Parallel algorithm of identification of similar chemical compounds

A.P. Sergeyev

JSC «Dialektika», 2 Glushkov ave., building 6, room 214, Kyiv, Ukraine

sergeev@dialektika.com

Abstract. *The statement about similarity of the chemical compounds presented by isomorphic molecular graphs is formulated and proved. On the basis of this statement the optimized algorithm of search and identification of similar chemical compounds is created and tested. This algorithm is presented by serial and parallel versions. Recommendations about practical use of algorithm are provided in various fields of knowledge. The estimation of computing complexity of the serial and parallel version of algorithm is executed. Recommendations about development and further improvement of algorithm are made.*

Keywords

Chemical compound, HPC-UA, molecular graph, isomorphism, parallel programming.

1 Introduction

In the world millions chemical compounds are synthesized. Therefore it is often simpler to synthesize compound with the necessary properties, than to look for the relevant information. Certainly, there are tens chemical databases which are available on a free basis. Nevertheless, the search of the necessary chemical compound which is carried out by means of the unique interface of chemical database, usually occupy many working days.

For acceleration of search of the necessary information, and also the analysis and selection of available data on chemicals databases, the parallel algorithm of search and identification of similar chemical compounds is offered. This parallel algorithm identifies the isomorphic molecular graphs. Isomorphic molecular graphs present the similar chemical compounds with compatible properties.

This algorithm has other applications, in cryptography or crystallography. Algorithm is changed and scalable easily, because it is object-oriented and parallel.

2 Related works

In articles «Fast algorithm identification of similar chemical compounds» (Proceedings of conference PDCS 2013, Ukraine, Kharkiv, March 13-14, 2013) and «Optimization of data processing of chemical databases» (Third International Conference «High Performance Computing» HPC-UA 2013, Ukraine, Kyiv, October 7-11, 2013) the fast algorithm of search and the analysis of structural components of chemical compounds were considered. In current article the optimized parallel version of this algorithm is described and tested. This algorithm is tested on the array of molecular graphs with dimensions between 10 and 100. The estimation of computing complexity of the parallel version of algorithm is gained.

3 Main part

Molecules of chemical compounds are unambiguously represented by molecular graphs [1]. In turn, molecular graphs are unambiguously represented by adjacency matrixes. In language of properties of an adjacency matrix it is possible to express various properties of molecular counts. If molecular graphs are isomorphic, then appropriated chemical compounds can be similar (figures 1 and 2). It should be noted, similarity isn't equivalent to identity. Similar chemical compounds can be completely identical or similar, i.e. *stereoisomers*. Stereoisomers of chemical compounds divide on three following categories:

- optical isomers;

- diastereomers;
- geometrical isomers.

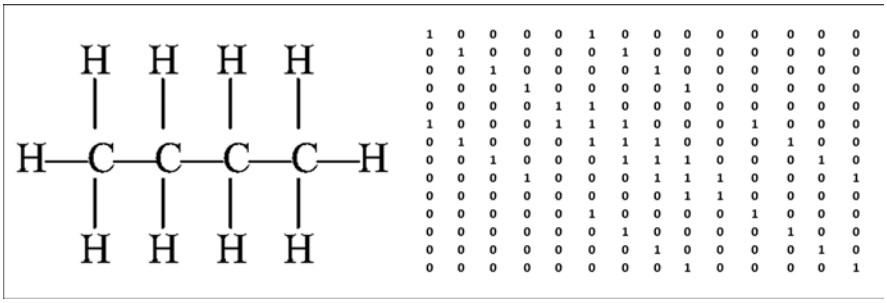

Fig. 1. Molecule of butane and adjacency matrix of corresponding molecular graph

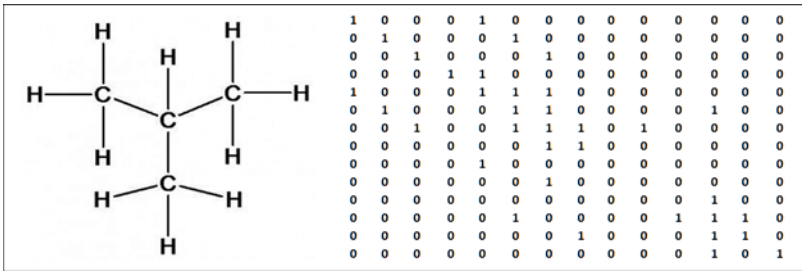

Fig. 2. Molecule of isobutene and adjacency matrix of corresponding molecular graph

These categories are given in an order of reduction of chemical similarity. Geometrical isomers have the smallest degree of chemical similarity. Physical and chemical characteristics of these compounds are strongly different. It takes place following statement.

