сообщения
объединенного
института
ядерных
исследований
дубна

E19-91-62

V.Michalik

CLUSTER ANALYSIS OF TRACK STRUCTURE
Theoretical Background and Computing Techniques

1991

Introduction

Most models of biological radiation action now agree that for the further study of the problem of biological effectiveness of radiation the knowledge of very local spatial properties of the radiation track structure in relation to the biological macromolecules, especially DNA, is necessary. It is proved by many experiments with low energy heavy particles, ultrasoft X-ray and the radioactive decay of ^{125}I evidencing clearly the increase in the biological effectiveness per unit of deposited energy with increase in the energy concentration (Goodhead 1987).

The detailed information about the microscopic radiation track structure in the nanometer regions can be achieved using the modern Monte Carlo programs simulating the particle tracks in matter. The ideal analysis of the radiobiological data should come from the full description of the track structure in the realistic cellular environment with the detailed accounting for subsequently developing physicochemical and chemical processes and following processes of biological reparation and modification that determine the final cellular state. But up to now the level of knowledge of these individual stages has been insufficient for this analysis and it would not be much practical. Therefore we have to search for some simplified approaches.

Already at the physical stage of radiation action it is useful to summarize the complex information about the radiation track structure. One of the possible ways to solve the problem is to introduce a classification of the radiation track structure based on the cluster concept. In the literature several different approaches to the track structure classification based on the cluster concept are known (Mozumder and Magee 1966, Paretzke 1983, Pitkevitch 1989). The approach described below stems from the suggestion made by Paretzke (1983) to use the conventional Cluster algorithms designed for the analysis of more dimensional data for the track structure classification.

K-means algorithm in the track structure analysis

Let us have a track segment of a charged particle with a defined initial energy producing N ionizations along this track segment. Let $\mathbf{x}_i$ be the cartesian coordinates of the i-th ionization, C_j is the j-th cluster, m_j is the number of ionizations in the cluster C_j, and $\bar{\mathbf{x}}_j$ is the virtual center of the j-th cluster, which can be simply calculated according to

$$\bar{\mathbf{x}}_j = \frac{1}{m_j} \sum_{i \in C_j} \mathbf{x}_i \tag{1}$$

Let us further define

$$e_j = \sum_{i \in C_j} |\mathbf{x}_i - \bar{\mathbf{x}}_j|^2 \tag{2}$$

In the present analysis **x** is always a three dimensional coordinate vector. To divide all N ionizations into k clusters in the K-means method (Späth 1980) we look for such an arrangement of clusters in which the total sum d of all e_j is minimized

$$d = \sum_{j=1}^{k} e_j \tag{3}$$

If the cluster C_j contains m_j ionizations, we speak about a cluster of the m_j-th order. In the process of practical realization of the method described above we start with N clusters of order one and then the number of clusters is reduced to N-1. Subsequently an arrangement of clusters with the minimum total sum of d given by (3) is sought.Then we can reduce the number of clusters to N-2 and repeat the process described above. This procedure of reducing the number of clusters step by step can be continued until the number of clusters is equal to one. In the analysis of multidimensional data using algorithms for partitioning a set of objects into K clusters the methods providing evaluation of clustering validity and selecting an "appropriate" number of clusters are employed. One of these methods (Rousseeuw 1989) was used in the track structure analysis and the result is in Fig.1. The variable at the ordinate called the "overall average silhouette width" must have a maximum for the "appropriate" number of clusters. In our case there is no distinct maximum because the cluster structure of tracks is not sufficiently strong for these methods to give an optimal number of clusters. Even when the cluster structure of tracks is sufficiently strong and the method of "silhouettes" gives an optimal value of K, the clusters obtained in this way will have neither firm energetic nor firm geometric borders and their biophysical interpretation would be difficult.

