An artificial immune system for the identification of polymer materials based on near infrared (NIR) spectra[*]

Małgorzata Rabiej[1), **)], Włodzimierz Biniaś[1)]

DOI: dx.doi.org/10.14314/polimery.2014.575

Abstract: This work presents an immune algorithm elaborated for the identification of polymers based on their NIR spectra. It uses the mechanisms and rules typical of natural immune systems. The identification of a polymer consists of a comparison of its NIR spectrum with reference spectra contained in a database. The algorithm acts in two stages. In the first stage, it compares the positions of the main absorption bands in the investigated spectrum with those of spectra from the database. Based on this comparison, the most similar reference spectra are selected. In the second stage, the shape of the numerical derivative of the investigated spectrum is compared with the shapes of the derivatives of the reference spectra selected in the first stage. Our investigations have shown that the algorithm is very effective and reliable. The algorithm can be used both for the identification of polymers in large databases and for the protection of such databases from an uncontrolled expansion.

Keywords: artificial immune system, immune algorithm, near infrared spectra, NIR, polymer identification.

Sztuczny system immunologiczny do identyfikacji materiałów polimerowych na podstawie ich widm w bliskiej podczerwieni (NIR)

Streszczenie: W pracy przedstawiono sztuczny system immunologiczny, należący do metod sztucznej inteligencji, przeznaczony do identyfikacji polimerów na podstawie ich widm w bliskiej podczerwieni (NIR). Analiza widm polega na porównaniu zgodności nieznanego widma z widmami zapisanymi w bazie danych, przy użyciu odpowiedniego algorytmu. Struktura algorytmu i działanie poszczególnych procedur naśladują strukturę naturalnego układu immunologicznego. Podobnie jak w naturalnych systemach, identyfikacja jest dokonywana na dwóch etapach. W pierwszym etapie przeprowadzana jest wstępna selekcja widm, realizowana za pomocą procedury odpowiadającej działaniu limfocytu T. Do wykonania tego zadania zastosowano filtry cyfrowe i pochodne numeryczne. W drugim etapie uruchamiana jest procedura odpowiadająca działaniu limfocytu B, której zadaniem jest wybranie spośród wyselekcjonowanych widm takiego, którego pochodna ma kształt najbardziej zbliżony do kształtu pochodnej nieznanego widma. W tym celu algorytm dzieli porównywane pochodne widm (antygen i przeciwciało) na małe fragmenty, czyli paratopy i epitopy, dla każdego z nich oblicza wskaźniki podobieństwa, takie jak: współczynnik determinacji (R^2), współczynnik Kendalla (τ), stosunek pól pod krzywymi (A), średni błąd względny (W) oraz sumę kwadratów różnic pochodnych. Na podstawie tych miar ocenia stopień dopasowania kolejnych paratopów i epitopów, a następnie stopień dopasowania antygenu i przeciwciała oraz poziom stymulacji. Zamiast klonowania i hipermutacji algorytm wykonuje lokalne przeszukiwanie każdego pasma widma. Widma, dla których stopień dopasowania przekroczy zadany próg są rozwiązaniem zadania. Przeprowadzone testy wykazały wysoką efektywność i niezawodność opracowanego algorytmu. Algorytm może służyć do identyfikacji polimerów w dużych bazach danych, a także do zabezpieczania baz danych przed wprowadzeniem kilku widm tego samego polimeru.

Słowa kluczowe: sztuczny system immunologiczny, algorytm immunologiczny, bliska podczerwień, NIR, identyfikacja polimeru.

Infrared spectroscopy covers electromagnetic radiation from a wavelength of 12 500 cm^{-1} (800 nm) to 10 cm^{-1} (1000 μm) [1, 2]. This wide range is divided into 3 subranges:

[1)] University of Bielsko-Biała, Willowa 2, 43-309 Bielsko-Biała, Poland.

[*)] Material contained in this article was presented at the IX International Conference „X-Ray investigations of polymer structure", Zakopane, Poland, 3—6 December 2013.

