Journal of Mathematical Chemistry 20 (1996) 385–394

Cluster analysis from molecular similarity matrices using a non-linear neural network

R. Cruz [a], N. López [b], M. Quintero [b] and G. Rojas [a]

[a] *Institute of Nuclear Sciences and Technology, Ave Salvador Allende y Luaces, Apdo 6163, C.P. 10600, Havana, Cuba*

[b] *Center of Mathematics and Theoretical Physics, Calle E#309 esq 15, Vedado, C.P. 10400 Havana, Cuba*

Received 21 October 1996

A non-linear neural network model to perform cluster analysis is presented. It provides an efficient parallel algorithm for solving this pattern recognition task, consisting, from the mathematical point of view, of a combinatorial optimization problem. A new classification technique is discussed in order to visualize clustering patterns within a molecular set, by means of numerical analysis of the similarity matrix. As an example of the application of the reported neural network model, a quantum molecular similarity study in the field of structure-activity relationships is reported. A molecular set made of eighteen quinolones is used as an example. The resultant cluster distribution showed a good qualitative correlation between similarity data and biological activity.

1. Introduction

Molecular similarity measures permit to obtain useful information on the relationships between members of any molecular set. Applications of molecular similarity studies can be found in the field of structure-activity relationships (SAR) and in the related domain of structure-property relationships (SPR).

As originally introduced by Carbó [1–5], the quantum similarity index is based on the comparison of electron density distributions derived from wavefunction calculations and takes the form

$$R_{AB} = \frac{\int D_A D_B \, dV}{\left(\int D_A^2 dV\right)^{1/2} \left(\int D_B^2 dV\right)^{1/2}}, \tag{1}$$

where D_A and D_B are the electron densities of the two molecules being compared. The Carbó index is sensitive to the shape of the molecular charge distribution and contains information on the interrelation between quantum probability distributions attached to molecules of any molecular set. This index has been extensively used in several structure-relationship studies and the technique has since been extended to cover electrostatic potentials and shape [6,7].

Once a set of molecules $M = \{M_I\}$ to be compared by means of the similarity index has been chosen and a numerical representation of this molecular set has been stored in the similarity matrix R, one can apply a large variety of numerical analysis and visualization techniques in order to analyze similarity computational results [8]. Similarity matrices provide a finite m-dimensional coordinate vector representation for the molecular set, where m is the cardinality of the set M. Simultaneous visualization of all the molecules can be achieved through a projection into some two-dimensional space. Neural networks have been used [7] to solve the dimensionality reduction problem.

The extraction of information from the similarity matrix R can also be achieved by inducing a certain order in the set M, which can have interesting connections with the properties of the compared set. If $P = \{P_I\}$ is the set of properties of the elements of M and $D = \{D_I\}$ the set of known density functions, then the elements of the sets M, D, P are forced to be in a one-to-one correspondence. Molecular similarity representations permit to order the elements of function set D; hence the sets M and P can be considered ordered by induction using the above correspondence. To order the set D, it is only necessary to agree on a predefined rule [4]. The classification of the clustering pattern's appearance within a known molecular set using the information stored in the similarity matrix can be used as the ordering rule. In this work, a neural network model to solve this cluster analysis problem from molecular similarity matrices is presented. This approach permits us to obtain cluster distributions of the studied molecules for different levels of similarity, without defining "a priori" the number of clusters.

Cluster analysis is a main task in unsupervised pattern recognition and it is known to be a combinatorial optimization problem (COP) [9–11]. In this work we introduce a non-linear neural network (NN) to solve it. This approach is inspired by the idea that artificial non-linear neural networks have the capability of solving optimization problems by the best-known algorithms and methods if they exist [12,13]. The introduced NN is described by a non-linear autonomous system of differential equations [14]. This system has an associated Liapunov energy function that is a sum of indefinite quadratic forms. Perturbed similarity matrices are used as matrices of energy functions. The solution of the system of differential equations corresponds to the NN solutions for the discussed cluster analysis problem.

