The VEGA suite of programs: a versatile molecular modeling platform for cheminformatics and drug design projects

Alessandro Pedretti*, Angelica Mazzolari, Silvia Gervasoni, Laura Fumagalli, and Giulio Vistoli

Dipartimento di Scienze Farmaceutiche, Università degli Studi di Milano, Via Luigi Mangiagalli, 25, I-20133 Milano, Italy

Abstract

The purpose of the paper is to offer an overview of the latest release of the VEGA suite of programs. This software has been constantly developed and freely released during the last 20 years and has now reached a significant diffusion and technology level as confirmed by the about 22500 registered users. While being primarily developed for drug design studies, the VEGA package includes cheminformatics and modeling features, which can be fruitfully utilized in various contexts of the computational chemistry. To offer a glimpse of the remarkable potentials of the software, some examples of the implemented features in the cheminformatics field and for structure-based studies are discussed. Finally, the flexible architecture of the VEGA program which can be expanded and customized by plug-in technology or scripting languages will be described focusing attention on the HyperDrive library of highly optimized functions.

Availability: The VEGA suite of programs is available free of charge for non-profit organizations at http://www.vegazz.net.
Contact: alessandro.pedretti@unimi.it

1 Introduction

While many freeware visualization tools have been proposed in the computational chemistry field by enjoying the progresses experienced by computer graphics, (Martinez et al, 2019) very few freeware pieces of software include features which go beyond the simple graphical analysis. Indeed the available resources rarely comprise tools by which one may perform every phase of a molecular modeling study from the design and preparation of the molecular structures to the analysis of the generated results. For more than twenty years, our scientific group has been involved in the never-ending development of the VEGA suite of programs (Pedretti et al, 2002) which includes features to perform a wide variety of simulations and to extensively analyze the obtained results. While being primarily developed for drug design studies, the VEGA platform includes cheminformatics resources which can be useful in a wide range of computational researches. Notably, the VEGA program has an open and flexible architecture by which custom features can be easily implemented by plugin design and scripting language (Pedretti et al, 2004).

2 The VEGA suite of programs

The VEGA program is entirely written using the C and C++ language thus assuring a remarkable stability and portability. To date, only the Windows executable is available and the package can be easily installed by freely download the setup file. The first VEGA public command line version (v. 1.1) was released in 1998, while the program has a graphic interface with a 3D OpenGL output from release 1.3 (2001). The current VEGA 3.2.1 is the 42nd major release freely distributed during these 20 years. Such a remarkable track record witnesses the noteworthy efforts spent to constantly update this software by paralleling the progresses experienced by the molecular modelling and cheminformatics fields. Thus, the current VEGA suite of programs has reached a significant technology level and includes a huge number of features by which one can perform most of the common activities required during an *in silico* project.

Starting from 2004 with the distribution of release 2.0 (the so-called VEGA ZZ program), the download of the software is free but requires an user registration based on Activation and Product keys which allow a precise monitoring of the VEGA users. The analysis of the registration data (updated at March 13, 2020) reveals that VEGA possesses 22466 registered users with 32205 active licenses. The users are worldwide distributed since they belong to 168 different countries. Along with scientific applications, the VEGA platform plays a notable role for educational purposes as demonstrated by the significant abundance of registered students (≅ 40%). Overall, the users belong to 3611 different academic/non-profit organizations and 323 commercial companies thus emphasizing that the VEGA software is primarily utilized within the academic community.

While avoiding a systematic description of the VEGA features, the following sections will focus on some relevant implemented tools arranged into cheminformatics features, tools for structure-based studies and tools to expand and customize its features.

