VARIOUS TECHNIQUES IN SOFTWARE APPLICATIONS FOR 3D VISUALIZATION AND EVALUATION OF BIOISOSTERIC MOLECULES

Desislava Y. Ivanova
Institute of Information and Communication Technologies Bulgarian Academy of Sciences, Acad. Georgi Bonchev Str. Block 2, 1113 Sofia, Bulgaria

divanovaacomin@gmail.com

***Abstract**:The best - known software applications for 3D visualization and evaluation of bioisosteric molecules will be compared. The software applications are based on parameters, which can be divided mainly into two directions - emphasizing electrostatic properties, wave functions and presence of electrons in different energy states, and QSAR - based - quantitative methods – empirical examination of molecules, which use mainly statistical data collection, thus evaluating the properties of molecules through regression models to investigate similarity. The first group of innovative methods are superior, because of their analytical purposes that are not mining blindly for bioisosterism through empirical statistical data, but analytically detect similarities according to molecules' shapes. In conclusion - the use of GPUs for parallel processing of matrix-based operation - for the calculations of quantum vectors - would lead to repeated acceleration of the speed of these calculations. Finally, this study demonstrates the GPU as another factor, which is important for the improvement of the evaluation methods of studying and visualization of molecular similarity.*

Key words: 3D visualization, bioisosteric molecules, software applications, innovative methods, bisisosterism, GPUs

1. Bioisosteres In Drug Discovery

Bioisosterism has a crutial role in the development of drug molecules almost from the origin of the pharmaceutical industry. The aim of bioisosterism is that the properties of a compound can be fine-tuned without affecting its fundamental biological activity. However this comes with its challenges. Successfully applying bioisosterism to achieve the intended molecular outcome is difficult because of the fundamental problem that chemical structure is an unreliable indicator of biological activity. A slight change in a molecule can have a far-reaching impact on a compound's activity, specificity and toxicity, in the same time completely different chemotypes may have near identical biological activity profiles. More exact and reproducible methods for suggesting relevant, non-obvious and yet synthetically intuitive bioisosteres would have wide applicability.

Bioisosteres are used by researchers throughout the pharmaceutical industry to find new hits and leads by modifying known actives or substrates, to develop leads by modifying physicochemical properties and protecting their knowledge using patents. Having identified an interesting target, researchers often had little choice in finding an active inhibitor or antagonist, except through bioisosteric modification of the natural ligand in a systematic and thoughtful manner. The modern HTS era has provided a lot of potential leads, but still the need for bioisosteres stays actual as structures found through HTS can have undesirable properties (either physical or biological) and often lack novelty.

The requirement to protect research positions through patent applications is crucial for the development of new medicines. In this respect, IP protection is probably the most important use of bioisosteres in the modern drug discovery project. Replacement of core groups in the lead series with new scaffolds that introduce better selectivity or physical properties. Scaffold hopping is a computational technique of replacing portions of molecules to create novel drug-like compounds with similar activity to the original. The method involves choosing a portion of the starting molecule, often the central scaffold, for replacement. The scaffold-hopping software searches a database of hundreds of thousands of fragments for the best replacements. The worth of the software depends on the algorithm used to evaluate the "best" matches. Different software tools use different approaches. Some use simple geometrical considerations and/or the presence of simple pharmacophore points while others use ligand similarity to rank the replacements. In all cases, the best matches are returned as possible candidates for synthesis. Choosing the right compounds to progress is important as it can frequently take up to a week or more of lab time to synthesize a new compound. Results must be imaginative, yet realistic suggestions that enable users to advance the compounds that are most likely to succeed.

High-throughput screening (HTS) is a method for scientific experimentation especially used in drug discovery and relevant to the fields of biology and chemistry. Using robotics, data processing and control software, liquid handling devices, and sensitive detectors, High-throughput screening allows a researcher to quickly conduct millions of chemicalor pharmacological tests. Through this process one can rapidly identify active compounds that modulate a particular biomolecular pathway. The results of these experiments provide starting points for drug design and for understanding the interaction or role of a particular biochemical process in biology. It still takes a highly specialized and expensive screening lab to run an HTS operation, so in many cases a small- to moderate-size research institution will use the services of an existing HTS facility rather than set up one for itself.