The methodology was tested using a family of eighteen antibacterial quinolones. The resultant pattern recognition analysis of the studied molecular set has originated a clustering of compounds according to their biological activity.

2. Neural network to perform cluster analysis from molecular similarity matrices

The cluster analysis used here places each pattern found in a given collection in one or several clusters, yielding the class (cluster) distribution of the pattern collection, so that:

1. the cluster number must be minimum,
2. the similarity between the patterns must be minimal if they are placed at different clusters and maximal if they are placed inside the same cluster,
3. every pattern must be placed at least at one cluster.

In some sense, this task is similar to the map-colorability problem [7], which consists of coloring the regions of a map in such a way that two adjacent regions do not have the same color and the number of colors must be minimal. Taking this into account we introduced a non-linear neural network to solve the cluster analysis task. This model is a modification of the NN proposed by Takefuji [13], in order to solve the map-colorability problem.

A NN with m^2 neurons was considered, where m is the pattern number of the studied collection. Each neuron is connected with the rest. The differential equation system expressing the state of the network at the time t is

$$\begin{cases} \dfrac{dx_{ij}}{dt} = A\displaystyle\sum_{\substack{k=1 \\ k\neq j}}^{m}(r_{jk}-q)y_{ik} + B\sum_{k=1}^{m}(y_{kj}-1) + Ch_j(y_{1j},\ldots,y_{mj}) \\ y_{ij} = f(x_{ij}), \quad i,j = 1,m. \end{cases} \tag{2}$$

In this system $R = (r_{ij})$ is the similarity matrix of the studied set of molecules, q is a similarity level, hence q is a function of the R matrix elements. For example, q can be taken as the average of the elements of the R matrix or the mean value between the maximum and the minimum similarity values. The value of y_{ij} is the state of the ijth neuron at a determined time; $y_{ij} = 1$ if the jth molecule is placed at the ith cluster and $y_{ij} = 0$ otherwise. The function $f(x_{ij})$ is the transfer function of the neural network. In this model the Takefuji maximum transfer function was used:

$$f(x_{ti}) = \begin{cases} 1 & \text{if } x_{ti} = \max\{x_{1i},\ldots,x_{mi}\}, \\ 0 & \text{otherwise}, \end{cases} \tag{3}$$

but the following functions can also be implemented:

- Sigmoid $$f(x_i) = \tfrac{1}{2}(1+\tanh(x_i)), \tag{4}$$

- McCulloch-Pitts $$f(x_i) = \begin{cases} 1 & \text{if } x_i \geqslant 0, \\ 0 & \text{otherwise}. \end{cases} \tag{5}$$

In the differential equations (2) the first term guarantees that intra-cluster similarities would be large and inter-cluster ones would be small. This fact is better explained in the corresponding term of the energy function. The second term restricts clusters or classes to be disjoint. The third term allows the state of the system to escape from the local minimum and to converge to a global minimum. This last term is the hill-climbing term and consists of the following discrete two-valued function:

$$h_j(y_{1j},\ldots,y_{mj}) = \begin{cases} L & \text{if } \sum_{k=1}^{m} y_{kj} = 0, \\ 0 & \text{otherwise}, \end{cases} \tag{6}$$

where L is a real positive number. This term guarantees that every pattern must be placed at least in one cluster. Each term is modulated by the non-negative constants A, B, C.

The Liapunov energy function associated with the system (2) is

$$E = -\frac{A}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{\substack{k=1\\k\neq j}}^{m}(r_{jk}-q)y_{ik}y_{ij} - \frac{B}{2}\sum_{j=1}^{m}\sum_{i=1}^{m}(y_{ij}-1)^2. \tag{7}$$

The first term in the above function is closer to a minimum if $y_{ij}y_{ik} = 1$ and $r_{jk} > q$ (the similarity value between molecules j and k is greater than a certain threshold). This means that the objects j and k are placed at the same cluster. When $r_{jk} < q$, the minimum value is reached for $y_{ij}y_{ik} = 0$. This means that objects j and k belong to different classes. So, in the global minimum of the energy function, the fact that intra-cluster similarities are large and inter-cluster are small is guaranteed. Hence, the problem of cluster analysis in this model is reduced to a combinatorial optimization problem of the function E over the set $Y = \{y_{ij}, y_{ij} = 0, 1, i, j = 1, m\}$.

