Open Access Library Journal
2020, Volume 7, e7049
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Selection and Analysis of Protein Circular Dichroism Spectra Using an Expansion of Spectral Factors

David A. Haner

California State Polytechnic University, Pomona, USA
Email: dahaner3830@gmail.com

How to cite this paper: Haner, D.A. (2020) Selection and Analysis of Protein Circular Dichroism Spectra Using an Expansion of Spectral Factors. *Open Access Library Journal*, **7**: e7049.
https://doi.org/10.4236/oalib.1107049

Received: November 28, 2020
Accepted: December 28, 2020
Published: December 31, 2020

Abstract

Techniques are presented to develop spectroscopic factors directly from circular dichroism spectra of proteins using singular value decomposition on a small database. Four spectra of maximum spectral variability are chosen to characterize the database. These selected protein spectra are then factored by singular values into component spectra, which are collected as comparative vector characteristics used as factor fractions. The necessary standardization for comparison is achieved using unit normalized spectra. Those spectra are used to quantify the parameter uncertainties as a means for comparison. The difference between the fit spectrum and the data spectrum for each protein is analyzed by least square to obtain parameter uncertainties due to the model. The sum of the factor fractions over the database is within the theoretical predictions.

Subject Areas

Biochemistry

Keywords

Circular Dichorism, Singular Value Decomposition, Ordering Spectra, Residual Error, Data Compaction

1. Introduction

It is generally accepted that the secondary structures of a protein are indicative of spectral variability characteristic of circular dichroism (CD) spectra [1] [2] [3]. The use of CD spectroscopy to determine the secondary structures of proteins has been active for 50 years. The earliest efforts postulated known struc-

tures to be characterized by specially prepared protein samples. The data was difficult to develop and some spectral uncertainties were indicated, but not over the whole bandpass. This deficiency has not been improved, though it is essential in describing the precision of the model parameter. There have been many technological advances in optical spectrometers that could present the mean value spectrum and the accompanying error spectrum.

A large number of reports have been presented detailing methods of spectral analysis that support the accepted secondary structure fractions obtained from x-ray crystallographic data. These produced varying degrees of success, but were rarely based on experimental/theoretical precision indicators of the model parameters, but rather on correlation to the x-ray data. The linear correlation coefficient however is not a quantifier of the adequacy of the mathematical model to duplicate the spectral measurements.

This report's objective is to address the question: What information is embedded in the spectroscopic features of a collection of spectra? The answer to this question must use rigorous techniques that involve only quantities that characterize the spectroscopic data. Thus the nomenclature of x-ray analysis does not apply directly. That connection possibly lies in the domain of computational quantum chemistry and extensive laboratory experimentation and measurement.

A range of techniques used to demonstrate the prominent spectroscopic factors found in a collection of spectra will be shown. The set chosen [1] [4] [5] will be for globular proteins and a polypeptide, *i.e.* Compton, Johnson Database (CJDB) with sixteen CD-spectra in all over the bandpass 178 nm - 260 nm, calibrated by the group and published in detail [Table 1 of reference 1]. The analysis requires a complete normalized data set of the published protein spectrum, [6] and an empirical error spectrum for each protein spectrum. The principle of the analysis is to model the experimental measurements in this special measurement condition. The data will be subjected to treatment to order and select the spectra in accordance with specific criteria. This will allow the solution of the model to follow the fully determined relationship between the number of knowns and unknowns. This is obtained by selecting the number of spectra from the top of the ordered list.

The algorithm employs routines in common use in linear algebra or matrices. One technique of central importance is singular value decomposition (SVD) [7] [8] [9] including its use in Legendre least squares.

The advantage of the use of SVD is that the number of singular values will correspond to the number of spectroscopic factors present in the database. The number of primary singular values (there are no secondary singular values) determines the number of spectroscopic parameters of the mathematical model for a fully determined solution. An application of SVD to this selected limited set of protein spectra gives the factors present to be expanded into components. It is the particular manipulation and collection of the component parts that produce

the matrix of fractions or a matrix of basis functions. These two matrices coupled with the selected spectral data set gives two possible initializations for analysis that can characterize any CD spectrum of the same size and calibration as the original database spectra.

When the error spectrum is treated by the algorithm, [9] the uncertainty in the parameters is given. By collecting the residual function, *i.e.* absolute value of the difference between the data spectrum and the model fit spectrum, and analysing it by the algorithm, the parameter uncertainty due to the modeling is presented.

