Adaptive Multistage Method of Anomalies Detection in ECG Time Series

Igor Baklan[1[0000-0002-5274-5261]] and Yurii Oliinyk[1[0000-0002-7408-4927]] and Iryna Mukha[1[0000-0002-4423-5106]] and Kateryna Lishchuk[1[0000-0002-9902-0065]] and Olena Gavrilenko[1[0000-0003-0413-6274]] and Oleksandr Ocheretianyi[1[0000-0001-9455-4781]] and Anna Tsytsyliuk[1[0000-0001-5527-4154]]

[1]Igor Sikorsky Kyiv Polytechnic Institute, Kyiv, 03056, Ukraine
iaa@ukr.net

Abstract. This work proposes a new approach for identifying heart anomalies on electrocardiograms data using adaptive multistage method of anomalies detection. The method includes: search of exact match pattern with fragments in linguistic chain, analysis by average distance, analysis by fuzzy distance, reduction to a common time grid. Used various matrix of linguistic distances. Described source data and database process filling.

Keywords: ECG, Linguistic modeling, Linguistic model, Linguistic chain.

1 Introduction

Cardiovascular disease is one of the leading causes of disability and mortality worldwide. However, due to proper diagnostics, it is possible to detect possible abnormalities of the heart activity in a timely manner and take all necessary steps to restore its normal functioning as soon as possible. One of the main instrumental methods for diagnosing cardiovascular diseases is electrocardiography - a graphical recording of fluctuations in the potential difference that occurs during cardiac function in the form of an electrocardiogram (ECG). An ECG analysis for abnormalities - areas on the curve that do not fit well-defined concepts of normal behavior - reveals pathological changes in the functioning of the heart muscle.

The modern approach to ECG analysis is to automate the process of identifying possible anomalies, which improves the accuracy and reliability of diagnostic results. Anomaly detection is one of the most important tasks of data mining technology. Computerized ECG processing and analysis involves the use of mathematical methods, most of which are based on the interpretation of ECG time series data. Statistical models (AR, ARIMA, Kalman filters), fuzzy models (FRCM, SSOD, FUCOT algorithms), hidden Markov processes, clustering and classification (c-mean, k-mean clustering algorithm, sliding method) are used to detect compression ratio (CR) anomalies. [22]

Anomalies Detection Approach in Electrocardiogram Analysis Using Linguistic Modeling [13] proposes a new approach - linguistic modeling of ECG data. This ap-

proach involves constructing a linguistic ECG model by converting ECG data into symbolic (linguistic) sequences, based on which a formal language can be constructed to solve the applied problems of analysis. This article proposes an algorithm for finding and detecting anomalies in the ECG based on the analysis of the linguistic ECG model, built on the selected parameters of linguistics. This algorithm involves the analysis of the received language circuits for the detection of atypical patterns (values) and fuzzy local trends.

The structure of the article is as follows. The first section discusses the problem of anomalies search and detection in CR, as well as the approaches used. The second section describes the process of constructing a linguistic ECG model. The third section presents an algorithm for finding and detecting anomalies in an ECG based on linguistic modeling, taking into account anomalies in fuzzy local trends. The fourth section presents the results of evaluating the effectiveness of the proposed solutions.

1.1 Related Works

ECG analysis is of great importance in medicine and is beginning to attract more attention and become involved in various fields. For example, there are studies [1] on the identification of a person in a security system through analysis of his ECG. There, a method was proposed for cropping PP intervals in which ML training was carried out using each sample of slice data as an input parameter. Also, work was done [5] to produce a web-based tool for distributed ECG annotations. This cross-platform system can be used to jointly view and annotate ECG, which will simplify communication between specialists. Another real-time approach for public health is the real-time approach based on multi-layer perceptron classifier [7]. Based on it, the distribution of incoming ECG cardiac contractions into one of 23 classes occurs using ECG sparse distributed signatures.

