Express. This development opens a new phase in chemical structure searching that should be the subject of another publication.

ACKNOWLEDGMENT

We thank Chemical Abstracts Service for their collaboration and help throughout this experiment. The number of ICI information staff involved with the experiment is large. Without their support the experiment could not have taken place, and their involvement is most gratefully acknowledged. In particular, we would thank Denise Ledgerwood, Colin MacBean, Malcolm Wilkins, Graham Cousins, and Duncan Adshead. Finally, we thank the end-user chemists and the chemistry management for their enthusiasm and interest in the project.

REFERENCES AND NOTES

(1) Haygarth Jackson, A. R. "Online Information Handling—the User Perspective". *Online Rev.* **1983**, *7*(1), 25-32.
(2) Meadow, C. T. "Online Searching and Computer Programming, Some Behavioral Similarities (Or ...Why End-Users Will Eventually Take Over the Terminal)". *Online (Weston, Conn.)* **1979**, *3*(1) 49-52.
(3) Richardson, R. J. "End-User Online Searching in a High Technology Engineering Environment". *Online (Weston, Conn.)* **1981**, *5*(4), 44-57.
(4) Faibisoff, S. G.; Hurych, J. "Is There a Future for the End-User in Online Bibliographic Searching?" *Spec. Libr.* **1981**, *72*, 347-355.
(5) Haines, J. S. "Experience in Training End-User Searchers". *Online (Weston, Conn.)* **1982**, *6*(6), 14-19.
(6) Adamson, G. W.; Bird, J. M.; Palmer, G.; Warr, W. A. "Use of MACCS within ICI". *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 90-92.
(7) Buntrock, R. E.; Valicenti, A. K. "End-Users and Chemical Information". *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 203-207.
(8) Buntrock, R. E. "Chemical Searching for People Who Hate Chemical Searching". *Database* **1985**, *8*(2), 82-83.
(9) Attias, R. "DARC Substructure Search System: A New Approach to Chemical Information". *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 102-108.
(10) Dittmar, P. G.; Farmer, N. A.; Fisanick, W.; Haines, R. C.; Mockus, J. "The CAS ONLINE Search System. 1. General System Design and Selection, Generation and Use of Search Screens". *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 93-102.
(11) Warr, W. A. "Online Access to Chemical Information: A Review". *Database* **1987**, *10*(3), 122-128.

Method for Clustering Proteins by Use of All Possible Pairs of Amino Acids as Structural Descriptors

SHIN-ICHI NAKAYAMA,* SATOKO SHIGEZUMI, and MASAYUKI YOSHIDA

University of Library and Information Science, Tsukuba-city, Ibaraki, 305 Japan

Received September 21, 1987

Proteins were represented as vectors, of which components were all possible pairs of amino acids. From a distance matrix between any pairs of proteins thus represented, several clusters corresponding to connected components were generated. Application of this method to three different sets of proteins showed that it was suitable for clustering closely related proteins with respect to the sequential similarity defined by Dayhoff.

INTRODUCTION

Since sequence data of proteins are believed to have been retained dynamically over evolutionary processes, much attention has been paid to exploring the evolutionary relationships among biological species by sequential similarity between proteins. As a result, various methods of measuring the sequential similarity have been designed.[1] Most of these methods, however, include laborious steps of aligning all or parts of protein sequences to measure the sequential similarity, and thus, a similarity matrix for a large quantity of proteins is not easily obtainable.

Meanwhile, Nishikawa and Ooi expressed proteins as points in a composition space of amino acids and classified them into four groups of intra- and extracellular enzymes and nonenzymes according to the analysis of distribution of points.[2] The method is simple, but composition of amino acids alone is not sufficient to represent structural features of proteins. We expressed proteins using all possible pairs of amino acids as structural descriptors and clustered them on the basis of an easily obtainable distance matrix.

METHOD OF CLUSTERING

If a protein of chain length n were divided to $n - 1$ binary fragments, a set of occurrence counts for each species of binary fragments would form a specific pattern to the protein. The pattern of the protein, however, should differ from that of another one unless the two proteins have the same structure. In some instances the fragments found in one protein may not be included in another one. Thus, all possible pairs of amino acids, which numbered 400, were taken as descriptors, and protein i (P_i) was then represented by a set of descriptor values as $P_i = (x_{i1}, x_{i2}, ..., x_{i400})$, where x_{ik} is the occurrence count for the kth descriptor of the ith protein and is readily derived from the one-dimensional structure of protein i.

Although many different methods are available to cluster a set of proteins represented above, most known methods use distance measurements between each pair of proteins in the set. Thus, for a data set comprising n proteins, a symmetric $n \times n$ distance matrix was generated, the elements of which, d_{ij}, were the distance values between each pair of proteins i and j. In the present work, the Euclidean distance measure was chosen because of its wide use in many areas.[3]

The Euclidean distance measure is considerably affected by scaling factors, and standardization of data is common practice. However, since the descriptors used here were similar in property and standardization was apt to reduce between-group discrimination, the distance measured was used without further standardization.