[**)] Corresponding author; e-mail: mrabiej@ath.bielsko.pl

— near infrared (NIR): 12 500—4000 cm^{-1} (800 nm—2.5 μm)

— mid (classical) infrared (MIR): 4000—400 cm^{-1} (2.5 μm—25 μm)

— far infrared (FIR): 400—10 cm^{-1} (25 μm—1000 μm).

Near infrared radiation is much more penetrating than classical medium infrared (MIR) and visible (VIS) radiation. NIR spectroscopy provides information on the chemical composition, physical state and various properties of the investigated materials. As a result, it can be used for qualitative and quantitative analyses. The investigated samples may have different states and shapes: solids, liquids, emulsions, powders, foils etc. For these reasons, NIR is used for the rapid spectroscopic investigation of many samples of polymers with little or no sample preparation. NIR is also frequently employed for the classification and selection of polymer wastes. The identification of a polymer starts with a comparison of its NIR spectrum with reference spectra contained in a database by means of suitable mathematical algorithms. However, it should be taken into account that the shape of a polymer NIR spectrum depends not only its chemical and physical structure but also on the thickness of the sample and acquisition parameters. The spectra of the same polymer may slightly differ in baseline, absorption of individual bands and shifted positions of the band maxima. The effects connected with shifts and distortions of the base line can be removed by transforming the spectra to their first derivatives. For this, the Savitzky-Golay [3] derivation algorithm is most frequently used.

Commercial NIR spectrometers are usually supplied with computer programs for a rapid comparison of the recorded spectra with those contained in the database. Unfortunately, these programs do not always identify the correct materials. Most probably, a low success rate results from the fact that, in comparing spectra, the programs use only one measure of similarity, e.g.: correlation coefficient, Euclidean distance etc., missing the simultaneous comparison of the locations of bands characteristic of a given polymer and other factors characterizing the shape. To make the results more reliable, the comparison should be more comprehensive, i.e. various elements of the shape of spectra should be quantified and analyzed.

In this work we propose an immune algorithm to identify polymers based on their NIR spectra. The algorithm acts according to the schemes met in natural immune systems called artificial immune systems (AIS). The elaborated system treats an analyzed spectrum as a picture and uses the mechanisms of immune picture recognition [4] for its identification.

DESCRIPTION OF A NATURAL IMMUNE SYSTEM

The immune system of a living organism has to protect against different risks, i.e. pathogens such as viruses, bacteria, parasites and fungi. The cells of an organism, participating in the detection and destruction of pathogens are called lymphocytes. The pathogens recognized by lymphocytes are called antigens. A general scheme according to which the natural immune system acts comprises two stages. In the first stage, all cells of the organism are monitored and antigens are recognized. This task is performed by T lymphocytes, which are the main detectors and responsible for the identification of antigens. The T lymphocytes distinguish the cells of the organism from alien cells.

In the second stage, B lymphocytes precisely recognize and destroy alien cells. On the surfaces of B lymphocytes, there are thousands of receptors (antibodies) that recognize ones own and alien cells. The antigen is recognized when the shapes of the receptor and antigen fit well. A specialized part of a receptor used for the identification of other investigated cells is called the paratope. The part of an antigen to which a paratope can be attached is called the epitope. The precision of matching and strength of the connection between a given paratope and epitope is characterized by the degree of fitting. When the degree of fitting exceeds a threshold value (stimulation level), the B lymphocyte is activated.

Stimulated lymphocytes are intensively cloned so that all antigens can be destroyed. Next they mutate in order to make the immune system capable of destroying not only the initial antigen but also molecules containing similar epitopes.

ORGANIZATION OF THE ARTIFICIAL IMMUNE SYSTEM

Artificial immune systems (AIS) [4—6] are not reproductions of natural systems. They only use some rules and principles characteristic of such structures.

The AIS system elaborated in this work aims to identify a polymer on the basis of its NIR spectrum. The spectrum is treated as a picture. The system analyses a picture taken from the database and compares it with the spectrum of the unknown polymer.

