Bull. Inst. Chem. Res., Kyoto Univ., Vol. 66, No. 4, 1989

A System for Protein Sequence Analysis Constructed on Workstation

Tetsunori FUJIMOTO*, Hironobu TAKAHASHI*, Yasushi KUBOTA*[1], Hiroshi NAKASHIMA** and KEN NISHIKAWA***

Received August 16, 1988

We constructed a system for protein sequence analysis, coupling the primary structure database with application programs on workstation SUN 3/260. By this system we can retrieve any sequence data stored in the database and analyse it to get available information on the secondary or tertiary structures of proteins from their amino acid compositions and sequences. The basic algorithm of the analysis is based on the correlation function which is widely used for the analysis of random data. Since the system was designed with efficient man-machine interface in mind, users can easily operate the system by using a pointing device (mouse) and display various results of analysis simultaneously by virtue of the bit-mapped multi-window system.

KEY WORDS: Protein sequence analysis/ Protein sequence database/ Sequence homology/ Prediction of folding type/ Prediction of secondary structure/ Correlation function/ Engineering workstation/

I. INTRODUCTION

Proteins play primary roles in all life processes such as catalyzing biochemical reactions, immunizing against diseases and molecular recognition. The essential function of a protein is intrinsically linked to its three-dimensional structure, which is determined eventually from its amino acid sequence. It is, therefore, of interest to extract information on the three-dimensional structure of a protein from its amino acid sequence from the viewpoint of not only protein eigineering, but also of information processing. Because proteins are linear co-polymers of amino acids, numerical expression of a protein sequence obtained by replacing amino acid residues by some parameters might give a basis to apply the methods of statistics and information theory. In addition, in the late 1970's much progress in the field of molecualr biology has been made in determining nucleotide sequences of DNA. Therefore, data on DNA sequence and protein sequence translated from it have been rapidly accumulated. Thus, it is indispensable to use computers for the numerical analysis and data processing of protein sequences.

On the other hand, recent rapid development of microprocessor technology has

* 藤本 哲知, 高橋 裕信, 窪田 綏: Tsukuba Res. Cen. SANYO Electric Co., Ltd. 2-1 Koyadai, Tsukuba, Ibaraki 305, Japan

** 中島 広志: The School of Allied Medical Professions, Kanazawa University, Kanazawa, Ishikawa 920, Japan

*** 西川 建: Institute for Chemical Research, Kyoto University, Uji, Kyoto 611, Japan

[1] To whom correspondence should be sent at (present address): Biosystems Lab., NOVA Inc., Suzusho Bldg. 301, Araki-cho 23, Shinjuku-ku, Tokyo 160, Japan

made it possible to widely utilize EWS(Engineering Workstation), the performance of which equals that of a mainframe. A stand alone EWS which runs on the UNIX[TM1] operating system, it realizes not only a high speed cimputing capability, but also a prominent man-machine interface supported by bitmap display and multiple overlapping window system. Moreover, network software, such as the Remote File Sharing (RFS[TM2]) and the Network File System (NFS[TM3]), provides networking capabilities that make distributed resoureces easily accessible by any type of computing systems. Therefore, under this software engineering environment, EWS might provide a new means for the study of protein structures. For this reason, we attempted to construct an integrated system for protein sequence analysis coupled with the protein sequence database on a workstation Sun3/260. GENAS proposed by Kuhara and co-workers[1,2] is one attempt to integrate database and application programs, although they constructed their system on a mainframe.

In this paper, first, we will profile a protein sequence database supported by a retrieval system, which can pick up the candidates of a protein name given from a keyboard and display appropriate sequence data on the candidates. The spellings of the candidates do not necessarily coincide exactly with that of the given protein name; *e.g.*, when a protein name "hemoglobin" is entered from a keyboard, not only "hemoglobin" but also "haemoglobin" are picked up as candidates by the system. This retrieval system was written in C language which seems to be the most adequate for this purpose. The system is also accessible to the database on three-dimensional structures of proteins of the Protein Data Bank[3], and it can represent a molecular structure by 3-D graphics under the library SunCore[TM4]. Secondly, we describe a system for the protein sequence analysis using the correlation function. The programs were written in FORTRAN 77, which ran on a protein sequence selected by the retrieval system. According to this system, we can obtain some information on the secondary or tertiary structures of proteins from primary structures such as sequence repetitions, secondary structure and structural homology.