The first term of the energy function can be written as

$$E^1 = \frac{A}{2}\sum_{i=1}^{m} E_i = \frac{A}{2}\sum_{i=1}^{m} \bar{y}_i^T W \bar{y}_i, \tag{8}$$

where $W_{mxm} = (w_{jk}) = (q - r_{kj})(1 - \delta_{kj})$ and $\bar{y}_i = (y_{i1},\ldots,y_{im})$; δ_{kj} is the Kronecker delta.

Therefore, in the case of $B \neq 0$ (disjoint cluster distribution), the optimization problem can be written as the following constrained quadratic 0–1 programming problem, which is known to be NP-hard [14]:

$$\min_{\bar{y}_i \in \bar{Y}} E^1 = \frac{A}{2}\sum_{i=1}^{m} \bar{y}_i^T W \bar{y}_i,$$

$$\bar{Y} = \left\{ \bar{y}_i = (y_{i1},\ldots,y_{im}); y_{ij} \in Y; \sum_{k=1}^{m} y_{kj} = 1,\ j = 1, m \right\}. \tag{9}$$

The quadratic form E_i in eq. (8) is the energy function for the ith cluster. If the similarity matrix R is positive definite and $q = 0$, the matrix W is negative definite. Hence the minimum value of function E_i is reached when all the molecules gather in the cluster i, and due to that, for the constraints: $y_i \in \bar{Y}$, the rest of the terms fullfil $E_j = 0$ for all $j \neq i$. This is the same as to say that for a very low level of similarity all the molecules belong to the same class. On the other hand, for the values of q such

that W becomes positive definite, a cluster distribution occurs with one single object in each class. Then, when the similarity matrix is positive definite, the number of clusters increases with the increment of q, as well as the number of molecules per cluster decreases.

3. Application to a structure-activity relationship study

3.1. CALCULATION OF THE SIMILARITY MATRIX

A family of eighteen antibacterial quinolones has been chosen to illustrate the application of the previously described NN model to solve the cluster analysis problem based on quantum similarity measures. The compounds, ordered according to their biological activity values [15], are listed in Table 1. The minimum energy conformations and density functions of molecules were obtained using AM1 MOPAC calculations [16]. The resultant structures were superimposed by a least squares fit of the carbon atoms marked with numbers 3, 4, 5, 10 and 12 [15].

The quantum similarity matrix was calculated analytically using the Carbó similarity index (1). First-order density functions in the LCAO-MO framework can be easily written as

Table 1
Compounds and biological activities of the quinolone set.

No.	Compounds [a]	Biological activity [b]
1	nalidixico (2b)	4.57
2	(1h)	4.68
3	(2c)	4.91
4	(1j)	4.93
5	(2d)	4.99
6	(1k)	5.28
7	(1i)	5.79
8	norfloxacin (1a)	6.50
9	enoxacin (2a)	6.51
10	pefloxacin (1e)	6.52
11	8F-norfloxacin (1g)	6.53
12	8F-pefloxacin (1f)	6.55
13	ofloxacin (1l)	6.56
14	fleroxacin (1d)	6.57
15	temafloxacin (1t)	6.92
16	ciprofloxacin (1b)	7.12
17	amifloxacin (1c)	7.13
18	tosufloxacin (2h)	8.00

[a] Quinolone antibacterial agent derivative. 1: quinolone derivatives and 2: naphthyridine derivates.
[b] Biological activity expressed as log(1/molar MIC), where MIC is the minimum inhibitory concentration against *E.* coli H650. Compounds and activity data were taken from ref. [13].