To cut down the process of cluster growing we can introduce a cluster parameter p and for every two ionizations with coordinates $\mathbf{x}_k$ and $\mathbf{x}_l$ belonging to the cluster C_j the following condition must hold

$$|\mathbf{x}_k - \mathbf{x}_l| \leq p \qquad (4)$$

Practical realization is as described above, but when an arrangement of clusters with the minimum total sum of d given by (3) is found, condition (4) has to be fulfilled for all j at the same time. The process of reducing the number of clusters step by step is continued until the limit beyond which condition (4) can no longer be fulfilled for each cluster. In Fig.2. there is a two-dimensional example of application of the K-means algorithm to a set of points for parameter values 1, 2 and 10.

Monte Carlo transport code

The input data for the K-means algorithm are three-dimensional coordinates of individual ionizations along the particle track. They were generated by Monte Carlo transport code TRION written by Lappa *et al* (1989). The code TRION is suitable for track simulation for heavy charged particles with Z up to 10 in the energy range from 0.3 $Z^{4/3}$ MeV/u to 200 MeV/u and for electrons from 10 eV to 2 MeV. In the process of particle transport through matter, the ionization of outer shells as well as of inner shells, excitations and elastic scattering of electrons are simulated individually by sampling with representative cross sections. The cross sections used are briefly listed in the following. Depending on the energy transferred to the ejected electron, the outer shell ionization is simulated either in the first Born approximation using optical oscillator strengths or in the binary encounter theory. The differential cross section in energy transfer for the inner shell ionization is obtained in the same way, but in the first Born approximation the generalized dipole oscillator strengths are used. The inner shell ionization is approximated as a collision of the charged particle with the atomic electron with respect to the motion and binding of the atomic electron. The excitation cross sections of ions are scaled from those of electrons taken from the published data. For elastic electron scattering the Rutherford differential cross section with the modified screening parameter is used. The set of cross sections was carefully tested by comparison with the theoretical and experimental data available in the literature. The detailed information can be found elsewhere (Lappa *et al* 1989).

In the case of heavy charged particles the track segments are simulated and all secondary electrons are followed until the lower cut-off energy is reached. For the present analysis this cut-off was set at 13 eV because uncertainties in the cross sections become much larger at lower energies and the prime interest in the present analysis is in the spatial distribution of individual ionizations.

Absolute frequency distribution of clusters

Using this clustering method we can compute an absolute frequency distribution of clusters h(j,p) giving us the mean number of clusters of order j produced by radiation per unit of deposited energy when the cluster parameter is p. If the distributior h(j,p) is computed from a sufficiently large number of tracks, it can be considered a characteristic of the radiation reflecting its ability to form clusters of ionizations. In some cases it can be useful to know the first or higher moments of the distribution h(j,p) or the summed distribution of clusters H(j,p) for which

$$H(j,p) = \sum_{k=j}^{\infty} h(k,p) \qquad (5)$$

and which gives the mean number of clusters of order higher or equal to j produced by the radiation per unit of deposited energy if the cluster parameter is p.

The distribution h(j) for heavy ions is computed for the track segment with the length given by several factors including the cluster parameter, δ-rays range, CPU time requirements of K-means for the given number of ionizations and changes of spatial patterns of energy deposition along the particle track. In the most cases the track segment length corresponds to the 5 keV energy loss of a primary particle. The h(j) for the full track of heavy ions cannot be computed due to the insufficient cross section data for low energy heavy ions.

For electron radiation, which is produced by interactions of all ionizing radiations with matter, we have to know h(j) for the full track. For low energy electrons with the energy up to tens of keV we can compute h(j) directly from the spatial distribution of ionizations, but for higher energies, when the whole number of ionizations would exceed several hundreds, it is neither possible due to the time consumptions of the K-means method, nor necessary. For electrons with higher initial energy E_0 the $h(j,E_0)$ can be reconstructed from the $h^1(j,E)$ of electron track segments and full low energy electron tracks. Then

$$h(j,E_o) = \frac{1}{E_o} \sum_{E=E_{min}}^{E_o} h^T(j,E)\ \Delta E . \qquad (6)$$

This is fully justified with respect to the fact that the possibility of selfoverlapping of electron path decreases with increasing initial energy (cross section of elastic scattering decrease). Already for electron energy 10 keV there are no significant differences between the distributions computed directly from the full track structure or from relation (6).

