Bull. Inst. Chem., Academia Sinica
Number 44, 45-59, March 1997

An Analysis of SELCON - A Self-consistent Method That Estimates Protein Secondary Structure from Circular Dichroism Data

Wei-Chen Lin and Lou-sing Kan*

Institute of Chemistry, Academia Sinica, Taipei, Taiwan 115, R.O.C.

A self-consistent method named SELCON that has been written by Sreerama and Woody (Anal. Biochem., **209**, 32-44 (1993)) in FORTRAN is made for the determination of the secondary structure of proteins by circular dichroism (CD) spectral data . Studies showed that SELCON is the best program for analyzing the secondary structure of proteins among its class. This paper presents a detailed analysis of it. First, the best range of the multiplication factor (*f*) in the calculation is found between 0.95 and 1.05. Second, all three data bases (Kabsch/Sander [Biopolymers, **1983**, *22*, 2577] [KS], Levitt/Greer [J. Mol. Biol., **1977**, *114*, 181] [LG], and Hennessey/Johnson [Biochemistry, **1981**, *20*, 1085] [HJ]) used in the program are good for correlating the estimated and found structures in globular proteins. Third, the estimation of the percentages of β-sheets is slight off for all three data bases. Fourth, there is a tendency to overestimate for both the α-helix and β-sheet of a protein Finally, the format of input of the program was revised to be user friendly. A demonstration is collected as an Appendix.

INTRODUCTION

Pauling and Corey discovered the α-helix and β-sheets as components of protein structure in 1951.[1] In the same year, Sanger and Tuppy published the amino acid sequence of insulin with 51 residues.[2] The merits of these two reports awakened studies of primary and secondary structures of proteins. One protein sequence after another has been presented in conferences on proteins since then. In consequence the physical properties of proteins in solution were characterized by intrinsic viscosity, light scattering, osmotic pressure, flow birefringence, sedimentation velocity and equilibrium, hydrogen-deuterium exchange, infrared and ultraviolet spectroscopy, optical rotatory dispersion, circular dichroism, and, nuclear magnetic resonance. Today, many of them have fallen into disuse. However, circular dichroism (CD), which can monitor the conformation of biopolymers, remains the most sensitive and convenient way to study proteins. As attractive as CD measurements are, however, it is worth mentioning that, unlike two other powerful techniques, X-ray diffraction of protein crystals and NMR for protein solutions, CD cannot determine the three-dimensional structure of a protein.

CD is defined as the difference between the absorption of left circularly polarized light and right circularly polarized light.[3]

$$\Delta A = A_l - A_r = (\varepsilon_l - \varepsilon_r)Cl = \Delta\varepsilon Cl \quad (1)$$

where C is the concentration of proteins in mole/l, and l is the distance of medium for light passing through, namely, path length of cuvette in cm. Thus, ΔA is the measured CD absorbance, and $\Delta\varepsilon$ is the difference of left and right circular light extinction coefficients at a wavelength.

The theory of the chromophores in an asymmetrical environment that generate the CD phenomenon has been presented elsewhere and will not be discussed in this paper.[3] However, the highly ordered structures (α-helix, β-sheet, turns, and unordered form) of proteins and peptides are extremely sensitive to CD spectra. The CD spectrum of a protein can be treated as a summation its component secondary structures with respect to wavelength.

$$\Delta A(\lambda) = \Sigma_i f_i c_i(\lambda) \quad (2)$$

Where f_i is the fraction ($\Sigma_i f_i = 1$) of secondary structures of α-helix, β-sheet, turns, unordered form, etc. In theory, the analyses of CD absorbance $c_i(\lambda)$ with respect to wavelength λ in Eq. (2) can determine percentages of α-helix, β-sheet, turns, and unordered form of a protein in (aqueous) solution. Amide bonds have optical transitions in the ultraviolet region below 260 nm. Thus, the recommended range to collect data is between 260 and 180 nm. Due to the difficulty of obtaining data at wavelengths lower than 200 nm, it is recommended that the Eq. (2) can be solved between 240 and 200 nm.

In order to build up the database of reference spectra in Eq. (2), proteins are further presumed (i) to have the same structure in the crystalline state and in aqueous solution; (ii) the result CD spectrum is considered a linear function of the component's secondary structure. The CD spectrum ($c_i(\lambda)$) of the ith secondary structure of proteins can be obtained while $f_i = 1$. Fig. 1 shows the so-called reference CD spectra of α-helix, β-sheet, turns, and unordered forms from proteins with known structures determined by X-ray diffraction studies.[4-11] All α-helices have two negative bands and one positive band at 222, 208, and 191-193 nm (Fig. 1).[12] All β-sheets have a single negative band in the region of 210 to 225 nm, and a positive band between 190 and 200 nm (Fig. 1).[8] However, the amplitudes of these bands are smaller than those of α-helix. All turns have a negative band at 190 nm and a positive one around 208 nm (Fig. 1). The unordered structure shows a strong negative CD band near 200 nm, and some weak bands between 220 and 230 nm, which can have either positive or negative signs (Fig. 1).[8] Then the f_i of unknown protein can in turn be calculated as a combination of spectra for weighted secondary structure. It is the ultimate goal for a CD measurement of proteins. The obvious major obstacle is that the above assumptions (i) and (ii) are rarely entirely satisfied by any existing protein. In addition, the amide bond is not the only contributor of CD absorbance in proteins. For this reason, the absorbance in Eq. (2) is not entirely contributed by the secondary structures of protein. Therefore, the convergence of Eq. (2) is a challenge task as long as the existing of CD studies on proteins.