$$D(\bar{r}) = \sum_{\mu} \sum_{\nu} D_{\mu\nu} \chi_{\mu} \chi_{\nu} , \tag{10}$$

where $D_{\mu\nu}$ is the first-order density matrix, as obtained from the output of MOPAC program, and χ_{μ}, χ_{ν} are the AO's. In the calculation of the Carbó index, some kind of overlap integrals emerges involving four atomic orbitals which, in the worst situation, are centered on four different atomic sites. This kind of integrals has usually been computed employing a CNDO-like approach [1]. Using this idea, the overlap quantum similarity measure can be written as follows:

$$\int D_I D_J dV \approx \sum_{a} \sum_{b} Q_a^I Q_b^J \int S_a^2(\bar{r}) S_b^2(\bar{r}) \, dV , \tag{11}$$

where the sum is performed over all the atoms of molecules I and J. Q is the Mulliken gross atomic population and the function S in the Slater s-type orbital centered at each atomic nucleus [2]. The integrals involved in eq. (11) are then easily computed [17]. This approach allows a rapid analytical integral evaluation, greatly enhancing the speed of similarity calculations.

3.2. COMPUTATIONAL EXPERIMENT

The NN differential equation system (2) was solved using the quantum similarity matrix R. Cluster analysis was performed for q equal to each different element (similarity level) of the similarity matrix, taken in increasing order. This means solving $m(m-1)/2$ differential systems in the worst case, where m is the number of molecules in the set. The solution was updated for every different performance of the clustering pattern. As the transfer function, the "Takefuji maximum" function was used, as given in eq. (3). In this case, the constant $B = 0$ was fixed because this function determines implicitly a disjunction of classes or clusters. The other system constants were fixed equal to unity. Computational solution of the differential equation system was performed using the first order Euler method, which is the simplest among the existing numerical methods. In order to solve the system (2) for each value of q, the initial values of $x_{ij}(t = 0)$ were randomly generated and the solution was iteratively updated until the system state reaches the equilibrium. The artificial neural network provides a parallel gradient descent method to minimize the energy function (7). Since a parallel machine was not available, a simulation of parallel computing model using a sequential machine was performed [13]. All calculations were carried out on a PC-486 DX2 66 MHz.

The effectiveness of the method for each similarity level was evaluated by using clusters for simulated property prediction and then comparing the observed and predicted property values [18]. Given a molecules I in a cluster J, the predicted property value for I can be estimated as the mean of property values for all the other structures in the cluster J. This value is calculated for each of the m compounds in a studied set, but sensible results will be obtained only for those molecules that occur

in a cluster containing at least three molecules. The correlation between the sets of observed (x) and predicted (y) values were calculated by means of the product moment correlation coefficient (PMCC). This is given by the expression

$$\text{PMCC} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}}, \tag{12}$$

where $\bar{x}$ and $\bar{y}$ are the means of the observed and predicted values and summations are defined over all of the molecules occurring in clusters containing at least three members.

3.3. RESULTS

The cluster distribution of molecules for each similarity level is given in Fig. 1. The NN pattern recognition technique, for all the different values of q, gave an excellent clustering for most of the eighteen studied compounds. It can be observed that active molecules belonging to the same activity class gather in the same cluster. For several values of q the method fails in the classification of compounds 5, 16 and 17, which are grouped with molecules that have activities not too close to their values. It seems to be due to the fact that the molecules are structurally very similar, although their activity is very different. Actually, molecules 1 and 5, for example, are different only by one atom in the position number 8 [15], but this difference does not vary very much the value of the similarity index. For all other compounds, clusters contain molecules with very close biological activity values.

For each value of q an individual partition of a set was obtained. By performing calculations for each different similarity value in increasing order, a classification in a top-down manner was obtained (Fig. 1). With minor oscillations when the similarity values were too close, a hierarchical divisive clustering was obtained. For the values of q near to the minimal element of the similarity matrix, all the compounds belong to the same cluster. The number of clusters increases with the increment of q, keeping the order according to biological activity values for most of the compounds. For the values very close to the maximal element of the similarity matrix, each cluster contains exactly one compound.

