Predicting Protein Subcellular Location Using Digital Signal Processing

Yu-Xi PAN\textsuperscript{1,2}, Da-Wei LI\textsuperscript{1,2}, Yun DUAN\textsuperscript{2,1}, Zhi-Zhou ZHANG\textsuperscript{1,2}, Ming-Qing XU\textsuperscript{1,2}, Guo-Yin FENG\textsuperscript{1,2}, and Lin HE\textsuperscript{2,3}\textsuperscript{*}

\textsuperscript{1}Bio-X Life Science Research Center, Shanghai Jiaotong University, Shanghai 200030, China; \textsuperscript{2}Institute for Nutritional Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Science, Shanghai 200030, China; \textsuperscript{3}Neuropsychiatric \& Human Genetics Group, Bio-X Center, Shanghai Jiaotong University, Shanghai 200030, China

Abstract

The biological functions of a protein are closely related to its attributes in a cell. With the rapid accumulation of newly found protein sequence data in databanks, it is highly desirable to develop an automated method for predicting the subcellular location of proteins. The establishment of such a predictor will expedite the functional determination of newly found proteins and the process of prioritizing genes and proteins identified by genomic efforts as potential molecular targets for drug design. The traditional algorithms for predicting these attributes were based solely on amino acid composition in which no sequence order effect was taken into account. To improve the prediction quality, it is necessary to incorporate such an effect. However, the number of possible patterns in protein sequences is extremely large, posing a formidable difficulty for realizing this goal. To deal with such difficulty, a well-developed tool in digital signal processing named digital Fourier transform (DFT) \cite{1} was introduced. After being translated to a digital signal according to the hydrophobicity of each amino acid, a protein was analyzed by DFT within the frequency domain. A set of frequency spectrum parameters, thus obtained, were regarded as the factors to represent the sequence order effect. A significant improvement in prediction quality was observed by incorporating the frequency spectrum parameters with the conventional amino acid composition. One of the crucial merits of this approach is that many existing tools in mathematics and engineering can be easily applied in the predicting process. It is anticipated that digital signal processing may serve as a useful vehicle for many other protein science areas.

Key words

sequence order effect; digital signal processing; digital Fourier transform (DFT); frequency domain; covariance discriminant algorithm; bioinformatics; proteomics

Stimulated by the biotechnology revolution and the achievement of the Human Genome Project, proteomic data in databanks have been rapidly increasing over the last two decades. For example, the number of sequences entering into the Swiss-Prot protein sequence database (http://au.expasy.org/sprot/) has increased from 3939 in release 2.0 in 1986, to 90,000 in release 38.0 in 1999, and to 144,731 in release 42.8 in 2004. Confronted with the enormous datasets in the area of proteomics, it is urgent and far more challenging to develop approaches to analyze the complex biological functions of proteins. The cell attributes of a protein, such as which compartment of a cell it belongs to, are closely correlated with its biological functions and play a crucial role in molecular biology, cell biology, pharmacology and medical science. Although the subcellular location of a protein could be obtained by conducting various experiments, those determination procedures are usually time-consuming and costly. In view of this, it is highly desirable to develop a computational method to efficiently and accurately predict the subcellular location of a protein to expedite the process of deducing its functions.
Many efforts have already been made in this regard [2–7]. However, most of the existing algorithms are based solely on the amino acid composition. Although the amino acid composition pattern of a protein is closely correlated with its cell attributes theoretically [8] and predicting algorithms based on it did yield some encouraging results [8], this method is unable to distinguish between two protein sequences with the same amino acid composition but different arranging orders. The sequence order effect should not be ignored as a factor relating to protein subcellular location. Studies have produced encouraging results [9,10] that provide evidence for the effect of sequence arranging order.

The present study was initiated in an attempt to address this problem through a novel approach. To improve the predicting algorithms by incorporating sequence order effect, we employed the digital signal processing (DSP) method through which the arranging order features of a protein sequence were extracted. A significant enhancement was observed in prediction accuracy.