To produce clusters among proteins the d_{ij} values were ordered by an algorithmic two-dimensional sorting operation to give a rearranged distance matrix.[4]

The clustering process generally consists of fixing a threshold T value in the d_{ij} values and grouping all pairs of objects whose d_{ij}s are less than a chosen threshold. Obviously, the d_{ij} values measured here are just numbers that are the complex function

0095-2338/88/1628-0072$01.50/0

Table I. Set of Proteins from Human Origins

superfamilies, families, entries (abbrev)	chain length
Hormones	
thyrotropin α chain related	
thyrotropin, follitropin, lutropin, and choriogonadotropin α chains	
lutropin α chain, human (LH-a,Hu)	89
choriogonadotoropin α chain, human (CG-a,Hu)	92
thryotropin β-chain related	
thyrotropin β chain	
thyrotropin β chain, human (TSH-b,Hu)	112
follitropin β chain	
follitropin β chain, human (FSH-b,Hu)	117
lutropin and choriogonadotropin β chains	
lutropin β chain, human (LH-b,Hu)	109
choriogonadotropin β chain, human (CG-b,Hu)	149
proinsulin related	
insulin	
proinsulin, human (Pi,Hu)	86
insulin-like growth factors	
insulin-like growth factor I, human (IGF-I,Hu)	70
insulin-like growth factor II, human (IGF-II,Hu)	67
Immunoglobulin-Related Proteins	
immunoglobulin variable regions	
Ig κ chain V regions	
Ig κ chain V region, human Ag (Ig-k,HuA)	108
Ig κ chain V region, human Cum (Ig-k,HuC)	105
Ig κ chain V region, human Pom (Ig-k,HuP)	109
Ig κ chain V region, human Len (Ig-k,HuL)	114
Ig lamda chain V regions, human	
Ig λ chain V region, human Ha (Ig-l,HuH)	112
Ig λ chain V region, human Bo (Ig-l,HuBo)	111
Ig λ chain V region, human Sh (Ig-l,HuS)	108
Ig λ chain V region, human Bau (Ig-l,HuBa)	106
Ig λ chain V region, human Del (Ig-l,HuD)	108
Ig heavy chain V region, human subgroup II	
Ig heavy chain V region, human Newm (Ig-h,HuN)	117
Ig heavy chain V region, subgroup III	
Ig heavy chain V region, human Bro (Ig-h,HuB)	120
Heme Carrier Proteins	
globins	
hemoglobin α chain	
hemoglobin α chain, human (Hb-a,Hu)	141
hemoglobin β-type chains	
hemoglobin β chain, human (Hb-b,Hu)	146
hemoglobin δ chain, human (Hb-d,Hu)	146
hemoglobin γ chain, human (Hb-g,Hu)	146
myoglobin	
myoglobin, human (Mg,Hu)	153
Lipid-Associated Proteins	
animal lipid-binding proteins	
lipid-binding protein A-II	
lipid-binding protein A-II, human (LP-AII,Hu)	77
lipid-binding protein C-I	
lipid-binding protein C-I, human (LP-CI,Hu)	57
lipid-binding protein C-III	
lipid-binding protein C-III, human (LP-CIII,Hu)	79

Table II. Set of Proteins from Various Sources

superfamilies, families, entries (abbrev)	chain length
Heme Proteins of Electron Transport	
cytochrome *c* related	
cytochrome *c*	
cytochrome *c*, sunflower (Cyt-*c*,Su)	111
cytochrome c_2	
cytochrome c_2, *Rhodopseudomonas palustris* (Cyt-c_2,Rhp)	114
cytochrome c_2	
cytochrome c_2, *Rhodospirillum rubrum* (Cyt-c_2,Rhr)	112
cytochrome c_2 and c_{550}	
cytochrome c_2, *Rhodopseudomonas sphaeroides* (Cyt-c_2,Rhs)	124
cytochromes *c'*	
cytochrome *c'*	
cytochrome *c'*, *Alcaligenes sp.* (Cyt-*c'*,Al)	127
cytochrome *c'*	
cytochrome *c'*, *Rhodospirillum rubrum* (Cyt-*c'*,Rhr)	126
Ester Hydrolases	
Phospholipases A_2	
phospholipase A_2, mammalian	
phospholipase A_2, pig (Pl,Pi)	124
phospholipase A_2, viper	
phospholipase A_2, gaboon adder (Pl,Ad)	118
phospholipase A_2, elapid	
phospholipase A_2, ringhals (Pl,Ri)	119
phospholipase A_2, insect	
phospholipase A_2, honey bee (Pl,Be)	129
bacterial and fungal ribonucleases	
ribonuclease (barnase)	
ribonuclease, *Bacillus amuloliquefaciens* (RNase,Ba)	110
ribonuclease U_2	
ribonuclease, *Ustilago sphaerogena* (RNase,Us)	113
pancreatic ribonuclease related	
ribonucleases	
ribonuclease, bovine (RNase,Bo)	124
Immunoglobulin-Related Proteins	
immunoglobulin variable regions	
Ig λ chain V region, human	
Ig λ chain V region, human Ha (Ig-l,HuH)	112
Ig λ chain V region, mouse	
Ig λ chain V region, mouse MOPC315 (Ig-l,Mo)	110
Ig heavy chain V region, human subgroup I	
Ig heavy chain V region, human Eu (Ig-h,HuE)	114
Ig heavy chain V region, human subgroup II	
Ig heavy chain V region, human He (Ig-h,HuH)	118
Ig heavy chain V region, human subgroup II	
Ig heavy chain V region, human Newm (Ig-h,HuN)	117
Ig heavy chain V region, subgroup III	
Ig heavy chain V region, human Bro (Ig-h,HuB)	120
Ig heavy chain V region, rabbit	
Ig heavy chain V region, rabbit BS-5 (Ig-h,Ra)	116

of the sequential similarity, amino acid composition, and chain length between pairs of proteins, and their magnitude does not afford any index for absolute extent of dissimilarity. However, it is possible to evaluate statistically how a given d_{ij} value differs significantly from other observed ones. In case the given d_{ij} value is smaller than the value predetermined by subtracting the standard deviation (σ) of d_{ij} values from their mean (m) ($d_{ij} \leq m - \sigma$), the probability of observing the given d_{ij} value is less than 16%,[5] and the distance is safely said to be significantly close. Thus, the $m - \sigma$ value was tentatively settled as the threshold T value.