The simplest method was to fit the data to-be-analyzed to the spectra of references by the

least squares method. The condition is $\Sigma f_i=1$.[4] In order to satisfy this condition, the CD spectra were deconvoluted into more and more reference structures (i.e., parallel and antiparallel β-sheets, etc. in addition to original α-helix, β-sheet, etc.) as structures of more and more proteins have been determined by the X-ray diffraction method. A computer program, LINCOMB, was written for this method.[13]

The nonconstrained least-squares fits used a multilinear regression method. It gives a reasonable estimate of α-helical content, but poorly on β-turns using the spectra of the polypeptide models suggested by Brahms and Brahms as reference.[5]

The singular value decomposition (SVD) is an eigenvalue method of multicomponents analysis which is superior to least squares. After deconvolution, each basis curve of the protein CD data is related to reference of secondary structures. The basis spectra are then used to analyze the conformation of unknown proteins.[14] In SVD the Σf_i doesn't have to equal one. When a set of 17 proteins were analyzed by SVD, the fits improved for α-helix compared to the nonconstrained fit, remained unchanged for total β-sheet, and were poorer for the turns.[14]

Other analytical methods for the secondary structure of proteins based on CD spectra have been developed and reported.[15,16] Those wanting a summary may read a recent review written by Greenfield.[17]

The self-consistent method (SELCON) was developed by Sreerama and Woody by making modifications of the variable selection methods.[18] In SELCON the spectrum of the protein to be analyzed is included in the basis set and an initial guess is made for the unknown structure as a first approximation. The spectra that are least like the spectrum of interest are systematically de-

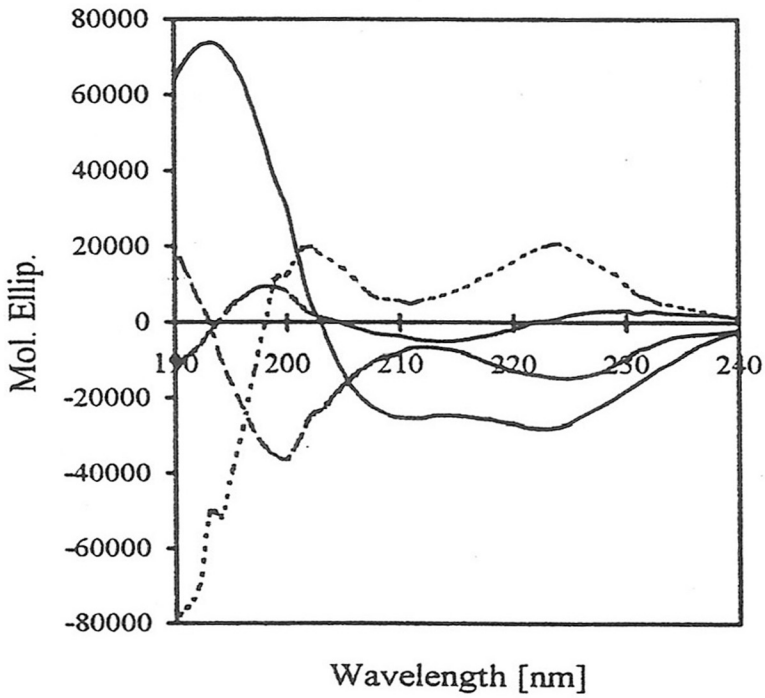

Fig. 1. The CD spectra of α-helix (solid line), β-sheet (broken line), turns (dot line) and unordered structures (broken-dot line) of proteins.

leted to increase the speed of finding the best solutions. The resulting matrix equation is solved using the SVD algorithm, and the initial guess is replaced by the solution. The process is repeated until self-consistency is attained. The best features of the variable selection and the locally linearized methods are incorporated in this procedure. This method has been evaluated as the best among its class by Greenfield and by our laboratory.[17,19] In addition, Sreerama and Woody considerately provide a FORTRAN program of SELCON that is distributed without charge. Thus, it is very accessible to interested parties. However, the weakness, if any, of the SELCON program may be the laborious input procedure. It would be more efficient to take the CD spectrum directly from a data file. Furthermore, SELCON doesn't specify the criteria of environment and methods of calculations.[18] These inspired us to analyze the program and rewrite the input portion to make it easier to use.

METHODS

Sreerama and Woody adopted 17 proteins with known structures as their data base. The CD spectral data of these proteins shown in Table 1 were provided by Dr. W. C. Johnson, Jr.[20] The abbreviations used are CHYT: chymotrypsin, CYTC: cytochrome C, ELAS: elastase, HBN: hemoglobin, LDH: lactate dehydrogenase, LYSM: (eggwhite) lysozyme, MGLB: myoglobin, PARN: papain, SUBB: subtilisin BPN, FLVD: flavodoxin, GPD: glyceraldehyde-3-phosphate dehydrogenase, PRAL: prealbumin, TPI: triosephospate isomerase, THML: thermolysin, BNJN: Bence-Jones protein, RUBR: rubredoxin, and PGLU: poly(L-glutamic acid) in Table 1. The range of wavelength is from 260 to 178 nm. These reference proteins can be summarized in the following categories. (i) all α-helix: PGLU, CYTC, HBN, and MGLB; (ii) β-proteins: BNJN, PRAL, RUBR, CHYT, and ELAS; (iii) α + β proteins: PARN, THML, and LYSM; (iv) α/β proteins: SUBB, GPD, FLVD, LDH, and TPI. There are three X-ray data sets used in the calculation for the above 17 proteins. The Levitt and Greer (LG),[21] Kabsch and Sander (KS),[22] and Hennessey and Johnson (HJ)[14] methods are based on C_α coordinates; hydrogen bond patterns, and visual inspection of the structure, respectively. The percentages of secondary structures are different among these three methods as shown in Table 2. The calculation is done similar to a SVD. The CD of a measured protein (P) is compared to all 17 references first under one (i.e., LG) of the three data sets shown in Table 2. Then the 17 proteins are arranged in an order of the magnitudes of the root mean square (RMS) deviations with the P. The unmatched reference proteins are eliminated to save computing time. The next step is to include the CD data of P into the first row of the reference matrix to obtain a new matrix. The proteins in the new matrix are then arranged in increasing order of the RMS deviations from P. Then an initial guess of the secondary structure of P is made. The new matrix is then reconstructed. Thus, these procedures are repeated to a self-consistent value that was set at the beginning of the calculation.[18]

RESULTS AND DISCUSSION

A. The determination of the modification factor *f*: CD measurements should be made on samples with a maximum absorbance. Therefore, a possible error may be introduced by the de-

Table 1. CD Data From 260 to 178 nm of 17 Proteins. The Details are in the Text.