In addition, ECG analysts have found application in the study of sleep and its stages. Thus, the use of ECG and respiratory signals for determining the stage of sleep using deep neural networks was described in [11]. Five ultra-precise neural networks were created, each for different types of respiration. This study can help in general when studying sleep and processes in the body at its different stages. The main problem scientists faced in this study was a decrease in accuracy due to age and / or more severe sleep apnea.

Nowadays, a lot of research is underway to improve the analysis and processing of ECG data. In general, neural networks are mostly used for analytics because of its ability to process complex and fuzzy data.

There are several problems when studying ECG results - extraneous noise, truncating a data set without losing information for training networks, and analyzing the information itself.

Firstly, Complex Deep Learning Models can be useful not only for data analysis but also for removing various types of noise (i.e. baseline wander, muscle artefact, electrode motion artefact) [10]. Based on the results of noise reduction tests between two DL models (CNN, LSTM). CNN had the best indicator both in synthetic and real data.

Therefore, the study of the best and most productive CNN model in the same study becomes relevant.

Secondly, for the analysis of data in most cases records of the same length are necessary - accordingly there are several methods for bringing records to the same size.

- Filling all records to the longest length.
- Shortening all records to the shortest.
- Grouping records by length.
- Trimming or padding data to a specific length [8].
- Heuristic-based crop [6].

For the analysis itself, a large number of new methods are described. Thus, in [2] a time series clustering algorithm is described that is suitable for multidimensional inputs and outputs of variable lengths and time offsets. This clustering is based on the distance between hidden linear dynamic systems (LDS). An example of the usage of the critical oscillation method for ECG analysis was carried out in [4], where the choice of ECG segments was made with the stationarity criterion. The heart was examined on the physical side and critical dynamics in frog's ECG has been detected in the high frequency.

In [12], a new tool was invented for obtaining new information, which is based on a new method of data analysis - Diverse learning. This technique extracts a structure from data and belongs to uncontrolled machine learning methods. So from the raw data, you can outline the dynamics of the process inside and the possibility of 3D visualization of DM appears. (Where each pulse is a point)

In another research [3], it was proved, using recurrence networks that the RNs from ECG f is bimodal and thus forms a separate class. In addition, this research shows that RNs from ECG have significantly higher value for the clustering coefficient and lower value for the average path length.

In the end, we can say that for ECG analytics it is important to preserve data when changing the recording time, eliminate unnecessary noise for corrective analysis, and optimize training algorithms and improve the analysis itself. the research results revealed that modern approaches for ECG analysis are convolutional neural networks, recurrence networks, time series, the critical oscillation method and new methods for training neural networks (for example, diverse training). The greatest advantage of neural networks is the ability to analyze new data based on a generalization of previous cases, but at the same time it is also a drawback because a lot of data and time are needed for training.

1.2 Researches Tasks

Main goal: to enhance the precision and confidence of diagnostic results about the level of the cardiovascular system as a result of the progress of computation approaches of ECG feature analysis.

The research purposes are:

- development of the anomaly detection algorithm;
- construction of anomaly patterns database.

2 Anomaly Detection Algorithm

We represent search algorithm having into linguistic row fragment linguistic pattern. It was important definition proximity criteria of linguistic rows and proximity level by expert judgment methods.

We consider next algorithm of search elements from pattern based in linguistic chain being investigated:

1. Search of exact match pattern with fragments in linguistic chain. If some pattern is found in linguistic chain, then this found substring marked as anomaly.
2. Analysis by average distance. If some linguistic chain substring has defined average distance level with pattern, then this found substring marked as anomaly.
3. Analysis using Analytic Hierarchy Process (AHP) method T.Saaty.
4. Reduction to a common time grid.
5. Pattern matching with input signal Availability, when it is shift operator. Shift operator used for shift linguistic chain or substring by input signal. For example, chain "CDE" transforms in "BCD" in case shift down on 1 step, and in "DEF" in case shift up on 1 step. If some pattern is found in linguistic chain with using shift operator, then this found substring marked as anomaly.
6. Pattern matching with input signal scaling operator. This operator scale linguistic chain by the time (x axis) or signal level (y axis).

