CMST 3(1) 7-17 (1997) DOI:10.12921/cmst.1997.03.01.07-17

COMPUTATIONAL METHODS IN SCIENCE AND TECHNOLOGY 3, 7-17 (1997)

PARAMETER ANALYSIS OF CLUSTERS OF MELTING TEMPERATURES OF DNA CHAINS

Jacek Błażewicz[1], **Anna Cysewska-Sobusiak**[2], **Agata Lerczak**[1], **Marta Kasprzak**[1], **Wojciech T. Markiewicz**[3]

[1]Institute of Computing Science, Poznań University of Technology,
Piotrowo 3a, PL-60965 Poznań, Poland, E-mail: blazewic@sol.put.poznan.pl
[2]Institute of Electronics and Telecommunications, Poznań University of Technology,
Piotrowo 3a, PL-60965 Poznań, Poland, E-mail: cysewska@et.put poznan.pl
[3]Institute of Bioorganic Chemistry, Polish Academy of Sciences,
Noskowskiego 12, PL-61704 Poznań, Poland. E-mail: markwt.@ibch.poznan.pl

Abstract

In the paper the problem of DNA sequencing by hybridization (SBH) is considered. With the developed software, MELTEM, several assembling procedures are used to ease a collecting a subset of oligonucleotides that would melt under practically identical conditions in a hybridization experiment. Some clustering approaches implemented in the program are compared. The algorithms of MELTEM are presented as well as the selected examples of a melting temperature analysis.

Keywords: DNA sequencing, clustering, hybridization experiment.

1. Introduction

Nucleic acids are a unique class of biomolecules that are capable of forming highly specific complexes due to a presence of heterocyclic bases forming intermolecular hydrogen bonds. Two types of nucleic acids, deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) form an informational and functional network of all biological processes. Investigations of nucleic acids allow to understand molecular mechanisms of life processes as well as they are a starting point of many analytical techniques of practical importance.

Sequencing of nucleic acids, especially sequencing of DNA (accessing genetic information) of many biological species as well as sequencing of human genome became tremendously important in recent years. All large sequencing projects carried out presently (e.g. HUGO) are based on sequencing procedures relying on electrophoretic analysis. Their output should be increased substantially if sequencing genomes of other organisms than of model species is envisaged and highly desirable in the nearest future. Thus, in order to improve the speed of DNA sequencing it is necessary to develop new methods such as e.g. sequencing by hybridization (SBH).

A sequence of nucleic acids encoding genes of living species is highly characteristic for each organism. Despite all the similarities in common genes, even highly related individual organisms can be distinguished due to the differences in their DNA sequences. These differences can be approached by direct sequencing as well as by hybridization experiments with appropriately designed oligonucleotides (nucleic acid fragments, allele specific oligonucleotides).

DNA isolated from different natural sources (organisms, tissues) can be analyzed by hybridization with allele specific oligonucleotides (ASO) and this method is widely used in basic research as well as in diagnostics. Usually, single nucleic acid sequences are used as hybridization probes. Longer nucleic acid probes that can be labeled with a larger number of marker groups (non-radioactive labeling) per molecule of a probe are often used. This allows to increase the sensitivity which is additionally secured to a long unique sequence of a probe. However, increasing the length of the nucleic acid probe diminishes its ability to distinguish single base differences (one or few bases differences). If two nucleic acid sequences to be compared contain several point mutations then a set of specific oligonucleotides could be used and thus allow to increase both the specificity and the sensitivity of a hybridization analysis. Such a set of ASOs can be obtained by chemical DNA synthesis.

One can consider using as ASOs also oligonucleotides produced during a combinatorial oligonucleotide synthesis. The latter approach was not yet realized experimentally but can be considered as an interesting alternative to prepare ASOs. Such an approach would be advised by analysis of melting temperatures of DNA subsequences (present in DNA sequence of interest) in the respect of their lengths. Using a set of ASOs forming complexes of a very similar thermodynamic stability, i.e. with a similar melting temperature T_m, would also help to increase the sensitivity of a detection of particular DNA sequences. This way of increasing a test sensitivity obviously does not preclude all the efforts to improve sensitivity by modification of reporter groups.