The main components of the AIS are:

— pathogens — data sets or data structures describing an individual reference spectrum, loaded from the database;

— standard — spectrum of the unknown polymer which is to be recognized, loaded as a text file;

— antigen — the pathogen recognized by the T lymphocyte as similar to the standard;

— T lymphocyte — numerical tools searching for, and reacting to, important characteristics of antigens. T lymphocytes compare the positions of the main absorption bands in the investigated spectrum with those in the reference spectra from the database;

— antibody — recognizes important characteristics of antigens, in this system a numerical derivative of a spectrum is an antibody;

— B lymphocyte — a procedure that finds the antigen best fitted to the antibody. Such an antigen is a solution of the identification task.

Both pathogens and antibodies are presented as real number vectors.

Cloning and hypermutation were substituted with a local search of common bands.

T lymphocyte

T lymphocyte compares the positions of the main absorption bands in the investigated spectrum with those in the reference spectra loaded from the database.

The T lymphocyte performs the following tasks:

— loads a text file with the spectrum that is to be recognized (standard);

— smoothes the spectrum by means of a variable-digital filter [7, 8];

— transforms the spectrum by means of a band-pass filter [8, 9] to enhance the absorption bands and reduce the remaining elements of the curve;

— determines the positions (wavenumbers) of the main bands by means of the Savitzky-Golay numerical derivative;

— loads the reference spectra (pathogen) from the database;

— compares the positions (wavenumbers) of the main bands in the investigated spectrum (standard) and reference spectra (pathogen);

— calculates the number of common bands in the standard and in all pathogens;

— classifies the pathogens; the pathogen in which the positions of the main bands are the same or similar is chosen as an antigen. Note that, several antigens may be chosen.

B lymphocyte

B lymphocytes have only one antibody, which is represented by a numerical derivative of the investigated NIR spectrum. The main task of this lymphocyte is to compare in detail the antibody and antigens, to isolate the antigens that are best fitted to the antibody and to memorize them as a separate set of data. To make such a comparison, the antigens chosen by the T lymphocyte (selected reference spectra) must be transformed into their numerical derivatives.

Next the B lymphocyte performs the following tasks:

1. Divides the wavenumber range 10 500—5500 cm^{-1} in which the curves representing the derivatives of investigated spectrum and selected reference spectrum are defined into small sub-ranges of identical widths. Each pair of segments related to the same sub-range of wavenumber is compared quantitatively. The method of comparison is the same as that used in our earlier work [10] dedicated to the identification of polymers based on their X-ray diffraction curves.

2. For each pair of spectra composed of a paratope (segment of the investigated spectrum) and an epitope (segment of an antigen) related to the same sub-range of wavenumber, the B lymphocyte calculates four well known measures of similarity [10]:

— Coefficient of determination R^2 — one of the basic measures describing the quality of fitting of two curves. Its value lie in the range [0; 1]. The closer the coefficient is to unity, the better fitted is the theoretical function.

— Kendall's tau coefficient τ [11] — describes the similarity of shapes of two curves in a given sub-range based on the conformity of the signs of all successive differences between the curves.

— Average relative error W — describes the relative differences between the two curves in a given sub-range.

— Area coefficient A —the approximated areas under the derivatives of the spectra are numerically calculated using a trapezoidal rule. For a given sub-range, an area coefficient is calculated as the ratio of the two areas mentioned above (smaller:larger).

These measures are commonly used in statistics for model verification. They are easily interpreted because their values range from 0 to 1 (coefficient of determination and area coefficient) or from -1 to 1 (Kendall's coefficient). Only in the case of extremely poor fitting will the average relative error W take values greater than 1. For this reason, these coefficients clearly reflect the quality of fitting.

The ranges of all measures have been divided into the same number of sub-ranges (8-th). Based on a local value of a given measure, a "mark" is assigned to the sub-range. The mark is an integer number (rank) describing the quality of fitting of the curves related to this sub-range. Marks can be positive or negative. A negative value means that the curves are poorly fitted and a positive value means good fitting. Table 1 shows marks corresponding to different values of the four measures of similarity. The marks were established arbitrarily [10], based on the results of the performed tests.

