Bulgarian Chemical Communications, Volume 54, Special Issue B1 (pp. 97-102) 2022 DOI:10.34049/bcc.54.B1.0454

The importance of biological databases in modeling of structure-activity relationship

F. I. Sapundzhi[1], T. A. Dzimbova[1, 2]

[1]*South-West University "Neofit Rilski", 2700 Blagoevgrad, Bulgaria*
[2]*Institute of Molecular Biology "Roumen Tsanev", Bulgarian Academy of Sciences, Sofia, Bulgaria*

Received: November 25, 2021; Accepted: April 15, 2022

Biological databases play a key role in bioinformatics research and applications. Many databases are known that contain different types of information: DNA, sequences of proteins, molecule structures and others. They give researchers access to a large amount of biological data. The aim of the current research is to present a brief overview of major sequence databases and portals currently available and to underline open problems and future trends. The article presents examples for the use of various biological databases for the study of opioid and cannabinoid compounds. This investigation gives a brief description of the importance of biological databases and sequence analysis in bioinformatics research.

Keywords: biological databases, sequences of proteins, bioinformatics, computer modeling, structure-activity relationship

INTRODUCTION

Biological databases play a key role in bioinformatics research. They can be represented as stories of biological information. They contain information from research areas including genomics, proteomics, metabolomics, microarray gene expression, and phylogenetics. The databases give researchers access to a large amount of biological data [1, 2].

Biological data may refer to compounds or information obtained from living organisms and their products. There are many forms of biological data, including sequence data, protein structure, genomic data, and amino acids and the relationships between them [3, 4].

Biological data are connected very closely with bioinformatics, which is an interdisciplinary science that focuses on analyzing and interpreting a huge volume of biological data about DNA, RNA, protein sequences, protein structures, gene expression profiles, and protein interactions [5,6]. Bioinformatics includes two evolving disciplines such as biology and information technology, because solving modern biological problems requires a lot of computational methods such as database management, data modeling, pattern recognition, data extraction, query processing and biological data visualization. The most important task of bioinformatics is the understanding of correlations, structures and patterns in biological data, and next this knowledge can be used in drug discovery, genome analysis and biological control [1]. This science is associated with the development of databases to store and retrieve biological data; algorithms and statistics for analyzing and determining the relationships in biological data; and statistical tools for data identification and interpretation [7-11].

The aim of the current research is to present a brief overview of major biological databases and portals currently available and to underline open problems and future trends.

Effective management of the huge amount of biological data that is generated on a daily basis is crucial. A challenge for scientists is to integrate and manage these data into existing biological databases. They represent the libraries of biological sciences, collected from scientific experiments, published literature, high-throughput experiment technology, and computational analysis [1-4].

In biological databases, information on gene function, structure, localization, clinical effects of mutations, similarities of biological sequences and structures can be found.

BIOLOGICAL DATABASES

Biological databases can be broadly classified in sequence and design databases. They can be classified in primary, secondary and composite databases (Fig. 1.).

The primary biological databases are well-organized, user-friendly portals to the huge amount of biological data that is produced by researchers around the world.

The first primary databases appeared in the 1980s and 1990s in order to store experimentally determined DNA and protein sequences.