### Representation of a Protein Sequence

A protein sequence is composed of a series of amino acids represented by characters A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W and Y. A sequence composed of these characters can not participate in any mathematical computation because, as lingual symbols, none of the elements has a corresponding numerical value. To describe a protein sequence in a quantitative way, a kind of code needs to be generated to represent the corresponding amino acid. The hydrophobicity value is one of the most important chemical traits of amino acids and it plays an important role in selection of folding types and functions of proteins [11]. We assigned the hydrophobicity values [12], which are listed in Table 1, to all amino acids within the sequences. Through this encoding procedure, a protein sequence is translated to a series of digital signals. All existing tools of DSP can be easily used for the current study.

Digital signal processing is defined as a process of sampling, transforming, synthesizing, estimating and recognizing signals through numerical computing with computers and other specialized equipments to extract useful information [1,13]. Take a protein sequence

\[ \text{R}_1 \text{ R}_2 \text{ R}_3 \text{ R}_4 \text{ R}_5 \ldots \text{ R}_L \]  

where \( \text{R}_1 \) is the first residue, \( \text{R}_2 \) the second, and so forth. To use the technique of DSP, let us first represent the protein sequence by a digital signal

\[ \{x_i\}, 1 \leq i \leq L \] \hspace{1cm} (2)

where \( x_i \) is the hydrophobicity for \( \text{R}_i \) \( (i = 1, 2, \ldots, L) \), as given in Table 1. By conducting digital Fourier transform (DFT) [1], defined as

\[ X(k) = \sum_{i=1}^{L} x(i)e^{-j \frac{2\pi (k-1)i}{L}}, \quad 1 \leq k \leq L \] \hspace{1cm} (3)

where \( \{X(k)\} \) is a series of complex number which has the same length as \( \{x_i\} \), a digital signal is expressed as the sum of a series of resonant waves with the frequency \( \frac{k-1}{L} \) and \( \{X(k)\} \) is the coefficient of each resonant wave. The smaller \( k \) is, the lower the wave’s frequency turns to be, and vice versa [13,14]. Thus, \( \{X(k)\} \) represents the periodicity features and the compositional pattern by resonant waves with various frequencies of the digital signal \( \{x_i\} \). Accordingly, the arranging order effect of the corresponding protein sequence is partially and quantitatively reflected by \( \{X(k)\} \). The magnitude of \( X(k) \) can be easily computed by

---

**Table 1** Hydrophobicity of each amino acid

<table>
<thead>
<tr>
<th>Amino acid</th>
<th>Hydrophobicity</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.7375</td>
</tr>
<tr>
<td>C</td>
<td>0.7875</td>
</tr>
<tr>
<td>D</td>
<td>-0.6125</td>
</tr>
<tr>
<td>E</td>
<td>-0.4875</td>
</tr>
<tr>
<td>F</td>
<td>1</td>
</tr>
<tr>
<td>G</td>
<td>0.6625</td>
</tr>
<tr>
<td>H</td>
<td>0.1625</td>
</tr>
<tr>
<td>I</td>
<td>0.925</td>
</tr>
<tr>
<td>K</td>
<td>-0.5625</td>
</tr>
<tr>
<td>L</td>
<td>0.8875</td>
</tr>
<tr>
<td>M</td>
<td>0.9625</td>
</tr>
<tr>
<td>N</td>
<td>-0.0625</td>
</tr>
<tr>
<td>P</td>
<td>0.5125</td>
</tr>
<tr>
<td>Q</td>
<td>0.025</td>
</tr>
<tr>
<td>R</td>
<td>0.6125</td>
</tr>
<tr>
<td>S</td>
<td>0.6125</td>
</tr>
<tr>
<td>T</td>
<td>0.6875</td>
</tr>
<tr>
<td>V</td>
<td>0.8625</td>
</tr>
<tr>
<td>W</td>
<td>0.775</td>
</tr>
<tr>
<td>Y</td>
<td>0.45</td>
</tr>
</tbody>
</table>
\[ |X(k)|^2 = \left( \sum_{n=0}^{L} x(n) \cos(2\pi(k-1)(n-1)/L) \right)^2 + \left( \sum_{n=0}^{L} x(n) \sin(2\pi(k-1)(n-1)/L) \right)^2 \quad 1 \leq k \leq L \] (4)