The most significant aspect about this work is that the modelling is entirely restricted to the spectra of a limited collection. The most useful contribution is the emphasis of the precision indices of the model parameters obtained in the analysis to facilitate comparison of spectral changes resulting from designed chemical processes.

Retention of the selected model basis vectors, the database factor fractions, and the original spectrum norms facilitates a data compression of the database. A replication of the original database may be created directly by using these data.

2. Theory

Recently it has become the practice to set criteria in selecting [10] [11] which data are to be included in the collection to best suit the objectives of the analysis. Usually the emphasis is to select the data that presents the greatest range of variability in the measurement characteristics within a classified group. It seems reasonable to examine the information content of just the CD spectra of proteins. Since one spectrum contains information about that protein, examination of a group of protein spectra should yield information about the range of possibilities in that database. Any averaging process tends to reduce the high-resolution features of the individuals of the collection. The greater the number of individuals in the averaged collection, the more the high-resolution features is reduced. This fact suggests that analysis of a database should require that a selection process be employed to choose a limited number of spectra, which contain the primary features expressed in the database. This process would minimize duplication and emphasize spectral variability.

The selection process reduces the number of spectra with few spectral variations for inclusion in the databases, while retaining the spectra with the greatest spectral variability. While there may be subsidiary [12] criteria for selection, the spectroscopic characteristics are of paramount consideration in this usage. The selection process requires that the individual spectra be compared in order to decide the membership of the reduced database. The most valuable spectra are those with enough variability that suggests a mixture of spectroscopic signatures. The spectra should also be selected for their uniqueness of spectral shape to ensure the scope of possible structure identification. Thus, duplicates are not needed. The selection process is labor intensive, but is the key to the utility of the

ensuing analysis. By ordering the spectra according to specific criteria, the order can be used as a selection process. For example, the criteria used to order the spectra could be the pair wise distance between the unit normalized spectra or, alternatively, the "angle" between the spectra (vectors), ordered by degree of orthogonality. The theme here is quality, not quantity. The spectra can be rejected if its error spectrum is excessive or indicative of blunder.

3. Normalization

For any comparison of items, they need to have a common reference point. For vector quantities, the distance from the origin [13] or the length of the vector/norm provides a reference. Vector comparison requires some type of normalization, usually unit normalization. Unit normalization is achieved by dividing every component of the vector by its norm, *i.e.* the length of the vector. The length of the data vector is initially set by the measurement engine, which quantifies the magnitude of each component comprising its fine structure. If several measurement vectors are collected of the same material of different amounts that causes the amplitudes of the components to change proportionately; but the component pattern of the variation is unchanged. For the case where the measurement vectors are for different species of varying amounts, the relative amplitudes of the measurement vectors are unique and the component amplitudes are not identically related.

To form a basis for comparison of this diverse set of measurement vectors, the usual approach is to unit normalize [14] all the vectors to be compared. This approach ensures that the square difference of the measurement vectors depends on the fine structure of their components not on the individual amplitude differences. There may be other standardizations that are appropriate for particular problems, but they are case dependent.

Depending upon the mathematical model that is developed to treat vector data, frequently scalars are involved. The basis of scalar magnitude comparison is a question of scaling, *i.e.* the relative magnitudes of the scalars for model parameters.

Our interest is to analyze a collection of spectroscopic data vectors and their concomitant error vectors using a mathematical model. D = BF where D is the collection of binned spectroscopic measurements taken at regular intervals over a specific bandpass. B are the spectroscopic basis vectors that characterize the independent factors that signify the fine structure of the collection, D. F are the scalars that quantify the amount of the factors present in an individual data vector.

Use unit normalization of the database in the name of consistency for it removes the individual norm of the data vectors, leaving only the fine structure of the vectors to quantify differences. Normalization is also required for comparison of angles and distances.

After the spectra of the database have been prepared, the individual spectra

can be compared for unique spectral features. Now, the selection process can be initiated.

4. Ordering and Selection

The starting database needs to be examined to decide how many spectroscopic features/factors are expressed in this particular collection of protein spectra. Then the selection process should be applied to reduce the number of spectra to correspond to the number of primary factors. Choosing the number of spectra equal to the number of factors yields a unique fully determined mathematical solution, and sets the initialization of the algorithm.