* To whom all correspondence should be sent.
E-mail: sapundzhi@swu.bg

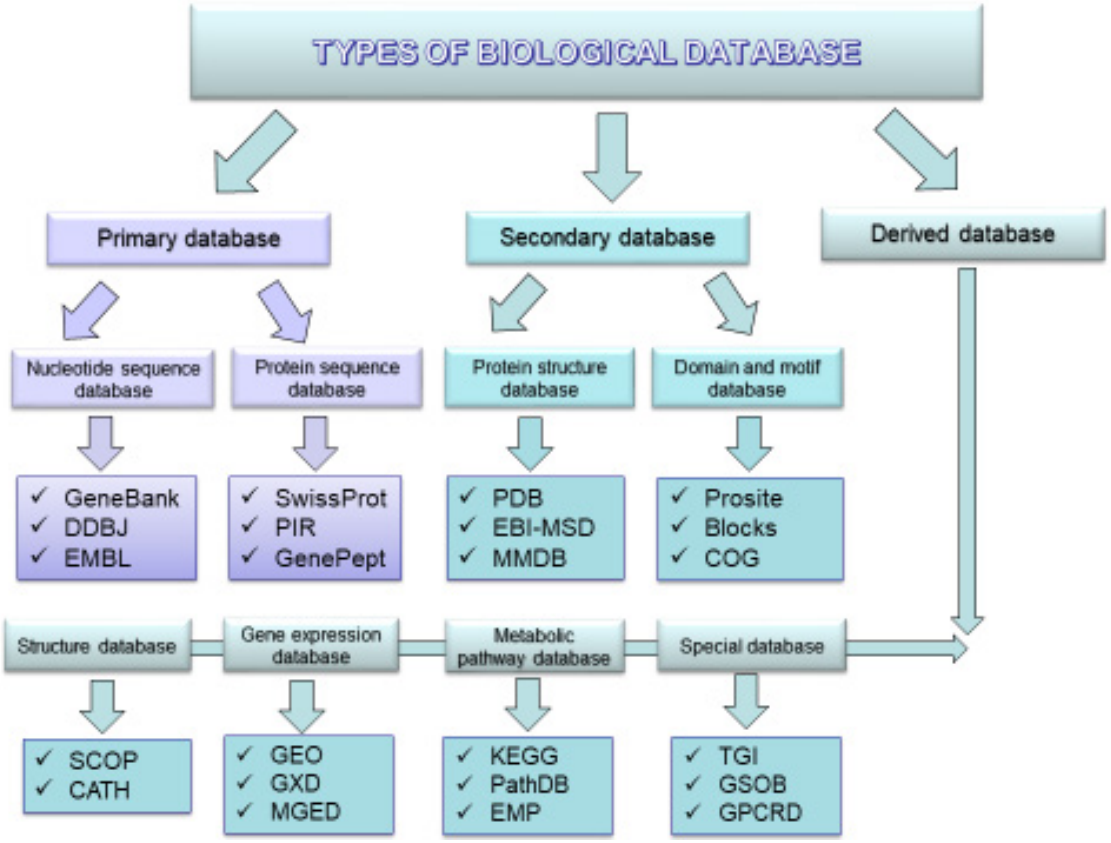

Fig. 1. Types of biological databases.

Today, most protein sequences that can be found in this databases are the product of conceptual translation of genes and genomes determined by DNA sequencing. They contain experimental data such as nucleotide sequences, protein sequences or macromolecular structures, which are submitted directly to the database by the researchers [12]. Examples of the primary databases are the following: GenBank (Genetic Sequence Databank) [13], ENA (European Nucleotide Archive) [14, 15], and DDBJ (DNA Data Bank of Japan) [16], ArrayExpress (Archive of Functional Genomics Data) [17], GEO (Gene Expression Omnibus) [18], PDB (Protein Data Bank) [19]. The nucleotide database was divided into three databases at NCBI: Core Nucleotide database, Expressed Sequence Tag (EST) and Genome Survey Sequence (GSS).

Primary Nucleotide Sequence Database

GenBank (Genetic Sequence Databank) [13]. The GenBank is one of the fastest growing repositories of known nucleotide sequences. The files from it contain data for sequence, access numbers, gene names, phylogenetic classifications, references to published literature, etc. It is developed and maintained at the NCBI. GenBank is a part of International Sequence Database Collaboration (INSDC) which includes ENA database [14], DDBJ database, and GenBank at NCBI. These organizations exchange data on a daily basis.

EMBL (European Molecular Biology Laboratory) [20]. The EMBL database of DNA and RNA sequences was established in 1980. It contains a collection of scientific literature, which is provided directly by researchers. EMBL is supported by EBI (European Institute of Bioinformatics) and is in close collaboration with GeneBank and DDBJ (Fig. 2).

Primary Protein Sequence Databases

There are a large number of protein sequence databases, from simple sequence stores to expertly selected universal databases. The databases that include protein sequences are GenPept, RefSeq, Swiss-Prot, PIR, PRF, and PDB.

SWISS-PROT (Swiss Institute of Bioinformatics, Geneva) [21] - a protein sequence and knowledge database established in 1986. It is a part of UniProt consortium and provides information about the functions of a protein, domain structure and post translational modifications, etc. It is characterized by a minimal level of excess and a high level of integration with other databases.

TrEMBL (translation of EMBL nucleotide sequence database) [22] consists of computer annotated entries derived from the translation of all coding sequences in the nucleotide databases. In this

database, the records are automatically annotated and unreviewed. It is a supplement of Swiss-Port database and contains all translations of EMBL nucleotide sequence entries which are not yet integrated in Swiss-Port.

