Published online 29 April 2017 *Nucleic Acids Research, 2017, Vol. 45, Web Server issue* ***W315–W319***
doi: 10.1093/nar/gkx337

GASS-WEB: a web server for identifying enzyme active sites based on genetic algorithms

João P. A. Moraes[1], Gisele L. Pappa[2], Douglas E. V. Pires[3,*] and Sandro C. Izidoro[1,*]

[1]Department of Computer Engineering, Advanced Campus at Itabira, Universidade Federal de Itajubá - UNIFEI, Itabira, 35903-087, Brazil, [2]Department of Computer Science, Universidade Federal de Minas Gerais - UFMG, Belo Horizonte, 31270-901, Brazil and [3]Centro de Pesquisas René Rachou, Fundação Oswaldo Cruz, Belo Horizonte, 30190-002, Brazil

Received February 18, 2017; Revised April 08, 2017; Editorial Decision April 17, 2017; Accepted April 27, 2017

ABSTRACT

Enzyme active sites are important and conserved functional regions of proteins whose identification can be an invaluable step toward protein function prediction. Most of the existing methods for this task are based on active site similarity and present limitations including performing only exact matches on template residues, template size restraints, despite not being capable of finding inter-domain active sites. To fill this gap, we proposed GASS-WEB, a user-friendly web server that uses GASS (Genetic Active Site Search), a method based on an evolutionary algorithm to search for similar active sites in proteins. GASS-WEB can be used under two different scenarios: (i) given a protein of interest, to match a set of specific active site templates; or (ii) given an active site template, looking for it in a database of protein structures. The method has shown to be very effective on a range of experiments and was able to correctly identify >90% of the catalogued active sites from the Catalytic Site Atlas. It also managed to achieve a Matthew correlation coefficient of 0.63 using the Critical Assessment of protein Structure Prediction (CASP 10) dataset. In our analysis, GASS was ranking fourth among 18 methods. GASS-WEB is freely available at http://gass.unifei.edu.br/.

INTRODUCTION

Active sites are regions usually on the surface of enzymes specially modelled by nature during evolution that either catalyse a reaction or are responsible for substrate binding. The active site can be, therefore, divided into two parts, which include the catalytic site and the substrate binding site (1). Active site amino acid residues are known to be more conserved during evolution than other enzyme regions, a useful information that has been used in function prediction tasks (2,3). A number of methods based on the structure of these active sites have been proposed over the years to infer protein function based on active site similarity (4–6). Given an active site template, these methods use different mathematical modelling and searching procedures to match the template to a given set of proteins (4–8).

Many of the current available methods present, however, limitations such as performing only exact matches on template residues (not accounting for conservative mutations), restricting the number of amino acids in the template or pruning the search space, using *ad-hoc* procedures and are usually not capable of finding inter-domain active sites.

In order to tackle these problems, we had proposed GASS (Genetic Active Site Search) (9), a search method based on genetic algorithms that aims to cope with the aforementioned issues. Since then, we have proposed MeGASS (10), a multi-objective version of GASS that also considers the depth of the residues when searching for active sites. This was important as, on general, active sites are closer to the protein surface to allow access of the substrate.

Here we propose a user-friendly web server implementation of our method, called GASS-WEB. Our method is now capable of, in addition to catalytic sites templates, perform searches based on binding sites templates. The web server interface has been improved and complementary information such as enzyme EC number, UNIPROT accession code and structure resolution are now presented on the results page. The method can be used under two different scenarios: (i) given a protein of interest, it matches the protein to a set of specific templates (i.e. known active sites) stored in a database; or (ii) given an active site template, it searches for it in a database of protein structures. GASS-WEB is freely available through an intuitive web interface at http://gass.unifei.edu.br/.

*To whom correspondence should be addressed. Tel: +55 3138390800; Email: sandroizidoro@unifei.edu.br
Correspondence may also be addressed to Douglas E. V. Pires. Tel: +55 3133497720; Email: douglas.pires@cpqrr.fiocruz.br

MATERIALS AND METHODS

Datasets

The GASS-WEB database consists of active site templates and their respective proteins structures obtained from the Protein Data Bank. GASS-WEB uses 1691 catalytic site templates based on the Catalytic Site Atlas (CSA) (11). Only literature entries were used. The database is also composed by 1819 binding sites templates from CSA (literature entries only) and 23 318 enzymes from the NCBI-VAST non-redundant database (12).