The distribution $h(j,E_\gamma)$ for photons can be computed if one knows the initial spectrum of electrons produced by a photon in the process of its interaction with matter. Then for monoenergetic photons with initial energy E_γ it holds

$$h(j,E_\gamma) = \frac{\int E\ S(E,E_\gamma)\ h(j,E)\ dE}{\int E\ S(E,E_\gamma)\ dE} , \qquad (7)$$

where $S(E,E_\gamma)$ is the initial spectrum of electrons, and $h(j,E)$ is the distribution of clusters for electrons with energy E. If there is a spectrum of photons, we have to integrate over the photon spectrum as well.

In Fig.3. there is an example of the absolute distribution of clusters h(j) for tracks of 2 MeV protons, when the cluster parameter is 3 nm. The borders of the dashed area are given by one standard deviation.

As mentioned above, the distribution h(j,p) or some moments of it could be used as suitable characteristics of radiation reflecting the spatial nonhomogeneity in energy deposition. However the value of the cluster parameter is still an open question. It has to be closely related to the dimensions of structures responsible for the radiation effects considered. The value of p has a lower limit stemming both from the quantum mechanics nature of the interaction processes and from the existence of collective plasmon-like states. A higher limit of p is given by the condition for the

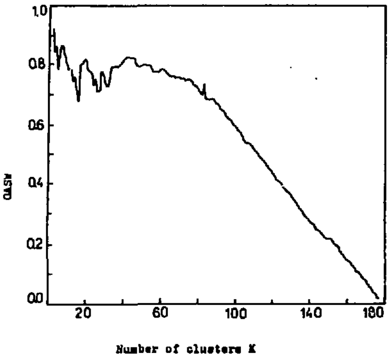

Fig.1 The K-means cluster analysis and the method of "silhouettes" applied to the track of a 5 keV electron.

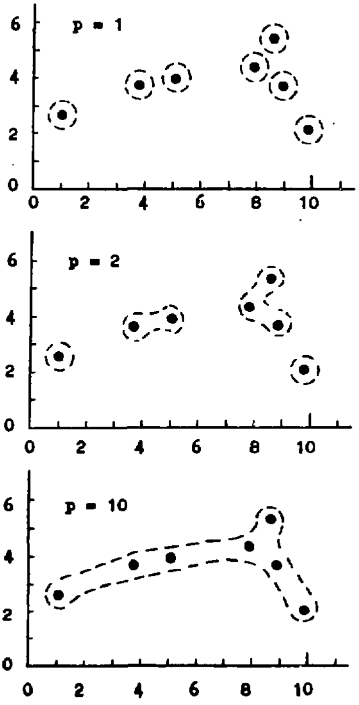

Fig.2. An example of the K-means cluster analysis with different cluster parameters applied to a set of points.