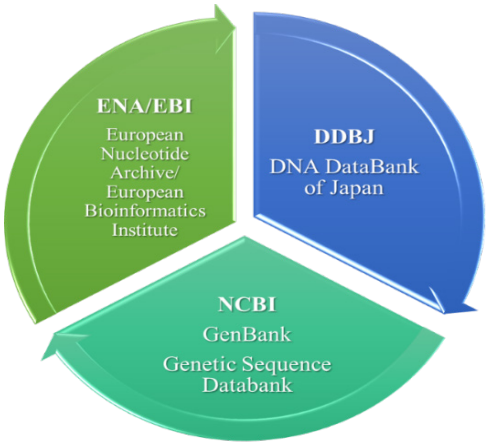

Fig. 2. Nucleotide Sequence Database linkage.

PIR (Protein Information Resource) [23] is a public bioinformatics resource to support genomic and proteomic research and scientific studies. It was established in 1984 by NBRF (National Biomedical Research Foundation) in USA. The database offers a variety of resources to support the distribution and consistency of protein annotations such as PIRSF, ProClass, and ProLINK.

PRF (Protein Research Foundation) [24] is the source of information related to amino acids, peptides and proteins. The information on synthetic compounds, protein sequence data, molecular aspects of proteins, and articles from scientific journals are available in this site.

Secondary Protein Structure Databases

PDB (Protein Data Bank) [19] is the main primary database for 3D structures of biological macromolecules determined by X-ray, crystallography and NMR. It was established in 1971 at Brookhaven National Laboratories. Since 1998 it is now supported by the RCSB (Research Collaboratory for Structural Bioinformatics) [25]. PDB provides tools and resources for studying the structures of biological macromolecules and their relationships and functions with other sequences. It accepts experimental data used to determine the structures and homology models [26, 27].

MSD/EMD (Macromolecular Structure Database) [28]. It is a European project for collection, management and distribution of macromolecular structures integrating current database and IT with a solid core of expertise in structural biology. It works closely with RCSB in the USA and PDBj in Japan [29].

EMD (Electron Microscopy Data Base) [30] is part of MSD database [3]. MSD manages, organizes and disseminates data on the structures of biological macromolecules solved by 3D electron microscopy.

MMDB (Molecular Modeling DataBase) [31] is a structural database of NCBI, containing experimentally determined 3D biomolecular structures. The database contains information on the biological function, the mechanisms associated with the function, and the evolutionary history of macromolecules and their relationships [32].

Secondary Domain Motif Database

The Domain Motif Database stores protein sequence motif data which are a set of preserved amino acid residues - important for protein function and located at a certain distance from each other.

PROSITE database [33] stores documentation describing protein domains, families and functional sites, related models and profiles for their identification [34]. Biologically significant sites, models and profiles can be found in it, which will allow the identification of the new sequence to which family of proteins it belongs.

PRINTS [35] is a database for protein fingerprints which are a group of conserved motifs used to characterize a protein family. They can encode protein folds and functionalities more flexibly and powerfully than single motifs [36].

ProDom [37] is a protein domain database automatically generated from the Swiss-Port and TrEMBL sequence database [38,39]. It contains automatic clustering of homologous domains, which is a rational way of organizing protein sequence data.

BLOCKS [40] is a server for sequence analysis at the Fred Hutchinson Cancer Research Center in Seattle, USA. The Blocks Database is a collection of blocks with known protein families that can be used to compare a protein or DNA sequence with documented protein families.

COGs (Clusters of Orthologous Groups of proteins COG) [41,42] is a tool for genome-scale analysis of protein functions and evolution. It is designed to classify proteins from fully sequenced genomes based on the concept of orthology.

3D Structure Databases

SCOP (Structural Classification of Protein database) [43] classifies protein 3D structures in a hierarchical scheme of structure classes. This database contains detailed information on the structural and evolutionary relationships of the proteins from PDB [44].

The **DBAli Database** [45] is related to SCOP which contains pairwise structural alignments generated by different methods [46].

CATH (Class, Architecture, Topology, Homologous) [47] contains a hierarchical classification of protein domains based on their folding patterns. They are obtained from protein structures deposited in PDB [48].

Gene Expression Databases

GEO (Gene Expression Omnibus) [49] **-** a public functional genomic data repository that accepts data based on arrays and sequences and provides tools to help users search for and download experiments and curated gene expression profiles [50].