A connection pattern graph between proteins was then drawn from the rearranged matrix by connecting lines between all pairs of proteins whose d_{ij} values were less than the T value. Connected components where an arbitrary path is found between each protein in the component and each other protein are defined as clusters.

RESULTS AND DISCUSSION

A large quantity of sequential information of proteins has been accumulated by Dayhoff in the *Atlas of Protein Sequence and Structure*,[6] where proteins are organized into protein superfamilies, families, subfamilies, and entries on the basis of detectable sequential similarity. To compare the results obtained by the present method with those of Dayhoff and to examine the availability of the method, three sets of proteins listed in Tables I–III were prepared. Table I includes a set of proteins from human origins. Table II includes a set from various sources. Table III comprises a set of only heme-carrier proteins. The sequence data for each protein were collected from the *Atlas of Protein Sequence and Structure*.

The distance measurements between all pairs of proteins, followed by transformation of distance matrixes, resulted in

Table III. Set of Heme Carrier Proteins

superfamilies, families, entries (abbrev)	chain length
globins	
hemoglobin α chains	
hemoglobin α chain, human (Hb-a,Hu)	141
hemoglobin α chain, dog (Hb-a,Do)	141
hemoglobin α chain, gray kangaroo (Hb-a,Ka)	141
hemoglobin α chain, echidna (Hb-a,Ec)	141
hemoglobin α chain, platypus (Hb-a,Pl)	141
hemoglobin α chain, chicken (Hb-a,Ch)	141
hemoglobin α chain, viper (Hb-a,Vi)	141
hemoglobin α chain, newt (Hb-a,Ne)	142
hemoglobin α chain, carp (Hb-a,Ca)	142
elasmobranch hemoglobin α chain	
hemoglobin α chain, Port Jackson shark (Hb-a,Sh)	147
hemoglobin β-type chains	
hemoglobin β chain, human (Hb-b,Hu)	146
hemoglobin δ chain, human (Hb-d,Hu)	146
hemoglobin β chain, dog (Hb-b,Do)	146
hemoglobin γ chain, human (Hb-g,Hu)	146
hemoglobin β chain, gray kangaroo (Hb-b,Ka)	146
hemoglobin β chain, echidna (Hb-b,Ec)	146
hemoglobin β chain, platypus (Hb-b,Pl)	146
hemoglobin β chain, chicken (Hb-b,Ch)	146
hemoglobin β chain, frog (Hb-b,Fr)	140
myoglobins	
myoglobin, human (Mg,Hu)	153
myloglobin, dog (Mg,Do)	153
myoglobin, red kangaroo (Mg,Ka)	153
myoglobin, platypus (Mg,Pl)	153
myoglobin, chicken (Mg, Ch)	153
lamprey globins	
lamprey grobin, lamprey (LG,La)	146
lamprey grobin, sea lamprey (LG,sLa)	146
gastropod mollusc globin (opisthobranchs)	
gastropod mollusc globin, *Aplysia limacina* (GG,Ap)	145
gastropod mollusc globin (prosobranchs)	
gastropod mollusc globin, *Busycon canaliculatum* (CG,Bu)	146
annelid globin	
annelid globin, bloodworm (AG,Bl)	146
insect globin	
insect globin, CTT-II β midge larva (IG,CTTII)	143
insect globin	
insect globin, CTT-III midge larva (IG,CTTIII)	135
leghemoglobins	
leghemoglobin, broad bean (Lg,bBe)	144
leghemoglobin, kidney bean (Lg,kBe)	145
leghemoglobin, soybean (Lg,So)	142
leghemoglobin, yellow lupin (Lg,Lu)	153

rearranged distance matrixes as shown in Tables IV–VI.

Fixing the T values as 12.8, 13.4, and 12.8 from the distribution of the d_{ij} values gave the connection pattern graphs for the three sets as shown in Figures 1–3, respectively. In the graphs the nodes designate proteins, and the lines between the nodes illustrate that the corresponding distances are less than the T value.

Figure 1 shows that one large cluster exists together with some small clusters and two independent proteins. The large cluster involves proteins in the superfamilies of animal lipid-binding proteins (LP-AII,Hu, LP-CI,Hu, and LP-CIII,Hu), proinsulin related proteins (Pi,Hu, IGF-I,Hu, and IGF-II,Hu), thyrotropin α chain related proteins (CG-a,Hu and LH-a,Hu), and most of the immunoglobulin variable regions, which appear to be independent of the other superfamilies at a glance and consist of the two families of immunoglobulin κ chain V regions (Ig-k,HuA, Ig-k,HuC, Ig-k,HuP, and Ig-k,HuL) and immunoglobulin λ chain V regions (Ig-l,HuD, Ig-l,HuS, Ig-l,HuBa, Ig-l,HuH, and Ig-l,HuBo). Proteins belonging to the same superfamily but to different families of immunoglobulin heavy chain V regions, human subgroups II (Ig-h,HuN) and III (Ig-h,HuB), form one independent small group. The third cluster consists of the family of hemoglobin β-type chains (Hb-g,Hu, Hb-b,Hu, and Hb-d,Hu) in the globins superfamily. Other globins in the families of hemoglobin α chains (Hb-a,Hu) and myoglobin (Mg,Hu) are independent of each other. Proteins in the last two clusters come from the families of thyrotropin β chain (TSH-b,Hu) and follitropin β chain (FSH-b,Hu) and of lutropin and choriogonadotropin β chains (LH-b,Hu and CG-b,Hu) in the superfamily of thyrotropin β chain related.