II. DATABASES AND RETRIEVAL

2.1 Protein sequence data

Since essentially all information on structure and function of a protein must reside in the primary structure, which is coded by DNA, we have collected sequence data and constructed a protein sequence database. The format of our protein sequence database is in accordance with the common format as shown in Fig. 1. The mentioned items are protein name and the number of amino acid residues(name), species from which the protein was extracted (source), entry name of NBRF[5] of the same protein sequence data(nbrf), references which describe the sequence data(re-

[1]UNIX is a registered trademark of AT & T.
[2]RFS is a trademark of AT & T.
[3]NFS is a trademark of Sun Microsystems, Inc.
[4]Sun Core is a registered trademark of Sun Microsystems, Inc.
[5]NBRF: the National Biomedical Research Foundation.

```
Protein sequence database.
name  759    p3 protein   human influenza a virus    updated   12/07/84
source       human influenza a virus (strain a/pr/8/34)
nbrf         p3iv34
reference    1   (sequence translated from the genomic rna sequence)
  authors    fields, s., and winter, g.;
  journal    cell 28, 303-313, 1982
comment      this protein is probably one of the three rna-dependent
             rna polymerases.
sites           from       to      description
    matp           1      759      mature protein
sequence      759 aa
     merikelrnlmsqsrtreiltkttvdhmaiikkytsgrqeknpalrmkwmmamkypitad    60
     kritemiperneggqtlwskmndagsdrvmvsplavtwwnrngpmtntvhypkiyktyfe   120
     rverlkhgtfgpvhfrnqvkirrrvdinpghadlsakeaqdvimevvfpnevgariltse   180
     sqltitkekkeelqdckisplmvaymlerelvrktrflpvaggtssvyievlhltqgtcw   240
     eqmytpggevknddvdqsliiaarnivrraavsadplasllemchstqiggirmvdilkq   300
     npteeqavgickaamglrisssfsfggftfkrtsgssvkreeevltgnlqtlkirvhegy   360
     eeftmvgrratailrkatrrliqlivsgrdeqsiaeaiivamvfsqedcmikavrgdlnf   420
     vnranqrlnpmhqllrhfqkdakvlfqnwgvepidnvmgmigilpdmtpsiemsmrgvri   480
     skmgvdeysstervvvsidrflrvrdqrgnvllspeevsetqgtekltitysssmmwein   540
     gpesvlvntyqwiirnwetvkiqwsqnptmlynkmefepfqslvpkairgqysgfvrtlf   600
     qqmrdvlgtfdtaqiiklpfaaappkqsrmqfssftvnvrgsgmrilvrgnspvfnynk   660
     atkrltvlgkdagtltedpdegtagvesavlrgflilgkedrrygpalsinelsnlakge   720
     kanvligqgdvvlvmkrkrdssiltdsqtatkrirmain                       759
//
```

Fig. 1. Data format of the protein sequence database.

ference), sites which give the position of signal peptide and mature protein(sites), and sequence data by one letter code(sequence). One data set terminates in a mark "//". All proteins are classified into 41 files from a biological viewpoint (Table 1). For efficient retrieval, a file of protein names was provided in alphabetic order (Fig. 2). When a protein name is given from the keyboard, its candidates are searched on the list file which gives information on the name of the file in which the sequences of the candidates are stored, and the order of the sequence data stored in the file. Therefore, the retrieval system can have access to the file, which stores the protein sequence data, and pick up the sequence data of the candidates from the list file.

2.2 Physico-chemical parameters

We have collected the 53 physico-chemical parameters inheremt in amino acids such as hydrophobicity, propensity to form α-helix and β-structure. It is noted that these parameters, however, are not always independent, *i.e.*, correlated more or less with each other.