To get a fully-determined, unique solution to D = BF, when taking D from the database, the number of CD spectra used must be equal to the number of primary structures to be determined. The question then arises: Can a technique be found that filters the CD spectra from the database to find those that are most useful? The answer is yes; the spectra will be those with the maximum information, *i.e.* those with the greatest spectral variability. At least three ways of selecting these spectra are apparent:

1) Take the inner-product (vector dot product) of all unit normalized spectra, pairwise over the database. This yields a symmetric square matrix with each element an average "cosine" between the two "vectors". Sum the matrix elements by row (protein) to find the aggregate smallest value. This identifies the most diverse protein spectrum in the collection. Remove this protein, "number one" from further consideration by removing the corresponding elements of the inner-product matrix. Repeat this aggregate sum for the next smallest sum to indicate the next most diverse protein spectrum, "number two". Continue the process until all spectra are ordered to the last two proteins. (Indeed, this is ordering independent vectors for orthogonalization).

2) Find the square difference calculation [13] (distance) for all pairs of unit normalized protein spectra to generate the symmetric square matrix. Again, use a process that selects the protein with the most aggregate diversity (largest sum, by row, of matrix elements). After it is selected and removed from the matrix continue the ordering process similar to 1). The selection process for 1) and 2) are equivalent.

3) Find the singular values derived by using Singular Value Decomposition (SVD) for all pairwise CD spectra over the database to create a symmetric square matrix whose elements contain a summary of the variances for the two singular values. Variance is equal to ratio of the maximum singular value squared to the sum of the squared singular values. When the components are nearly identical, one singular value is much smaller. When the components are most independent, the singular values are more nearly equal. Our SVD algorithm does not order the singular values. Thus, our treatment for selecting the aggregate most diverse protein was chosen to accommodate this. Again, the protein spectra are ordered in a similar way to processes 1) and 2) above. A subroutine was pro-

grammed to compute spectral ordering using these three processes on the database (CJDB). Examples obtained using these techniques are presented in the Results.

Once the protein spectra have been ordered, the number of spectra selected from the top of the list is equal to the number of factor components for the fully determined spectral analysis. None of the spectra of the database has been modified by an averaging process.

5. Useful Factors

The analysis starts with taking a collection of candidate spectra to select, which are to be incorporated based on biochemical considerations. To reduce the number of primary spectra to equal the number of spectroscopic factors, the use of SVD is suggested.

SVD is a matrix factorization method for rectangular matrices. The factorization is expressed as:

$$D = USV^{\mathrm{T}} \tag{1}$$

where:

D = the rectangular matrix (usually the database).

U = orthogonal column matrix.

S = diagonal matrix containing the singular values, S_i.

V^{T} = orthogonal row matrix, transposed.

The singular values, S_i, characterize the database as factors, and they should have a small range of values. The variance derived from the singular values, S_i, can help to understand the capability of the analysis of this database.

$$\mathrm{Viz}: \mathrm{Variance}\left(S_i\right) = \left(s_i^2 / \sum s_i^2\right) \tag{2}$$

If the variance is a rapidly decreasing sequence [15], the database lacks resolution. Following this parameter will help in a selection process and, if the variance is a slowly decreasing sequence, this indicates the resolution is optimal. While some of the criteria for the selection process to reduce the primary spectra to comprise the restricted database are indicated, the details are case application dependent.

To begin the general case, the singular values will be such that they will be divided into two groups: primary and secondary.

The primary group includes singular values that have variance ≥ 0.01 (this is a design choice). Thus:

$$D = \underline{D} + E \tag{3}$$

$$D = \underline{D}\left(S_1\right) + \underline{D}\left(S_2\right) + \cdots + \underline{D}\left(S_j\right) + E\left(S_j + 1\right) + \cdots + E\left(S_m\right) \tag{4}$$

Database = (Basis group) + (uncertainty group).

Primary $(1, \cdots, j)$; Secondary $(j+1, \cdots, m)$.

If the matrices $\underline{D}(S_i)$ are averaged by summing its component spectra for each protein for the factor, the result is an average over the database. One spectrum

characterizing that factor would be a basis vector (function):

$$\text{Spectra}(j) = \Sigma\text{spectra}(\text{factor}, j)/m(\#\text{of proteins}) \tag{5}$$

Using the same reasoning, one could contract the uncertainty group into a characteristic noise component. (This technique may be useful in the selection process).