Methods for describing molecules in a manner more related to their biological activity would have the potential to enable modifications and research activities to progress in a more sufficient way. The goal for finding relevant, non-obvious, accurate bioisosteres is not lucking an interest. These methods broadly fall into two categories: knowledge-based approaches and computational techniques.

2. QSAR software applications (legacy)

Knowledge-based approaches tend to use data mining techniques to find component parts that have previously been substituted for each other without a significant change in the activity under study. This approach has broad appeal; it can highlight changes that are known to be successful together with detailed examples. However, many replacements are specific to a particular protein and take no account of which parts of the moiety to be replaced are most important for activity. Equally, if the moiety to be replaced is not present in the literature then no suggestions are possible. The variety and availability of computer algorithms to suggest bioisosteric replacements has increased significantly in recent years. Most methods attempt to excise a chosen moiety from a molecule and replace it with a fragment from a fragment database. These fragments are typically scored against the moiety to be

replaced using shape or electrostatic measures of similarity, or by using the presence or absence of pharmacophore points.

The concept of "Structure-Activity Relationship" (SAR) is that the biological activity of a chemical can be related to its molecular structure. When quantified, this relationship is known as "QSAR". A QSAR model makes use of existing experimental toxicity data for a series of chemicals to build a model that relates experimentally observed toxicity with molecular descriptors in order to predict the toxicity of further chemicals. Quantitative structure–activity relationship models (QSAR models) are regression or classification models used in the chemical sciences and engineering. Like other regression models, QSAR regression models relate a set of "predictor" variables (X) to the potency of the response variable (Y), while classification QSAR models relate the predictor variables to a categorical value of the response variable. In QSAR modeling, the predictors consist of physio-chemical properties or theoretical molecular descriptors of chemicals; the QSAR response-variable could be a biological activity of the chemicals. QSAR models first summarize a supposed relationship between chemical structures and biological activity in a data-set of chemicals. Second, QSAR models predict the activities of new chemicals.

Principal steps of QSAR include a selection of Data set and extraction of structural/empirical descriptors, variable selection, model construction and a validation evaluation, based on the principle - Structure–Activity Relationship (SAR) - that similar molecules have similar activities.

VEGA software

Several institutes contributed to the development of the platform, called VEGA - QSAR, including regulators and public bodies in Europe and USA. VEGA freely offers models for properties.

The steps of the workflow include insert the list of the molecules identifiers, choose where to send the prediction output, ask prediction, and get results. The input can be given in different standard formats used in the chemical domain, including SMILES and SDF. To avoid the well - known problems about the non-unique representations VEGA transforms all the chemical structure into a unified internal string format.

The overall reliability of the prediction is measured by combining statistical values, elements of case based reasoning, and possibly presence of active substructures; the possible reasons of concern are underlined. All those considerations are weighted and summed up in an index (in 0 – 1) that is called Applicability Domain Index (ADI).

All of these methods seem to suffer the same problem: they rarely suggest truly novel scaffolds to the researchers. The reason for this is not clear but the one commonality is that these computational methods, like the literature methods, rely on fragment-to-fragment comparison. In this method, the replacement fragment is scored as an isolated molecule in itself, and not in the context of the final molecule. This is a subtle but critical problem, which Innovations in Pharmaceutical Technology means that there is no possibility for the properties of the fragment to influence the properties of the final molecule. Moreover, as the final molecular context of the fragment is not considered, fragments that might represent only a small change to the final molecule in its entirety may be scored poorly because they represent a large change when scored at the fragment level.

3. Analytical visualization software applications (innovative)

The Field Point approach is describing molecules by encoding the electrostatic environment surrounding a ligand. Drawing out the full field down to a series of 'hot spots' around the molecule – termed 'Field Points' – provided both a powerful insight into the behavior of molecules and a mechanism by which molecules could be compared in a computationally efficient and therefore rapid manner. Field Point descriptions of molecules have been used extensively to provide richer, more informative views of the way in which ligands interact with proteins, to interrelate compounds from different chemical series that act at the same protein site, to find novel chemical series through virtual screening, and to decode Structure Activity Relationships (SAR) by comparing molecules as proteins 'see' them. Field Point technology can also provide a much more accurate basis on which to identify novel, chemically relevant bioisosteres.