Statement 1. The chemical graphs corresponding to geometrical isomers aren't isomorphic.

Correctness of this statement follows from determination of isomorphism and geometrical isomers [3].

Optical isomers (enantiomers) have closest "affinity". They are a specular reflection of each other possess and aren't combined in space (property of a *hiralnost*). Following statement is true.

Statement 2. The chemical graphs corresponding to optical isomers, are isomorphic.

The validity of this statement follows from determination of isomorphism of graphs and an essence of an optical isomerism. Molecules of optical isomers have identical structure, and chemical graphs corresponding to them are identical despite of renumbering of vertices (are isomorphic).

In fig. 3 the molecule ofloxacin (which represents a combination of two optical isomers - left-handed and right-handed forms) is shown.

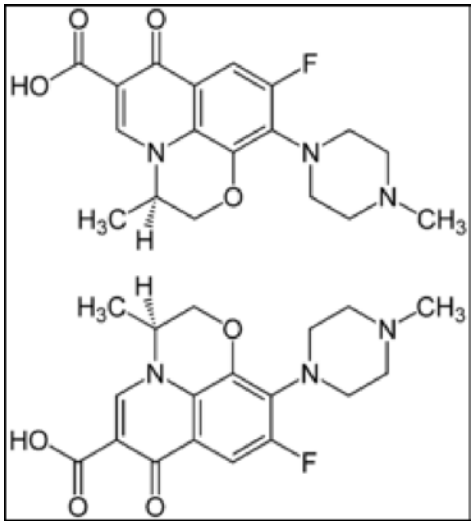

Fig. 3. Molecule of ofloxacin, consisting of two optical isomers

The parallel algorithm of search and identification of similar chemical compounds is based on the algorithm described in [5]. The structure of algorithm is presented in fig. 4.

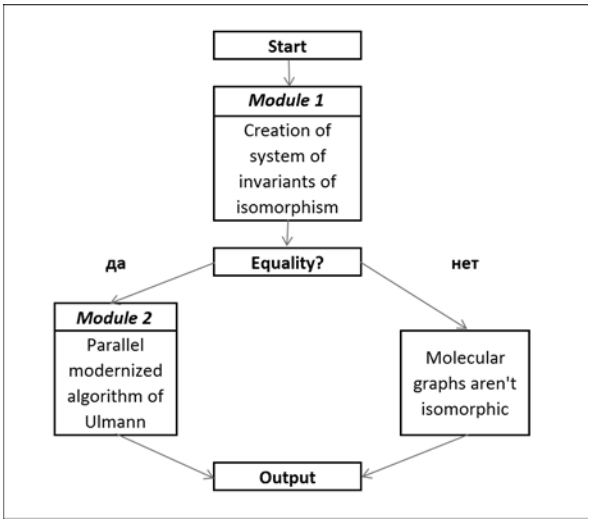

Fig. 4. Block diagram of algorithm of search and identification of similar chemical compounds

The algorithm of search and identification of similar chemical compounds checks adjacency matrixes of the corresponding molecular graphs of chemical compounds. If matrixes are equal despite of permutation of rows and columns, then molecular graphs are isomorphic. Also, chemical compounds are optical isomers (i.e. are similar). The algorithm has two modules.

Module 1. Creation of system of invariants of isomorphism and check on compliance.

For each adjacency matrix, representing the molecular graph, is created *system of invariants of isomorphism*. This system includes quantity of vertices of the graph (i.e. dimension of adjacency matrix), quantity of edges of the graph (quantity of nonzero elements of a matrix) and sequence of degrees of vertices of the graph. Degree of vertex *i* is a sum of nonzero elements in row *i* or in column *i* of an adjacency matrix.

Statement 3. If for graphs of A and B the system of invariants of isomorphism doesn't equal, graphs doesn't isomorphic.

The proof of this statement follows from definition of an invariant of isomorphism which remains stable after performing isomorphic transformations.