Further pursuit of this avenue to analysis is abandoned because it leads to just more averaging without any real improvement of resolution. Clearly one needs to reduce the size of the database to a minimum size that has optimal resolution. This is achieved by making the number of spectra equal to the number of singular values. In order to achieve the reduction in the number of primary spectra, a process of ordering them according to quantifiable criteria needs to be employed. The following is a description of the technique used in this case.

After the selection process, the application of SVD gives:

$$D = D(S_1) + D(S_2) + D(S_3) + \cdots + D(S_m) \tag{6}$$

where:

S_m = the particular singular values for the selected data.

m = the number of selected spectra = the number of factors = the number of singular values.

In short, all singular values are primary and there are no secondary singular values to consider and the selected spectra of database, D, are considered as characterizing the original ensemble of target protein spectra. Where:

$$\text{Variance}(S_m) = \left(s_m^2 / \sum s_m^2\right) \geq J \tag{7}$$

with J = the limit of resolution in the experimental data.

After selection of the characteristic minimum number of protein spectra, the development of the mathematical model incorporates the factor components into the algorithm and prepares the analysis for future target CD spectra.

There are two ways to proceed. These are:

A/ Using the component spectra for each factor, find a norm for each component spectra for a protein, thus the factor expansion is returned as fractional coefficients for the data set.

The first alternative A/ is the method of finding the fractions by norms. This expansion technique is to change the collection of information contained in the component spectra of each factor. Recall each factor is expanded into component spectra of each of the primary proteins. The norm of each component spectra (the magnitude of the component) is computed and stored. This process is repeated for each factor. Since the number of factors is equal to the number of protein spectra, the number of factors when collected is a square matrix of size, $m \times m$ (where m = the number of factors). Since D = BF and F is now defined for D, the selected spectral data, B the basis are to be obtained. What makes this approach interesting is that historically F was given by x-ray secondary structure fractions, which was the direct mixing of spectral data and geometric structural fractions; whereas now the fractions are uniquely of spectral origin.

B/ Average the component spectra for each factor making a spectrum for the factor.

The second alternative B/ would be to take the matrix for each singular value component and reduce it into a single component spectrum as an average representation for that factor. The techniques required to scale these factor basis functions are left for future research. At this point, the mathematical form of the data analysis has merged:

$$D = USV^t = D(S_1) + D(S_2) + \cdots + D(S_m) = BF\,(\text{component norms}) \tag{8}$$

where B, the basis vectors, is now computed using D and F as known.

6. Fully-Determined Solution

The CJBD protein list (see list in results) was ordered based on the greatest pairwise diversity using the inner-product, the squared difference or the singular value decomposition and was found to be: 3, 1, 13, 8, 9, 16, 6, 2, 11, 12, 10, 7, 14, 4, 5*, 15* (*either order) in all three cases. A database with this order was constructed and submitted to the program yielding an integer sequence. This study shows that the data should be analyzed for four factors and the first four proteins are the optimum protein data.

7. Initialization

Initialization is the process of setting the solution parameters of the analysis algorithm. The analysis algorithm is a program that will take a target protein spectrum as input and give as computed results the factor fractions present with uncertainties.

$$D = BF \tag{9}$$

where D is the target spectrum, B represents the basis functions characteristic of the factors contained in the limited set of selected protein spectra and F represents the fraction for each factor that is uniquely quantified by this basis.

The following quantities are given as an aid to the researchers who want to adapt their programs to increase the scope of their laboratory analysis. The singular values for the original database are given in the original publication [1]. The fundamental subroutines used and results are independent of coding format; however, double precision was used.

Note that the variance for the whole database suggests that there are at most four major factors. A selected set with four factors should be expected. After the selection, techniques are applied to this database, the ordering sequence is 3, 1, 13, 8, 9… etc. for the three methods used. If the 16th, polyglutamic acid, is removed, the order remains unchanged.

There were no spectral error functions given in the spectral database; however an empirically formulated error function (12% at 194 nm and, linearly, 9% at 224 nm) was used to tailor an error spectrum for each protein spectrum. The experimental error propagates uncertainty into the norms that control the initialization of the mathematical model. The use of measurement uncertainty in the

propagation of parameter uncertainty is described in the next section, linear least squared theory.

The scaled results for the factor expansion producing the component spectra norms are presented in Table 1.

Table 1. A summary of the scaled factor component spectra norms for the selected proteins.