The principle behind fragment replacement methods to identify bioisosteres is simple: remove a portion of an active molecule, search a fragment database for a replacement moiety that will physically fit into the vacated space, and score the replacement for similarity to the original. In practice, a number of factors contribute to the effectiveness of the method. Primary amongst these are the accurate scoring of potential replacements, the relevance of the fragment database, and the originality and synthesisability of the suggested bioisosteres.

In scoring replacement fragments, it is essential to remember that the molecular fields around them are a property of the whole molecule and not of the isolated fragment. Replacement fragments cannot be assessed accurately in isolation from the whole molecule as they can have a significant effect on the retained portions of the molecule. Not only will fragments that are strongly electron donating have different effects from ones that are electron withdrawing, but the context of the molecule into which they are placed will determine the extent and character of those effects. To this end, we chose to join replacement fragments into the retained portions of the target, minimising the energy of the result to ensure sensible geometry, before scoring the whole of the proposed new molecule against the original molecule using Field Points. This approach, using the fields of the whole molecule, is only tractable because of the significant computational advantages provided by Field Point representation.

Because the score is based on the whole molecule, any effects that the new fragment may have on the original molecule are automatically considered. This process gives a results list that is significantly richer in non-obvious bioisoteres than would be the case had we only considered the isolated fragment. Using the whole molecule has an additional benefit in that the medicinal chemist is presented with a list of potentially active molecules rather than partial fragments. This allows them to select molecules for synthesis more easily without mentally having to construct and retrosynthesise the final molecule.

SPARK SOFTWARE

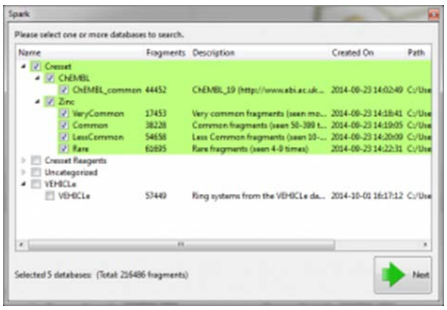
Delivers software to help to discover, design and optimize small molecules. The scientific methods use 3D molecular electrostatics and shape to shed light on the properties and behaviors of chemical structures and, crucially, to understand the key interactions which underpin biological activity.

Spark finds biologically equivalent replacements for key moieties a molecule for R-group exploration, patent busting or scaffold hopping. Allows visualizing of the results in detail side-by-side, or cluster similar chemical scaffolds and Search for moieties from real, published or unexplored compound databases. Spark is available on Windows®, Os X® and Linux®. It can be accessed as desktop applications and command lines and KNIME™ and Pipeline Pilot™ nodes.

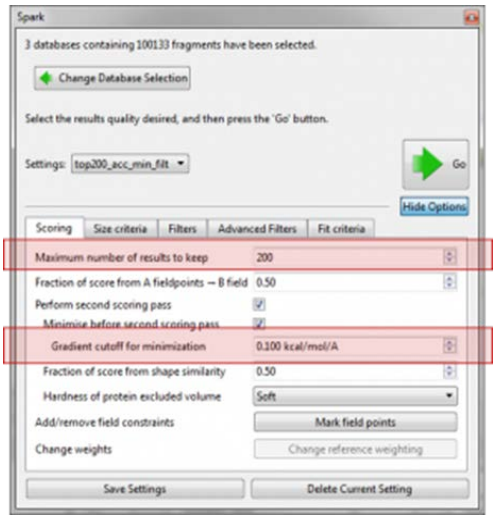
The Spark CSD Fragment Database is a collection of fragments derived from the small molecule crystal structures in the Cambridge Structural Database (CSD).

Experimental Setup starts with a load of the pdb file of the molecule is loaded into Spark and split into ligand and protein using the new protein import facility. The head group of the ligand for replacement is chosen and then selected the search of the

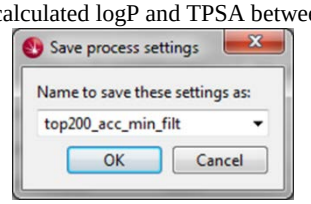
Chembl_common, VeryCommon, Common, LessCommon and Rare databases.



Before starting the experiment some of the parameters can be altered, for example on the 'Scoring' tab the number of results can be limited and the minimization of the final molecules can be improved by decreasing the gradient cutoff during the minimization.