Table 1. Measures of similarity

Mark (M_i)	R^2 determination coefficient		W average relative error		τ Kendall's coefficient		A area coefficient	
	values of measures							
	from	to	from	to	from	to	from	to
-4	0	0.4	2	∞	-1	-0.6	0	0.4
-3	0.4	0.5	1	2	-0.6	-0.3	0.4	0.5
-2	0.5	0.6	0.8	1	-0.3	0	0.5	0.6
-1	0.6	0.7	0.6	0.8	0	0.2	0.6	0.7
0	0.7	0.8	0.4	0.6	0.2	0.4	0.7	0.8
1	0.8	0.85	0.3	0.4	0.4	0.5	0.8	0.85
2	0.85	0.9	0.2	0.3	0.5	0.6	0.85	0.9
3	0.9	0.95	0.1	0.2	0.6	0.7	0.9	0.95
4	0.95	1	0	0.1	0.7	1	0.95	1

3. Based on these measures of similarity defined above, the B lymphocyte calculates a local degree of simi-

larity (*LDS*) for each segment as the sum of four marks (M_i) assigned to the sub-range:

$$LDS = \sum_{i=1}^{4} M_i \tag{1}$$

4. Calculates a total degree of similarity (*TDS*) for the antibody and each antigen. This parameter must take into account not only the values of *LDS* determined for individual sub-ranges but also the lengths of chains formed from sub-ranges in which the curves are well fitted and the lengths of chains in which the curves are poorly fitted. The longer the chains are composed of well fitted sub-ranges, the higher should be the *TDS*. On the other hand, long chains of poorly fitted sub-ranges should act in the opposite way. This is why the *TDS* is constructed as a sum of three elements [see eq. (2)]. The first one is the sum of *LDS*'s for all sub-ranges into which the wavenumber range: 10 500—5500 cm^{-1} has been divided. The second takes into account the lengths of chains formed from well fitted segments (with $LDS \geq 3$), and the third element refers to the lengths of chains formed from poorly fitted segments (with $LDS \leq 0$)

$$TDS = \sum_{i=1}^{n} LDS_i + \sum_{j=2}^{n} g_j 2^j - \sum_{j=2}^{n} b_j 2^j \tag{2}$$

where: n — number of all sub-ranges in the wavenumber range 10 500—5500, g_j — number of chains formed of j sub-ranges with $LDS \geq 3$, b_j — number of chains formed of j sub-ranges with $LDS \leq 0$.

The *TDS* is calculated for all antigens chosen by the T lymphocyte, i.e. for all reference spectra selected in the first stage of identification.

5. Calculates a similarity level. To choose the antigens that are most similar to the investigated spectrum, the algorithm calculates the Euclidean distances between the derivatives of the investigated spectrum and derivates of all antigens with positive *TDS*. Based on the obtained results, another parameter called a similarity level (*SL*) is calculated. It is given by the following formula:

$$SL = TDS + N \cdot 10 + P \tag{3}$$

where: N — the number of common bands found by the T lymphocyte in the first stage of the process.

This parameter includes three components. The first one — *TDS*, depends on the quality of fitting of the sub-ranges. The second — $N{\cdot}10$, depends on the conformity of the band positions (N is multiplied by 10 to make the weights of both criteria close to each other). The third component — P is a bonus amounting to 50 points. It is given to only one antigen — the one for which the Euclidean distance from the experimental curve is the smallest.

As one can see, the final factor which quantitatively expresses the similarity of a reference spectrum to the investigated one, takes into account the total degree of similarity (*TDS*), the sum of absolute values of differences between the derivatives (point by point) and the number of common bands. Thanks to such a comprehensive comparison, the algorithm unambiguously and quickly identifies the polymer in the database.

EXAMPLES OF APPLICATION

The AIS system described in this paper was implemented by a computer program and tested for several spectra of various polymers. A test database was created to carry out the assessment. It contained smoothed spectra of various polymers and their blends. FT-NIR spectra were recorded by means of a Nicolet Magna-IR 860 spectrophotometer in the wavenumber range of 10 500—5500 cm^{-1} with a resolution of 16 cm^{-1}. Each spectrum was an average of 250 measurements.