Factor	1	2	3	4	Total
Protein 3	0.623	0.075	0.010	0.291	1.0
Protein 1	0.912	0.017	0.020	0.051	1.0
Protein 13	0.317	0.362	0.003	0.318	1.0
Protein 8	0.601	0.032	0.007	0.361	1.0

The scaled factor component norms for each protein are summed to 1.0. Each protein spectrum is a linear combination of the factor basis.

8. Linear Least Square Theory

When linear least square fitting (LSQ) is employed [9], it is seeking the linear coefficients of basis functions to closely replicate experimental measurements at intervals of the independent variable. These coefficients are the model parameters and the basis functions represent the mechanistic relationship between the observations and the independent variable.

The observations are measurements and obtained by devices and recorded as numbers and are the data. The measurement devices are of finite precision and reproducibility. Thus, the measurements are repeated a number of times and the averages are used. The deviations in the accumulations of the average are important and valuable information, for it is a measure of the uncertainty of the average number.

In short, the data are two components: the average value and the uncertainty. If the phenomenon being measured uses some interval or range of independent variables such as time/wavelength, then the average data is mapped as a function over the interval and the uncertainty is also mapped as a function over the same interval. The map of the component pair is a line with thickness equal to the deviation or uncertainty.

When the LSQ method is applied to the data, the data is differenced to the linear model to find coefficients that minimize the difference squared, etc. The same LSQ method is applied by combining the data and the uncertainty to make a new data function and yielding slightly different coefficients. The difference in the values of the coefficients for the two cases gives the uncertainty in the coefficients due to the uncertainty in the data.

One can use the data error components as a data function and the coefficients in this case are only the uncertainties in the coefficients for the data pairs.

Residual, $R(x_i)$ has the same domain, [9] as the experimental measurement

error, $E(x_i)$. However, it relates to the difference between measurements and the model calculation. When this function is treated by the analysis algorithm, the results are parameter uncertainties due to precision of the modeling.

Thus, the complete data analysis of a CD protein spectrum yields two sets of parameter uncertainty. One is generated using the experimental measurement uncertainty and the other is generated by the residual from the model fit to the measurement spectrum. Comparison of the pairs of uncertainties gives a quantitative way to adjudicate the success of the analysis for this spectrum.

If the model uncertainty overlaps the measurement uncertainty, the analysis is valid and credible. Use the field of residual uncertainties for all the techniques considered to decide which technique appears to be preferred for this database.

In quantifying the changes due to altering a protein structure, one needs to have a measure of the parameter variability between the beginning and ending states to be sure that a change has taken place.

9. Results

Due to the ordering and selection process used, the resulting factor fractions should sum to one or less, and decrease in order of decreasing variability, within the limits of residual uncertainty.

Because the protein spectra are selected and listed according to the amount of spectral variability, the most variable are assigned to the summation of one. It follows that any other spectra under consideration should scale to values equal to one or less, depending on the degree of variability.

In the tables of results, the proteins are listed by numbers corresponding to their chemical name [1]. These are: CJBD Protein List.

1) α-Chymotrypsin, 2) cytochrome, 3) elastase, 4) hemoglobin, 5) lactate dehydrogenase, 6) lysozyme, 7) myoglobin, 8) papain, 9) ribonuclease, 10) subtilisin BPN, 11) flavodoxin, 12) glyceraldehyde-3-phosphate dehydrogenase, 13) prealbumin, 14) subtilisin NOVO, 15) triosephosphate isomerase, 16) poly (L-glutamic acid).

9.1. Normalized Spectra Results

The normalized spectra results are seen in **Table 2**.

Table 2. Singular value decomposition results for the selected spectra 3, 1, 13, 8.

Factor	Singular Value	Variance
F_1	1.602	0.6415
F_2	1.0276	0.2639
F_3	0.15135	0.0057
F_4	0.5960	0.089

The singular value decomposition results are the singular values and the vari-

ances. The ordering of singular values is unique to the subroutine used to produce the matrix factors.

The first factor is prominent, however the fourth factor is more prominent than the third. Recall it is the norms of the factor components that control the details of the analysis, not the singular values.

After the factor component fractions are set using the norms of Table 1, and the basis vectors are computed, then the linear factor fractions are obtained for each protein spectrum.

In Table 3, the four factor fractions and total are listed by line for each protein of the collection.