There are many programs that help to analyze DNA sequence according to melting temperatures of oligonucleotide duplex subsequences. Such programs are designed in order to assist in choosing most convenient oligonucleotide sequences that can be used as polymerization primers either in a dideoxy DNA or in a PCR amplification of DNA. One of the most popular software of such type is a program written by W. Rychlik [1,2], However, all these programs help to scroll along a DNA sequence looking for a local T_m and cannot collect oligonucleotide subsequences of a given DNA sequence depending on their melting temperatures.

The required program should realize clustering procedures on a full set of oligonucleotides which sequences should be then ordered. Such software could be very useful to reveal any particular clusters of T_m if present in DNA sequences of interest. We decided to develop this one and use it to analyze several DNA sequences derived from GenBank database and characteristic either for prokaryotic or eukaryotic source. Furthermore, the developed software, MELTEM, allows to use several assembling procedures to ease collecting a subset of duplex oligonucleotides that would melt under practically identical conditions in a hybridization experiment. Several clustering approaches implemented in MELTEM were compared. The algorithms of MELTEM are presented below as well as the selected examples of T_m analysis.

In the paper, Section 2 defines the basic biochemical model of melting temperatures of DNA duplexes. Section 3 describes the basic clustering algorithms and the developed software used for cluster analysis. An implementation of such program is given in Section 4, and Section 5 presents two examples of cluster analysis.

2. Calculation of melting temperatures of DNA duplexes

The stability of complexes of nucleic acids (duplexes) was thoroughly studied. The collected data allowed to propose the nearest neighbor model that describes quite precisely the thermodynamic parameters of DNA sequences. The nearest neighbor model is based on the assumption that the stability of a helical DNA (RNA) duplex is influenced to the greatest extent by the molecular interaction within the nearest neighborhood of each nucleotide unit - this assumption allows to reduce the number of parameters used in thermodynamic calculations. Four different deoxyribonucleotides, deoxyadenosine (A), deoxycytidine (C), deoxyguanosine (G) and thymidine (T), forming all DNA, are involved in intermolecular hydrogen pairing which holds two antiparallel strands of the DNA helix. Thus, due to this base pairing (Watson-Crick base pairing) two types of pairs are involved: A-T (equal to T-A) and G-C (equal to C-G). This reduces the number of 16 possible dimers to 10 nearest neighbors and the same number of thermodynamic parameters. Although the nearest neighbor model does not allow to calculate a thermodynamic stability of nucleic acid duplexes with a very high precision, its accuracy is good enough for the use in predicting experimental conditions for most of the hybridization experiments of oligonucleotides of moderate length (few tenths of nucleotide units).

The formula used to calculate a melting temperature of DNA duplex (T_m) and appropriate experimentally estimated parameters according to a modified van't Hoff's equation [3] were used by Breslauer et al. [4] (Formula 1, Table 1). The functional relationship can be written as

$$T_m = \frac{\Delta H}{\Delta S + R\ln(c/4)} + k, \tag{1}$$

where:

ΔS - a sum of entropy for all neighbor dimers,

ΔH - a sum of enthalpy for all neighbor dimers,

R - gas constant,

c - oligonucleotide concentration [mole/1],

$k = 16.6 \log c_s$ - the experimental constant necessary to correct for a salt concentration c_s [mole/1].

In the program described in the next Sections, five algorithms of clusters analysis have been implemented. What is presented hereinafter concerns:

- the selected basic algorithms performed with the use of methods described in [5-8],
- the modifications carried out in interpretation of particular steps of the basic algorithms,
- the structure of the program implemented.