The first item of this series, \(|X(1)|^2\), is the constant contribution of the series. As it is just the sum of the rest of the elements, we omitted it. Starting from the second element \(|X(2)|^2\), we selected the former \(M\) elements, which are \(|X(2)|^2\), \(|X(3)|^2\), \cdots, \(|X(M+1)|^2\), as the representing factors of the protein sequence order effect. Because these \(M\) elements are the coefficients of the frequency spectrum [14] of \(\{x_i\}\), we call them frequency spectrum parameters. In some cases, the protein sequence is very short, where \(L < M+1\), we set the lacking elements in the end to zero. By using the frequency spectrum parameters expressed above together with the amino acid composition of the protein sequence, we obtained the complete representing vector of a protein in this study, which is a \((20+M)\)-dimensional \((20+M)\)-D vector,

\[ Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{20} \\ y_{20+1} \\ \vdots \\ y_{20+M} \end{bmatrix} \] (5)

where,

\[ y_t = \frac{a_t}{\sum_{j=1}^{20} a_j + w \sum_{j=2}^{M+1} |X(j)|}, \quad (1 \leq t \leq 20) \] (6)

\[ y_t = \frac{w |X(t-20+1)|}{\sum_{j=1}^{20} a_j + w \sum_{j=2}^{M+1} |X(j)|}, \quad (21 \leq t \leq 20+M) \]

where \(a_t (1 \leq t \leq 20)\) is the amino acid composition of the protein. In other words, the normalized occurrence frequency of the 20 amino acids in the protein \(R_1, R_2, R_3, \ldots, R_L\), where \(|X(j)|\) is the \(j\)-th element of the series defined by Equation (4), and \(w\) is the weight factor for the frequency spectrum parameters. In the current study, we chose \(w=25\). Thus, through the \((20+M)\)-D vector defined above, not only is amino acid composition considered but also the sequence order effect of the protein is, to a certain extent, reflected with the help of DSP tools.

### Predicting Algorithms

In this section we will explain how the existing predicting algorithms are reformulated to be applied to the new protein representing form given by Equation (5). Suppose there are \(N\) proteins composing a dataset \(S\), which is the union of \(m\) subsets

\[ S = S_1 \cup S_2 \cup S_3 \cup \cdots \cup S_m \] (7)

Each of the subsets consists of proteins with the same cellular attribute. The size of each of them is \(n_i (i=1,2,3,\ldots,m)\), that is, the number of proteins within the subset \(S_i\) is \(n_i\) and \(N = n_1 + n_2 + \cdots + n_m\). In order to incorporate the sequence order effect, we replace the traditional amino acid composition, which is expressed as a 20-D vector, by the new representing form containing the frequency spectrum parameters, which is a \((20+M)\)-D vector given by Equation (5). According to such a baseline, the \(k\)-th protein in the subset \(S_i\) should now be

\[ y'_k = \begin{bmatrix} y'_{k,1} \\ y'_{k,2} \\ \vdots \\ y'_{k,20+M} \\ y'_{k,20+M} \end{bmatrix}, \quad (k=1,2,\ldots,n_i; i=1,2,\ldots,m) \] (8)

where \(y'_{k,j} (j = 1,2,\ldots,20+M)\) has the same meaning as \(y_j\) in Equation (5), but belongs to \(Y'_i\) instead of \(Y\). For each subset \(S_i\), we define the central vector as

\[ \bar{Y}_i = \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \vdots \\ \bar{y}_{20+M} \end{bmatrix}, \quad (i=1,2,\ldots,m) \] (9)

where,

\[ \bar{y}_j = \frac{1}{n_i} \sum_{k=1}^{n_i} y'_{k,j}, \quad (j = 1,2,\ldots,20+M) \] (10)

Considering the central position of \(\bar{Y}_i\), it can be viewed as the central protein for the subset \(S_i\).