Although there can be no absolute measure of the correctness of a classification, the cluster structure is not uniform and inconsistent with that obtained by Dayhoff. It can be altered depending on T values and graph-theoretical grouping strategies.[7] In the present case, however, selection of a smaller T value such as 10.8 ($=m - 2.0\sigma$) gave seven smaller clusters (Ig-k,HuC, Ig-k,HuP, and Ig-k,HuL; Ig-l,HuBa and Ig-l,HuS; Ig-l,HuH and Ig-l,HuBo; CG-a,Hu and LH-a,Hu; IGF-I,Hu and IGF-II,Hu; Hb-b,Hu and Hb-a,Hu; LH-b,Hu and CG-b,Hu), which are expressed by bold lines in Figure 1. All other proteins became independent of each other. Choice of the cliques as clusters also gave an unsatisfactory clustering structure.

An inspection of Figure 1 indicates that proteins with relatively smaller chain lengths gather in the large cluster. Thus, it is suspected that the cluster structure is influenced by the distribution of protein chain lengths in the set, and the distance measurement in the present method gives smaller distance values than those expected from the sequential similarity by Dayhoff for proteins of smaller chain lengths.

In Table II, proteins with similar chain lengths are collected from various sources. The results shown in Figure 2 indicate a somewhat clear structure. The four clusters are generated from the three superfamilies of cytochrome c related (Cyt-c_2,Rhp, Cyt-c_2,Rhs, Cyt-c_2,Rhr, and Cyt-c,Su), phospholipase A_2 (Pl,Ad, Pl,Ri, and Pl,Pi), and immunoglobulin V regions (Ig-h,HuN, Ig-h,HuE, Ig-h,HuB, and Ig-h,Ra, and Ig-l,HuH and Ig-l,Mo). The superfamilies of bacterial and fungal ribonucleases (RNase,Us and RNase,Ba) and cytochrome c' (Cyt-c',Rhr and Cyt-c',Al) form no clusters. Pl,Be and Ig-h,HuH are not grouped to the superfamilies of phospholipase A_2 and immunoglobulin V regions, respectively. RNase,Bo in the superfamily of pancreatic ribonuclease related proteins exists independently. These results suggest that proteins can be clustered fairly well to superfamilies as long as such proteins have similar chain lengths.

The descriptors used in the present methods are all possible pairs of amino acids and include only a bit of information on sequence. Thus, it is not unusual even if the different results are obtained in the clustering of distantly related proteins with respect to the sequential similarity. In other words, this speculation indicates that the present method gives clusters similar to those obtained by Dayhoff for closely related proteins. Indeed, the globins superfamily is well separated in each family as shown in Figure 3. Exceptions are only Hb-a,Ca in the family of hemoglobin α chains and Lg,Lu in that of leghemoglobin.

In the above clustering the T value was settled as $m - \sigma$ from a statistical point of view. If a T value were fixed smaller than $m - \sigma$, clusters of more closely related proteins would be produced. Indeed, new subclusters represented by bold lines appeared as shown in Figure 3 by fixing a T value as $m - 2.5\sigma$ and then linking together all pairs of proteins whose $d_{ij} \leq m - 2.5\sigma$. The resulting subclusters correspond to the respective subfamilies in the globin family except for Hb-a,Ch.

From these results it is concluded that the present clustering method is quite successful in grouping closely related proteins to families or subfamilies by the selection of proper T values.