The performed tests showed that the algorithm correctly identifies unknown polymers. To estimate the effectiveness of the method of comparison and selection of spectra used in this algorithm, the program was equipped with three additional options in which the experimental spectra were compared to the reference spectra based on other measures of similarity. These options are: the correlation coefficient, the sum of absolute differences between derivatives and the sum of squared differences between derivatives. These measures were chosen because they are used in the commercial software OMNIC. A user of this software can choose one of them to identify an unknown polymer on the basis of its NIR spectrum. It turned out, that in contrast to the AIS system, the three alternative measures of similarity mentioned above do not always give the correct solution.

For example, the spectrum of polyacrylonitrile (PAN) sample shown in Fig. 1, loaded as an "unknown" into the program, was wrongly recognized using the first (correlation coefficient) and the second (sum of absolute differences between derivatives) measure of similarity. Fig. 2 presents this "unknown" spectrum compared with the reference spectrum of PAN taken from the database. As one can see, both spectra and their derivatives differ considerably. This was the reason for which the "unknown" spectrum was wrongly recognized as poly(ethylene terephthalate) (PET) on the basis of the first measure (correlation coefficient) and as a blend of Wool 20 % PAN 80 % on the basis of the second measure.

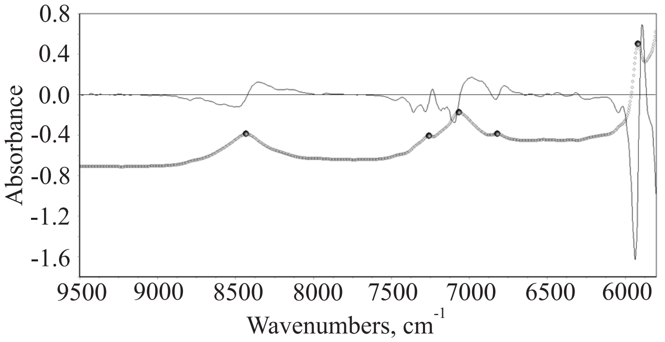

Fig. 1. Spectrum of PAN sample (thick line) loaded as "unknown" into the program and its derivative (thin line); the algorithm determines the maxima of the main bands (circles) by means of band filters

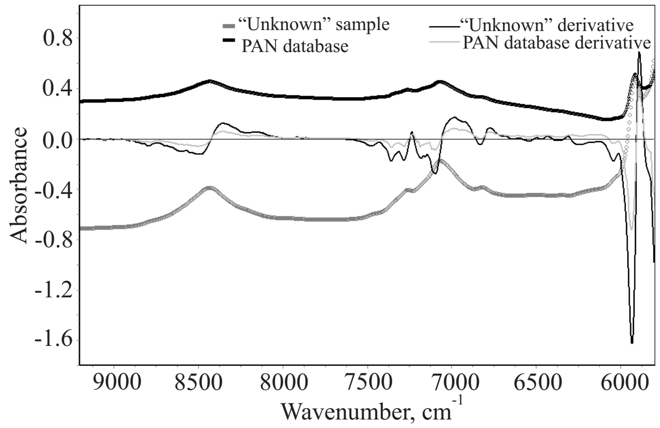

Fig. 2. Spectrum of the "unknown" PAN sample and its derivative compared with the reference spectrum of PAN taken from the database and its derivative

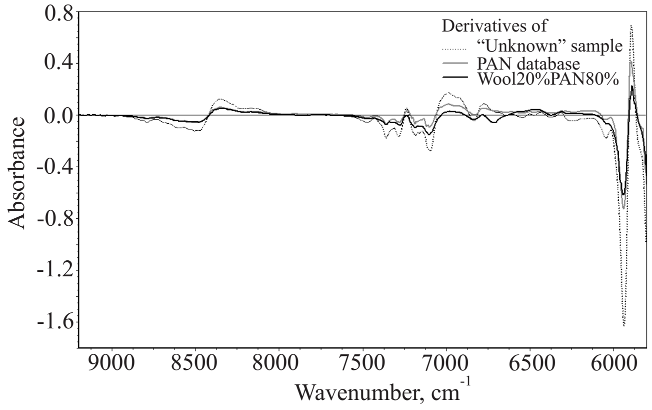

Fig. 3. Derivatives of "unknown" PAN spectrum, reference spectrum of Wool 20 % PAN 80 % blend and reference spectrum of pure PAN