2.1 Cheminformatics features

Along with the various features to download structures directly from on-line resources (such as PDB, PubChem and Zinc), to draw and modify molecules (by 2D and 3D editors), to build peptides (by sequence and secondary structure) and to convert a plethora of molecular

file formats, the features to manage databases represents one of the most important VEGA strengths. The database explorer is able to manage databases with different formats, such as accdb, sdf, mol2, mdb, MySQL, SQLite and compressed archives, and can also handles set of SMILES strings as stored in csv format. To give a taste of the various functions implemented, the database explorer can perform for each molecule inserted into a database a set of preliminary tasks such as 1) converting SMILES or 2D to 3D structures, 2) adding the missing hydrogens, 3) generating the tautomers, the stereo and geometric isomers, 4) ionizing at a given pH, 5) assigning atom types and atomic charges and 6) optimizing the resulting structures. Again, for each collected molecule, the database explorer can calculate a set of molecular properties which can be exported for further analyses (excel, ARFF and CSV formats are supported). The databases can be searched by similarity in respect to a given input molecule or can be filtered by applying a set of filters based on molecular properties or on the occurrence of specific functional groups. The implemented features to manage databases can be used to develop targeted applications as exemplified by the recently published MetaQSAR (Pedretti et al, 2018), a database engine developed purposely to collect, to manage and to analyze metabolic data. MetaQSAR represents the engine for a set tools for the metabolism prediction implemented in the VEGA program as exemplified by the plug-in for the FAME approach (Šícho et al, 2019). Finally, the implemented functions to manage and convert SMILES strings were utilized to develop the CombiSMILES tool, which generates set of molecules by systematically adding libraries of possible substituents at defined positions of a given scaffold. The molecular properties calculated for the databases can then be used 1) for correlative studies using an included tool which develops the predictive models by exhaustively combining the given properties; 2) for classification analysis by using the recently published EFO algorithm (Mazzolari et al, 2018) or 3) by external programs such as WEKA (https://www.cs.waikato.ac.nz/ml/weka).

2.2 Features for structure-based studies

The VEGA program includes several features to handle and to analyze protein structures and comprises tools to solvate and neutralize complex systems using different solvents including membrane models. The so prepared systems can be then easily simulated by exploiting the implemented graphical interface for NAMD 2 (Phillips et al, 2005). Clearly, all above described features for databases find fruitful applications in preparing ligand datasets for docking simulations and virtual screening campaigns. To this end, the VEGA suite of programs includes graphical interfaces for some well-known docking engines such as AutoDock 4 (Morris et al, 2009) for which a parallelized version was implemented (GriDock, Pedretti et al, 2010), AutoDock Vina (Trott et al, 2010) and PLANTS (Korb et al, 2009). Furthermore, the program includes some relevant tools for the post-processing of the docking results such as ReScore+, (Pedretti et al 2016) which rescores the computed poses by using a set of well-known scoring functions, including the MLPInS (Vistoli et al, 2010) to describe the hydrophobic interactions and the number of stabilized contacts, as well as a set of scripts to compare the poses or to perform statistical analyses on the obtained score values (Vistoli et al, 2017). Notably, the above cited EFO algorithm proved successful in developing consensus models to maximize the predictive power of virtual screening simulations (Pedretti et al, 2019). Finally, the VEGA platform includes an engine (WarpEngine, Pedretti et al, 2018) to distribute almost all performable calculations on the CPUs available also on local networks.

2.3 Flexible and customizable architecture

The above mentioned possibility to implement new features in the VEGA platform by using plug-in architectures or scripting languages has ben previously described (Pedretti et al, 2004). The list of VEGA commands which can be used to develop scripts in C (a built-in C compiler is provided in the package), DOS batch script, JavaScript, PHP, Python, R and REBOL is constantly updated and can be found in https://www.ddl.unimi.it/manual/pages/gl_index.htm. While requiring more skilled users in programming, the software also includes a SDK and the relative documentation to develop new plug-ins. Here, attention is focused on the HyperDrive technology which is the core library including several time-critical functions required for high speed computing. The highly optimized and parallel code, especially designed for the modern CPUs, allows an easy optimization of the program performances. While being targeted for developing molecular modelling programs, the library also comprises functions useful for generic applications. In detail, the key characteristics of HyperDrive can be summarized, as follows: 1) Hardware independent: the same application can be developed for Linux (ARM, x86 and x64) and Windows (x86 and x64) without changing the source code. 2) Same software for single or multiprocessor systems: HyperDrive checks the number of available CPUs and switches from sequential to parallel mode. 3) Simultaneous multithreading (SMT) ready: HyperDrive can exploit the full power of modern multiscalar CPUs with hardware multithreading. 4) OpenCL support: some routines are written to run on the GPU through the OpenCL abstraction layer (https://www.khronos.org/opencl/) 5) SIMD optimization: The most common functions are written in assembly and optimized using the SSE/SSE2 SIMD instruction sets (https://software.intel.com/sites/landingpage/IntrinsicsGuide/). Though developed to be used within the VEGA environment, HyperDrive can be used in external applications without requiring specific C/C++ compilers by including the C++ wrappers in the headers. In detail, the implemented functions range from specific molecular modeling functions (such as the surface calculation) to mathematical function (such as discrete Fourier transformation) until to basic low-level functions (such as file management or string manipulation).