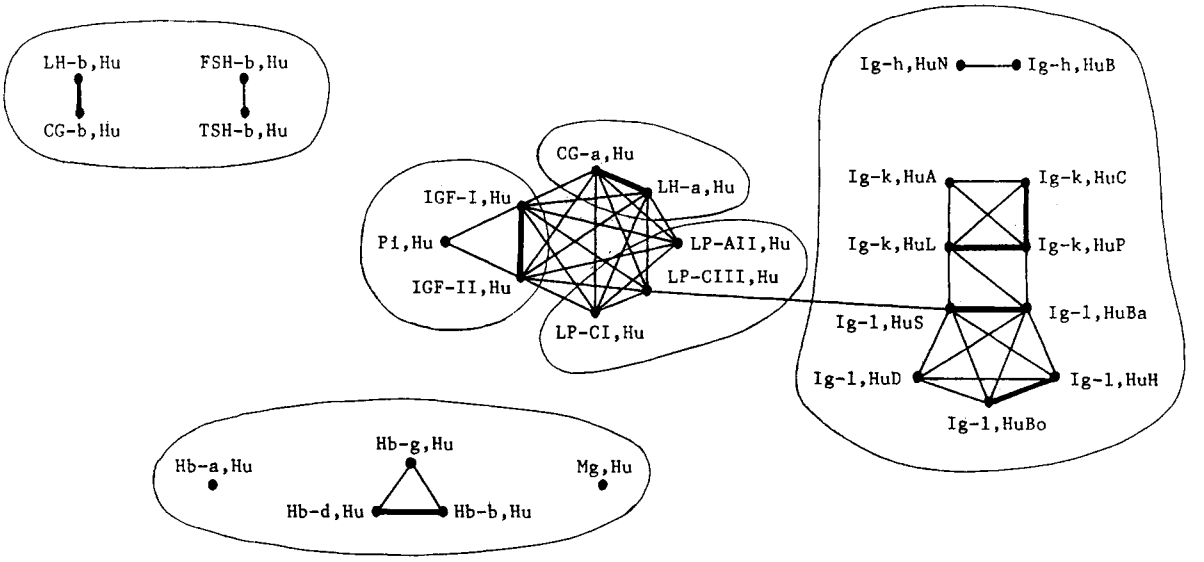


Figure 1. Connection pattern graph drawn from the rearranged distance matrix of Table IV. Proteins in the same superfamily are enclosed by a solid line. Bold lines indicate links whose $d_{ij} \leq m - 2.0\sigma$.

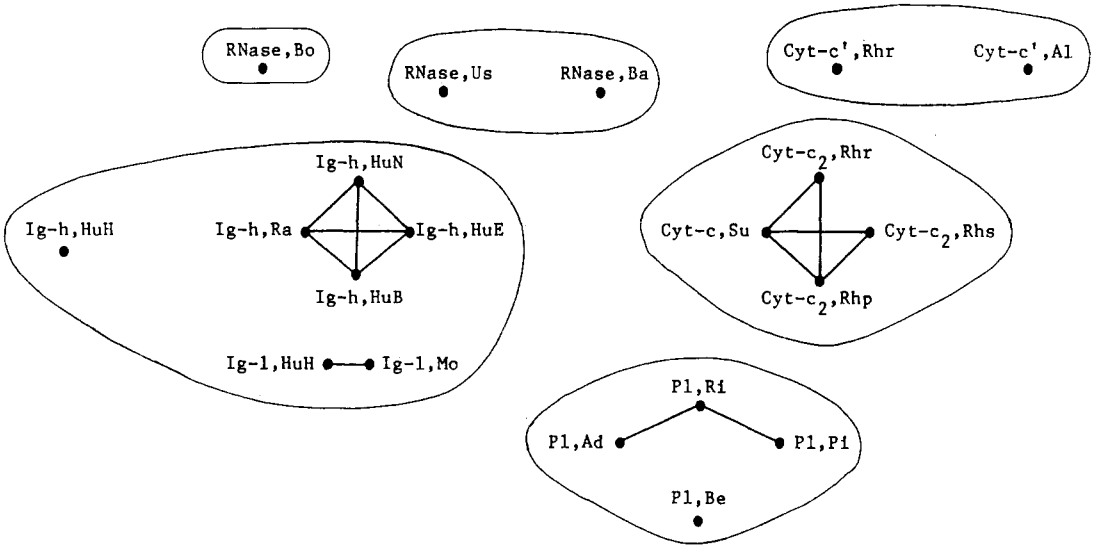


Figure 2. Connection pattern graph drawn from the rearranged distance matrix of Table V. Proteins in the same superfamily are enclosed by a solid line.

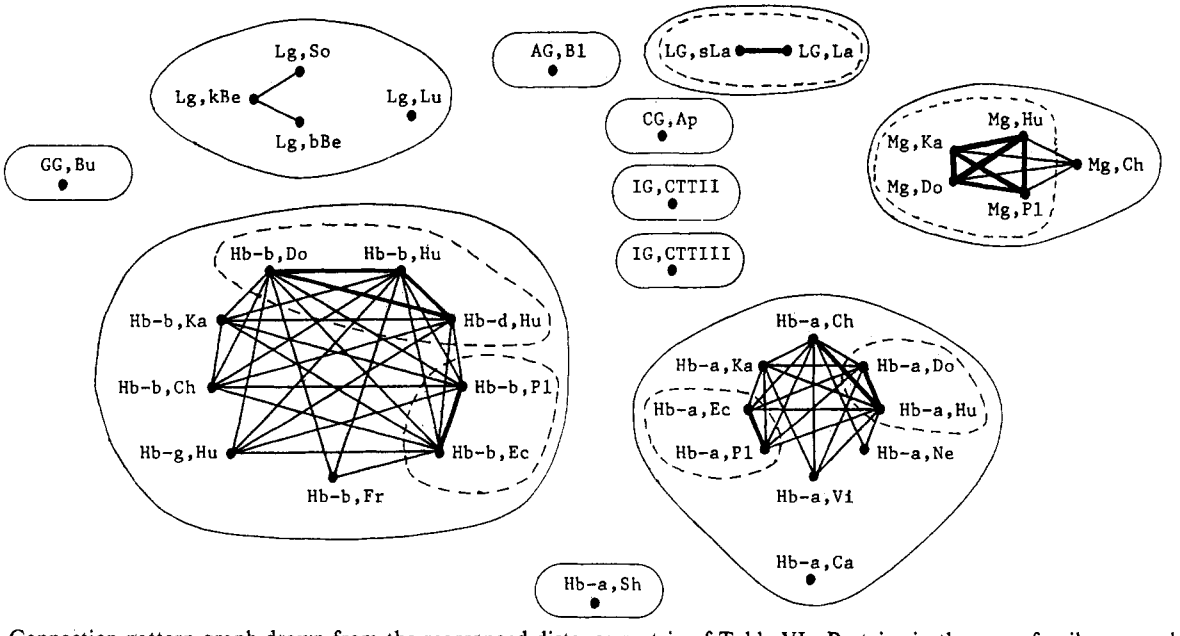


Figure 3. Connection pattern graph drawn from the rearranged distance matrix of Table VI. Proteins in the same family are enclosed by a solid line and those in the same subfamily by a dotted line. Bold lines indicate links whose $d_{ij} \leq m - 2.5\sigma$.