Wave length (nm)	Proteins: CHYT	CYTC	ELAS	HBN	LDH	LYSM	MGLB	PAPN	SUBB	FLVD	GPD	PRAL	TPI	THML	BNJN	RUBR	PGLU
	CD degree																
260	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
259	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
258	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
257	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
256	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
255	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
254	0	0	0	0	0	-0.01	0	0	0	0	0	0	0	0	0	0	0
253	0	0	0	0	0	-0.02	0	0	0	0	0	0	0	0	0	0	0
252	0	0	0	0	0	-0.04	0	0	0	0	0	0	0	0	0	0	0
251	0	0	0	0	0	-0.05	-0.01	0	0	0	0	0	0	0	0	0	0
250	0	0	0	0	0	-0.07	-0.01	-0.01	0	0	0	0	0	0	0	0	0
249	0	0	0	0	0	-0.09	-0.01	-0.02	-0.01	-0.02	-0.01	0	0	0	0	0	0
248	0	0	-0.01	0	0	-0.11	-0.01	-0.03	-0.02	-0.06	-0.04	0	0	0	0	0	0
247	-0.01	0	-0.03	-0.02	-0.03	-0.12	-0.04	-0.04	-0.02	-0.11	-0.07	0	-0.03	-0.04	0.01	0	0
246	-0.02	0	-0.05	-0.05	-0.07	-0.14	-0.09	-0.06	-0.04	-0.17	-0.11	0	-0.08	-0.1	0	0	0
245	-0.03	0	-0.07	-0.1	-0.12	-0.17	-0.15	-0.09	-0.08	-0.23	-0.15	0	-0.14	-0.15	-0.04	0.01	0
244	-0.05	-0.02	-0.09	-0.17	-0.18	-0.2	-0.22	-0.12	-0.13	-0.28	-0.2	0	-0.2	-0.2	-0.1	0	0
243	-0.07	-0.06	-0.11	-0.26	-0.25	-0.24	-0.31	-0.16	-0.17	-0.34	-0.26	0	-0.28	-0.25	-0.16	-0.04	-0.05
242	-0.1	-0.11	-0.13	-0.38	-0.33	-0.3	-0.42	-0.22	-0.2	-0.39	-0.33	0	-0.37	-0.3	-0.2	-0.1	-0.16
241	-0.13	-0.17	-0.15	-0.52	-0.43	-0.37	-0.57	-0.29	-0.24	-0.44	-0.41	0	-0.48	-0.35	-0.21	-0.14	-0.34
240	-0.16	-0.23	-0.17	-0.7	-0.54	-0.45	-0.77	-0.36	-0.3	-0.5	-0.5	0	-0.6	-0.4	-0.2	-0.2	-0.59
239	-0.2	-0.31	-0.18	-0.92	-0.67	-0.55	-1.01	-0.44	-0.38	-0.56	-0.6	-0.02	-0.73	-0.43	-0.2	-0.29	-0.89
238	-0.24	-0.4	-0.19	-1.19	-0.82	-0.68	-1.3	-0.53	-0.47	-0.63	-0.7	-0.04	-0.9	-0.5	-0.2	-0.4	-1.27
237	-0.3	-0.5	-0.2	-1.51	-0.99	-0.87	-1.66	-0.64	-0.57	-0.71	-0.82	-0.06	-1.11	-0.63	-0.17	-0.49	-1.75
236	-0.37	-0.64	-0.2	-1.89	-1.19	-1.08	-2.08	-0.76	-0.67	-0.8	-0.95	-0.09	-1.35	-0.8	-0.1	-0.6	-2.31
235	-0.46	-0.82	-0.2	-2.34	-1.41	-1.3	-2.56	-0.9	-0.79	-0.9	-1.09	-0.12	-1.61	-0.98	0	-0.74	-2.94
234	-0.58	-1.04	-0.21	-2.83	-1.65	-1.53	-3.09	-1.05	-0.92	-1.02	-1.25	-0.17	-1.88	-1.2	0.1	-0.9	-3.62
233	-0.76	-1.28	-0.22	-3.34	-1.92	-1.78	-3.66	-1.23	-1.06	-1.15	-1.42	-0.23	-2.18	-1.49	0.21	-1.04	-4.34