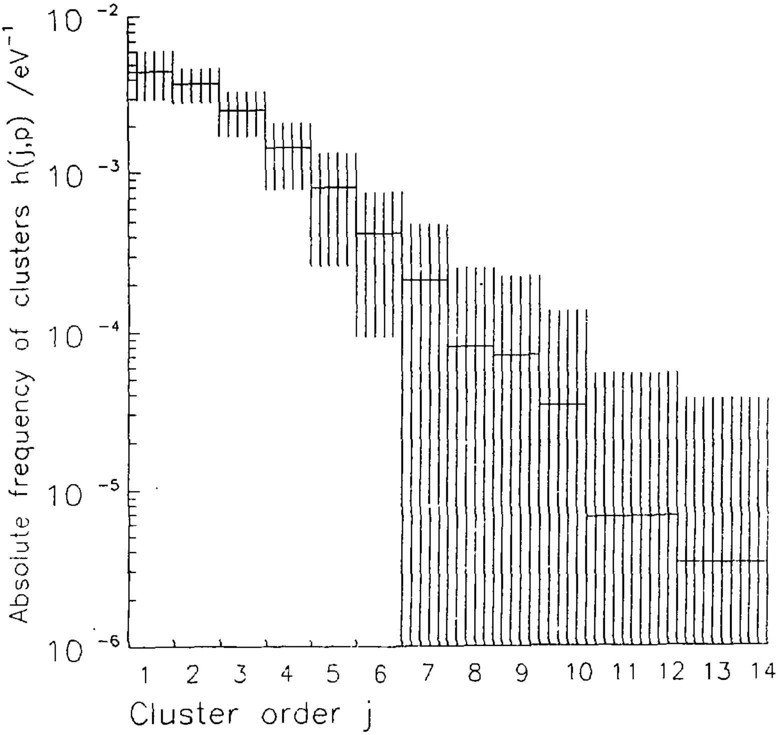

Fig.3. The absolute frequency distributions of clusters in 2 MeV proton tracks when the cluster parameter is 3 nm.

cluster parameter to be much smaller than the dimensions of the simulated track segment. For the given practical applications we have to find this parameter heuristically and we can suppose it will be closely bound to the geometrical dimensions of the structures sensitive to the radiation. The form of the distribution h(j,p) will depend on the ratio of the parameter p to the mean free path between ionizations λ_i.

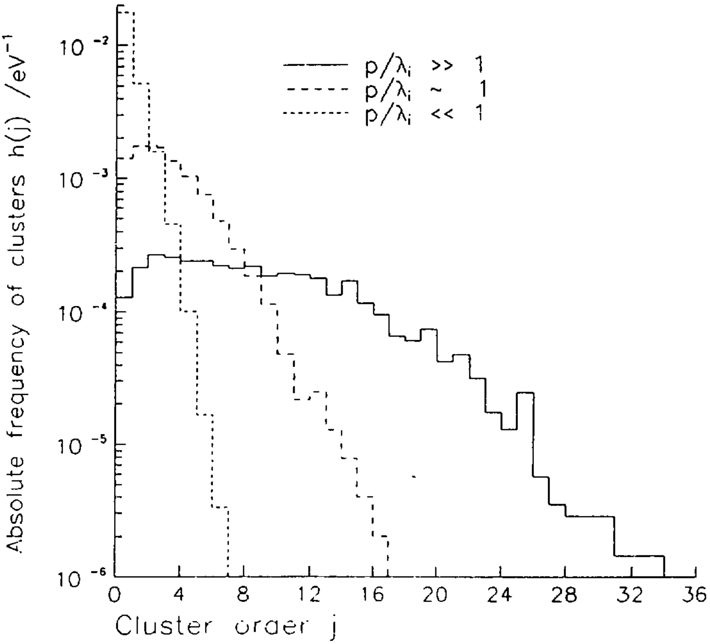

Fig.4. The absolute frequency distribution of clusters for $p/\lambda_1 \ll 1$ and $p/\lambda_1 \gg 1$.

For $p/\lambda_1 \ll 1$ the cluster of order one will be the most probable and the probability of occurrence of higher order clusters will decrease rapidly. With increasing ratio p/λ_1 the spectra will be wider and the greater the p/λ_1 the more probable will be the occurrence of higher order clusters. See Fig.4.

Comparing the normalized distribution $h(j)/\sum h(j)$ with the Poisson one, we can conclude that for small values of the mean cluster size $\bar{j}$ the distribution of clusters can be approximated by the Poisson distribution with mean λ that is the solution of the relation $\lambda = (1-e^{-\lambda})\bar{j}$. For high values of $\bar{j}$ there are large differences between the two distributions. See Fig.5.