2.3 Method of retrieval

When a keyword(a string of letters) is given, the candidates(a set of the strings of the letters which coincide completely or incompletely with the given string) can be selected in the following way: let us consider two strings, $s_1 s_2 \ldots s_m$ and $t_1 t_2 \ldots t_n$ which should be compared. As a measure of the extent of difference between two lettees s_i and t_j, we adopt the quantity $f(i,j)$;

$$f(i,j) = \min\{f(i-1,j)+1, f(i,j-1)+1, f(i-1,j-1)+d(s_i,t_j)\}, \qquad (1)$$

where

$$f(0,0) = 0$$

and

Table I. The 41 data files classified from a biological viewpoint.

	member	*content*
1)	antigen	antigen
2)	azupla	azurin, plastocyanin
3)	base 1	various kinds of protein from a to k
4)	base 2	alphabetical order from 1 to p
5)	base 3	from r to z
6)	coat	coat protein
7)	colla	collagen, keratin
8)	cyt	cytochrome protein
9)	dnabind	dna binding protein
10)	doxin	ferredoxin, rubredoxin, adrenodoxin, etc.
11)	ec 11	enzymes ec number 1.1.-.- 1.5.-.-
12)	ec 16	enzymes ec number 1.6.-.- 1.-.-.-
13)	ec 21	enzymes ec number 2.1.-.- 2.6.-.-
14)	ec 27	enzymes ec number 2.7.-.-
15)	ec 31	enzymes ec number 3.1.-.- 3.3.-.-
16)	ec 32118	enzymes ec number 3.2.1.18 neuraminidase
17)	ec 34	enzymes ec number 3.4.-.- 3.-.-.-
18)	ec 456	enzymes ec number 4.-.-.- 5.-.-.- 6.-.-.-
19)	haglu	hemagglutinin
20)	hb	hemoglobin
21)	hist	histone
22)	hormone 1	hormone 1
23)	hormone 2	hormone 2
24)	ig	immunoglobulin
25)	inhib	ihibitor
26)	insulin	insulin
27)	inter	interferon, interleukin
28)	lens	lens protein, crystallin etc.
29)	light	light harvesting protein, phycocyanin etc.
30)	lipo	lipoprotein
31)	mb	myoglobin, leghemoglobin, etc.
32)	muscle	muscle protein, myosin, actin, etc
33)	onco	onco gene encoding protein
34)	polypep	polypeptide
35)	ribo	ribosomal protein
36)	toxin	toxin, venom protein
37)	undef 1	undefined protein 1
38)	undef 2	undefined protein 2
39)	undef 3	undefined protein 3
40)	undef 4	undefined protein 4
41)	virus	virus protein

$$d(s_i, t_j) = \begin{cases} 0 & s_i = t_j \\ 1 & s_i \neq t_j \end{cases}$$

This procedure provides a way to evaluate quantitatively a difference between two strings[4]. An example of the matrix $f(i,j)$ for two strings, hemoglobin (10 letters) and haemoglovin (11 letters) is given in Table 2. The value 2 of $f(i,j)$ at (10,11)

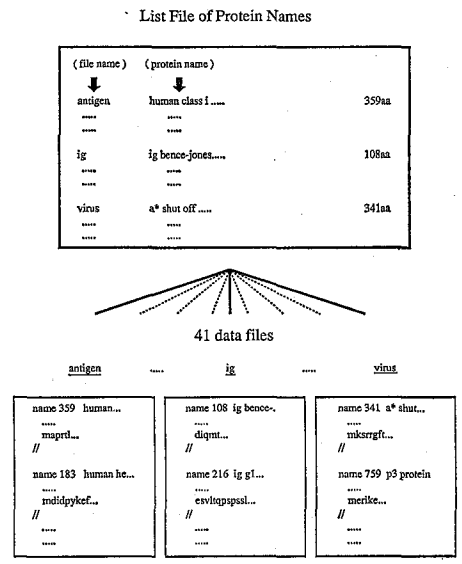

Fig. 2. Constitution of the database: the retrieval system has access to the list file of protein names, and can get information on the name of the file in which the sequence data of the candidates are stored, as well as the order of the data stored in the file.