232	-0.96	-1.54	-0.23	-3.88	-2.2	-2.01	-4.27	-1.44	-1.21	-1.3	-1.6	-0.3	-2.5	-1.8	0.3	-1.2	-5.1
231	-1.16	-1.84	-0.25	-4.44	-2.49	-2.21	-4.92	-1.67	-1.36	-1.46	-1.79	-0.38	-2.84	-2.06	0.36	-1.39	-5.9
230	-1.31	-2.17	-0.27	-5	-2.79	-2.37	-5.59	-1.91	-1.52	-1.63	-2	-0.47	-3.2	-2.3	0.4	-1.6	-6.73
229	-1.39	-2.5	-0.3	-5.51	-3.11	-2.52	-6.23	-2.15	-1.69	-1.81	-2.21	-0.58	-3.55	-2.55	0.42	-1.81	-7.57
228	-1.39	-2.8	-0.34	-5.98	-3.43	-2.65	-6.79	-2.36	-1.86	-2	-2.42	-0.7	-3.88	-2.8	0.4	-2	-8.45
227	-1.34	-3.05	-0.39	-6.41	-3.71	-2.76	-7.25	-2.54	-2.03	-2.19	-2.61	-0.82	-4.16	-3.01	0.32	-2.13	-9.37
226	-1.28	-3.26	-0.44	-6.78	-3.93	-2.86	-7.62	-2.69	-2.19	-2.38	-2.78	-0.94	-4.4	-3.2	0.2	-2.2	-10.28
225	-1.24	-3.43	-0.5	-7.09	-4.1	-2.95	-7.91	-2.8	-2.32	-2.55	-2.92	-1.06	-4.61	-3.37	0.05	-2.22	-11.12
224	-1.23	-3.57	-0.57	-7.34	-4.22	-3.04	-8.13	-2.88	-2.43	-2.7	-3.03	-1.18	-4.8	-3.5	-0.1	-2.2	-11.82
223	-1.22	-3.67	-0.64	-7.51	-4.29	-3.12	-8.26	-2.91	-2.52	-2.82	-3.11	-1.3	-4.96	-3.57	-0.21	-2.16	-12.32
222	-1.23	-3.73	-0.72	-7.6	-4.32	-3.19	-8.3	-2.92	-2.6	-2.92	-3.16	-1.41	-5.08	-3.6	-0.3	-2.1	-12.56
221	-1.25	-3.73	-0.8	-7.6	-4.31	-3.25	-8.25	-2.91	-2.66	-2.98	-3.19	-1.52	-5.16	-3.64	-0.4	-2.01	-12.5
220	-1.28	-3.7	-0.89	-7.54	-4.28	-3.31	-8.13	-2.9	-2.7	-3	-3.2	-1.62	-5.2	-3.7	-0.5	-1.9	-12.26
219	-1.32	-3.65	-0.98	-7.47	-4.22	-3.36	-7.98	-2.88	-2.72	-2.96	-3.17	-1.71	-5.22	-3.76	-0.56	-1.76	-11.97
218	-1.36	-3.57	-1.07	-7.37	-4.15	-3.4	-7.83	-2.85	-2.72	-2.9	-3.12	-1.77	-5.2	-3.8	-0.6	-1.6	-11.7
217	-1.41	-3.47	-1.17	-7.25	-4.06	-3.44	-7.7	-2.83	-2.72	-2.84	-3.06	-1.8	-5.12	-3.8	-0.62	-1.44	-11.5
216	-1.48	-3.35	-1.27	-7.11	-3.97	-3.47	-7.58	-2.81	-2.7	-2.8	-3	-1.8	-5	-3.8	-0.6	-1.3	-11.35
215	-1.56	-3.23	-1.37	-6.98	-3.87	-3.51	-7.45	-2.8	-2.66	-2.8	-2.95	-1.76	-4.85	-3.84	-0.54	-1.18	-11.23
214	-1.65	-3.13	-1.48	-6.89	-3.78	-3.57	-7.35	-2.82	-2.61	-2.8	-2.9	-1.7	-4.7	-3.9	-0.45	-1.1	-11.15
213	-1.75	-3.04	-1.59	-6.86	-3.71	-3.66	-7.31	-2.87	-2.56	-2.77	-2.86	-1.65	-4.58	-3.96	-0.33	-1.07	-11.13
212	-1.86	-2.98	-1.7	-6.88	-3.67	-3.77	-7.35	-2.95	-2.51	-2.7	-2.84	-1.6	-4.5	-4	-0.2	-1.1	-11.18
211	-1.98	-2.94	-1.81	-6.93	-3.64	-3.91	-7.46	-3.04	-2.45	-2.6	-2.83	-1.55	-4.46	-4	-0.06	-1.21	-11.29