If \(Y\) is the protein whose cellular location is to be predicted, it can be represented as a \((20+M)\)-D vector as given by Equation (5). Now the problem is how to effectively measure the similarity between the query protein \(Y\) and the central vectors \(\bar{Y}_i\). To avoid drawing unilateral conclusions, we conducted two predicting algorithms in this study which have different definitions for the similarity between two vectors. Each is discussed below.

©Institute of Biochemistry and Cell Biology, SIBS, CAS
ProtLock algorithm [2]

The measurement of similarity between vectors \( Y \) and \( \overline{Y} \) in the ProtLock algorithm was defined as

\[
D^0_i(Y, \overline{Y}) = \left( Y - \overline{Y} \right)^T C^{-1} \left( Y - \overline{Y} \right), \quad (i = 1, 2, \ldots, m)
\]

where \( C \) is a matrix given by

\[
C = \begin{bmatrix}
    c_{1,1} & c_{1,2} & \cdots & c_{1,20+M} \\
    c_{2,1} & c_{2,2} & \cdots & c_{2,20+M} \\
    \vdots & \vdots & \ddots & \vdots \\
    c_{20+M,1} & c_{20+M,2} & \cdots & c_{20+M,20+M}
\end{bmatrix}
\]

The superscript \( T \) in Equation (11) is the transposition operator and \( C^{-1} \) is the inverse matrix of \( C \). Each element of the matrix \( C \) is defined as

\[
c_{p,q} = \sum_{i=1}^{n_i} \sum_{k=1}^{n_k} \left( y^i_{k,p} - \overline{y}^i_p \right) \left( y^i_{k,q} - \overline{y}^i_q \right), \quad (p, q = 1, 2, \ldots, 20+M)
\]

The predicting rule was given by

\[
D_p^i(Y, \overline{Y}^i) = \text{Min}_{\alpha} \{ D^0_1(Y, \overline{Y}^i), D^0_2(Y, \overline{Y}^i), \ldots, D^0_m(Y, \overline{Y}^i) \}
\]

where \( \alpha \) can be 1, 2, 3, \ldots, or \( m \), the operator Min means taking the minimum one among those elements in brackets, and the superscript \( \alpha \) indicates the cellular attribute predicted for the query protein \( Y \). If there is a tie, \( \alpha \) can not be uniquely determined, although this rarely occurs.

Covariance discriminant algorithm [3,5]

In the covariance discriminant algorithm, a function was used as a scale to measure the similarity between proteins \( Y \) and \( \overline{Y} \), instead of geometrical distance used in the ProtLock algorithm mentioned above. The smaller the value of the function, the more similar the two proteins were estimated to be. For the \((20+M)\)-D new protein-representing vector, the so-called discriminant function can be expressed as

\[
F(Y, \overline{Y}^i) = \left( Y - \overline{Y}^i \right)^T C_i^{-1} \left( Y - \overline{Y}^i \right) + \ln |C_i|, \quad (i = 1, 2, \ldots, m)
\]

where the former item \( (Y - \overline{Y}) C^{-1} (Y - \overline{Y}) \) is the squared Mahalanobis distance [15–17] between \( Y \) and \( \overline{Y} \), and

\[
C_i = \begin{bmatrix}
    c_{1,1} & c_{1,2} & \cdots & c_{1,20+M} \\
    c_{2,1} & c_{2,2} & \cdots & c_{2,20+M} \\
    \vdots & \vdots & \ddots & \vdots \\
    c_{20+M,1} & c_{20+M,2} & \cdots & c_{20+M,20+M}
\end{bmatrix}
\]

is the covariance matrix of the subset \( S_i \). Each of its elements is defined as

\[
c_{p,q} = \frac{1}{n_i - 1} \sum_{k=1}^{n_i} \left( y^i_{k,p} - \overline{y}^i_p \right) \left( y^i_{k,q} - \overline{y}^i_q \right), \quad (p, q = 1, 2, \ldots, 20+M)
\]

and \( |C_i| \) is the determinant of the matrix \( C_i \). Similarly, the predicting rule was given by

\[
D_p^i(Y, \overline{Y}^i) = \text{Min}_{\alpha} \{ F(Y, \overline{Y}^i), F(Y, \overline{Y}^i), \ldots, F(Y, \overline{Y}^i) \}
\]