Those proteins summing to 1.0 will be seen as those in Table 1. Proteins 1, 3, 8, 13 have totals equal to 1.0 and proteins 6, 12, 11, 10 have totals near 1.0, which indicates that half of the sixteen spectra, are well characterized by four factors.

Table 3. Four factor fractions and total for each protein.

Protein Number	F_1	F_2	F_3	F_4	Total
1	0.912	0.017	0.020	0.051	1.0
2	0.491	0.110	0.006	0.249	0.856
3	0.623	0.075	0.010	0.291	1.0
4	0.387	0.116	0.005	0.149	0.657
5	0.418	0.135	0.006	0.166	0.725
6	0.696	0.125	0.013	0.137	0.971
7	0.424	0.071	0.005	0.167	0.667
8	0.601	0.032	0.007	0.361	1.0
9	0.587	0.197	0.006	0.431	1.22
10	0.537	0.149	0.008	0.224	0.918
11	0.388	0.240	0.003	0.318	0.950
12	0.577	0.153	0.009	0.217	0.955
13	0.317	0.362	0.003	0.318	1.0
14	0.299	0.168	0.002	0.241	0.709
15	0.362	0.124	0.004	0.208	0.698
16	0.457	0.029	0.006	0.198	0.690

It is useful to note that there are no negative fractions in the table; thus indicating the adequacy of the basis to represent the spectroscopic variability in the collection. The results of the table are summarized by listing the factor fraction total in descending order by protein number: 3, 1, 13, 8, 9, 6, 12, 11, 10, 2, 5, 14, 15, 16, 7, 4. Compare this sequence to the ordering by variability shown on page 8 in the section "fully determined solution".

The adequacy of the model is determined by the uncertainty in the model parameters. That uncertainty is quantified using the difference between the experimental data and the model fit. Those results are shown in Table 4.

Table 4. Factor fraction variation for model-data residual errors.

Protein Number	ΔF_1	ΔF_2	ΔF_3	ΔF_4
1	0	0	0	0
2	0.058	0.068	0	0.086
3	0	0	0	0
4	0.080	0.039	0.002	0.035
5	0.061	0.002	0.002	0.017
6	0.126	0.014	0.003	0.016
7	0.118	0.002	0.003	0.038
8	0	0	0	0
9	0.124	0.062	0.002	0.058
10	0.088	0.018	0.003	0.058
11	0.085	0.014	0.002	0.021
12	0.027	0.019	0	0.010
13	0	0	0	0
14	0.043	0.016	0.001	0.001
15	0.063	0.016	0.001	0.010
16	0.177	0.020	0.006	0.118

The model-fit residual errors are the calculated error resulting from the difference between the experimental data and the model values for each protein spectrum. The values for ΔF_3 are small due to the minor importance of the third factor. There are no residual errors for protein 3, 1, 13, 8 as they are part of the exact solution for the basis.

By combining the factor fraction for a particular protein with the residual variation for that protein, a complete description of the factor fractions and this precision indicator are given *e.g.*

$$F_1 \pm \Delta F_1; \cdots; F_4 \pm \Delta F_4$$

While modern instrumentation for spectral measurements has greatly improved, it is desirable to quantify the precision of the measurements at the time of measurement. It is the variation in the primary data that controls variation in the model parameters as well as the adequacy of the model.

The uncertainty of the model parameters due to the measurement uncertainty, shown in Table 5, is useful to compare to the residual error, shown in Table 4.

Table 5. Factor fraction variation for empirical data error.

Protein Number	ΔF_1	ΔF_2	ΔF_3	ΔF_4
1	0.094	0.005	0.002	0.016
2	0.034	0.024	0 (0.00015)	0.047
3	0.067	0.009	0.001	0.024
4	0.024	0.015	0	0.042
5	0.015	0.010	0	0.034
6	0.042	0.014	0.001	0.026
7	0.018	0.017	0	0.030
8	0.058	0.021	0.001	0.003
9	0.043	0.025	0	0.053
10	0.020	0.014	0	0.033
11	0.030	0.016	0	0.050
12	0.30	0.014	0	0.035
13	0.019	0.034	0	0.069
14	0.023	0.009	0	0.034
15	0.018	0.013	0	0.035
16	0.023	0.019	0.001	0.014

The factor fraction values for the empirical data error in this table are results from just the primary data variations.

The magnitudes of the variations shown in Table 5 are small in comparison to those of Table 4, which would indicate the primary data variations (empirical, not measured) are adequate for the analysis.