210	-2.11	-2.91	-1.92	-6.97	-3.6	-4.09	-7.57	-3.12	-2.39	-2.47	-2.8	-1.5	-4.4	-3.9	0.1	-1.4	-11.45
209	-2.24	-2.86	-2.03	-6.92	-3.53	-4.31	-7.59	-3.18	-2.31	-2.31	-2.71	-1.43	-4.26	-3.62	0.3	-1.68	-11.63
208	-2.38	-2.79	-2.13	-6.67	-3.36	-4.46	-7.43	-3.17	-2.21	-2.13	-2.58	-1.33	-4	-3.2	0.5	-2	-11.87
207	-2.52	-2.66	-2.22	-6.11	-3.05	-4.45	-7.02	-3.06	-2.08	-1.91	-2.41	-1.19	-3.6	-2.68	0.7	-2.34	-12.1
206	-2.65	-2.4	-2.3	-5.26	-2.6	-4.3	-6.26	-2.84	-1.9	-1.6	-2.2	-1	-3	-2.1	0.9	-2.7	-11.93
205	-2.77	-1.94	-2.37	-4.12	-2.04	-4.04	-5.11	-2.55	-1.68	-1.17	-1.95	-0.77	-2.19	-1.45	1.12	-3.12	-11
204	-2.86	-1.35	-2.43	-2.6	-1.4	-3.63	-3.67	-2.28	-1.41	-0.6	-1.65	-0.48	-1.2	-0.7	1.3	-3.6	-9.49
203	-2.9	-0.7	-2.5	-0.63	-0.68	-3.03	-2.04	-2.09	-1.11	0.08	-1.28	-0.12	-0.1	0.22	1.38	-4.14	-7.65
202	-2.9	0.02	-2.6	1.58	0.22	-2.24	-0.09	-1.99	-0.78	0.8	-0.8	0.3	1.1	1.4	1.4	-4.6	-5.5
201	-2.86	0.84	-2.74	3.85	1.38	-1.29	2.32	-1.96	-0.42	1.48	-0.19	0.78	2.4	2.87	1.38	-4.86	-3.01
200	-2.76	1.72	-2.91	6.28	2.68	-0.22	5.03	-1.94	0.07	2.1	0.5	1.3	3.75	4.3	1.3	-4.9	-0.25
199	-2.58	2.59	-3.08	8.9	4	0.91	7.84	-1.88	0.77	2.68	1.19	1.82	5.14	5.36	1.13	-4.74	2.75
198	-2.33	3.37	-3.15	11.31	5.32	2.03	10.6	-1.7	1.61	3.2	1.9	2.3	6.6	6.2	0.9	-4.4	6.25
197	-2.02	3.99	-3.06	13.07	6.63	3.06	13.13	-1.36	2.49	3.66	2.64	2.67	8.12	7.02	0.65	-3.92	10.41
196	-1.65	4.37	-2.85	14.18	7.74	3.92	15.21	-0.85	3.31	4	3.3	2.9	9.4	7.8	0.4	-3.3	14.75
195	-1.22	4.48	-2.57	14.7	8.46	4.57	16.66	-0.19	4	4.18	3.78	2.97	10.17	8.42	0.16	-2.56	18.76
194	-0.75	4.37	-2.23	14.7	8.91	4.98	17.61	0.6	4.54	4.2	4.1	2.9	10.5	8.7	-0.1	-1.8	22.5
193	-0.28	4.13	-1.85	14.22	9.2	5.17	18.19	1.49	4.89	4.07	4.32	2.73	10.54	8.5	-0.39	-1.15	25.96
192	0.16	3.84	-1.45	13.37	9.27	5.2	18.39	2.37	5.07	3.8	4.4	2.5	10.3	8	-0.7	-0.6	28.42
191	0.55	3.59	-1.06	12.25	9.02	5.12	18.14	3.14	5.06	3.43	4.31	2.23	9.78	7.39	-1.01	-0.15	29.29
190	0.88	3.36	-0.66	11	8.45	4.94	17.46	3.73	4.85	3.02	4.1	1.9	9.1	6.7	-1.3	0.2	29.2
189	1.12	3.14	-0.25	9.75	7.6	4.65	16.41	4.08	4.42	2.61	3.8	1.49	8.37	5.92	-1.54	0.44	28.88
188	1.26	2.95	0.12	8.52	6.63	4.26	15.17	4.22	3.89	2.21	3.4	1	7.6	5.1	-1.75	0.6	28.09
187	1.3	2.8	0.37	7.35	5.69	3.81	13.91	4.2	3.36	1.81	2.88	0.45	6.8	4.3	-1.97	0.68	26.52
186	1.26	2.67	0.52	6.34	4.8	3.32	12.65	4.06	2.84	1.4	2.35	-0.1	5.97	3.6	-2.2	0.7	24.39
185	1.16	2.55	0.58	5.58	3.95	2.82	11.4	3.82	2.32	0.97	1.89	-0.61	5.13	3.07	-2.41	0.64	22.06
184	1.01	2.44	0.55	5	3.15	2.32	10.16	3.51	1.79	0.55	1.5	-1.1	4.3	2.6	-2.6	0.5	19.74
183	0.82	2.33	0.44	4.52	2.4	1.81	8.97	3.17	1.27	0.15	1.14	-1.59	3.51	2.09	-2.78	0.27	17.58
182	0.59	2.22	0.26	4.1	1.71	1.33	7.9	2.82	0.79	-0.22	0.8	-2.05	2.8	1.6	-2.95	0	15.53
181	0.32	2.1	0.01	3.73	1.11	0.89	7.01	2.47	0.38	-0.54	0.49	-2.44	2.23	1.19	-3.12	-0.27	13.53
180	0.05	1.83	-0.29	3.4	0.58	0.49	6.25	2.12	0.03	-0.8	0.2	-2.76	1.8	0.8	-3.3	-0.5	11.84
179	-0.19	1.34	-0.61	3.09	0.12	0.13	5.55	1.77	-0.27	-1	-0.08	-3.04	1.48	0.37	-3.5	-0.67	10.65
178	-0.4	0.83	-0.95	2.82	-0.3	-0.18	4.88	1.43	-0.55	-1.17	-0.34	-3.27	1.16	-0.1	-3.7	-0.8	9.75