Based on the obtained coincidence, an expert makes conclusion about the revealed abnormality in the heart.

The task of fuzzy string matching will be listed next: considering a (extensive) text T of range n, and a (small) sample P of length m. They both are series of symbols from an alphabet Σ of power σ, and a supreme amount of variances allowed k, discover all the subsequences of T which edit size to P is no more than k. [24,25] Those fragments are named "instances" and it is usual to account only their begin or finish marks. The change interval between two strings x and y is the least amount of variances that would modify x into y or in reverse. The allowed variances are deletion, insertion and substitution of symbols. The task is important for $0 < k < m$. The variance rate is determined as $\alpha = \frac{k}{m}$. Finally, if "change interval" among current strings (*ed*()) implies the edit interval, we can to account begin points (i.e. {$|x|$; *T=xP'y; ed(P; P')*≤*k*}) or finish points (i.e. {$|xP'|$; *T=xP'y; ed(P; P')*≤*k*}) of instances. [20]

For the average distance, we used some classic estimates of the closeness of linguistic chains. Which are:

1. Hamming distance is the amount of points where the equivalent numbers of two binary words of the identical length are dissimilar. In general, Hamming distance is used for lines of the same length in an alphabet consisting of q characters and serves as a difference metric (a function that determines the distance in metric space) from objects of the same dimension. In other words, Hamming distance measures the minimum number of replacements required to change one line to another or the minimum number of errors that could convert one tape to another. In a more general

context, Hamming distance is one of the chain metrics for measuring the machining distance between two sequences. [14].

2. Levenstein distance is the minimum number of character deletions, inserts, and replacements required to convert one line to another. [15].
3. Another way to formalize the difference between words is using the Jarro-Winkler distance. Each line character is compared to all other corresponding line characters. The number of matching but different atomic numbers is divided by 2 and determines the number of transpositions. [16].
4. Damerau-Levenstein distance is a chain metric for measuring the processing distance between two sequences. In short, the Damerau-Levenstein distance between two words is the minimum number of operations required to change one word to another. [17][18][19].
5. Mahalonobis distance is a measure of the distance between vectors of random variables. This measure is often used to determine the similarities between an unknown and a known sample.

Average distance calculated by normalized this distance:

$$Adist = \sum_{i=1}^{5} Dist_i,$$

where $Dist_i$ are distances of Hamming, Levenshtein, Damerau–Levenshtein, Jaro–Winkler, Mahalonobis.

As a result, we find the patterns that exceed the expert level determined using appropriate methods.

In the process of teaching the system to recognize the types of anomalies, the data of various datasets were used. It turned out that the same linguistic chain in different datasets can correspond to different types of anomalies. We propose to include as one of the step of our algorithm the usage of method of multycriteria choice.

A large number of mathematical methods are proposed to solve the tasks of multycriteria choice. Applied decision theory methods differ in the way they present and process expertise, but they are all designed to help make an effective decision. Among the existing methods, the analytical hierarchy process (AHP) proposed by T. Saati [22] is quite universal and theoretically grounded. This method is simple, the method of solving the problem corresponds to the intuitive presentation of the problem being solved. The advantages of this method can also be attributed to the ease of its implementation at the software level. But this method has several major disadvantages: first, the limitation on the number of alternatives that are compared at the same time, and secondly, there is the problem of hierarchy coherence, and third, a large amount of expert information. In addition, when applying this method, there is a problem that the real matrices of pairwise comparisons are usually not completely consistent. Therefore, all these facts give rise to the problem of the limitations of the application of this method, but in [23] methods for solving these problems and removing restrictions on its application are proposed.

Therefore, in order to determine which type of anomaly this or that linguistic chain most closely corresponds, it was decided to use the AHP by T. Saati.