Note that the sum of the \(20+M\) component in Equation (5) is equal to 1 (imposed by the normalization condition), that is, of the \(20+M\) components, only \(20+M-1\) are independent. Accordingly, the covariance matrix \( C_i \) as defined in Equation (17) must be a singular matrix [15]. This implies that the Mahalanobis distance given in Equation (16) and the corresponding covariance discriminant function defined in this equation would be divergent and undefined. To overcome this difficulty, let us take the following dimension-reducing procedure [15]. Instead of defining a protein in a \((20+M)\)-D space, let us define it in a \((20+M-1)\)-D space by leaving out one of its elements. The remaining elements thus obtained would be completely independent and the corresponding covariance matrix \( C_i \) no longer singular. In such a \((20+M-1)\)-D space, the Mahalanobis distance in Equation (16) and covariance discriminant function can be defined without the difficulty of divergence. Furthermore, according to the invariance theorem given in Appendix A of Chou [15], the values of the Mahalanobis distance and covariance discriminant function remains the same regardless of which one of the \(20+M\) components is left out. Accordingly, the values of the Mahalanobis distance and covariance discriminant function can be uniquely defined through such a dimension-reducing procedure. The same procedure can also be used to solve the divergence problem occurring in Equation (11) of the ProtLock algorithm.
Results and Discussion

The data used in this study were selected from the proteins with annotated subcellular locations in release 42.8 of the Swiss-Prot databank. All samples with ambiguous annotations such as "by similarity", "probable", "potential" and "possible" and proteins with more than one cell attribute were excluded. Those proteins were also excluded that the number of samples within the database which belong to their subcellular locations is less than 100. For example, Golgi apparatus proteins and vacuole proteins have very few samples within their location classes, that is, they lack statistical significance. The remaining 20,172 proteins compose our raw dataset, which consists of 10 subcellular location classes named nuclear, cytoplasm, mitochondria, extracellular, peroxisome, endoplasmic reticulum, lysosome, membrane, chloroplast and periplasm. The number of proteins and their distributions in the 10 categories are listed in Table 2.

We did not remove the homological samples from our dataset for the following reasons. (1) The number of proteins with the same name but from different species is extremely small compared to the size of the whole dataset. Thus, their effect on prediction quality was minimal. (2) In the current study, the main focus was the improvement in prediction quality due to the incorporation of sequence order effect. Moreover, all the results discussed in this study were obtained on the basis of a uniform dataset. The improvement discussed had nothing to do with the homologous samples.

The prediction quality was examined by three various test methods: the self-consistency test, the jackknife test and the independent test. In the self-consistency test, the subcellular location of each protein in the dataset was predicted using the rules derived from the same dataset. In the jackknife test, each protein in the dataset was singled out in turn as a ‘test’ protein and the rule parameters were determined from the remaining N-1 proteins, which build up the training dataset. Jackknife tests are considered one of the most effective and objective methods for cross-validation in statistics [18]. For the demonstration of practical application, the independent test was also conducted. In this test, two independent datasets, the training dataset and the testing dataset, were set up. None of the proteins contained in one dataset is included in the other. To construct these two datasets, each of the subsets composing the whole dataset was meanly and stochastically separated into two groups. One of them was sent to the training dataset and the other to the testing dataset. Accordingly, each subset of the training dataset and the testing dataset had a size that was just half of that of the corresponding subset in the whole dataset.