termination of concentration. In SELCON, the CD spectra were multiplied by a factor f to compensate for experimental error. The best range of f values are between 0.9 and 1.05 as shown in Fig. 2. Fig. 2 shows the structure analytical results of CHYT as a function of f (0.9 - 1.5). The predicted percentages of α-helix and β-sheet are nearly constant when f is in the range of 0.95 and 1.05, but increase and decrease, respectively, quickly as $f > 1.05$. The turns and unordered structures seemed invariant in the whole f range studies. These results can be applied to all their data bases (data not shown).

B. The selection of data base: We used CHYT, one of the 17 reference proteins, to verify the data base (LG, HJ, and KS) of SELCON. The results are shown in Table 3 (as well as in Fig. 2). One can easily spot that the predicted secondary structures of CHYT are different depending on whether the LG, KS, or HJ data base was used. KS has the best prediction on turns but the worst on β-sheet (8% off). The α-helix and unordered are slightly more (3 and 5 %, respectively) in predicted than in X-ray diffraction data. Similarly, the predictions from HJ and LG also have the large discrepancies in β-sheet. The LG method produces the highest predicted β-sheet among the three. Similar tests were performed on other proteins in the reference list (data not shown). The

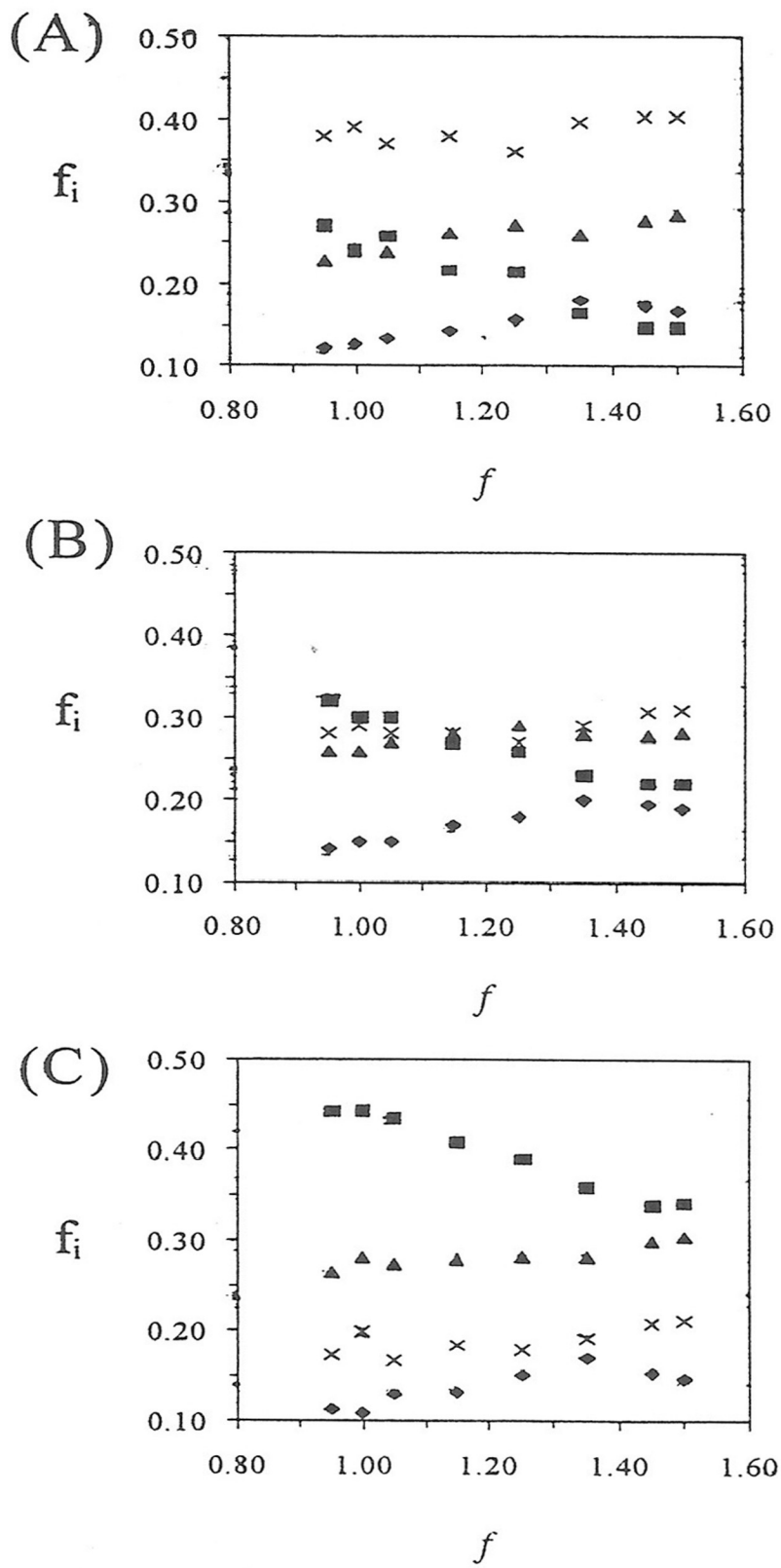

Fig. 2. The plots of percentages (f_i) of α-helix (diamond), β-sheet (square), turn (triangle), and unordered (cross) structures of CHYT as function of f values. The data bases were used (a) HJ, (b) KS, and (c) LG.

Table 2. The Percentages of Secondary Structures of 17 Proteins Under Reference Data Sets, HJ, LG, and KS

	Reference data sets											
	HJ				LG				KS			
Protein	α	β	Turn	Unorder	α	β	Turn	Unorder	α	β	Turn	Unorder
CHYT	0.10	0.34	0.20	0.36	0.11	0.50	0.25	0.14	0.10	0.38	0.26	0.26
CYTC	0.38	0.00	0.17	0.45	0.49	0.11	0.22	0.18	0.43	0.02	0.22	0.33
ELAS	0.10	0.37	0.22	0.31	0.10	0.46	0.28	0.16	0.10	0.38	0.26	0.26
HBN	0.75	0.00	0.14	0.11	0.86	0.00	0.08	0.06	0.76	0.00	0.12	0.12
LDH	0.41	0.17	0.11	0.31	0.43	0.26	0.19	0.12	0.37	0.14	0.25	0.24
LYSM	0.36	0.09	0.32	0.23	0.46	0.19	0.23	0.12	0.39	0.11	0.34	0.16
MGLB	0.78	0.00	0.12	0.10	0.88	0.00	0.07	0.05	0.78	0.00	0.10	0.12
PAPN	0.28	0.09	0.14	0.49	0.28	0.29	0.18	0.25	0.25	0.19	0.26	0.30
SUBB	0.30	0.09	0.21	0.40	0.32	0.37	0.21	0.09	0.31	0.20	0.25	0.24
FLVD	0.38	0.24	0.16	0.22	0.46	0.34	0.13	0.07	0.38	0.25	0.24	0.15
GPD	0.30	0.22	0.14	0.34	0.30	0.37	0.23	0.10	0.26	0.23	0.26	0.25
PRAL	0.07	0.45	0.14	0.34	0.06	0.61	0.19	0.14	0.07	0.49	0.23	0.21
TPI	0.52	0.14	0.11	0.23	0.53	0.24	0.11	0.12	0.45	0.18	0.15	0.22
THML	0.32	0.18	0.20	0.30	0.40	0.31	0.19	0.10	0.34	0.16	0.30	0.20
BNJN	0.00	0.59	0.10	0.31	0.00	0.89	0.07	0.04	0.03	0.50	0.24	0.23
RUBR	0.08	0.12	0.38	0.42	0.00	0.43	0.35	0.22	0.17	0.19	0.28	0.36
PGLU	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00

SELCON program appears to give very good correlation between estimated and found structures in globular proteins.

In conclusion, the predication of the secondary structure of proteins from CD data is still the most convenient method. The discrepancy between the estimated and the found structures was narrowed but not eliminated after many years of effort. An important future goal is to increase the reference proteins.[18] Finally, there are a few technical tips that may facilitate the accuracy of SELCON analysis. First, the sample must be pure. Impurities should be absolutely avoided.