In our case, various linguistic chains act as alternatives, types of anomalies (for example, tachycardia, arrhythmia, etc.) act as criteria, datasets as experts.

3 Data Sources for Filling Anomalies Linguistic Chain Database

In the present study, the ECG signals are obtained from MIT-BIH Arrhythmia dataset [21]. The MIT-BIH Dataset was primary set of traditional experimental samples to estimate EKG deviations. Since 1980, this set of data was used for the initial study for heart motion all over 500 researches globally [20]. This dataset contains 48 half-hour extracts of dual path, 24-hour ECG records obtained from 47 patients observed at the BIH Arrhythmia Laboratory. Our study is based on six samples from the arrhythmia dataset containing files 101, 106, 112, 138, 200 and 222 that includes sufficient ordinary beats, PVC and APC arrhythmia for the research. It is discovered 3 sorts of the arrhythmias included here containing normal beat (NORMAL), premature ventricular contraction (PVC) and arterial premature contraction (APC). The given ECG signal is measured with 360Hz. The categorization of foregoing sorts is examined considering that they are more potential to be misunderstood by calculation device against other signal profiles. ECG signals were outlined by binary annotation file (.atr), a binary file (.dat) and a text header file (.hea),. Most of samples contain a binary file, which has scanned examples of at least one signal saved in 212 templates, and most samples contain at least one annotation file. Annotation files hold group of marks. Each of them depict a component of at least one signal at a particular time in the sample. In our research, each model, ECG beats are chosen by chance out of six files so that database is collected of 1,000 beats of every sort. On **Fig. 1** shows "MNPQSTUWXXVRNLLMNNNN" Arrhythmia anomalies linguistic chain.