Table II. Matrix f(i, j) for "hemoglobin" and "haemoglovin".

n	10	9	8	7	6	5	4	4	3	**2**
i	9	8	7	6	5	4	3	3	**2**	3
v	8	7	6	5	4	3	2	**2**	3	4
o	7	6	5	4	3	2	**1**	2	3	4
l	6	5	4	3	2	**1**	2	3	4	5
g	5	4	3	2	**1**	2	3	4	5	6
o	4	3	2	**1**	2	3	4	5	6	7
m	3	2	**1**	2	3	4	5	6	7	8
e	2	**1**	2	3	4	5	6	7	8	9
a	1	**1**	2	3	4	5	6	7	8	9
h	**0**	1	2	3	4	5	6	7	8	9
	h	e	m	o	g	l	o	b	i	n

represents the difference between the two strings. Thus, the matrix $f(i,j)$ gives us the basis for retrieval of the candidates for a given keyword under the tolerance which is set in advance.

III. NUMERICAL ANALYSIS OF SEQUENCE

We provide the following softwares for protein sequence analysis in this system;

(1) detection of sequence repetition[5,6]

(2) prediction of folding type and intra-and extracellular proteins[7,8]
(3) prediction of secondary sturcture by homology method[9]
(4) detection of sequence homology by correlation method[6,10]

Here, we will briefly describe these algorithms for calculation from sequences or amino acids compositions, respectively, although these are described in detail in the references mentioned above.

3.1 Detection of seqeunce repetition[5,6] and sequence homology by correlation method[6,10]

Since 20 amino acids are expressible by physico-chemical parameters such as hydrophobicity, a given amino acid sequence can be converted to a numerical sequence of such values. Therefore, we can introduce the autocorrealtion function as a measure of the extent of sequence repetition (*i,e.*, periodicity in a primary sequence). That is, the autocorrelation function $A(\tau)$, for τ residues apart can be calculated for a numerical sequence of a protein X of n residues long as follows;

$$A(\tau)=\frac{\sum_{i=1}^{n-\tau}(x(i)-\bar{x})(x(i+\tau)-\bar{\bar{x}})}{[\{\sum_{i=1}^{n-\tau}(x(i)-\bar{x})^2\}\{\sum_{i=1}^{n-\tau}(x(i+\tau)-\bar{\bar{x}})^2\}]^{1/2}} \qquad (2)$$

$$\bar{x}=\frac{1}{n-\tau}\sum_{i=1}^{n-\tau}(x(i))$$

$$\bar{\bar{x}}=\frac{1}{n-\tau}(\sum_{i=1}^{n-\tau}x(i+\tau))\,,$$

where $\{x(i)\}$ is a string of numerals represented by appropriate physico-chemical parameters. If the sequence has any repetition of τ_0 residues long, $A(m\tau_0)$ ($m=0,1,2,\ldots$) must exhibit a high value. Similary, as a measure of the extent of homology between two amino acid sequences(or parts of the sequences), X and Y, are given by introducing the cross correlation function described below. The cross correlation function $C(j)$ at the position j of the sequence Y is expressed by comparing a certain fixed partial sequence n residues long, which starts at the u-th residue and ends at the $(u+n-1)$-th residue in the sequence X, with the part of the sequence Y from the j-th residue to the $(j+n-1)$-th residue;

$$C(j)=\frac{\sum_{i=1}^{n}(x(u+i-1)-\bar{x})(y(j+i-1)-\bar{y})}{[\{\sum_{i=1}^{n}(x(u+i-1)-\bar{x})^2\}\{\sum_{i=1}^{n}(y(j+i-1)-\bar{y})^2\}]^{1/2}} \qquad (3)$$

$$\bar{x}=\frac{1}{n}(\sum_{i=1}^{n}x(u+i-1))$$

$$\bar{y}=\frac{1}{n}(\sum_{i=1}^{n}y(j+i-1))$$

The method can be effectively extended to two dimensions as described elsewhere[6,10].