The modifications result from two properties of analysis of clusters of oligonucleotides melting temperatures. Firstly, such an analysis is one-dimensional. Secondly, usually there exist a few or more sequences forming duplexes of the same melting temperature. Thus, it is convenient to divide all the sequences into groups instead of considering any one as a separate object. Each of these groups consists of the sequences of the same melting temperature, creating the object to be considered in an analysis of clusters. Specificity of all objects under analysis is taken into account.

3. Basic algorithms and their modifications

Basic algorithms include non-hierarchical as well as hierarchical agglomeration methods [5-8], The latter, which are most often in use in the cluster analysis, can be described by the same scheme called the central agglomerating procedure. This procedure depends on an analysis of the distance matrix between objects to link particular clusters up to the moment that all the objects make only one cluster. For n objects with a distance matrix $\{d_{ij}\}$ between them, there are following five steps to be realized:

Step 1. Assume each object is a one-element cluster.

Step 2. Search for a pair of clusters p and q $(p < q)$ which distance is minimal, calculating $d_{pq} = \min_{i<j} d_{ij}$.

Step 3. Link the searched out clusters p and q into a new cluster, numbering that one by p and removing the cluster of a number q. Total number of clusters as well as these numbers of clusters which are greater than q are reduced by 1.

Step 4. Transform distances d_{pr} $(r \neq q)$, according to a given method of analysis in use.

Step 5. Repeat Steps 2-4 to obtain only one cluster which includes all objects.

In each iteration, the clusters just linked and their distances as well as an allocation of objects into clusters must be stored.

In our studies, three methods (i.e. the single linkage method, the complete linkage method and the pair group method) were selected from a whole combinatorial group of hierarchical methods [5,6,8], A number of clusters has been denoted by *lg*. During a merging process of two clusters *p* and *q* into a new cluster *p*, elements in the distance matrix were transformed using the formula:

$$d_{pr} := a_1 d_{pr} + a_2 d_{qr} + b d_{pq} + c|d_{pr} - d_{qr}|, \tag{2}$$

where *r* is equal to each value different than *p* and *q*. There is $(1 \leq r \leq lg, r \neq p, r \neq q)$,
while *c* is a parameter specifying a given method considered.

In the **single linkage method,** where $c = -1/2$, the clusters *p* and *q* are linked when $d_{pq} = \min d_{ij}$ $(1 \leq i, j \leq lg)$. Thus, storing minimal distances between clusters, elements of the distance matrix $\{d_{ij}\}$ are transformed as follows:

$$d_{pr} := \min(d_{pr}, d_{qr}). \tag{3}$$

In the **complete linkage method,** where $c = 1/2$, the clusters *p* and *q* are linked when $d_{pq} = \min d_{ij}$ $(1 \leq i, j \leq lg)$. Maximum distances between clusters are stored, and elements of the distance matrix $\{d_{ij}\}$ are transformed in accordance with

$$d_{pr} := \max(d_{pr}, d_{qr}). \tag{4}$$

In the **pair group method,** where $c = 0$, the clusters *p* and *q* are linked when $d_{pq} = \min d_{ij}(1 \leq i, j \leq lg)$ and elements of the distance matrix $\{d_{ij}\}$ are transformed according to

$$d_{pr} := \frac{1}{2}(d_{pr} + d_{qr}) \tag{5}$$

Each of these methods can also be realized using the general combinatorial procedure with suitable values of a_1, a_2, *b* and *c*, respectively.

Using the combinatorial methods mentioned above, it was not necessary to make any modification connected with a complex structure of objects. Linking of clusters does not depend on their size. Furthermore, any clusters *p* and *q* are linked only when $d_{pq} = \min d_{ij}(1 \leq i, j \leq \mathrm{lg})$. In such a case, analysis of clusters is one-dimensional what causes in turn that *p* and *q* are the clusters next to each other. Referring to Step 2 in the central agglomerating procedure, it can be sufficient to consider only (*lg* -1) adjacent pairs instead of all the pairs, i.e. *lg*(*lg*-1)/2.