With this procedure, an order in a set of studied molecules is created using computed density functions, and by induction, the set of attached properties (in this case biological activity) is also ordered. Therefore, one has the possibility of introducing into the set of molecules a new element, which has a known computed density function, but with an unknown property value. Therefore, the similarity index for the new compound could be computed, and the new similarity matrix could be also qualitatively analyzed using this NN classification algorithm. If the introduction of the new element does not alter the obtained order, the relative value of the unknown property can be estimated.

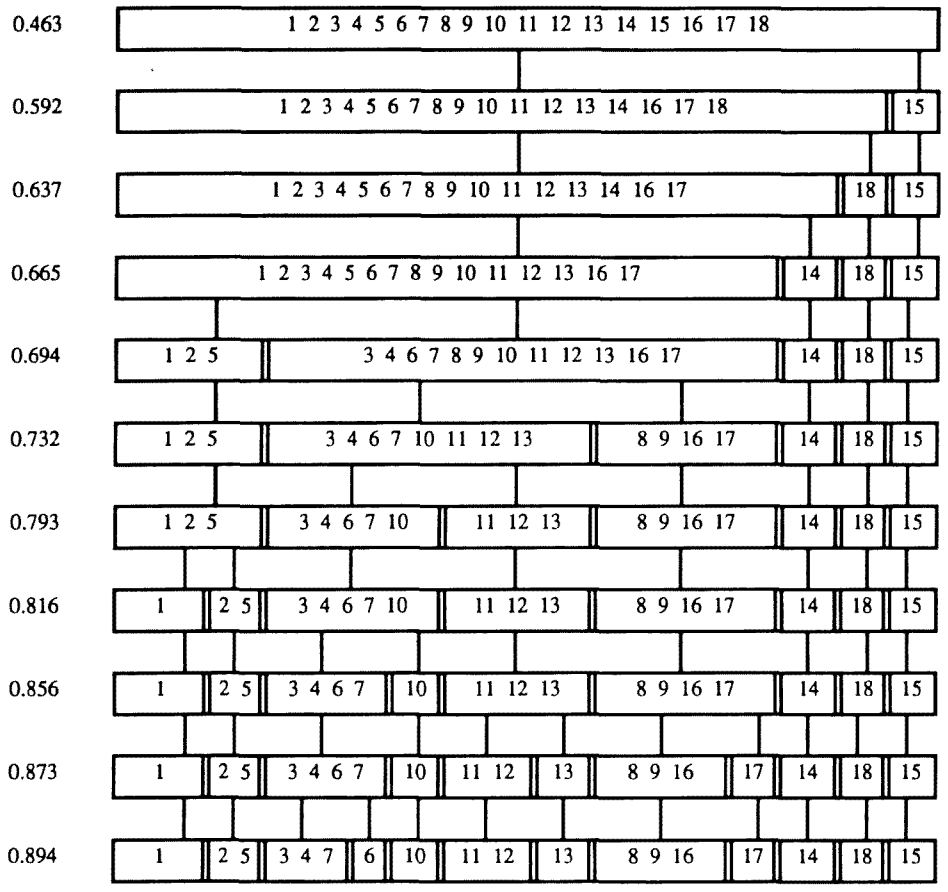

Fig. 1. Cluster distribution of compounds at different levels of similarity.

The PMCC results for each similarity level are listed in Table 2. For the first four levels, the PMCC results gave a perfect inverse correlation. This is due to the calculation procedure of predicted properties described above (these values are ranked in an inverse order), and to the fact that there is only one cluster with more than one member. For similarity levels with PMCC greater than 0.75, there is a good agreement between the predicted and observed activity values in clusters with more than two members. This fact can be used for the relative estimation of the activity value for a new molecule with known electronic density and unknown property. In this case, we can perform cluster analysis with the new similarity matrix at this similarity level and, depending on the cluster the new element belongs to, we can estimate the predicted value of the activity.