GXD (Gene Expression Database) [51] is a community resource with gene expression information for the laboratory mouse. It provides information on which transcripts and proteins are produced by which genes, where and in what amounts their products are expressed and how their expression varies in different murine strains and mutants. GXD is integrated with the Mouse Genome Database (MGD).

MGED (Microarray Gene Expression data) [52] contains microarray data generated by functional genomics and proteomics experiments.

Metabolic Pathway Databases

KEGG PATHWAY database (Kyoto Encyclopedia of Genes and Genomes) [53,54] contains graphical pathway maps for all known metabolic pathways from various organisms. It is a collection of databases dealing with genomes, biological pathways, diseases, drugs, and chemical substances.

EcoCyc [55] a bioinformatics database - E. coli database, stores information regarding the genome and biochemical machinery of *E. coli* K-12 MG1655.

LIGAND [56] is a chemical database for enzyme reactions at the Institute for Chemical Research, Kyoto. It is a composite database currently consisting of the COMPOUND, DRUG, GLYCAN, REACTION, RPAIR and ENZYME databases.

PathDB [57] is a relational database that can store detailed and specific metabolic information and visualization methods that aid in the *in silico* detection process. It allows the storage of quantitative information on enzymatic and spontaneous reactions and transport steps.

EMP (Database of enzymes and metabolic pathways) [58] is a public server. It is an encoding of the contents of over 10 000 original publications on the topics of enzymology and metabolism. This information is transformed into a queryable database. It plays an important role in the interpretation of genetic sequence data.

Special Databases

TGI (TIGR Gene Indices) [59] database uses all publicly available expressed sequence tags (EST) and data known on gene sequence stored in GenBank for each target species.

GPCRdb [60] contains reference data, interactive visualization and experiment design tools for G protein-coupled receptors (GPCRs). It controls the alignment of sequences, structures, and receptor mutations in the literature.

OPEN PROBLEMS AND FUTURE TRENDS

In our work we rely very much on biological databases. The publication of the crystal structures of MOR (mu-opiod receptor), DOR (delta-opioid receptor), CBR1 (cannabinoid receptor type 1) and CBR2(cannabinoid receptor type 2) helps in the targeted drug design:

- MOR (RCSB, PDBid:4DKL) [61-64];
- DOR (RCSB, PDBid:4ej4) [65-70];
- CBR1 (RCSB, PDBid:5tgz) [71-74];
- CBR2 (RCSB, PDBid:2hff) [75-78].

The possibilities provided by computer methods together with biological databases are related to determining the intimate mechanism of interaction between ligands and receptors, including determining binding sites, binding energies, the presence of specific groups in the structure of ligands that promote their binding to receptor and many others. Determining the relationship between the structure of the ligand and its action using computer methods shortens the time to find potentially active compounds.

CONCLUSIONS

The biological databases are an important tool to help scientists study and explain biological phenomena from the structure of biomolecules to their interaction and to understand the progression of species. The current research presents examples from the use of various biological databases for the study of deferent compounds. This investigation gives a brief description of the importance of biological databases and sequence analysis in bioinformatics research.

Acknowledgement: *This paper is partially supported by Project of the National Science Fund of Bulgaria, BNSF H27/36; National Scientific Program "Information and Communication*

Technologies for a Single Digital Market in Science, Education and Security (ICTinSES)", financed by the Ministry of Education and Science.