As mentioned earlier, M is the number of frequency spectrum parameters obtained through DSP tools representing the sequence order effect. The greater M is, the more order effect information is incorporated. However, M has an upper limit. For the dataset with which Mahalanobis distance is to be conducted, the number of samples within the training dataset must be greater than the dimension of the vectors. Otherwise, the covariance matrix of the training dataset C would be singular and the inverse of this matrix would be undefined [19]. In the variance discriminant algorithm, because Mahalanobis distance is applied for each subset S as given by Equation (16), the number of proteins within the training dataset of each S must be greater than 20+M or the covariance matrix C would be divergent and the covariance discriminant function would be undefined. This means that the optional value for M is limited by the size of the smallest training dataset. Looking at the sample size of each subset listed in Table 2, the smallest subset, peroxisome, is composed of 126 proteins. For the self-consistency test and the jackknife test, the training dataset of this subset contains 126 and 125 samples, respectively. In the independent test, however, the size of the training dataset for this smallest subset is only 63. Therefore, 20+M should be less than 63. We set 20+M to 60 in this study, so M was set to 40.

For the convenience of comparison between the tradi-

### Table 2 Number of proteins in each cellular location

<table>
<thead>
<tr>
<th>Cellular location</th>
<th>Number of proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nuclear</td>
<td>3714</td>
</tr>
<tr>
<td>Cytoplasm</td>
<td>6993</td>
</tr>
<tr>
<td>Mitochondria</td>
<td>798</td>
</tr>
<tr>
<td>Extracellular</td>
<td>378</td>
</tr>
<tr>
<td>Peroxisome</td>
<td>126</td>
</tr>
<tr>
<td>Endoplasmic reticulum</td>
<td>192</td>
</tr>
<tr>
<td>Lysosome</td>
<td>150</td>
</tr>
<tr>
<td>Membrane</td>
<td>5764</td>
</tr>
<tr>
<td>Chloroplast</td>
<td>1670</td>
</tr>
<tr>
<td>Periplasm</td>
<td>387</td>
</tr>
<tr>
<td>Total</td>
<td>20,172</td>
</tr>
</tbody>
</table>

* The protein sequences and annotations were extracted from release 42.8 of the Swiss-Prot databank. The data were selected according to the rule expressed in “Results and Discussion”.

©Institute of Biochemistry and Cell Biology, SIBS, CAS
tional amino acid composition representing form and the new representing form including frequency spectrum parameters, each of the predicting algorithms mentioned above was conducted based on the two forms of input. The results of the three test methods using the ProtLock algorithm and the covariance discriminant algorithm are listed in Table 3.

Using the covariance discriminant algorithm, the results based solely on amino acid composition were 58.8% for the self-consistency test, 57.9% for the jackknife test and 54.8% for the independent test. Using the same algorithm but based on the new input form incorporating the frequency spectrum parameters, the results were 79.2%, 74.3% and 75.8% for the self-consistency test, jackknife test and independent test, respectively, an improvement of between 16% and 21%.

A similar improvement was observed for the ProtLock algorithm. With the amino acid composition as the input, the results were 47.9% for the self-consistency test, 47.6% for the jackknife test and 48.4% for the independent test. Based on the new input form proposed in this study, the results were increased to 50.5% for the self-consistency test, 49.8% for the jackknife test and 50.4% for the independent test. The success rate increased by an average of 2.4%.

These results show that, regardless of whether the ProtLock algorithm or the covariance discriminant algorithm was used, or which of the three test methods was conducted, higher prediction quality was achieved with the frequency spectrum parameters incorporated. We can conclude that sequence order effect, as well as amino acid composition, is a crucial factor, which has a close relationship with the protein cell attributes. Moreover, as a set of variables which were expected to represent the sequence order effect, the frequency spectrum parameters proposed in this study were effective in extracting some significant information correlated with the protein’s cellular attributes from the protein sequence and delivering the information to the predicting algorithms.

The increasing magnitudes derived from the covariance discriminant algorithm were much greater than those derived from the ProtLock algorithm. This confirms that the covariance discriminant algorithm is more powerful than the ProtLock algorithm in dealing with the coupling effect, as discussed in earlier research [5]. The novel protein representing form proposed in this study, combined with the covariance discriminant algorithm, is an effective tool in the practical application of protein subcellular location prediction.