On the contrary, some modifications are needed, if non-hierarchical methods such as the *k*-means method and the loader method, are used.

The ***k*-means method,** which is a version of the MacQueen's method [5, 7], assumes that its user has to do preliminary division of objects into clusters, based on a determined criterion (e.g. to divide a selected characteristic - the temperature scale is taken herein - into equal intervals). This method is not convergent. Three steps in its basic algorithm are as follows:

Step 1. Compute the cluster centrode and the intracluster variability, taking into account preliminary data.

Step 2. Assign all objects to assign to a nearest centrode. In a case the object will be relocated in other cluster, both centrodes are corrected and intracluster variability is updated.

Step 3. Repeat Step 2 up to either stabilization (i.e. when there are no changes in allocation of objects into clusters) or achievement of maximum allowable number of algorithm iterations.

For sets of objects of a very large size, the **loader method** [7] is most often used for a preliminary assignment into clusters. Taking a number of objects equal to l and a parameter t as a threshold distance, the basic algorithm is given by two steps:

Step 1: First object makes a nucleus of the first cluster while a total number of clusters is equal to $k = 1$.

Step 2: Compute the second power of $d(i,j)$ between objects and nuclei j of clusters ($j=1,2, ..., k$). If $d(i,j) \leq t$, the object i will be in the cluster j. In any other case, the object i makes the nucleus of a new cluster - number of clusters k will grow by 1. All objects $i = 2, 3,\ldots, l$ are in turn the subject of computations.

One disadvantage of this method is that the obtained number of clusters is unknown. The number of clusters can be tuned to some extent by using an appropriate value of the parameter t. Modification concerns Step 2. Objects are agglomerated in accordance with an increase in melting temperatures and then considered in succession. Then, in Step 2 object i will be assigned into cluster k if a square of distance $d(i, k)$ between the object and a nucleus of cluster k is not greater than t. Otherwise, object i makes the nucleus of a new cluster and the whole number of clusters will be increased by 1.

4. Implementation of the program

The program was written using Borland C++ for Windows with *ObjectWindows* library. There are three parts in its structure (Fig. 1), connected with such functions as calculations, communication with users and coordinating operations of all functional parts. The computation part contains algorithms of clusters analysis of a set of functions to calculate melting temperatures of given oligonucleotides. The communication part contains classes and functions which make it easy to operate on files, to arrange configurations, and to display and store the results obtained. The central part allows to start computations, to do a given change in the configuration, and to realize file communications. After the start of the program, a current configuration will appear in the main window which performs management of the whole program, making configuration set-up and data storage, forming and removing data structures, debugging, viewing and storing in results.

Particular files are classified depending on their function in the program. The central part contains an information about the project, i.e. all data needed to form the realizable file. The information about the results of clustering (Fig. 2) is determined by quantities:

- a number of objects, *lobs,*
- a number of clusters, *lg,*
- a size of clusters, *lsg,*
- a composition of clusters, *nrg.*