Table 3. The Impact of Reference Data Set to the Secondary Structures of CHYT

	Secondary structure	Reference data set		
		KS	HJ	LG
SELCON	α	0.15	0.13	0.13
	β	0.30	0.26	0.43
	Turn	0.26	0.24	0.27
	Unordered	0.29	0.37	0.17
X-RAY	α	0.10	0.10	0.11
	β	0.38	0.34	0.50
	Turn	0.26	0.20	0.25
	Unordered	0.26	0.36	0.14

However, the polymorphous of the protein, which is hard to exclude, in solution is a source of inaccuracy. Second, the buffer should be transparent. The scattering of light should be avoided. Third, the cuvette should be short. It is advisable to use a 1 mm cuvette if both solubility and availability of samples are not a problem. SELCON is a reliable calculation method to predict the secondary structures of proteins based on good CD spectral data.

ACKNOWLEDGMENTS

This work was supported in part by National Science Council grant NSC-86-2113-M-001-035. WCL is a NSC postdoctoral fellow. We are indebted Dr. T.-S. Hwang for many valuable suggestions and discussions.

REFERENCES

1. Pauling, L.; Corey, R. B. *Proc. Natl. Acad. Sci. U.S.A.* **1951**, *39*, 205.
2. Sanger, F.; Tuppy, H. *Biochem. J.* **1951**, *49*, 463 & 481.
3. In "Circular Dichroism and the Conformational Analysis of Biomolecules", **1995**; Fassman, G. D., eds., Plenum Press, Chapter 2.
4. Greenfield, N.; Fasman, G. D. *Biochemistry* **1969**, *8*, 4108.
5. Brahms, S.; Brahms, J. *J. Mol. Biol.* **1980**, *138*, 149.
6. Chen, Y.-H.; Yang, J. T. Biochem. Biophys. Res. Commun. **1971**, *44*, 1285.
7. Johnson, W. C., Jr. *Proteins: Struct. Funct. Genet.* **1990**, *7*, 205.
8. Venyaminov, S. Y.; Vassilenko, K. S. *Anal. Biochem.* **1994**, *222*, 176.
9. Yang, J. T.; Wu, C.-S.; Martinez, H. M. *Methods Enzymol.* **1986**, *130*, 208.
10. Johnson, W. C., Jr. *Annu. Rev. Biophys. Chem.* **1988**, *17*, 145.
11. Woody, R. W. in *Peptides, Vol. 7*; Hruby, V. J., Ed.; Academic Press, New York, **1985**; pp 15-114.
12. Chang, C. T.; Wu, C.-S. C.; Yang, J. T. *Anal. Biochem.* **1978**, *91*, 13.
13. Percel, A.; Park, K.; Fasman, G. D. *Anal. Biochem.* **1992**, *203*, 83.
14. Hennessey, J. P.; Johnson, W. C., Jr. *Biochemistry* **1981**, *20*, 1085.
15. Hennessey, J. P., Jr. *Anal. Biochem.* **1981**, *125*, 177.
16. Yang, J. T.; Wu, C.-S. C.; Martinez, H. M. *Methods Enzymol.* **1986**, *30*, 208.
17. Greenfield, N. J. *Anal. Biochem.* **1996**, *235*, 1.
18. Sreerama, N.; Woody, R. W. *Anal. Biochem.* **1993**, *209*, 32.
19. Tsay, L. M.; Hu, S.-M.; Wang, C.; Lee, W. S.; Lin, L. J.; Kan, L.-S. *J. Chin. Chem. Soc. (Taipei)* **1995**, *42*, 493.
20. Johnson, W. C., Jr. *Proteins: Struct. Funct. Genet.* **1990**, *7*, 205.
21. Levitt, M.; Greer, J. *J. Mol. Biol.* **1977**, *114*, 181.
22. Kabsch, W.; Sander, C. *Biopolymers* **1983**, *22*, 2577.

APPENDIX:

An input protocol for SELCON

The document accompanied by SELCON, kindly distributed by Drs. Sreerama and Woody (Department of Biochemistry, Colorado State University, Fort Collins, Colorado 80523, USA) is self-explanatory. However, it requires a typing input of all CD data that is not very convenient for most users. We hereby revised the program so that it could read the data files generated by CD spectropolarimeter.

The data file name must have an extension "IN". For instance, if you want to name the data file generated by a JASCO 720 as KAN, the input file must be KAN.IN.

The first line is a comment line.

The second line lists # of data points, # of secondary structures, *f* value, # of iterations and convergence value in sequence. There is no format for these numbers, but they are separated by commas.

The third line contains the starting wavelength (the larger one), ending wavelength, and interval (in nm) in sequence, respectively. The format is the same as above.

The fourth line is again a comment line but is limited to four spaces.

The fifth line contains the CD data set. These data can be transferred directly from the file generated by a spectropolarimeter (i.e., a JASCO 720). The data set ends with "0"

Save the file under name with extension of IN as described previously. Then move the file under the SELCON directory and run the program.

The following is an example of an ".IN" file made by our program.

(an example)

```
47     4     1.05     0.05  -0.03   100     0.002
250      240   1
test
0.01
0
-0.01
-0.03
-0.05
-0.08
-0.12
-0.18
-0.24
-0.32
-0.42
-0.54
-0.68
-0.84
-1.02
-1.21
```

-1.42
-1.64
-1.86
-2.09
-2.3
-2.49
-2.65
-2.8
-2.91
-3
-3.07
-3.1
-3.1
-3.06
-3
-2.94
-2.89
-2.84
-2.8
-2.8
-2.87
-3
-3.14
-3.21
-3.14
-2.89
-2.44
-1.81
-1.07
-0.38
0.07
0

TO RUN SELCON (Attention: lower case letters are written by the program and upper case letters are inputs by user.)