3 Conclusions

By combining the VEGA package with a set of free of charge cited programs, one may build a reasonably complete molecular modeling suite which is primarily developed for drug design studies but can find fruitful application in the different contexts of the computational chemistry. In this way, one may conduct in silico experiments with very limited economic resources so much so that the VEGA programs is very often used for educational purposes and for academic researches in the developing countries. Our commitment is then to offer an always updated and ever more powerful tool which allows anyone to carry out high level in silico studies at virtually no software cost.

Conflict of Interest: none declared.

References

Korb O. et al (2009) Empirical scoring functions for advanced protein-ligand docking with PLANTS. *J Chem Inf Model*. **49**, 84-96.

Mazzolari A et al (2018) Prediction of the formation of reactive metabolites by a novel classifier approach based on enrichment factor optimization (EFO) as implemented in the VEGA program. *Molecules* **23** pii: E2955.

Martinez X. et al (2019) Molecular graphics: bridging structural biologists and computer scientists. *Structure*, **27**, 1617-1623.

Morris, GM., et al (2009) Autodock4 and AutoDockTools4: automated docking with selective receptor flexiblity. *J. Comput. Chem*. **30**, 2785-91.

Pedretti, A. et al (2002) VEGA: a versatile program to convert, handle and visualize molecular structure on Windows-based PCs. *J Mol Graph Model.* **21**, 47-49.

Pedretti, A. et al (2004) VEGA--an open platform to develop chemo-bio-informatics applications, using plug-in architecture and script programming. *J Comput Aided Mol Des.* **18**, 167-73.

Pedretti A et al (2016) Structural effects of some relevant missense mutations on the MECP2-DNA binding: a MD study analyzed by Rescore+, a versatile rescoring tool of the VEGA ZZ program, *Mol Inform.* **35**, 424-33.

Pedretti, A. et al (2018) MetaQSAR: an integrated database engine to manage and analyze metabolic data. *J Med Chem*, **61**, 1019-1030.

Pedretti A et al (2018) WarpEngine, a flexible platform for distributed computing implemented in the VEGA program and specially targeted for virtual screening studies. *J Chem Inf Model*. **58**, 1154-1160.

Pedretti A et al (2019) Rescoring and linearly combining: a highly effective consensus strategy for virtual screening campaigns. *Int J Mol Sci.* **20** pii: E2060.

Phillips JC et al (2005) Scalable molecular dynamics with NAMD. *J. Comput. Chem*. **26**, 1781-1802.

Šícho, M. et al, (2019) FAME 3: predicting the sites of metabolism in synthetic compounds and natural products for phase 1 and phase 2 metabolic enzymes. *J Chem Inf Model.* **59**, 3400-3412.

Trott, O. and Olson, AJ. (2010) AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. *J Comput Chem*. **31**, 455–461.

Vistoli G et al (2010) In silico prediction of human carboxylesterase-1 (hCES1) metabolism combining docking analyses and MD simulations. *Bioorg Med Chem*. **18**, 320-9.

Vistoli G et al (2017) Binding Space Concept: A new approach to enhance the reliability of docking scores and its application to predicting butyrylcholinesterase hydrolytic activity. *J Chem Inf Model.* **57**, 1691-1702.