Let X and Y be two different (or same) protein sequences, then, the correlation function, $C_p(i,j)$, of parameter p at the position (i,j) in a square array is defined as

$$C_p(i,j) = \frac{\sum_{l=-k}^{k} (x_p(i+l)-\bar{r}_p)(y_p(j+l)-\bar{r}_p)}{[\{\sum_{l=-k}^{k}(x_p(i+l)-\bar{r}_p)^2\}\{\sum_{l=-k}^{k}(y_p(j+l)-\bar{r}_p)^2\}]^{1/2}}, \quad (4)$$

where $(2k+1)$ is equal to the length of segments to be compared ("frame" or "window"-length of (2k+1)) and $\bar{r}_p$ is an average of the value of parameter p over 20 amino acids. In order to reduce the signal-noise ratio the average correlation function $\overline{A(\tau)}$, $\overline{C(j)}$ and $\overline{C(i,j)}$ are introduced;

$$\overline{A(\tau)} = \frac{1}{n}\sum_{p=1}^{n} A_p(\tau) \quad (5)$$

$$\overline{C(j)} = \frac{1}{n}\sum_{p=1}^{n} C_p(j) \quad (6)$$

$$\overline{C(i,j)} = \frac{1}{n}\sum_{p=1}^{n} C_p(i,j), \quad (7)$$

where n is the number of parameters of amino acids. Since Eqs. (5),(6) and (7) are the arithmetic average, n kinds of parameters should be selected so as to be independent of each other as much as possible in order to avoid artificial weighting on those of the parameters which are correlated. As described in detail elsewhere[6,10], by using the technique of factor analysis, we selected the following six parameters to compute $\overline{A(\tau)}$, $\overline{C(j)}$ and $\overline{C(i,j)}$(the appropriate window-length to compute $C_p(i,j)$ was eleven *i,e.,* $(2k+1)=11$ in Eq. (4).);

(1) partial specific volume[11]
(2) propensity to form reverse turn[12]
(3) pK value of the α-amino group[13]
(4) polarity[14]
(5) relative mutability[15]
(6) pK value of the α-carboxyl group[13]

This set of six parameters provides a good structural homology[6,10].

3.2 Prediction of folding type and intra-and extracellular proteins[7,8]

The amino acid composition of a protein is expressible as a point in the 20 dimensional space, taking the fractions of amino acids along 20 axes. Hence, a number of proteins of known composition will be distributed as points in this composition space, each representing the amino acid composition of a protein. In order to avoid unbalanced sampling, the collected proteins cover a wide range of molecular weight and are from various sources. For the present analysis, first we must convert the real amino acid composition to the normalized composition, in order to adjust scales along all the coordinate axes.

As the first step of grouping, proteins within a certain radius are gathered as the

central group. In this case, the radius was chosen in such a way that about one-tenth of the total proteins were included in the region. Next, the rest of the proteins located outside the central region are grouped on the basis of angular dependence; taking one protein as a reference, the number of proteins within a solid angle of 60° from the direction of the reference point was computed, and then the direction of the highest density was sought by shifting the reference protein. Proteins belonging to the highest density direction were collected as one group. The same process of seeking a direction of maximum density was continued for the rest of the proteins until the number of proteins gathered became smaller than a cut-off value (*i,e.*, 10). Thus, not the radial, but the angular distribution of points in the space show distinct separation into the groups having strong correlations to the location (inside or outside the cell), biological function (enzyme or nonenzyme), and folding type.

3.3 Prediction of secondary structures by homology method[9]

This procedure is based on the assumption that homologous segments in different proteins may share a similar conformation. This assumption is applied to the prediction of secondary structures in proteins. Sequences homologous to a target protein are searched, without allowing any gap, and compared with a number of refernece proteins of known three-dimensional structure, and then we count the number of occurrences (n_α, n_β and n_c) by looking at the secondary structure (α,β and coil states) of the corresponding site of a homologous segment.

In this section, we adopted the following criterion as the sequence homology: when central residue pairs of the eleven-residue "window" having $\bar{C}$ greater than 0.3 are sequentially consecutive over eight residues long, those residue pairs are homologous. Here, we introduce two kinds of weighting factors. One of them is a factor (ν) which depends on the value of $\bar{C}$. We have made a simple definition of $\nu=1$ for $0.3\leq\bar{C}<0.4$, $\nu=2$ for $0.4\leq\bar{C}<0.5$, and so on. With this modification, the number of homologies n_k ($k=\alpha$, β or c) is replaced by the sum of ν. Another kind of factor is to express the relative weights among the three conformational states. This is necessary because the number of occurrences, n_k, depends on the fractions of α, β and coil states (*i.e.*, f_α, f_β and f_c) averaged over the reference proteins. This factor, W_k, is expected to be proportional to the inverse of the average fraction, *i.e.*, $1/f_k$. In this study, however, we treat them as adjustable parameters so as to optimize the results of prediction.