REFERENCES

1. Y. P. P. Chen (ed.), Bioinformatics technologies. Springer Science & Business Media, 2005.
2. N. Toomula, A. Kumar, S. Kumar, V. Bheemidi, *J. Comput. Sci. Syst. Biol.,* **4,** 87 (2011).
3. D. Jelic, D. Toth, D. Verbanac, *Food Technol. Biotechnol,* **41**, 269 (2003).
4. J. Dziuba, A. Iwaniak, in: Y Mine, F Shahidi (eds.), Nutraceutical Proteins and Peptides in Health and Disease, CRC Press, Boca Raton, 2006, p. 543.
5. J. Wren, A. Bateman, *Bioinformatics*, **24** (19): 2127 (2008).
6. S. Mir, Y. Alhroub, D. Armstrong, J. Berrisford. A. Clark, M. Conroy, J. Dana, M. Deshpande, D. Gupta, A. Gutmanas, P. Haslam, L. Mak, A.Mukhopadhyay, N. Nadzirin, T. Paysan-Lafosse, D. Sehnal, S. Sen, O. Smart, M. Varadi, G. Kleywegt, S. Velankar *Nucleic Acids Research,* **46** (D1), D486 (2018).
7. A. Kinjo, G. Bekker, H.Suzuki, Y.Tsuchiya, T. Kawabata, Y. Ikegawa, H. Nakamura, *Nucleic Acids Research,* **45** (D1), D282 (2017).
8. A. Bayat, *BMJ,* **324** (7344), 1018 (2002).
9. K. Herbert, J. Spirollari, J. Wang, W. Piel, J. Westbrook, W. Barker, Z. Hu, C. Wu. Bioinformatics databases. Wiley Encyclopedia of Computer Science and Engineering, 2007.
10. M. Traykov, I. Trenchev, *Russian Journal of Genetics*, **52**(9), 985 (2016).
11. M. Arita, I. Karsch-Mizrachi, G. Cochrane, *Nucleic Acids Res.,* **49**(D1)**,** D121 (2021).
12. M. Lazarova, S. Markov, A. Vassilev, *AIP Conference Proceedings,* **2302**, 080004 (2020).
13. GenBank: http://www.ncbi.nlm.nih.gov/genbank.
14. G. Hamm, G. Cameron, *Nucleic Acids Research,* **14** (1), 5 (1986).
15. ENA:https://www.ebi.ac.uk/ena/browser/home.
16. DDBJ: https://www.ddbj.nig.ac.jp.
17. ArrayExpress: https://www.ebi.ac.uk/arrayexpress.
18. GEO: https://www.ncbi.nlm.nih.gov/geo/.
19. PDB: http//www.pdb.org.
20. EMBL: http://www.ebi.ac. uk/embl/index.html.
21. Swiss-Prot: http://www.expasy.org.
22. UniProt: https://www.uniprot.org.
23. PIR:https://proteininformationresource.org/.
24. PRF: https://www.proteinresearch.net/.
25. RCSB: http://www.rcsb.org/pdb/.
26. H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, P. Bourne, *Nucl. Acids Res.,* **28**, 235 (2000).
27. O. Ogasawara, Y. Kodama, J. Mashima, T. Kosuge, T. Fujisawa, *Nucleic Acids Res.,* **48**, 45 (2020).
28. MSD: http://www.ebi.ac.uk/msd/.
29. PDBj: https://pdbj.org/.
30. EBI:http://www.ebi.ac.uk/msd/MSDProjects/ IIMS3D_EMdep.html.
31. MMDB:http://www.ncbi.nlm.nih.gov/Structure/MMDB/mmdb.shtml.
32. Y. Wang, I. Geer, T. Madej, A. Marchler-Bauer, D. Zimmerman, S. Bryant S.H., *Nucl.Acids Res.* **30** (1), 249 (2002).
33. Prosite: http://us.expasy.org/prosite/.
34. L. Falquet, M. Pagni, Ph. Bucher, N. Hulo, Ch. Sigrist, K. Hofmann, A. Bairoch, *Nucl. Acids Res.* **30,** 235 (2002).
35. PRINTS:http://www.bioinf.man.,ac.uk/dbbrowser/PRINTS/.
36. T. Attwood, M. Blythe, D. Flower, A. Gaulton, J. Mabey, N. Maudling, L. McGregor, A. Mitchell, G. Moulton, K. Paine, P. Scordis, *Nucl. Acids Res*., **30** (1), 239 (2002).
37. Prodom:http://prodes.toulouse.inra.fr/prodom/ 2002. 1/html/home.php.
38. F. Servant, C. Bru, S. Carre`re, E. Courcelle, J. Gouzy, D. Peyruc, D. Kahn, *Briefings in Bioinformatics*, **3(3),** 246 (2002).
39. F. Corpet F. Servant, J. Gouzy, D. Kahn, *Nucl. Acids Res*., **28** (1), 267 (2000).
40. Blocks: http://www.blocks.fhcrc.org/.
41. COG: http://www.ncbi.nlm.nih.gov/COG.
42. R. Tatusov, M. Galperin, D.Natale, E. Kooninet, *Nucleic Acids Research,* **28** (1), 33 (2000).
43. SCOP: http://scop.mrc-lmb.cam.ac.uk/scop/.
44. A. Murzin, S. Brenner, T. Hubbard, C. Chothia, *J. Mol. Biol.,* **247**, 536 (1995).
45. D. B. Ali, http://guitar.rockefeller.edu/DBAli/.
46. M. Marty-Renome, *Bioinformatics,* **17,** 746 (2001).
47. CATH:http://www.biochem.ucl.ac.uk/bsm/cat_ new/ index.html.
48. C. Orengo, A. Martin, G. Hutchinson, S. Jones, D. Jones, A. Michie, M. Swindells, J. Thornton, *Structure,* **5** (8), 1093 (1997).
49. GEO:https://www.ncbi.nlm.nih.gov/guide/ genes-expression/.
50. R. Edgar, M. Domrachev, A. Lash, *Nucleic Acids Res*., **30** (1), 207 (2002).
51. http://www.informatics.jax.org/.
52. www.mged.org.
53. https://www.genome.jp/kegg/pathway.html.
54. M. Kanehisa, S. Goto, *Nucleic Acids Res*., **28** (1) 27 (2000).
55. I. Keseler, J. Collado-Vides, A. Santos-Zavaleta, M. Peralta-Gil, S. Gama-Castro, L. Muñiz-Rascado, C. Bonavides-Martinez, S. Paley, M. Krummenacker, T. Altman, P. Kaipa, A. Spaulding, J. Pacheco, M. Latendresse, C. Fulcher, M. Sarker, A. Shearer, A. Mackie, I. Paulsen, R. Gunsalus, P. Karp. D.*Nucl. Acids Res.*, **39,** 583 (2011).
56. S. Goto Y. Okuno, M. Hattori, T. Nishioka, M. Kanehisa, *Nucleic Acids Res.,* **30(1)**, 402 (2002)
57. PathDB: http://www.ncgr.org/pathdb.
58. EMP: http://emp.mcs.anl.gov/.
59. TIGR: http://www.tigr.org/tdb/tgi.
60. GPCRdb: https://gpcrdb.org.