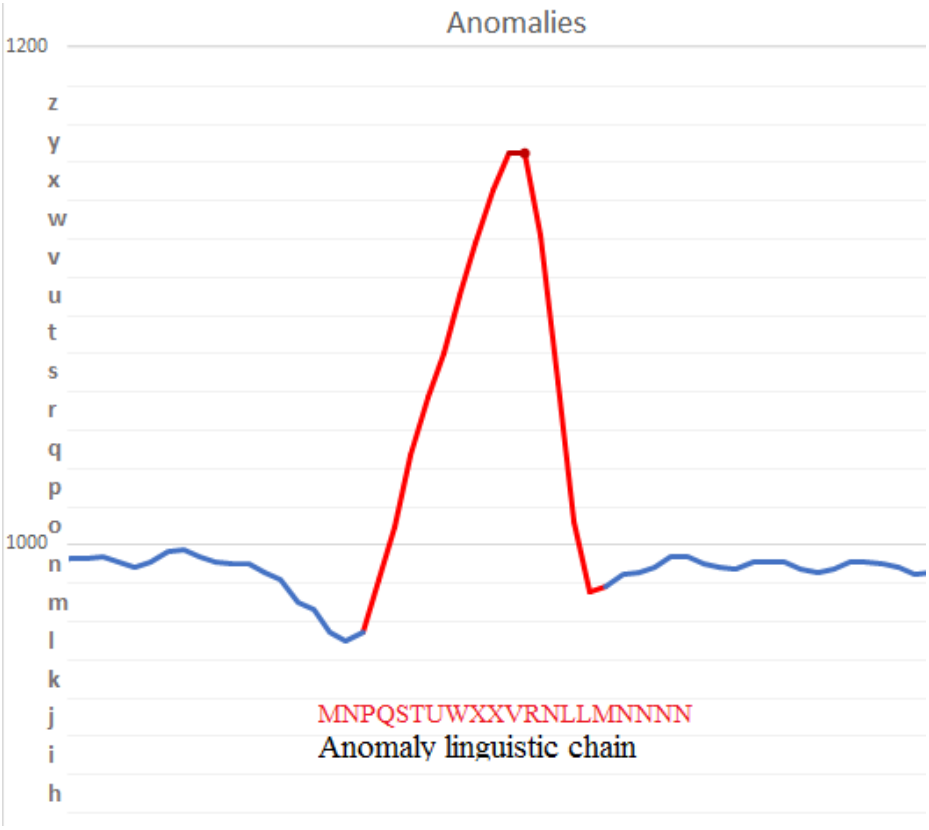

Fig. 1. Arrhythmia anomalies linguistic chain example

On **Fig. 1** shows 2 data series: series1 with anomaly linguistic chain "ggijlnopponjhhiiii", series 2 –contains unmarked linguistic chain "ggijlmopqpnjhhiiii". These 2 linguistic chains have big closeness, therefore series2 has anomaly with big probability.

The linguistic Alphabet1 [a-z] contains 26 symbols and doesn't allow create adequate linguistic chain. For example, many data series elements have only one value 'iii' (**Fig. 1**). For avoid this problem need to increase alphabet capacity. Linguistic Alphabet2 contains 100 unicode symbols (from 192 to 292).

On **Fig. 1** shows 2 data series: series1 with anomaly linguistic chain "èéîóûĀąċĐěęđĂôíïïïð", series 2 –contains unmarked linguistic chain "éêîõüāćčĔěĠğĕăôîîðïïð". These 2 linguistic chains have big closeness too, therefore series2 has anomaly with big probability.