9.2. Unnormalized Spectra Results

This section is presented as another test problem for those interested in the development of computer programs. To demonstrate the necessity of using normalized spectra, the un-normalized spectra were ordered by using the square difference procedure. The results, shown in Table 6, were: 16, 7, 4, 15, 3, 5, 1, 8, 13, 14, 6, 9, 2, 11, (12), (10). The norms of the database spectra are also given in Table 6 as a reference.

Table 6. Spectrum norms for the data protein spectra.

Ordered Protein	Spectrum Norm
16	84.8
7	49.4
4	39.5
15	29.1
3	8.63

Continued

5	23.9
1	7.97
8	13.6
13	9.04
14	17.4
6	18.0
9	10.8
2	14.8
11	12.9
12	14.1
10	13.7

The spectrum norms for the data proteins are computed for the published data. The ordered protein numbers are the results using the square difference procedure.

It is noted that the first four norms are about two to four times greater than those that follow in the table. The next five entries include those proteins selected using normalized spectra. These observations indicate that using the leading number of preferred spectra as representative will have low expectations of success. Hence, the selected proteins were 7, 4, 15, 3 (protein 16 was omitted).

The singular value decomposition results, shown in Table 7, are the singular values of all primary data and the variances. The ordering of the singular values is unique to the subroutines used to produce the matrix factors.

Table 7. Singular decomposition results for the selected protein spectra 7, 4, 15, 3.

Factor	Singular Value	Variance
F_1	69.166	0.972
F_2	9.649	0.019
F_3	6.258	0.008
F_4	2.161	0.0017

The variance of the first singular values is 0.97 indicating that the resolutions by factor analysis will be very limited and may produce distortion and instability.

In Table 8, the four factor fractions and total listed are by line for each protein of the collection. The proteins summing to 1.0 show the initializing component norms.

Table 8. Four factor fractions and total for each protein.

Protein Number	F_1	F_2	F_3	F_4	Total
1	0.016	0.424	0.385	0.000	0.825
2	0.129	0.247	0.284	−0.005	0.654
3	0.030	0.515	0.450	0.005	1.0
4	0.714	0.096	0.182	0.008	1.0
5	0.906	0.104	0.041	0.038	1.090
6	0.442	0.448	0.408	0.015	1.313
7	0.775	0.086	0.134	0.004	1.0
8	0.037	0.416	0.387	−0.005	0.836
9	0.149	0.384	0.375	0.002	0.910
10	0.588	0.212	0.139	0.028	0.967
11	0.701	0.220	0.150	0.036	1.107
12	0.528	0.287	0.232	0.023	1.071
13	0.806	0.206	0.090	0.049	1.151
14	0.760	0.051	0.003	0.034	0.849
15	0.876	0.059	0.034	0.030	1.0
16	0.761	0.334	0.367	−0.024	1.438

The general results are not bounded because of the lack of standardization of the spectra before comparison. Note there are three negative values of the factor fractions for proteins 2, 8, 16, showing the failure of the analysis.

10. Summary and Conclusion

This research is significant because it presents in detail the techniques required to describe the factors present in a spectral database using only the original spectra. The primary objective of the research was to clarify the spectroscopic factors present in a CD database of spectra without importing data from other experimental domains e.g. x-ray crystallography. This begins by scanning the database to find the number of primary spectral factors that are represented. A reference for the comparison of the spectra is established by unit normalization. For a spectrum to contain the signature of many factors, there should be a high degree of variability in the spectrum. Here the spectra are ordered in degree of variability. To facilitate a unique solution, the number of selected spectra is equal to the number of primary factors represented.

The factor component expansion of this selected set of spectra is compacted into their norms and scaled into factor fractions. The factor fractions with the selected spectra are used to calculate the concomitant basis functions to complete the model. Applying the least squares algorithm of the model to each entry of the database obtains its factor fractions and collects the residual variations between this model fit and the spectral data. These results, Table 3 and Table 4, are summarized as the factor fraction and their accompanying fraction variation

due to residual variation for the database.

The adequacy of the mathematical model and its initialization is decided by the degree of congruency between the model-generated spectrum and the experimental spectrum. When the congruency (residual) is quantified and analyzed by the algorithm, a measure of the model validity is quantified.