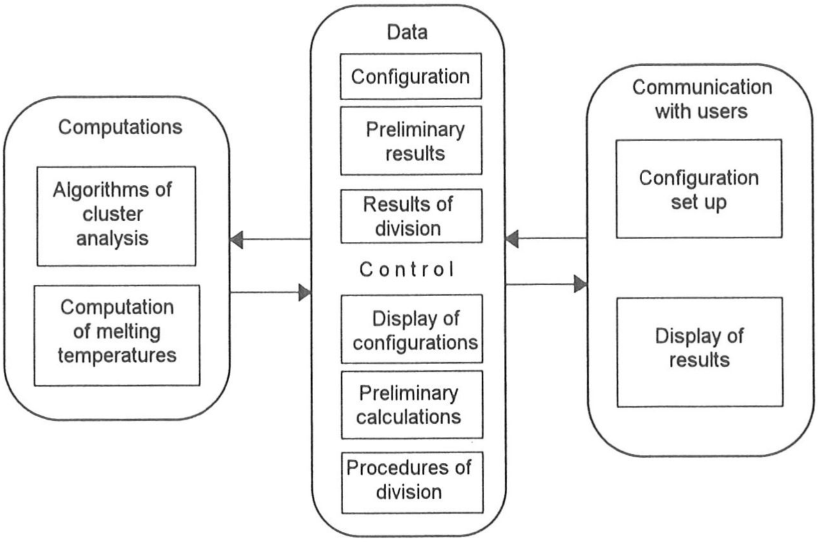


Fig. 1. The program structure.

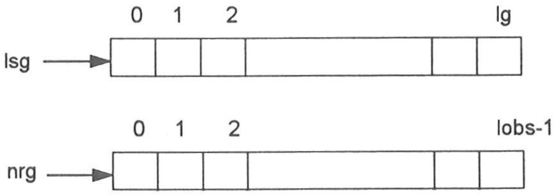


Fig. 2. Information about results of clustering.

Origins and ends of intervals including the particular melting temperatures T_ms are denoted by *spart* and *fpart,* respectively. All the objects to be considered were divided into equal intervals and the clusters by the use of the methods described above. In each case at least one group of tests has been done. If needed, additional groups of tests might be realized as well. An information about preliminary results is given by the following quantities (Fig. 3):

- an initial temperature, TabStart (expressed in K),
- a number of temperature values, TabCount (equal to MAXTEMP+1 at most),
- numbers of sequences at a particular temperature, NumTab,
- numbers of sequences at each temperature to be stored, AlcTab,
- tables of sequences for temperatures of interest, SeqTab.

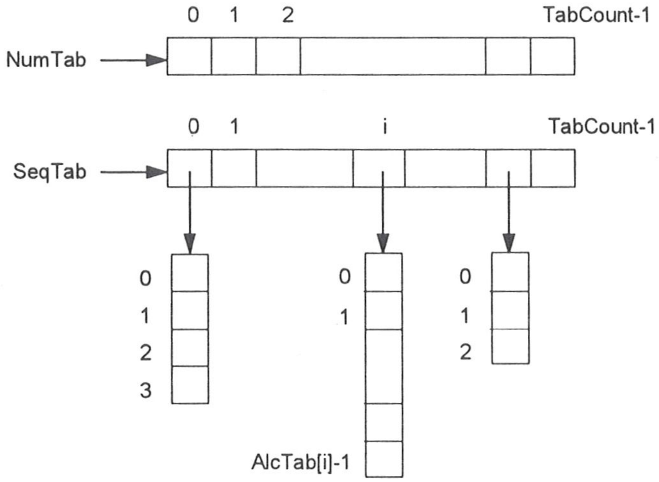


Fig. 3. Information about preliminary results.

During the hybridization experiment, parameters such as salt concentration, probe density and the length of an oligonucleotide were much the same as those which are utilized in real laboratory conditions. Furthermore, at given configurations, mean melting temperatures of oligonucleotides were approximately equal to the room temperature. All necessary data connected with a group of tests were specified, including the experiment configuration, natural sequence and agglomeration method considered. Two reliable measures, i.e. the intracluster variability *V* and Fortier and Solomon indicator *FS,* were used to estimate clustering quality. Clustering by a given method is better when these measures are lower. It was shown that the loader method can always be used as the most efficient one. The representative examples obtained with the program implemented are presented below.

5. Examples of MELTEM DNA analysis

The program allows to calculate melting temperatures (T_m) of oligonucleotide duplexes and oligonucleotide subsequences of a given longer DNA. The number of sequences with the same T_m can be plotted vs. temperature taking different ranges of temperature, and can be applied to different clustering analyses.