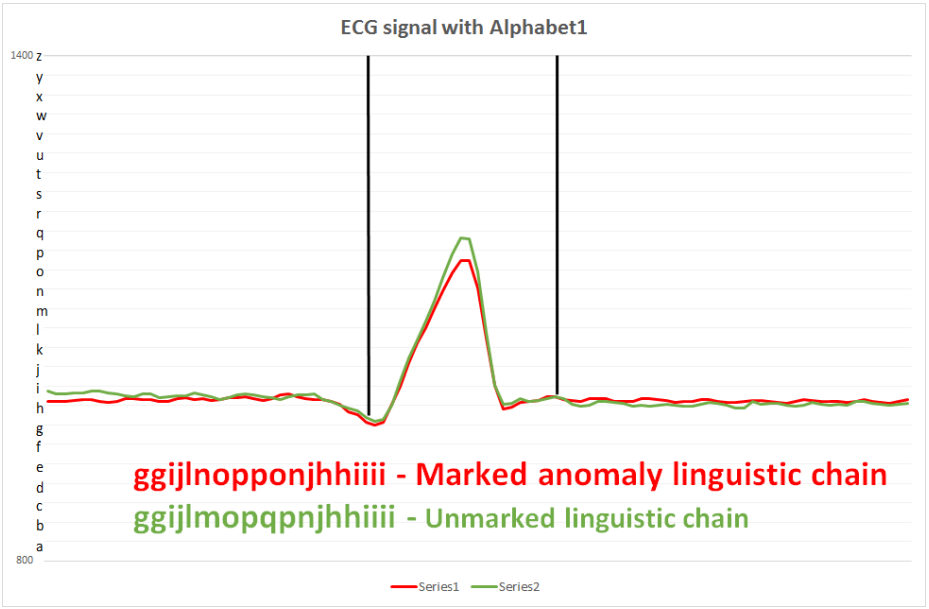

Fig. 2. Linguistic chains comparison for Alphabet1

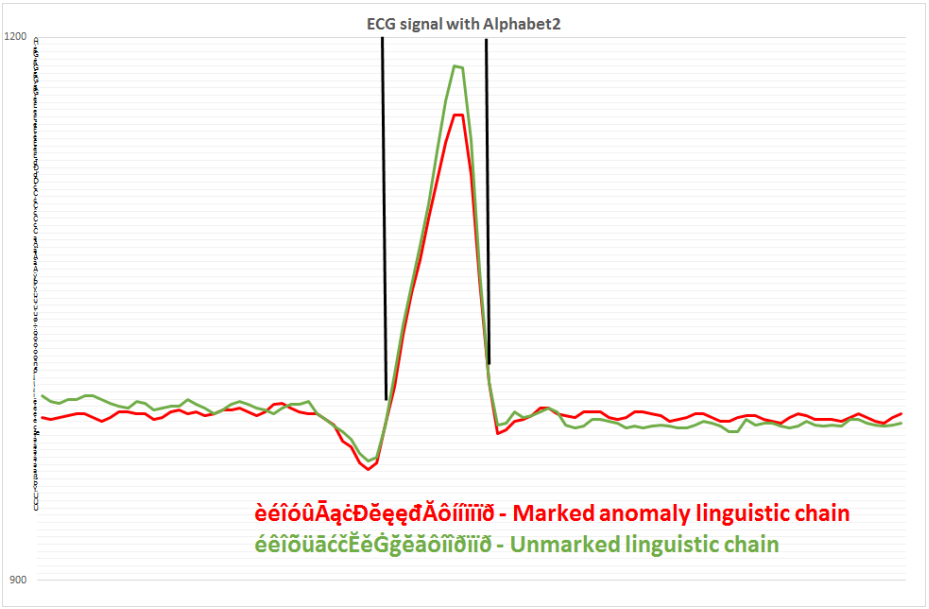

Fig. 3. Linguistic chains comparison for Alphabet2

Table 1. Translation ECG signals to linguistic chain

ECG signal level	961	965	986	1007	1036	1059	1077	1100	1121	1142	1157	1157	1124	1065	1009
Alphabet1	g	g	i	l	n	o	p	p	o	n	j	h	h	i	i
Alphabet2	è	é	î	ó	û	Ā	ą	ċ	Đ	ě	ę	ę	đ	Ă	ô

Table 1 contains translation of ECG signals with arrhythmia anomalies to linguistic chains. Linguistic chain based on Alphabet2 more complicated then based on Alphabet1.

4 Conclusion

We proposed adaptive multistage method of anomalies detection in ECG time series based on using linguistic modeling. For finding of the closeness of linguistic chains we used combination of classic estimates such as Hamming, Levenshtein, Damerau–Levenshtein, Jaro–Winkler, Mahalonobis with AHP. Our approach allows to find various anomalies related to many heart diseases.

References

1. Yoo, P. D. An Enhanced Machine Learning Based Biometric Authentication System Using RR-Interval Framed Electrocardiograms. arXiv preprint arXiv:1907.13517. (2019).
2. Hsu, C. C. Y., Hardt, M., & Hardt, M. Linear Dynamics: Clustering without identification. arXiv preprint arXiv:1908.01039. (2019).
3. Kachhara, S., & Ambika, G. Bimodality and Scaling in Recurrence Networks from ECG data. arXiv preprint arXiv:1908.01286. (2019).
4. Contoyiannis, Y., Diakonos, F., & Kampitakis, M. Applying the Method of Critical Fluctuations on Human Electrocardiograms. arXiv preprint arXiv:1908.06408. (2019).
5. Ding, Z., Qiu, S., Guo, Y., Lin, J., Sun L., Fu D., & Lv T. LabelECG: A Web-based Tool for Distributed Electrocardiogram Annotation. arXiv preprint arXiv:1908.06553. (2019).
6. Yuan, B., & Xing, W. Diagnosing Cardiac Abnormalities from 12-Lead Electrocardiograms Using Enhanced Deep Convolutional Neural Networks. arXiv preprint arXiv:1908.06802. (2019).
7. Das, A. K., Catthoor, F., & Schaafsma, S. Heartbeat Classification in Wearables Using Multi-layer Perceptron and Time-Frequency Joint Distribution of ECG. In 2018 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE) (pp. 69-74). IEEE. (2018).
8. Liu, Y., He, R., Wang, K., Li, Q., Sun, Q., Zhao, N., & Zhang, H. Automatic Detection of ECG Abnormalities by using an Ensemble of Deep Residual Networks with Attention. arXiv preprint arXiv:1908.10088. (2019).
9. Ding, X., Yan, B. P., Zhang, Y. T., Liu, J., Su, P., & Zhao, N. Feature Exploration for Knowledge-guided and Data-driven Approach Based Cuffless Blood Pressure Measurement. arXiv preprint arXiv:1908.10245. (2019).
10. Arsene, C. Complex Deep Learning Models for Denoising of Human Heart ECG signals. arXiv preprint (2019).