If invariants of isomorphism are equal, molecular graphs may be isomorphic. Module 2 checks the isomorphism of molecular graphs.

Module 2. This module performs direct check on isomorphism (it is based on Ullmann's algorithm [1]). If graphs are isomorphic, such equality is true: $A=PBP^T(1)$, where A and B – adjacency matrixes of checked graphs, P – matrix of substitution, P^T – transposed matrix of substitution.

There is classical (serial) algorithm of Ullmann [2].

1. Let $P = (p_{ij})$, a matrix of substitution of dimension of n×n, where n – quantity of vertices of molecular graphs (dimension of adjacency matrixes A and B).
2. It calls procedure ISO (A; B; P; 1), in this case 1 – initial value of the counter k.
3. If k >= n, the creation of a matrix of substitution by dimension of n×n is complete, also the molecular graphs presented by adjacency matrixes A and B are isomorphic.
4. If k < n, for i=1 to n, then three procedures are followed:

- $P_{ki} \leftarrow 1$ and for all j=1 $P_{kj} \leftarrow 0$.
- If $P_{k,k}(A) = P_{k,n}(P)B(P_{k,n}(P))^T$
- It calls the procedure ISO(A; B; P, k+1)

During check of isomorphism of graphs (by Ullmann's algorithm) consecutive (recursive) generation of submatrixes of substitution is made (dimensions of these matrixes change from 1 to n). Thus procedure of backtracking search (ISO)

is used. Then equality (1) for submatrixes is checked by dimension of n×n. If equality (1) is true, then graphs A and B are isomorphic. In worst variant computing complexity of this algorithm is estimated as $O(A^B B^2)$

The parallelization (made by MPICH2 library) is applied to acceleration of algorithm of Ullmann. The *m* various matrixes of substitution of P_{ij} (matrix's dimension is m×m, where m – number of processors), are created in parallel version of algorithm. Elements of matrixes of substitution are calculated by formula: $P_{1j} = 1$ (j=1 … m), other P_{ij} elements are equal 0. Then for k=2 the ISO procedure is executed on *m* processors at the same time.

Computing complexity of algorithm decreases to $O(A^B B)$ (at worst variant) in parallel version of algorithm of Ullmann .

In table 1 time of algorithm running is given (in seconds).

Tab.1. Results of algorithm running

Dimension of adjacency matrix	***Serial version, 1 processor (sec)***	**Parallel version, 2 processors (sec)**	**Parallel version, 4 processors (sec)**
10×10	0,23	8,12	4,95
20×20	0,41	11,83	6,84
30×30	0,61	13,13	7,55
40×40	0,95	15,13	11,59
50×50	6,94	19,27	14,35
60×60	17,37	27,39	21,42
70×70	22,43	34,62	29,11
80×80	28,74	49,06	35,41
90×90	112,95	80,84	65,36
100×100	385,31	276,64	205,93

4 Conclusion

The algorithm of identification of similar chemical compounds is offered. The estimations of computing complexity for the consecutive and parallel versions of this algorithm are gained. Results of testing of algorithm on the array of matrixes of dimension from 10×10 to 100×100 are represented.

This algorithm can be applied in a crystallography, biology and other fields of knowledge for search of the similar structures.

References

[1] L.B. Kier: Molecular orbital theory in drug research. *New York: Academic Press,* 1971, 164-169.
[2] J.R.Ulmann: An algorithm for Subgraph Isomorphism. *National Physical Laboratory, Teddington, Middlesex, England,* 1976.
[3] A.M. Jognosn, G.M. Maggiora. Concepts and Applications of Molecular Similarity. *New York: John Wiley & Sons,* 1990.
[4] A. Varnek, A. Tropsha. Chemoinfomatics: Approaches to Virtial Screening. *RSCPunlishing,* 2008.
[5] A.P. Sergeyev. Fast algorithm of identification of similar chemical compounds. *Proceedings of conference PDCS'2013, Ukraine, Kharkiv,* March 13-14, 2013, p. 299-300.
[6] T.W. Avdeeva, A.P. Sergeyev Optimization of data processing of chemical databases. *Proceedings of conference HPC'UA'2013, Ukraine, Kyiv,* October 7-11, 2013, p. 23-27.