Table 2
PMCC values for neural network clustering at different levels of similarity.

q	PMCC	NC[a]	NCC[b]
0.463	−1.0	1	1
0.592	−1.0	2	1
0.637	−1.0	3	1
0.665	−1.0	4	1
0.694	0.477	5	2
0.732	0.573	6	3
0.793	0.759	7	4
0.816	0.825	8	3
0.856	0.837	9	3
0.873	0.808	11	2
0.894	0.707	12	2

[a] Number of clusters.
[b] Number of clusters with more than two elements.

4. Conclusions

A global solution of the combinatorial pattern recognition problem (cluster analysis) is obtained using a non-linear neural network model. Due to the model characteristics, a parallel computational implementation would be very efficient. Nevertheless, numerical computational experiments in sequential machines have shown a good performance too.

The molecular similarity study shows the advantage of the used non-linear NN to perform cluster analysis. The reported classification technique was applied to a SAR study using the Carbó quantum similarity index, and showed a good qualitative correlation between clustering patterns and biological activity for the quinolone set used.

This model can be helpful in extracting information from similarity matrices and provides an alternative method to visualize the relationships and clustering patterns within the molecular set. Since numerical analysis of similarity data with this model is very efficient and can be applied to general matrix structures computed using different similarity measures and indices, the presented model provides a tool for SAR and SPR studies.

Acknowledgement

The authors thank Prof. R. Carbó for his helpful comments and suggestions.

References

[1] R. Carbó, M. Arnau and L. Leyda, Int. J. Quantum Chem. 7 (1980) 1185.
[2] R. Carbó and B. Calabuig, J. Chem. Inf. Comput. Sci. 32 (1992) 600.

[3] R. Carbó and B. Calabuig, Int. J. Quantum Chem. 42 (1992) 1681.
[4] R. Carbó and E. Besalú, in: *Molecular Similarity and Reactivity: From Quantum Chemical to Phenomenological Approaches*, ed. R. Carbó (Kluwer Academic, 1995) p. 3.
[5] E. Besalú, R. Carbó, J. Mestres and M. Solá, in: *Topics in Current Chemistry 173*, eds. J.D. Dunitz, K. Hafner, S. Ito, J.M. Lehn, K.N. Raymond, C.W. Rees, J. Thiem and F. Vögtle (Springer, 1995) p. 31.
[6] A.C. Good, E.E. Hodgking and W.G. Richards, J. Chem. Inf. Comput. Sci. 32 (1992) 188.
[7] A.C. Good, S. Sung-Sau and W.G. Richards, J. Med. Chem. 36 (1993) 433.
[8] R. Carbó and B. Calabuig, Int. J. Quantum Chem. 42 (1992) 1695.
[9] P. Brucker, in: *Optimierung und Operations Research, Lectures Notes in Economics and Mathematical System*, eds. R. Henn, B. Korte and W. Oletti (Springer, Berlin, 1978).
[10] M.R. Garey and D.S. Johnson, *Computers and Intractibility: A Guide to the Theory of NP-Completeness* (Freeman, 1979).
[11] C.H. Papadimitriou, SIAM J. Comput. 10 (1981) 542.
[12] J.J. Hopfield and D.W. Tank, Biolog. Cybern. 52 (1985) 141.
[13] Y. Takefuji, *Neural Network Parallel Computing* (Kluwer Academic, 1992).
[14] P.M. Pardalos and J.B. Rosen, in: *Lecture Notes in Computer Sciences 268*, eds. G. Goss and J. Hartmanis (Springer, Berlin, 1987).
[15] B. Llorente, N. Rivero, R. Carrasco and R.S. Martinez, Quant. Struct. Act. Relat. 13 (1994) 419.
[16] J.J.P. Stewart, QCPE 455.
[17] R. Carbó, B. Calabuig, E. Besalú and A. Martínez, Molec. Eng. 2 (1992) 43.
[18] G.M. Downs and P. Willet, J. Chem. Inf. Comput. Sci. 34 (1994) 1094.