11. Sun H, Ganglberger W, Panneerselvam E, Leone MJ, Quadri SA, Goparaju B, Tesh RA, Akeju O, Thomas RJ, Westover MB Sleep Staging from Electrocardiography and Respiration with Deep Learning. arXiv preprint arXiv:1908.11463. (2019).
12. Wang, S. C., Wu, H. T., Huang, P. H., Chang, C. H., Ting, C. K., & Lin, Y. T. Novel imaging revealing inner dynamics for cardiovascular waveform analysis via unsupervised manifold learning. arXiv preprint arXiv:1909.04206. (2019).
13. Baklan, I., Mukha, I., Oliinyk, Y., Lishchuk, K., Nedashkivsky, E., & Gavrilenko, O. Anomalies Detection Approach in Electrocardiogram Analysis Using Linguistic Modeling. In International Conference on Computer Science, Engineering and Education Applications (pp. 513-522). Springer, Cham. (2019).
14. Hamming, R. W. "Error detecting and error correcting codes" (PDF). The Bell System Technical Journal. 29 (2): 147–160. doi:10.1002/j.1538-7305.1950.tb00463.x. ISSN 0005-8580. (1950).
15. Levenshtein, VI. "Binary Codes Capable of Correcting Deletions, Insertions and Reversals." Soviet Physics Doklady 10, p.707 (1966).
16. Jaro, M. A. "Advances in record linkage methodology as applied to the 1985 census of Tampa Florida". Journal of the American Statistical Association. 84 (406): 414–20. doi:10.1080/01621459.1989.10478785.(1989).
17. Levenshtein, Vladimir I., "Binary codes capable of correcting deletions, insertions, and reversals", Soviet Physics Doklady, 10 (8): 707–710 (1966).
18. Damerau, Fred J. "A technique for computer detection and correction of spelling errors", Communications of the ACM, 7 (3): 171–176, doi:10.1145/363958.363994 (1964).
19. De Maesschalck, R.; D. Jouan-Rimbaud, D.L. Massart The Mahalanobis distance. Chemometrics and Intelligent Laboratory Systems 50:1–18 (2000).
20. Majorek, Karolina A.; Dunin-Horkawicz, Stanisław; et al., "The RNase H-like superfamily: new members, comparative structural analysis and evolutionary classification", Nucleic Acids Research, 42 (7): 4160–4179, doi:10.1093/nar/gkt1414, PMC 3985635, PMID 24464998. (2013).
21. P. Sellers, The theory and computation of evolutionary distances: Pattern recognition, J. Algorithms 1, 359–373, (1980).
22. W. Chang and J. Lampe. Theoretical and empirical comparisons of approximate string matching algorithms. In Proc. 3rd Combinatorial Pattern Matching (CPM'92), LNCS 644, pages 172-181, (1992).
23. Saaty, T.L. The Analytic Hierarchy Process, New York: McGraw Hill. International, Translated to Russian, Portuguese, and Chinese, Revised editions, Paperback (1996, 2000), Pittsburgh: RWS Publications. (1980).
24. O.A.Pavlov, K.I.Lishchuk, V.I.Kut Mathematical optimization models for substantiating and finding the weights of objects in the method of pairwise comparisons. System Research and Information Technologies, No.2, pp,13-21 (2017).
25. A. Apostolico, Z. Galil, Pattern Matching Algorithms, Oxford University Press, Oxford, (1997).
26. M. Crochemore, W. Rytter, Text Algorithms, Oxford University Press, Oxford, (1994).