prediction accuracy. First, with more and more sequences clearly annotated in databases, especially more sequences of phylogenetically distant species are added to databases, the prediction fluctuation will decrease. Due to the relatively small database of extracellular proteins, they are the worst to be predicted. From the variation of the self-consistency and Jack-knife test, the information loss could be observed. By the self-consistency test, each protein is predicted using the rules derived from all databases, without being excluded. Although it is not an objective test, it is useful for testing the potential of a new algorithm [23]. For example, the self-consistency test of extracellular proteins from prokaryotic cells is 85%, 7% higher than the Jack-knife test. But as for large datasets, the variations between self-consistency and the Jack-knife test are very small. Accordingly, by improving and expanding a database to reduce the information loss, proteins from a small dataset will be predicted more accurately. Second, this method may be complementary to other methods. For example, Markov models tend to predict cytoplasmic proteins more precisely, while neural networks predict extracellular or other proteins better. Properly combining all the methods should lead to more favorable prediction results. Finally, extracting location signals and reducing noise may be a better way to improve the prediction performance. Predicting with the entire sequence of a protein is much less sensitive than predicting with segments that are related with subcellular locations. Including these location signals such as signal peptides and the surface residues [12] and taking into account residue correlations suggest a way to improve the accuracy. Since the extracellular proteins are the worst to predict in this work, powerful methods to predict signal peptides [24] become very valuable for determining this type of proteins.

In conclusion, the quantitative assessment of the performance of our Markov models indicates that it performs better than existing methods based on the amino acid composition.

Table 2
Accuracy of the protein location prediction for three categories of prokaryotic sequences

Location	Neural network	Markov model	
	Cross-validation	Jack-knife	Coefficient
Cytoplasmic	80	93.6	0.83
Extracellular	77	77.6	0.77
Periplasmic	85	79.7	0.69
Overall accuracy	81	89.1	–

Neural network models predicted protein subcellular locations using the amino acid composition [11]. All prediction results except the overall accuracy were derived from Fig. 2 of [11] and might have marginal variations. Markov models were fourth-order Markov chains. The prediction results were given by a Jack-knife test and the coefficients were calculated by Eq. 5.

Table 3
Accuracy of the protein location prediction for four categories of eukaryotic sequences

Location	Neural network	Markov model	
	Cross-validation	Jack-knife	Coefficient
Cytoplasmic	55	78.1	0.60
Extracellular	75	62.2	0.63
Nuclear	72	74.1	0.68
Mitochondrial	61	69.2	0.53
Overall accuracy	66	73.0	–

See the legend of Table 2 for further explanation.

Table 4

Accuracy of the protein location prediction for three categories of eukaryotic sequences with fourth-order Markov chain models

Location	Jack-knife	Coefficient
Mixed (cyto+mito)	88.4	0.67
Extracellular	61.8	0.64
Nuclear	74.9	0.68
Overall accuracy	78.7	–

Cytoplasmic and mitochondrial proteins were considered as one group, the so-called mixed group. All the prediction results were given by Jack-knife tests and coefficients were calculated by Eq. 5.

This new approach provides a superior prediction performance with a relatively simple and easy formalism, which is more easily generalized to large databases. Furthermore, an appropriate classification of protein locations is needed taking into account the environmental similarity.

Acknowledgements: The author thanks A. Reinhardt for providing the database, Prof. Z.X. Wang and Dr X.H. Han for fruitful discussions and encouragement. This research was supported by 863 High Technology Foundation(103-13-03-02) and the Pandeng Project of the Ministry of Science and Technology of China.

References

- [1] Frishman, D. and Argos, P. (1997) *Proteins Struct. Funct. Genet.* 27, 329–335.
- [2] Thompson, M.J. and Goldstein, R.A. (1997) *Protein Sci.* 6, 1963–1975.
- [3] Person, B. and Argos, P. (1994) *J. Mol. Biol.* 237, 182–192.
- [4] Rost, B., Casadio, R. and Fariselli, P. (1996) *Protein Sci.* 5, 1704–1718.
- [5] Frishman, D. and Mewes, H.W. (1997) *Nat. Struct. Biol.* 4, 626–628.
- [6] Pearson, W.R. (1990) *Methods Enzymol.* 183, 63–98.
- [7] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) *J. Mol. Biol.* 215, 403–410.
- [8] Mewes, H.W., Hani, J., Pfeiffer, F. and Frishman, D. (1998) *Nucleic Acids Res.* 26, 33–37.
- [9] Nakashima, H. and Nishikawa, K. (1994) *J. Mol. Biol.* 238, 54–61.
- [10] Cedano, J., Aloy, P., Peez-Pons, J.A. and Querol, E. (1997) *J. Mol. Biol.* 266, 594–600.
- [11] Reinhardt, A. and Hubbard, T. (1998) *Nucleic Acids Res.* 26, 2230–2236.
- [12] Andrade, M.A., O'Donoghue, S.I. and Rost, B. (1998) *J. Mol. Boil.* 276, 517–525.
- [13] Nakai, K. and Kanehisa, M. (1992) *Genomics* 14, 897–911.
- [14] Nakai, K. and Kanehisa, M. (1991) *Proteins Struct. Funct. Genet.* 11, 95–119.
- [15] Borodovsky, M., McIninch, J.D., Koonin, E.V., Rudd, K.E., Médigue, C. and Danchin, A. (1995) *Nucleic Acids Res.* 23, 3554–3562.
- [16] Borodovsky, M., Koonin, E.V. and Rudd, K.E. (1994) *Trends Biochem. Sci.* 19, 309–313.
- [17] Krogh, A., Brown, M., Mian, I.S., Sjeander, K. and Haussler, D. (1994) *J. Mol. Biol.* 235, 1501–1531.
- [18] Chou, K.C. (1997) *Biopolymers* 42, 837–853.
- [19] Bayroch, A. and Boeckmann, B. (1992) *Nucleic Acids Res.* 20, 2019–2022.
- [20] Boyd, D., Schierle, C. and Beckwith, J. (1998) *Preprotein Sci.* 7, 210.
- [21] van Kampen, N.G. (1981) *Stochastic Processes in Physics and Chemistry.* North-Holland Publishing Company, New York, USA.
- [22] Matthews, B.W. (1975) *Biochim. Biophys. Acta* 405, 442–451.
- [23] Chou, K.C., Liu, W.M., Maggiora, G.M. and Zhang, C.T. (1998) *Proteins Struct. Funct. Genet.* 31, 97–103.
- [24] Nielsen, H., Engelbrecht, J., Brunak, S. and von Heijne, G. (1997) *Protein Eng.* 10, 1–6.