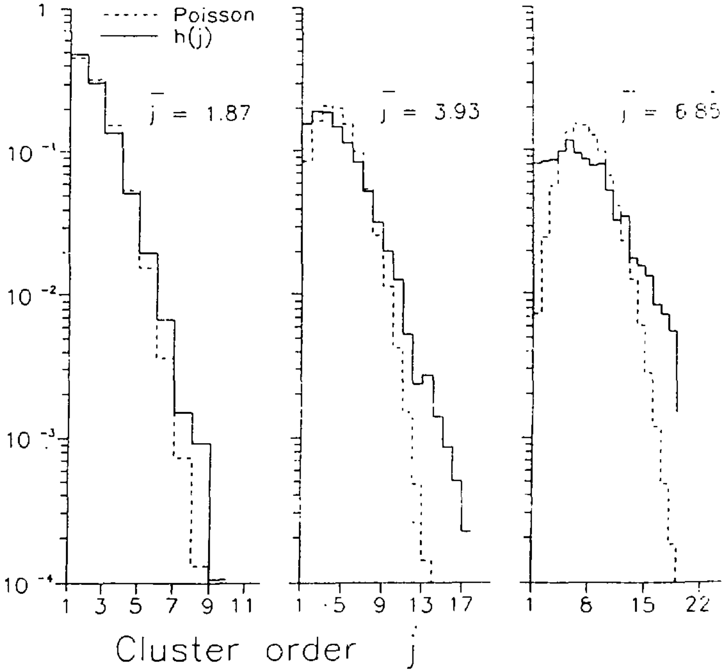

Fig.5. The normalized absolute frequency distributions of clusters with different values of $\bar{j}$ compared with the Poisson distribution.

Conclusion

One of the possible approaches to the track structure classification based on the conventional algorithm for partitioning of objects was described. The radiation is characterized by the distribution of clusters reflecting its ability to form clusters of ionizations along the radiation track. This distribution can be determined when one knows the spatial distribution of ionizations produced by radiation in question. The limitations lie in the inaccuracies of the Monte Carlo simulation and in reducing the cluster analysis to the ionizations alone.

Nevertheless it can be used to estimate radiation quality parameters and to improve selection among them. This analysis also allows one to estimate the yields of different DNA damages such as single strand breaks, double strand breaks, base damages, damages of sugar only and complex damages when simultaneous damage of the sugar backbone and the base are formed in the close vicinity or when a double strand break is accompanied by one or more extra strand scissions or by damaged bases or by both.

References

Goodhead D T 1987 Physical basis for biological effect In: Nuclear and atomic data for radiotherapy and related radiobiology (Vienna:IAEA) 37-53

Lappa A V, Bigildeev E A, Vasilev O N and Burmistrov D S 1989 Code TRION for calculation of characteristics of radiation action and applications in microdosimetry and radiobiology Proceedings of the Sixth All-Union Symposium on Microdosimetry Kanev USSR (Moscow:MIFI) 235-261 in Russian)

Mozumder A and Magee J L 1966 Model of tracks of ionizing radiations for radical reactions mechanism Radiat. Res. **28** 203-214

——————— 1966 Theory of radiation chemistry. VII. Structure and reactions in low LET tracks J. Chem.Phys. **45** 3332-3341

Paretzke H G Concepts of charged particle track structures Proc. 8th. Symp. on Microdosimetry EUR 8395 67-77

Pitkevich V A 1989 Cluster description of physical stage of biological radiation action Proc. of the Sixth All-Union Symposium on Microdosimetry Kanev USSR (Moscow:-MIFI) 169-190 (in Russian)

Rousseeuw P J 1987 Silhouettes: a graphical aid to the interpretation and validation of cluster analysis J.Comp.Appl.Math. **20** 53-65

Spath H 1980 Cluster Analysis Algorithms for Data Reduction and Classification of Objects (New York:Wiley)

Received by Publishing Department
on February 2, 1991.