61. F. Sapundzhi, T. Dzimbova, N. Pencheva, P. Milanov, *Bulg. Chem. Commun.*, **47**, 613 (2015).
62. T. Dzimbova, F. Sapundzhi, N. Pencheva, P. Milanov, *J. Pept. Sci.*, **18**, (S1) S84, P072, (2012).
63. F. Sapundzhi, T. Dzimbova, N. Pencheva, P. Milanov, *Bulg. Chem. Commun.*, **51**, 569 (2019).
64. F. Sapundzhi, K. Prodanova, M. Lazarova, *AIP Conference Proceedings*, **2172**, 100008 1-6, (2019)
65. F. Sapundzhi, T. Dzimbova, N. Pencheva, P. Milanov, *Der Pharma Chemica*, **8**, 118 (2016).
66. F. Sapundzhi, T. Dzimbova, P. Milanov, N. Pencheva*, Int. J. Bioautomation,* **17** (1), 5 (2013).
67. F. Sapundzhi, T. Dzimbova, N. Pencheva, P. Milanov, *J. Pept. Sci.*, **20** (S1), S294, (2014).
68. F. Sapundzhi, T. Dzimbova, N. Pencheva, P. Milanov. *Bulg. Chem. Commun.*, **49** (E), 23 (2017).
69. F. Sapundzhi, T. Dzimbova, N. Pencheva, P. Milanov. *Bulg.Chem. Commun.*, **49** (E), 768 (2017).
70. F. Sapundzhi, T. Dzimbova, N. Pencheva, P. Milanov, *Bulg. Chem. Commun.*, **50**, Special Issue B, 44 (2018).
71. F. Sapundzhi, T. Dzimbova, *Bulg. Chem. Commun.*, **50**, Special Issue B, 15 (2018).
72. F. Sapundzhi, T Dzimbova, *J. Chem. Technol.,* **55** (5), 959 (2020).
73. F. Sapundzhi, V. Slavov, *J. Chem. Technol.,* **55** (5), 935 (2020).
74. R. Topalska, T. Dzimbova, *J. Chem. Technol.,* **55** (5), 714 (2020).
75. F. Sapundzhi, T. Dzimbova, *J. Chem. Technol.,* **55** (5), 709 (2020).
76. F. Sapundzhi, T. Dzimbova, *Bulg. Chem. Commun.*, **52** (A), 197 (2020).
77. F. Sapundzhi, *IJOE*, **15** (11), 139 (2019).
78. F. Sapundzhi, T. Dzimbova, *IJOE,* **15** (15), 39 (2019).