The 'Advanced Filters' tab can be used to control the physical properties of the final molecules so that can be selected omly molecules with a calculated logP and TPSA between set borders.



In conclusion the advantages of Spark are a simple fast (less than an hour duration) experiment that is able to detect not only a large number of known scaffold replacements, but also structurally-novel suggestions in clear IP space. The new tile view allows the results to be rapidly triaged by eye, allowing the focus in on the most desirable replacements quickly and easily.

4. *Application of the Graphical processing unit in the field of Bioisosterism*

4.1 Parallel Computing

Even though supercomputers have kept a substantial lead over even the most sophisticated desktop machines in the continuous competition of hardware for speed, capacity, and robustness of computer platforms, the use of supercomputers has been considered a privilege to a limited number of people. In one important aspect — delivery of the technology to the fingertips of the largest number of people — parallel computing has always been at a disadvantage. Large academic institutions, large corporations, and government organizations can afford to execute computing tasks on customized supercomputers. As the name entails, algorithms executed on parallel computers and implemented in MPI (Message Passing Interface) have proven to be vastly superior in terms of time performance [2], which however comes at a high price: depending on the number of processors, and therefore computational potential, supercomputers can be priced anywhere upwards of several thousand dollars to build. Moreover, demand for computing frequently will be greater than the available processing time, requiring the process to be delayed in a scheduling queue. The resources and personnel required to establish and to keep a cluster operational are economically justifiable only in rare occasions.

To address the affordability problem, NVIDIA Corporation made available in the end of 2007 its proprietary platform for parallel programming, CUDA – Compute Unified Device Architecture [3]. The platform has two components: (1) hardware - the NVIDIA Graphics Processor Unit (GPU) on the graphics card; and (2) software - the programming interface to the GPU, provided by the CUDA language. As initially intended, the parallel computational capability would provide for efficient graphics rendering operations, where simple, independent algebraic calculations must be executed. The native carrier for graphics information is the **matrix**, stored in memory in the form of an *array*. Thus, graphics cards have gradually undergone a natural evolution to become robust platforms for matrix operations in parallel, a virtual requirement for all modern video-editing and gaming software.

An example of a problem that is suitable for a parallel implementation is the operation of matrix summation, A + B = C. The sum of elements in row *i* and column *j* of both matrices, A*ij* and B*ij* , is recorded in element C*ij* . This operation is *independent of the additions performed to compute the results of the remaining elements*, as demonstrated in Fig. 1:

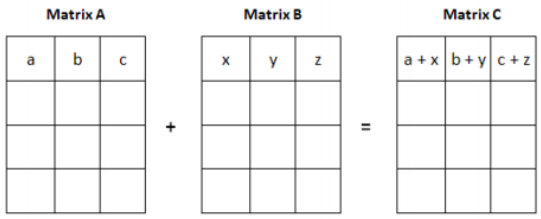


Fig. 1: Operations required for matrix addition

Theoretically, nothing prevents two operations from matrix addition from being executed at the same time, which is exactly what happens in practice when employing a parallel platform, where a distinct CPU is responsible for independently making each calculation and recording the result into memory [4].

A similar approach, although slightly more complex – due to the interconnected information that is required to gather and store (memory reads and memory writes), in terms of the matrix elements, as seen in Fig. 2:

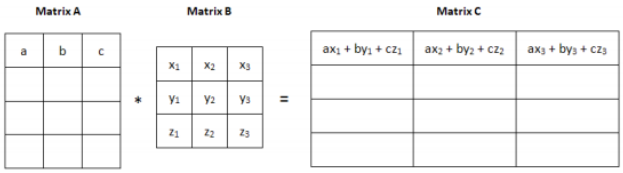


Fig. 2: Operations required for matrix multiplication

An element in row *i* column *j* in C is the sum of by-element products of row *i* from A and column *j* from B. Although in theory all elements in C can independently be calculated, for a large size matrix, even in multi-processor machines, there are not enough available processors to do all the necessary calculations in parallel. Instead, processors have to loop over the elements of the rows and columns of the matrices in order to produce the result of the multiplication. The following Fig. 3 is used to distinguish between the architecture of a CPU and a GPU:

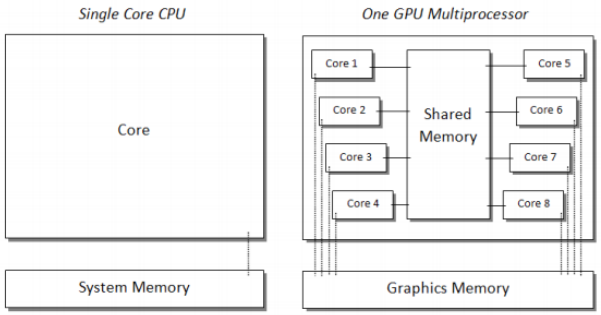
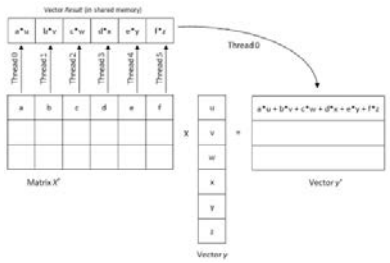


Fig. 3: Comparison betweenCPU / GPU architectures

When data is prepared for manipulation by the GPU, it is divided into blocks, which are then independently assigned for execution on one of the multiprocessors. Each block can have as many as 512 threads running simultaneously. Unless data is structured properly into blocks that are of size a multiple of a *warp* (32), a number of the threads in a warp may be inactive during a cycle, which is a drawback when the goal is maximum speed of execution. [3] However, there is still a great advantage in parallel calculations, such as matrix-vector multiplication, demonstrated in Fig. 4:



The speed-up of the multiplication for the first step, when done in parallel, as compared to a serial computation, will be the lesser of the two – the number of threads in a vCPU, or the width of the matrix (i.e. for a matrix that is 50 elements wide, we can expect a speed-up of 50x).

4.2 Bioisosterism by utilizing matrix functions and the GPU architecture

In conclusion, from the material presented in the previous sections, we can determine the importance of parallelism in bioisostere search algorithms – especially for molecules with complex structures. A rigorous analytical approach to examining bioisostere candidates by using any analytical descriptor (quantum wave-function electron field approach, symmetrical transformation approach, or other) could require trillions of linear algebra operations. The speed up of vector-matrix multiplication, matrix-matrix multiplication, and matrix inversion varies from 50-500x and above. To put this into perspective, any software that relies on the CPU for a specific bioisosteric algorithm (be it NERF or other), and takes one hour to deliver a full list of potential bioisosteres by utilizing the CPU, could deliver the same number of results in less than 10 seconds. When applying a "visual triage technique"- when the researcher needs to examine the bioisosteres visually, a difference between an 1-hour wait and a 10-second wait could prove to be of vast interest in making the process much more efficient.

Acknowledgments

This work has been supported by Program for career development of young scientists, BAS.

Bibliography:

[1] Blaise Barney. An introduction to parallel computing. In https://computing.llnl.gov/tutorials/parallel comp/, USA, 2009. Lawrence Livermore National Library.
[2] Mark Esman. Kind of machine to run Stata. In http://www.stata.com/support/faqs/win/specs.html, USA, 2006. StataCorp, LP
[3] NVIDIA. Cuda programming guide. Pp. 1–98, 2701 San Tomas Expressway, USA, 2008. NVIDIA Corporation.
[4] Harry F. Jordan and Gita Alaghband. Parallel machines and computations. In Fundamentals of Parallel Programming, Marcia J. Horton, editor, pages 2–16, New Jersey, USA, 2003. University of Colorado, Prentice Hall
[5] Vinter A and Rose S, Molecular Field Technology and its Applications in Drug Discovery, Innovations in Pharmaceutical Technology, 23, pp. 14-18, 2007
[6] Bellenie BR, Barton NP, Emmons AJ, et al, Discovery and optimization of highly ligand-efficient oxytocin receptor antagonists using structure-based drug, Bioorganic and Medicinal Chemistry Letters, 19, pp. 990-994, 2009
[7] Dart MJ, Wasicak JT, Ryther KB, et al, Structural aspects of high affinity ligands for the a4b2 neuronal nicotinic receptor, Pharmaceutica Acta Hevetiae, 74, pp. 115, 2000
[8] T. Cheeseright, The Identification of Bioisosteres as Drug Development Candidates, Cresset BioMolecular Discove, 2007
[9] Comparative Molecular Field Analysis (CoMFA) Hugo Kubinyi BASF AG, D-67056 Ludwigshafen, Germany