Table IV. Rearranged Distance Matrix for Proteins from Human Origins[a,b]

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
1	CG-a,Hu	0	#	*	*	*	*	*																					
2	LH-a,Hu	42	0	*	*	*	*	*																					
3	IGF-I,Hu	114	117	0	#	*	*	*	*																				
4	IGF-II,Hu	117	120	88	0	*	*	*	*																				
5	LP-CI,Hu	120	125	112	112	0	*	*																					
6	LP-CIII,Hu	124	128	122	127	113	0								*														
7	LP-AII,Hu	125	126	123	126	117	132	0																					
8	Pi,Hu	133	133	119	122	133	138	134	0																				
9	Ig-k,HuL	147	150	145	150	137	139	148	146	0	#	*	*	*	*														
10	Ig-k,HuP	145	148	141	143	138	142	147	150	104	0	#	*	*															
11	Ig-k,HuC	146	148	141	140	131	143	148	146	115	108	0	*																
12	Ig-k,HuA	143	144	137	138	132	138	146	150	117	117	114	0																
13	Ig-l,HuBa	146	150	148	149	143	148	140	150	123	122	132	131	0	#	*	*	*											
14	Ig-l,HuS	139	143	129	138	134	128	141	143	122	135	135	131	98	0	*	*	•											
15	Ig-l,HuH	151	156	153	153	147	150	154	159	130	134	131	135	111	122	0	#	•											
16	Ig-l,HuBo	143	146	139	148	140	144	147	151	129	133	137	134	112	114	107	0	•											
17	Ig-l,HuD	141	144	140	140	142	143	148	145	144	147	150	145	112	111	126	124	0											
18	Ig-h,HuB	150	151	152	153	153	153	158	153	148	147	146	150	148	146	152	146	147	0	*									
19	Ig-h,HuN	131	133	140	141	141	136	146	144	144	147	142	144	140	136	141	133	144	124	0									
20	TSH-b,Hu	134	135	131	137	135	141	141	144	151	153	151	150	155	147	164	156	146	170	158	0	*							
21	FSH-b,Hu	134	137	138	137	134	146	151	145	160	162	154	159	157	149	160	157	153	171	152	127	0							
22	CG-b,Hu	165	166	161	161	169	170	171	167	175	174	174	172	175	168	176	175	175	187	168	157	161	0	#					
23	LH-b,Hu	137	137	134	138	143	146	151	145	160	162	154	159	157	149	160	157	153	171	152	133	141	107	0					
24	Hb-g,Hu	161	162	160	162	141	147	154	167	167	170	162	160	166	151	166	167	165	175	170	172	163	192	175	0	*	*		
25	Hb-b,Hu	157	159	162	162	150	157	156	167	172	178	174	177	171	164	171	173	158	183	177	172	166	188	168	114	0	#		
26	Hb-d,Hu	162	163	163	165	151	156	158	167	173	180	175	177	173	162	174	175	161	185	177	172	169	189	170	120	56	0		
27	Hb-a,Hu	172	176	168	170	162	152	170	173	183	178	180	184	182	172	182	179	175	188	177	168	173	193	176	155	145	147	0	
28	Mg,Hu	170	172	163	166	152	157	165	167	181	176	184	171	174	162	187	180	169	187	183	177	178	200	183	165	170	175	176	0

[a] Asterisks and number signs show the pairs whose $d_{ij} \leq m - \sigma$ and $m - 2.0\sigma$, respectively. [b] Tenfold values of d_{ij} are given.

Table V. Rearranged Distance Matrix for Proteins from Various Sources[a,b]

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	Cyt-*c*,Su	0	*	*	*																
2	Cyt-c_2,Rhp	128	0	*	*																
3	Cyt-c_2,Rhr	131	126	0																	
4	Cyt-c_2,Rhs	134	130	144	0																
5	Ig-h,HuE	139	156	153	160	0	*	*	*												
6	Ig-h,HuN	150	162	164	170	126	0	*	*												
7	Ig-h,Ra	142	160	164	168	131	117	0	*												
8	Ig-h,HuB	157	163	172	174	131	124	133	0												
9	Ig-h,HuH	156	156	157	173	160	139	135	157	0											
10	Ig-l,HuH	157	164	165	175	138	141	144	152	162	0	*									
11	Ig-l,Mo	144	152	159	168	141	147	142	160	156	132	0									
12	RNase,Us	146	156	156	168	142	150	152	155	161	158	147	0								
13	Cyt-*c*′,Rhr	165	151	145	156	170	186	184	190	186	181	171	179	0							
14	Pl,Be	147	156	150	164	150	150	152	165	158	159	157	152	181	0						
15	RNase,Bo	149	156	147	165	148	153	162	170	163	166	154	147	162	144	0					
16	Pl,Pi	157	166	156	170	157	161	167	168	175	165	162	150	178	153	145	0		*		
17	Pl,Ad	151	161	159	169	150	161	159	165	168	159	155	151	183	150	147	144	0	*		
18	Pl,Ri	146	157	151	160	157	157	156	167	167	154	158	147	177	147	147	130	130	0		
19	RNase,Ba	147	153	156	159	156	158	159	167	167	159	153	150	175	149	149	166	152	155	0	
20	Cyt-*c*′,Al	160	157	153	169	167	186	180	187	182	180	175	175	152	176	167	178	179	170	183	0

[a] Asterisks show the pairs whose $d_{ij} \leq m - \sigma$. [b] Tenfold values of d_{ij} are given.