Due to the ordering of the spectra by degrees of variability and the scaling process, the sum of the factor fractions is expected to be equal to 1.0 or less. There are no negative factor fractions. The results of Table 3 and Table 4 show this to be true to within the sum of residual variations for ten proteins, except 4, 5, 7, 14, 15, 16* (*polypeptide); these are at the closing of the ordering list and have the least variability.

A secondary benefit of the analysis is that it can also be used to calculate compression of database storage, using the norms of the original database and the selected basis functions derived from the factor fraction component norms for all the proteins. In the case detailed in this paper, the compression of storage for the primary database would be 62%.

This presentation has been developed exclusively within the spectroscopic domain and the basis functions have some relationship to the spectroscopic signatures of the active chromophores. The description of those relationships in any spectroscopic domain is complex and will be left to future research.

Acknowledgements

The author wishes to acknowledge the following persons for their contribution to the continuity of the project: Michelle A. Haner and Stephanie Pastor for support typing and formatting the manuscript; Brent Norum for repairing the computer hardware; and Patrick W. Mobley for his criticism and many thoughtful suggestions.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Compton, L.A. and Johnson, W.C. (1986) Analysis of Protein Circular Dichroism Spectra for Secondary Structures Using a Simplified Matrix Multiplication. *Analytical Biochemistry*, **155**, 155-167. https://doi.org/10.1016/0003-2697(86)90241-1

[2] Mulkerrin, M.G. (1996) Protein Structure Analysis Using Circular Dichroism. In: Henry, A.H., Ed., *Spectroscopic Methods for Determining Protein Structures in Solution*, VCH Publishers, New York, 5-27.

[3] Kliger, D.S., Louis, J.W. and Randall, C.E. (1990) Polarized Light in Optics and Spectroscopy. Academic Press, San Diego, 237-274. https://doi.org/10.1016/B978-0-08-057104-1.50011-9

[4] Hennessey, J.P. and Johnson, W.C. (1981) Information Content in the Circular Dichroism of Proteins. *Biochemistry*, **20**, 1085-1094. https://doi.org/10.1021/bi00508a007

[5] Hennessey, J.P. and Johnson, W.C. (1982) Experimental Error in Circular Dichroism in Proteins. *Analytical Biochemistry*, **125**, 177-178. https://doi.org/10.1016/0003-2697(82)90400-6

[6] Haner, D.A. and Mobley, P.W. (2018) Error Analysis for Protein Conformation Quantities in Circular Dichroism Spectrum. *Open Access Library Journal*, **5**, e4966.

[7] Malinowski, E.R. (1991) Factor Analysis in Chemistry. 2nd Edition, Wiley Interscience Publications, John Wiley & Sons, New York, 111.

[8] Press, W.H., Flannery, P.P., Teukolsky, S.A., Vetterling, W.T., *et al*. (1986) Numerical Recipes, the Art of Scientific Computing. Cambridge University Press, New York, 489.

[9] Hildebrand, F.B. (1956) Introduction to Numerical Analysis. McGraw-Hill, New York, 258.

[10] Oberg, K.S., Ruysschaert, J.M. and Goormaghtigh, E. (2003) Rationally Selected Basis Proteins: A New Approach to Selecting Protein Spectroscopic Secondary Structure Analysis. *Protein Science*, **12**, 2015-2031. https://doi.org/10.1110/ps.0354703

[11] Oberg, K.S., Ruysschaert, J.M. and Goormaghtigh, E. (2004) The Optimization of Proteins Secondary Structure Determination with Infrared and Dichroism Spectra. *European Journal of Biochemistry*, **271**, 2937-2948. https://doi.org/10.1111/j.1432-1033.2004.04220.x

[12] Khrapunov, S. (2009) CD Spectroscopy Has Intrinsic Limits for Protein Secondary Structures. *Analytical Biochemistry*, **389**, 174-176. https://doi.org/10.1016/j.ab.2009.03.036

[13] Anton, H. (1973) Elementary Linear Algebra. 4th Edition, John Wiley & Son, New York, 138.

[14] Rozett, R.W. and Peterson, E.M. (1975) Methods of Factor Analysis of Mass Spectra. *Analytical Chemistry*, **47**, 1301-1308. https://doi.org/10.1021/ac60358a032

[15] Haner, D.A. and Mobley, P.W. (2015) Simulations Relating to the Determination of Protein Secondary Structure Fraction from Circular Dichroism Spectra. *Open Access Library Journal*, **2**, e1601. https://doi.org/10.4236/oalib.1101601