The final quantity used in the prediction is written as

$$q_k = W_k \sum_{i=1}^{n_k} \tau_j \qquad (k = \alpha, \beta \text{ or } c)$$

The adjustable factors were eventually set $W_\alpha=1.3$, $W_\beta=1.4$ with $W_c=1.0$.

Correctness of prediction for 22 sample proteins of known three-dimensional structure is about 60% on the average[9], a better value[16] in comparison with two other existing methods by Chou and Fasman[17], and Robson and co-workers[18].

IV. CASE STUDIES OF THE ANALYSIS

In this section we will show three typical examples of the following proteins by the analysis of this system. First is sheep keratin B2A[19], which seems to have some repetition of short segments, as the application of the method of the autocorrelation function. Second, the methods of prediction of the folding type and secondary structure are applied to human leukocyte interferon[20]. Third, we examine homologies among penicillopepsin[21], endothiapepsin[22] and human renin[23], the three-dimensional structures of which are known except for renin.

Fig. 3 shows that the molecule reveals a clear periodicity of the five residue intervals. The periodicity interrupts at the 50th residue lag, and appears again at the 73th residue lag, after the disturbance of a lag of 23 residues. It is inferred that some structural irregularity occurs in this region. In order to analyse the frequency of residues, we tried the Fourier transform of the autocorrelation. We can clearly find two peaks, the height of 3.26 at 5.12 residues lag and of 3.24 at 4.95 residues lag, respectively.

Fig. 4 shows the results of the predictions of the folding type and secondary structure for human interferon. The folding type of the protein is predicted as α-protein as shown in the lower left window. In the lower right window, the reliabilities of prediction for each folding type are tabulated. In the middle two windows, the result of the secondary prediction is symbolically depicted over two pages.

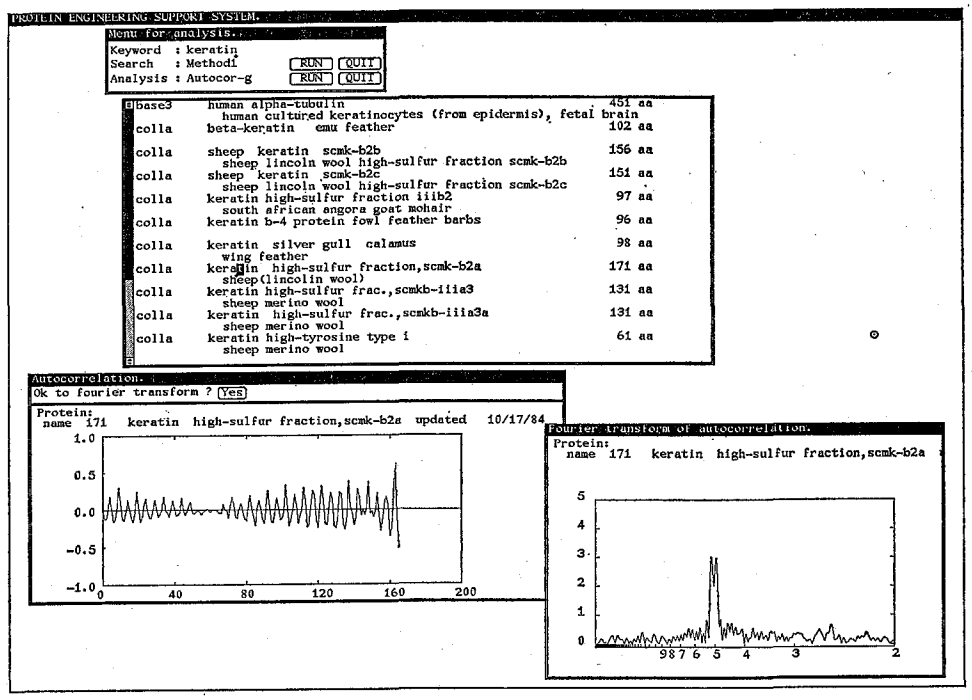

Fig. 3. The periodic pattern in the sequence of sheep keratin B2A calculated by the autocorrelation function $\overline{A(\tau)}$ and its Fourier transform.