Table VI. Rearranged Distance Matrix for Heme Carrier Proteins[a,b]

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
1	Hb-a,Hu	0	#	#	*	*	*	*	*																											
2	Hb-a,Do	93	0	*	*	*	*	*																												
3	Hb-a,Ch	98	120	0	*	*	*	*	*																											
4	Hb-a,Ka	101	106	112	0	*	*	*																												
5	Hb-a,Pl	105	121	127	113	0	#																													
6	Hb-a,Ec	104	128	122	123	83	0																													
7	Hb-a,Vi	119	124	128	124	145	145	0																												
8	Hb-a,Ne	121	131	128	133	140	138	135	0																											
9	Hb-a,Ca	140	136	141	141	160	162	131	139	0																										
10	Hb-b,Hu	145	150	151	144	145	157	150	150	158	0	#	#	*	*	*	*	*	*																	
11	Hb-d,Hu	147	153	153	149	151	164	151	152	159	46	0	#	*	*	*	*	*																		
12	Hb-b,Do	144	150	143	142	148	157	146	150	156	73	81	0	*	*	*	*	*	*																	
13	Hb-b,Pl	150	147	148	141	149	157	150	147	152	100	105	102	0	#	*	*	*																		
14	Hb-b,Ec	150	151	149	144	153	158	152	146	156	105	104	101	70	0	*	*	*	*																	
15	Hb-b,Ka	159	162	144	160	162	169	156	158	167	123	125	103	120	127	0		*																		
16	Hb-g,Hu	155	155	151	142	154	160	152	156	167	115	124	120	118	124	129	0																			
17	Hb-b,Ch	157	160	153	153	154	157	152	160	160	120	119	116	119	124	123	134	0																		
18	Hb-b,Fr	159	164	155	155	156	166	160	160	160	125	128	124	130	121	139	144	144	0																	
19	IG,CTTII	174	174	168	170	180	180	168	167	160	173	180	166	169	175	177	166	174	179	0																
20	IG,CTTIII	163	162	156	150	168	172	165	166	154	163	166	158	157	161	167	163	156	169	140	0															
21	Lg,bBe	172	169	164	171	161	164	169	162	173	155	159	151	151	156	168	153	155	168	165	160	0	*	*												
22	Lg,kBe	164	167	164	164	162	165	169	159	164	157	157	156	157	151	167	172	155	162	162	155	125	0	*												
23	Lg,So	166	166	170	165	164	165	172	162	164	167	169	164	157	157	167	170	167	169	162	163	128	104	0												
24	Lg,Lu	171	172	159	165	172	172	166	159	169	154	159	150	155	155	162	159	162	170	166	162	138	142	148	0											
25	LG,La	159	159	158	159	163	161	168	172	164	177	180	169	171	176	176	174	177	183	170	159	167	168	164	174	0	#									
26	LG,slA	157	157	155	156	162	160	165	172	163	173	174	163	172	173	170	174	168	178	169	155	165	159	162	171	52	0									
27	GG,Ap	170	176	172	172	172	175	177	179	173	179	181	172	180	185	174	179	183	189	155	151	174	166	161	174	153	151	0								
28	Hb-a,Sh	152	156	161	165	152	163	164	157	163	160	164	163	162	162	167	178	165	165	181	172	169	153	159	179	181	180	190	0							
29	Mg,Ch	172	171	159	163	166	172	171	170	171	174	177	170	174	175	185	173	180	181	171	163	171	172	183	164	181	176	187	175	0	*	*	*	*		
30	Mg,Hu	176	172	160	170	169	174	169	167	168	170	175	168	168	171	184	166	175	174	164	167	168	176	183	163	185	183	192	181	105	0	#	#	#		
31	Mg,Do	174	167	155	158	166	172	165	164	169	169	172	162	164	168	179	167	173	167	170	166	170	170	178	160	176	174	197	173	105	94	0	#	#		
32	Mg,Ka	175	174	162	163	164	170	169	167	178	168	172	164	168	171	179	162	175	169	169	168	174	170	176	161	180	178	189	180	108	88	84	0	#		
33	Mg,Pl	169	168	158	156	160	168	166	168	172	167	170	162	162	165	178	160	174	163	165	159	161	162	172	163	175	172	181	172	105	95	92	93	0		
34	GG,Bu	173	171	175	167	169	177	167	174	174	168	171	162	172	173	175	174	175	172	176	168	161	163	173	177	184	175	178	181	174	173	173	176	167	0	
35	AG,Bl	177	181	178	170	186	189	184	185	177	178	179	178	172	184	187	186	183	188	168	165	181	174	180	184	167	165	166	188	188	188	190	188	183	174	0

[a] Asterisks and number signs show the pairs whose $d_{ij} \leq m - \sigma$ and $m - 2.5\sigma$, respectively. [b] Tenfold values of d_{ij} are given.

Registry No. Pi, 9035-68-1; TSH, 9002-71-5; FSH, 9002-68-0; LH, 9002-67-9; CG, 9002-61-3; IGF-I, 67763-96-6; IGF-II, 67763-97-7; Cyt-*c*, 9007-43-6; Cyt-c_2, 9035-43-2; PI, 9001-84-7; RNase, 9001-99-4; Cyt-*c'*, 9035-41-0; insulin, 9004-10-8.