The MELTEM program was used to analyze several natural DNA sequences. To exemplify two sequences, one of plant origin (*Lupinus luteus*, length 681 b.p., 210 A, 136 C, 129 G, 206 T, 39% of G-C pairs) and one human gene sequence *(her2* gene, length 757 b.p., 169 A, 216 C, 250 G, 122 T, 62% of G-C pairs) were compared, as shown in Figure 4, by the loader method (Ml).

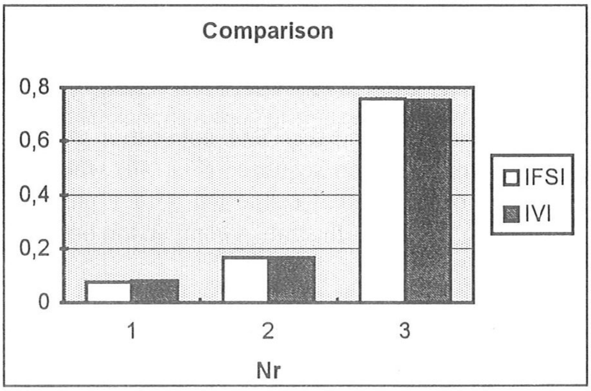
a) FILE: stest plant.lsq
(plant origin *Lupinus luteus*, length 681 b.p., 39% of G-C pairs)

CONFIGURATION:

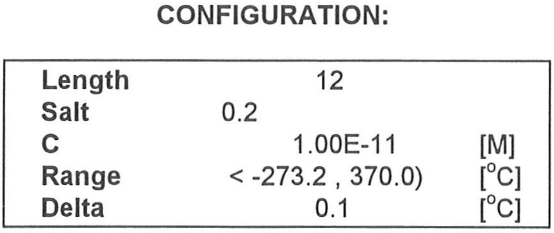
Length	12	
Salt	0.2	
C	1.00E-11 [M]	
Range	< -273.2 , 370.0)	[^{o}C]
Delta	0.1	[^{o}C]

MEASURES:

Nr	SEQUENCE	METHOD	PAR. 1	PAR. 2	FS	V	|FS|	|V|
1	stest plant.lsq	M1	2	100	1.74E+04	4.68E+02	0.0745	0.0810
2	stest plant.lsq	M1	3	100	3.88E+04	9.56E+02	0.1662	0.1655
3	stest plant.lsq	M1	6	100	1.77E+05	4.35E+03	0.7593	0.7535
R					**2.34E+05**	**5.78E+03**	**1.0000**	**1.0000**

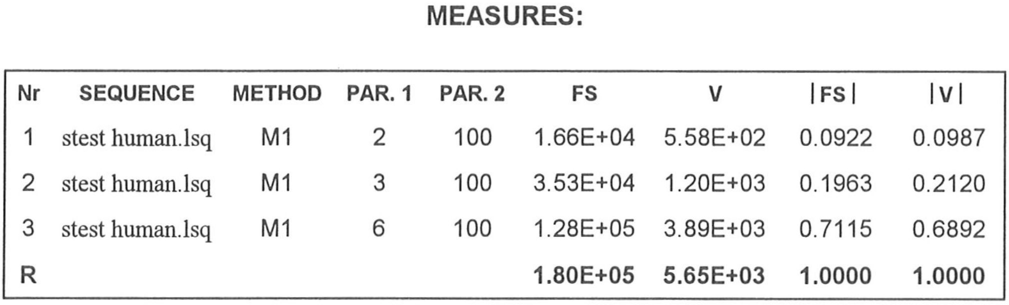


b) FILE: stest human.lsq
(human gene *her2 gene,* length 757 b.p., 62% of G-C pairs)



CONFIGURATION:

Length	12	
Salt	0.2	
C	1.00E-11	[M]
Range	< -273.2 , 370.0)	[°C]
Delta	0.1	[°C]



MEASURES:

Nr	SEQUENCE	METHOD	PAR. 1	PAR. 2	FS	V	|FS|	|V|
1	stest human.lsq	M1	2	100	1.66E+04	5.58E+02	0.0922	0.0987
2	stest human.lsq	M1	3	100	3.53E+04	1.20E+03	0.1963	0.2120
3	stest human.lsq	M1	6	100	1.28E+05	3.89E+03	0.7115	0.6892
R					**1.80E+05**	**5.65E+03**	**1.0000**	**1.0000**

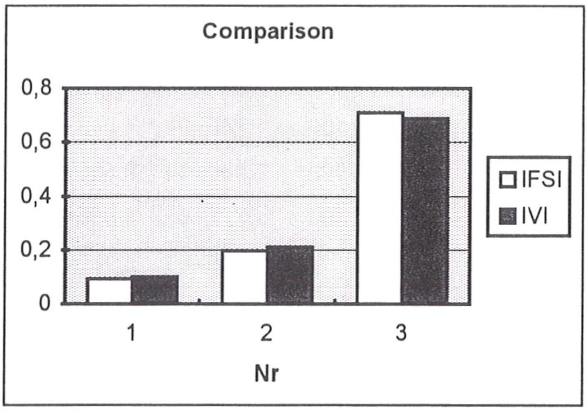


Fig. 4. Information files generated by MELTEM application for two natural sequences: a plant sequence (Fig. 4a) and a human sequence (Fig. 4b) where Ml denotes the loader method, PAR.l - mean size of an interval connected with a number of clusters, PAR. 2 - maximum number of clusters; the plots illustrate a comparison between results falling under a given group of tests.

Two characteristics specify these sequences: both are of the same length (a number of oligonucleotides equals 12) and extremely differ in percentage content of the G-C pairs. The maximum number of clusters (PAR. 2) is equal to 100 what allows to obtain a uniform clustering. The values of PAR. 1 were selected to meet the requirement that the number of clusters obtained by method Ml approximates such a number obtained by a division into equal intervals. In all groups of tests better results were obtained when a higher number of clusters (lower mean size of an interval) was allowed. Then, the intracluster variability *V* as well as the indicators *FS* achieve highest values at the test Nr 3.

6. Conclusion

The MELTEM allows to calculate melting temperatures (T_ms) of oligonucleotides of a given length and to generate subsets of oligonucleotides in a given DNA sequence. It was shown that the loader method is the best to realize the clustering of oligonucleotides. The numbers of oligonucleotides of a given T_m can be plotted as well as can be subjected to different clustering procedures resulting in several groups of T_m ranges. Such software might be very useful to reveal any particular clusters of T_m if present in DNA sequences of interest.

The G-C pairs content as well as a configuration of a hybridization experiment do not seem to have a significant effect on the quality of clustering.

Acknowledgments

This research has been supported by the KBN grant 8T11F01008p09.

References

[1] Rychlik, W., Rhoads, R.E., (1989) Nucleic Acids Res., **17**(21), pp. 8543-8551.
[2] Rychlik, W., *OLIGO - version 3.4-221, DNA/RNA Primer Selection Software,* 1989.
[3] Cantor, C.R., Schimmel, P.R., *Biophysical Chemistry. Part III,* W.H. Freeman and Co., San Francisco 1980.
[4] Breslauer, K.J., Frank, R, Blocker, H., Marky, L.A., (1986) Proc. Natl. Acad. Sci.U.S.A., **83**, pp. 3746-3750.
[5] Anderberg, M.R., *Cluster analysisfor applications,* Academic Press, New York 1973.
[6] Sneath, P.H.A., Sokal, R.R., *Numerical taxonomy,* W.H. Freeman, San Francisco 1973.
[7] Hartigan, J.A., *Clustering algorithms,* J. Wiley & Sons, New York 1975.
[8] Kucharczyk, J., *Algorytmy analizy skupień wjęzyku ALGOL 60,* PWN, Warszawa 1982.