As mentioned above, the weight factor $w$ in Equation (6) was set to 25. The purpose of using $w$ was to rearrange the magnitude of the frequency spectrum parameters in order to adjust the relative proportion between amino acid composition and the order effect factors. Although the relative proportion between them can be adjusted by choosing various values for $w$, the success rates are not sensitive to those changes, that is, the predicting results obtained by the two algorithms do not vary much for different values of $w$. The two algorithms expressed above are able to reshape the sample space with the usage of $C_i^{-1}$ in Equation (11) and $C_i^{-1}$ in Equation (16) [19]. Despite the different definition of the two matrixes, they have the common function of modifying the shape of the sample space and balancing the magnitude of each dimension of

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Input representing form</th>
<th>Test method</th>
</tr>
</thead>
<tbody>
<tr>
<td>ProtLock (Cedano et al., 1997)</td>
<td>Amino acid composition *</td>
<td>Self-consistency *</td>
</tr>
<tr>
<td></td>
<td>With frequency spectrum parameters incorporated</td>
<td>Jackknife *</td>
</tr>
<tr>
<td>Covariance-discriminant (Chou and Elrod, 1999)</td>
<td>Amino acid composition *</td>
<td>Independent test *</td>
</tr>
<tr>
<td></td>
<td>With frequency spectrum parameters incorporated</td>
<td></td>
</tr>
</tbody>
</table>

* conducted on 20,172 proteins classified into 10 location classes as described in “Results and Discussion”; † conducted based on the rule parameters derived from the 10,086 proteins training dataset for the 10,086 proteins test dataset. The training dataset and the test dataset were obtained by meanly and stochastically separating each subset of the whole dataset into two groups of the same size; †† the conventional amino acid composition consists of 20 components, each representing the occurrence frequency of one of the 20 native amino acids in a protein; †‡ the new representing form of a protein with frequency spectrum parameters incorporated consists of $20 + M$ components [Equation (5)]. The optimal number for $M$ is 40.
the space. Accordingly, within a wide range, the variation of \( w \) imposes little upon prediction quality. For instance, when \( w \) increases from 0.1 to 50, the rates of correct prediction derived from the covariance discriminant algorithm based on the self-consistency test change from 78.0% to 79.3%. Similar changing magnitudes were also observed for the jackknife test and the independent test. As the best prediction quality was obtained using the covariance discriminant algorithm, and because the jackknife test is considered to be one of the most objective cross-validation test methods [18], the optimal value for \( w \) was that resulting in a relatively better jackknife-tested rate.

The goal of the current study, in applying DSP tools for the prediction of protein subcellular location, was not limited to the enhancement of the success rate, we were also introducing a different approach to incorporating the sequence order effect. Furthermore, we were attempting to find the mechanisms or rules governing a protein’s biological functions, determined by its primary sequence, through the application of well-established mathematical and engineering tools. The attempt to predict protein subcellular location is only an initial step, but a significant one. We consider the following points to be potential obstacles to our improving the prediction quality.

The whole predicting process is carried out on the basis of the knowledge provided by the training dataset of proteins. Clearly, the more complete the training dataset, the stronger the prediction is supported. With the protein sequences and the knowledge of their cellular attributes available so far, it is too premature to construct a complete training dataset. This becomes the first obstacle to our predicting research. On further observation of the success rates of each subset, as listed in Table 4, we made an interesting discovery. Fig. 1 shows the scatter graph of the number of proteins within the training dataset of each subset and their corresponding success rates, according to the data given in Table 4. To facilitate comparison between the two testing methods, the sample size of the training dataset of each subset for the independent test was doubled (Fig. 1). In Fig. 1, the abscissa axis indicates the sample size of the training dataset of each subset while the ordinate axis indicates the rate of correct predictions. The curve with open circles indicates the independent test and the curve with filled circles indicates the jackknife test. From the figure we can get following points. (1) With the increase in the sample size of the training dataset, both curves are generally climbing up. (2) Taking the sample size of 1669 as a separating point, the points to the right side are all above 60% while those to the left are mostly beneath that level. (3) Taking the same point as a separating point, the curves to the right side tend to be smooth while those to the left go up and down sharply. (4) With the curve of the jackknife test singled out, the success rates of the two smallest subsets are extremely low (under 20%).