SELCON (enter)
"what is your input file? (test)", KAN (enter)
"your input file = kan.in"
"choose a reference base data set: hj, ks, or lg?", KS (enter)
Your are all set at this point. The computer will do the rest. The following is a typical output sheet.
(an example)
THE FRACTIONS OF SECONDARY STRUCTURES FROM THE SELF-CONSISTENT

```
METHOD
    Ref: Sreerama and Woody, Anal. Biochem. (1993), 209, 32
  Beginning wavelength: 250.00
    Ending wavelength: 204.00
      Total No of CD points 47
        Secondary structural elements= 4
          Multiplication FACTOR for CD spectrum= 1.050
            Constraints:
          Sum of Fractions - 1 =  .050;  Each fraction =  -.030
        No. of iterations for Self-Consistent Solution: 100
      RMS difference between successive solutions - Convergence  .002
    SAMPLE CD: FILE from test
   .01  .00 -.01  -.03  -.05  -.08  -.13  -.19  -.25  -.34  -.44  -.57  -.71
  -.88 -1.07 -1.27 -1.49 -1.72 -1.95 -2.19 -2.41 -2.61 -2.78 -2.94 -3.06 -3.15
  -3.22 -3.25 -3.25 -3.21 -3.15 -3.09 -3.03 -2.98 -2.94 -2.94 -3.01 -3.15 -3.30
  -3.37 -3.30 -3.03 -2.56 -1.90 -1.12  -.40  .07
 YOUR REFERENCE BASEDATA IS SSDATA.KS
 Ordered DELTA(CD) values :
    0.0000  .4393  .4542  .4852  .4909  .6167  .6233  .7312
    1.0500 1.2338 1.2428 1.2938 1.3240 1.6140 2.2301 2.9913
    3.5160 6.4532
 Number of Proteins in the database= 18
  The SAMPLE data from: test
  The Ordered sequence of PROTEINS:
  test GPD  FLVD CYTC THML SUBB PAPN LDH  LYSM TPI
 RUBR CHYT PRAL ELAS BNJN
  HBN  MGLB PGLU
 IGUESS = 0; The structure of the Protein with closest CD spectrum: GPD
  Initial Guess:  .260  .230  .260  .250
 Number of Iteration: 1
  SOLUTIONS:
  Selection Criteria: ABS(F-sum -1.0 ) <  .050
              f (alpha,beta etc) > =  -.030
  NS NBAS  Alpha  Beta  Turns  Other  F-Sum
   1  6  .333  .178  .258  .240 1.009
   3  6  .271  .206  .248  .233  .957
   4  6  .265  .229  .254  .231  .979
   1  7  .322  .180  .259  .249 1.010
   2  7  .315  .200  .256  .192  .962
   3  7  .286  .217  .263  .186  .952
   4  7  .247  .223  .254  .237  .961
   5  7  .247  .221  .254  .237  .958
```

```
 1   8  .315  .166  .246  .237  .964
 2   9  .316  .185  .230  .242  .973
 3   9  .274  .224  .260  .192  .951
 1  14  .302  .188  .242  .231  .963
 1  15  .302  .188  .242  .231  .963
Selected solution in each Basis set of differing no of Proteins
 1   6  .333  .178  .258  .240 1.009
 1   7  .322  .180  .259  .249 1.010
 1   8  .315  .166  .246  .237  .964
 2   9  .316  .185  .230  .242  .973
 1  14  .302  .188  .242  .231  .963
 1  15  .302  .188  .242  .231  .963
Average solution from above set
 6   6  .315  .181  .246  .238  .980
Number of Iteration:  2
SOLUTIONS:
Selection Criteria: ABS(F-sum -1.0 ) <   .050
          f (alpha,beta etc) > =  -.030
NS NBAS  Alpha  Beta  Turns  Other  F-Sum
 1   6  .343  .169  .255  .238 1.005
 4   6  .316  .183  .241  .220  .961
 1   7  .330  .173  .257  .247 1.007
 1   8  .322  .160  .244  .235  .962
 2   9  .332  .170  .226  .238  .967
 1  14  .306  .184  .241  .230  .961
 1  15  .306  .184  .241  .230  .962
Selected solution in each Basis set of differing no of Proteins
 1   6  .343  .169  .255  .238 1.005
 1   7  .330  .173  .257  .247 1.007
 1   8  .322  .160  .244  .235  .962
 2   9  .332  .170  .226  .238  .967
 1  14  .306  .184  .241  .230  .961
 1  15  .306  .184  .241  .230  .962
Average solution from above set
 6   6  .323  .173  .244  .237  .977
Number of Iteration:  3
SOLUTIONS:
Selection Criteria: ABS(F-sum -1.0 ) <   .050
          f (alpha,beta etc) > =  -.030
NS NBAS  Alpha  Beta  Turns  Other  F-Sum
 1   6  .344  .168  .255  .238 1.005
 4   6  .324  .176  .239  .219  .958
```

1	7	.332	.172	.256	.247	1.007
1	8	.323	.159	.244	.235	.961
2	9	.335	.168	.225	.238	.966
1	14	.306	.184	.241	.230	.961
1	15	.306	.184	.241	.230	.961

Selected solution in each Basis set of differing no of Proteins

1	6	.344	.168	.255	.238	1.005
1	7	.332	.172	.256	.247	1.007
1	8	.323	.159	.244	.235	.961
2	9	.335	.168	.225	.238	.966
1	14	.306	.184	.241	.230	.961
1	15	.306	.184	.241	.230	.961

Average solution from above set

6	6	.324	.172	.244	.236	.977

SELCON，一個以圓二色數據來估計蛋白質二級結構的程式的分析研究

林維政　甘魯生

蛋白質之所以有功能是它們的氨基酸鏈可以折疊成一定的構造，圓二色是在水溶液中測量蛋白質二級結構最方便的方法，唯一的缺點是一種不自我矛盾的方法，所以二十年來科學家致力於尋求更完美的自我標準。Sreerama 和 Woody 二氏所完成之 SELCON 程式是近年來比較完善的，本篇的目的是將 SELCON 仔細加以分析，將所有參數加以比較之後發現 SELCON 確是一有用之程式，因此之故我們也改良了此程式數據輸入方式，使利用 SELCON 時更加方便及快速。這個輸入之程式也收錄在附錄中。