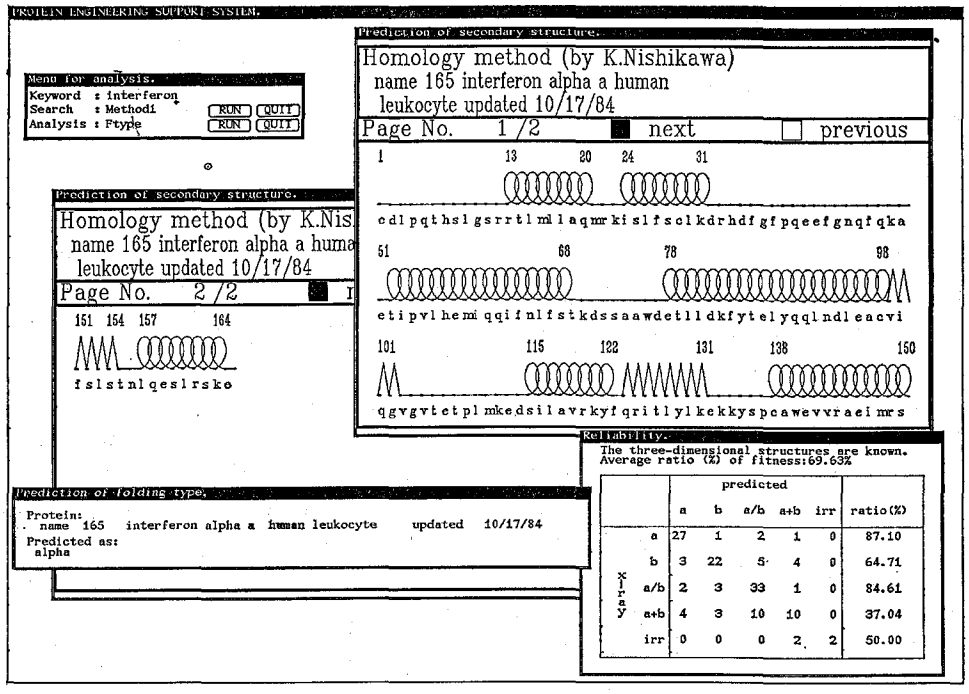

Fig. 4. Prediction of the folding type and secondary structure for human interferon.

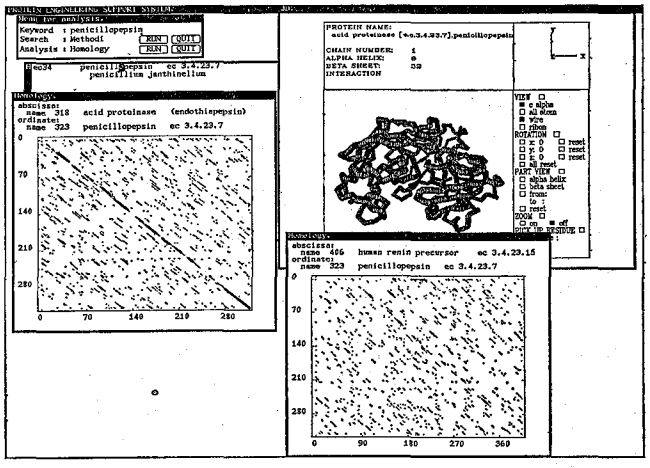

Fig. 5. Comparison of penicillopepsin against endothiapepesin and human reinn. 3–D graphics of penicillopepsin is also displayed in the upper right window.

In order to test how well the sequence homology estimated by the system suggests the structural homology, the comparison was made for penicillopepsin and endothiapepsin as shown in Fig. 5. Successive large dots which represent $\overline{C(i,j)}$s greater than 0.6 appear along the diagonal, indicating that these proteins are closely related to each other. In the upper right window, the three-dimensional

structure of penicillopepsin is represented by 3-D graphics. However, when penicillopepsin and renin are compared, the extent of homology between the proteins is not as high as that of homology between penicillopepsin and endothiapepsin, as seen in the lower right window.