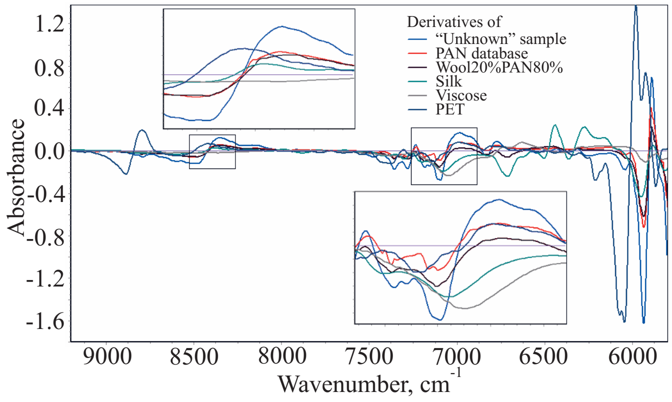

Fig. 4. Derivative of the "unknown" PAN sample and derivatives of the reference spectra chosen from the database as antigens, i.e. as most similar to the "unknown" spectrum

Fig. 3 shows that in fact, the derivative of the reference spectrum of PAN and that of the blend Wool 20 % PAN 80 % are very similar to one another and they clearly differ from the derivative of the "unknown" spectrum. Despite these difficulties, the AIS system has correctly identified the "unknown" sample as PAN. In Fig. 4, one can see the derivative of the "unknown" PAN spectrum and the derivatives of the reference spectra, chosen from the database by the T lymphocyte as antigens i.e. as the most similar to the first one. The final similarity level [see eq. (3)] calculated by the B lymphocyte for the reference spectrum of PAN amounted to 598 points (TDS = 508, number of common bands N = 4, N*10 = 40, P = 50), while the spectrum of blend Wool 20 % PAN 80 % gave 585 points (TDS = 505, N = 8, N*10 = 80).

SUMMARY

The AIS system described in this paper is an effective and reliable tool for the identification of polymers on the basis of their NIR spectra. The division of the identification task into two stages and a comprehensive evaluation of the shapes of derivatives by means of different measures of similarity considerably increases the chance of a correct solution. The system is fast because a detailed comparison of the shapes is performed for only a few, most similar spectra and less similar ones are eliminated in the first stage of the procedure. The performed tests showed that the algorithm works correctly even without data on the main band positions, i.e. when identification is based only on the comparison of the shapes of derivatives of the spectra. This is very important as it can be used for the protection of databases from an uncontrolled expansion. When the database already contains a spectrum of similar shape to one that you wish to add, it will be automatically rejected by the algorithm.

REFERENCES

[1] Pasquini C.: *J. Braz. Chem. Soc.* 2003, *14*, 198, http://dx.doi.org/10.1590/S0103-50532003000200006

[2] Hindle P.H.: *Proc. Control Qual.* **1997**, *9*, 105.

[3] Savitzky A., Golay M.J.E.: *Anal. Chem.* **1964**, *36*, 1627, http://dx.doi.org/10.1021/ac60214a047

[4] Wierzchoń S.T.: „Sztuczne systemy immunologiczne", Akademicka Oficyna Wydawnicza EXIT, Warszawa 2001, pp. 87—106.

[5] "Artificial Immune Systems and Their Applications" (Ed. Dasgupta D.), Springer Verlag, Berlin 1999.

[6] De Castro L.N., Timmis J.: "Artificial Immune Systems: A New Computational Intelligence Approach", Springer 2002.

[7] Biermann G., Ziegler H.: *Anal. Chem.* **1986**, *58*, 536, http://dx.doi.org/10.1021/ac00294a008

[8] Rabiej M.: *Solid State Phenomena* **2013**, *203—204*, 189, http://dx.doi.org/10.4028/www.scientific.net/SSP.203-204.189

[9] Jones R.: *Analyst* **1987**, *112*, 1495, http://dx.doi.org/10.1039/an9871201495

[10] Rabiej M.: *J. Appl. Cryst.* **2013**, *46*, 1136, http://dx.doi.org/10.1107/S0021889813015987

[11] Kendall M.G.: "Rank Correlation Methods", Charles Griffin & Company Limited, London 1948.