REFERENCES AND NOTES

(1) Davison, D.; Thompson, K. H. "A Non-Metric Sequence Alignment Program". *Bull. Math. Biol.* **1984**, *46*, 579-590, and references cited therein.
(2) Nishikawa, K.; Ooi, T. "Correlation of the Amino Acid Composition of a Protein to Its Structural and Biological Characters". *J. Biochem. (Tokyo)* **1982**, *91*, 1821-1824.
(3) Sneath, P. H. A.; Sokal, R. R. *Numerical Taxonomy*; W. H. Freeman: San Francisco, 1973.
(4) Ito, T.; Kodama, Y.; Toyoda, J. "A Similarity Measure between Patterns with Nonindependent Attributes". *IEEE Trans. Pat. Anal. Math. Intel.* **1984**, *PAMI-6*, 111-115.
(5) Hoel, P. G. *Introduction to Mathematical Statistics*, 4th ed.; Wiley: New York, 1971.
(6) Dayhoff, M. O. *Atlas of Protein Sequence and Structure*; National Biomedical Research Foundation: Washington, DC, 1972; Vol. 5 and subsequent supplements.
(7) Augston, J. G.; Minker, J. "An Analysis of Some Graph Theoretical Cluster Techniques". *J. Assoc. Comput. Mach.* **1970**, *17*, 571-588.

A New Algorithm for Selection of Synthetically Important Rings. The Essential Set of Essential Rings for Organic Structures

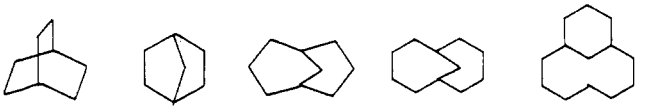
SHINSAKU FUJITA

Research Laboratories, Ashigara, Fuji Photo Film Co., Ltd., Minami-Ashigara, Kanagawa, 250-01, Japan

Received February 9, 1987

The concept of tied rings, multi-tied rings, and dependent rings is introduced, wherein transannular bonds and heterogeneity and abnormality of a ring are key classifiers. The essential set of essential rings (ESER) is defined as a set of rings other than tied, multi-tied, and dependent rings. An algorithm for detection of the ESER and its scope and limitations are discussed.

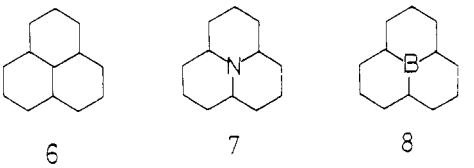
The perception of synthetically important rings is a crucial problem in the manipulation of organic structures by a computer. The smallest set of the smallest rings (SSSR) and its analogues have been widely adopted by computer systems for this purpose.[1] The SSSR is not unique in some cases when the equivalent sets are present in a given structural formula. For example, three 6-membered rings are equivalent in compound **1** and two rings are arbitrarily selected from the three. Corey's first criterion solved this difficulty by the concept of "collection of maximum proper covering sets of rings".[2] This approach is successful in obtaining all three rings of compound **1** but fails to select important rings for organic syntheses in some cases (e.g., **2**–**6**). The Corey's "synthetic subset" adopted



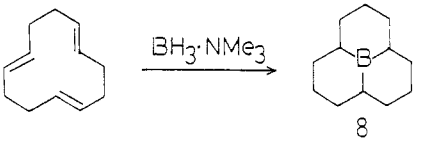
1 2 3 4 5

additional rings with six or fewer members.[3] This criterion is also successful in selecting a 6-membered ring along with two 5-membered rings (SSSR) from compound **1** and **2**. However, an 8-membered ring in compound **3** would be ignored by this procedure. Later, Wipke[4] chose the SSSR and all other rings with eight or fewer atoms. This principle, which is adequate for the purpose of abstracting 6- and 8-membered rings from **2** and **3**, respectively, is not fruitful in the cases of compounds **4** and **5**. A 10-membered ring in **4** and a 12-membered one in **5** are desirable to be adopted in a synthetic point of view. Since these rings in compounds **2**–**5** are in the same situation from the viewpoint of topology, they should be selected by a simple algorithm that meets our chemical sense. Fugamann's approach[5] gave satisfactory results in the above cases. But a more chemist-friendly algorithm is desirable.

A more delicate problem should be mentioned here. Three 6-membered rings should be selected from a carbocyclic compound (**6**) but a 12-membered one need not be chosen. However, the 12-membered rings of compounds **7** and **8** are



6 7 8



8

desirable to be selected, since the center atoms are a nitrogen and a boron atom, respectively. Let us consider that compound **8** is obtained from cyclododecatriene as follows. The 12-membered ring is important synthetically. Thus, a carbocyclic ring is to be preferred synthetically.

Although the importance of the concept of the SSSR is unchangeable now and in the future, a rational extension is desirable to solve the above-described problems. We propose here the essential set of essential rings (ESER), which is a simple algorithm to settle these problems.

DEFINITION AND ALGORITHM OF ESER

Rings are classified as essential rings and nonessential rings. First, we define nonessential rings, which are tied rings, multi-tied rings, or dependent rings. Then ESER is defined as a set of rings other than nonessential rings.

Tied Ring and Multi-Tied Ring. A tied ring is defined as a ring with one transannular bond that links directly two nonadjacent nodes of rings. For example, the 10-membered ring of compound **9** is a tied ring in which a bond between nodes 5 and 10 is a transannular bond defined as above. The tied rings are nonessential rings in any case, since they are

0095-2338/88/1628-0078$01.50/0