V. CONCLUDING REMARKS

The fundamental algorithm of numerical analysis of protein sequence in the system is based on the correlation function. The function is effectively applied by taking the arithmetic average of correlations, computed in terms of several physicochemical parameters, resulting in improved signal to noise ratio.

Since the present system is constructed on UNIX, we can easily link heterogeneous programing languages. According to the algorithms, therefore, we could incorporate the most suitable language (*e,g.*, *C* language for the retrieval and Fortran 77 for the numerical analysis) into the system. Another feature is to fully utilize the prominent man-machine interface supported by bitmapped display. Several results can be presented simultaneously on the display. Also, use of a pointing device (mouse) easily allows us to operate this system.

We are now improving the system in order to predict the tertiary structures of proteins.

REFERENCES

(1) S. Kuhara, F. Matsuo, S. Futamura, A. Fujita, T. Shinohara, T. Takagi and Y. Sakaki, *Nucleic Acids Research*, **12**, 89 (1984).
(2) S. Kuhara, T. Takagi, S. Futamura, Y. Sakaki, K. Hayashi and F. Matsuo, Fundamental Infology Report 86-3 (in Japanese), Information Processing Society of Japan (1986).
(3) F.C. Bernstein, T.F. Koetzle, G.J.B. Williams, E.F. Meyer, Jr., M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi and M. Tasumi, *J. Mol. Biol.*, **112**, 535 (1977).
(4) The Institute of Electronics, Information and Communication Engineers ed., "Pattern Information Processing," (in Japanese) CORONA Publishing Co., Ltd, 1983, pp. 134–135.
(5) Y. Kubota, S. Takahashi, K. Nishikawa and T. Ooi, *J. Theor. Biol.*, **91**, 347 (1981).
(6) Y. Kubota, *Bull. Inst. Chem. Res., Kyoto Univ.*, **60**, 309 (1982).
(7) K. Nishikawa, Y. Kubota and T. Ooi, *J. Biochem.*, **94**, 981 (1983).
(8) K. Nishikawa, Y. Kubota and T. Ooi, *ibid.*, **94**, 997 (1983).
(9) K. Nishikawa and T. Ooi, *Biochem. Biophys. Acta*, **871**, 45 (1986).
(10) Y. Kubota, K. Nishikawa, S. Takahashi and T. Ooi, *Biochim. Biophys. Acta*, **701**, 242 (1982).
(11) E.J. Cohn and J.T. Edsall, "Proteins, Amino Acids, and Peptides," Van Nostrand-Reinhold, Princeton, New Jersey, 1943.
(12) M. Levitt, *Biochemistry*, **17**, 4277 (1978).
(13) H.A. Sober Ed., "Handbook of Biochemistry, Selected Data for Molecular Biology," 2nd ed., The Chemical Rubber Co., Cleveland, Ohio, 1970.
(14) R. Grantham, *Science*, **185**, 862 (1974).
(15) M.O. Dayhoff Ed., "Atlas of Protein Sequence and Structure," Vol. 5, Suppl. 3, National Biomedical Research Foundation, Washington, D.C. 1978.
(16) K. Nishikawa, *Biochim. Biophys. Acta*, **748**, 285 (1983).
(17) P.Y. Chou and G.D. Fasman, *Adv. Enzymol.*, **47**, 45 (1978).
(18) J. Garnier, D.J. Osguthorpe and B. Robson, *J. Mol. Biol.*, **120**, 97 (1978).
(19) T.C. Elleman, *Biochem. J.*, **130**, 833 (1972).

(20) T. Taniguchi, N. Mantei, M. Schwarzstein, S. Nagata, M. Muramatsu and C. Weissmann, *Nature*, **285**, 547 (1980).
(21) J. Tang, M.N.G. James, I.N. Hsu, J.A. Jenkins and T.L. Blundell, *Nature*, **271**, 618 (1978).
(22) T.L. Blundell, B.L. Sibanda and L. Pearl, *Nature*, **304**, 273 (1983).
(23) T. Imai, H. Miyazaki, S. Hirose, H. Hori, T. Hayashi, R. Kageyama, H. Ohkubo, S. Nakanishi and K. Murakami, *Proc. Natl. Acad. Sci. USA*, **80**, 7405 (1983).