To sum up, these observations demonstrate that the greater the sample size of the training dataset, the more stable and reliable the predicting result. When the size of any subset is as low as to be close to the dimension of the

Table 4 Rates of correct prediction for each subcellular location of proteins using the jackknife test and the independent test

<table>
<thead>
<tr>
<th>Cellular location</th>
<th>Jackknife test a</th>
<th>Independent test b</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of proteins in training dataset</td>
<td>Rate of correct prediction</td>
</tr>
<tr>
<td>Peroxisome</td>
<td>125</td>
<td>45.2%</td>
</tr>
<tr>
<td>Lysosome</td>
<td>149</td>
<td>42.7%</td>
</tr>
<tr>
<td>Endoplasmic reticulum</td>
<td>191</td>
<td>48.3%</td>
</tr>
<tr>
<td>Extracellular</td>
<td>377</td>
<td>46.8%</td>
</tr>
<tr>
<td>Periplasm</td>
<td>386</td>
<td>63.3%</td>
</tr>
<tr>
<td>Mitochondria</td>
<td>797</td>
<td>36.8%</td>
</tr>
<tr>
<td>Chloroplast</td>
<td>1669</td>
<td>67.4%</td>
</tr>
<tr>
<td>Nuclear</td>
<td>3713</td>
<td>61.0%</td>
</tr>
<tr>
<td>Membrane</td>
<td>5763</td>
<td>88.7%</td>
</tr>
<tr>
<td>Cytoplasm</td>
<td>6992</td>
<td>77.1%</td>
</tr>
</tbody>
</table>

\( ^a, ^b \) refer to Table 3 footnote.

©Institute of Biochemistry and Cell Biology, SIBS, CAS
representing vector, the corresponding prediction of the subset is extremely unreliable. This was taken into account in the construction of the dataset used in this study, in that the location classes considered to be too small were excluded from the count.

We did not discuss the self-consistency test in this paper because, in this test method, each of the testing sample is included in the training dataset as well. The results obtained by the test cannot accurately reflect the completeness of the training dataset. In the jackknife and independent tests, however, the knowledge of the testing sample is removed from the training dataset.

Another potential obstacle is that the predicting algorithms used in this study restrict the representing form of a protein in many ways, limiting information about the primary sequence that could be delivered as input to the prediction. For instance, with regard to the Mahalanobis distance, the dimension of the input vector must be smaller than the sample size of the training dataset of each subset. For both algorithms, only vectors of a single column were accepted as input data, and matrixes with multi-columns were not. Through the current input form, it is hard to consider many kinds of physical, chemical and biological characteristics of amino acids simultaneously and separately. Therefore, useful information cannot be incorporated adequately.

Conclusion

The dataset of protein sequences with annotated subcellular locations is crucial for current research. The more complete the dataset is, the closer it comes to reflecting nature. In the current study, we constructed our dataset according to the latest release of the Swiss-Prot databank. After comprehensive analysis and strict selection, our dataset for protein subcellular location covered as many proteins as were available, which makes our work well-grounded and reliable.

To describe a protein sequence more precisely in mathematical language, we used frequency spectrum parameters as a complement to amino acid composition. Through the hydrophobicity of each amino acid, a protein sequence was translated into a digital signal series then analyzed using DSP tools. This is the first attempt to analyze a protein sequence within the frequency domain and to extract the sequence characteristics in the frequency domain through DFT. One of the significant merits of introducing the DSP approach is that many well-established and sophisticated mathematical and engineering tools can easily be applied to improve the prediction quality of various protein attributes. It is anticipated that DSP may serve as a useful vehicle for many other protein science areas. The results obtained indicate that the frequency spectrum parameters extracted through the DSP tools have effectively reflected the protein sequence order effect and are helpful for the prediction of protein subcellular location.

References


Edited by Jun YU