A complementary dataset composed of six subsets (DS) randomly chosen from CSA was also used to test the GASS-WEB. Each DS has a specific protein family (distinct EC number). All protein structures were extracted from the PDB, and their active sites validated by CSA. Each subset has its own literature-derived template and proteins annotated as CSA HOMOLOG (Supplementary Table S1). The dataset has 551 proteins and 6 templates LIT.

Genetic active site search

The heuristic search behind GASS-WEB uses the GASS method, which works in two steps: (i) a data pre-processing step, which generates the databases used by the server and (ii) a genetic algorithm (GA) (13,14), which performs the search itself (Figure 1). The pre-processing step finds the proteins of interest to the user and active sites templates in Protein Data Bank (PDB) (15), CSA and NCBI-VAST database and returns, for each amino acid, its name, chain, the last heavy atom and its coordinates, which composes the template. This information is stored in a relational database, and accessed by GASS to create its initial pool of candidate active sites. GASS then performs a heuristic search to find matching active sites in the selected proteins and outputs one or more candidate active sites. The method can also search in a database of proteins structures an active site template. In order to account for conservative mutations, GASS also has the option of consulting a substitution matrix (8).

The GA implemented by GASS evolves a population of individuals, where each individual represents a solution to the problem, i.e. a candidate active site match. These solutions are evaluated and ranked according to a fitness function (Supplementary Equation S1), which for GASS-WEB is a modified root-mean-squared deviation (RMSD) between the template and searched amino acid residues. The fitness function indicates the degree of structural similarity of the candidate active site with the template. After evaluation, individuals are selected to undergo crossover and mutation operations according to tuned probabilities. This process goes on until a stop criterion, which is usually based on a maximum number of generations (iterations), is met.

Different from other search methods, GASS returns a ranking of the n best solutions found. Since it is possible that the best solution (best fitness) is buried in the protein, instead of being in a pocket, it might be interesting for the user to analyse a set of solutions to filter potential false positive cases.

WEB SERVER

GASS-WEB was implemented using the Flask framework for Python (16), and the front-end design was created using the Bootstrap framework (17), a well-established tool for intuitive design. GASS-WEB back-end runs on top of an Apache server, and the communication with the server is made through a Web Server Gateway Interface, as determined by the Python Enhancement Proposal 333.

The server works by filtering and redirecting the user input to a C++ implementation of the GASS search method. All requests are queued using the Redis Queue data structure and handled using RQ workers, this allows for an easy control and parallel processing of the jobs running.

Input

All searches on GASS-WEB are performed based on a PDB file, which can be provided by the user through the four letter code (as used by the RCSB PDB), or by uploading a custom file. This custom file needs to be in accordance with PDB format standards. The file is then validated and converted into a binary file for the heuristic search method.

GASS-WEB offers three different types of resources: searching proteins for CSA active sites (that uses CSA catalytic sites templates), searching proteins for CSA binding sites (that uses binding sites templates generated from the CSA literature entry) and searching the NCBI-VAST database for specific active site templates.

Both resources, CSA sites search and CSA binding sites search, have the same input. To perform a search using CSA sites search, for example, it is necessary to provide a protein structure by either uploading your own file, which must comply with the PDB format, or supplying a four-letter PDB code (Figure 2A1 and A2). The next step is to choose the template size for matching, which is the number of residues of the active site (Figure 2A3). Then one is ready to submit your query to GASS-WEB (Figure 2A4). GASS-WEB takes about 1 min to show the results in both resources.

The protein search using the NCBI-VAST database also requires the protein structure by either uploading your own file (PDB format) or supplying a four-letter PDB code (Figure 2B1 and B2). In contrast, it requires an active site template (Figure 2B3). The format of the template is detailed on the website. The protein search using the NCBI-VAST database takes considerably longer to finish than the other two previously presented (about 50 min), as it compares the template to all other proteins in NCBI-VAST database. For this reason, an optional field was added, allowing the user to receive the results via email once the search finishes (Figure 2B4).

Output

After the search is complete, the user is automatically redirected to the results page where results are displayed and also available as a CSV file for download (Figure 2C8), which will be kept for a week and can be accessed using the job URL. The results are displayed on a table ordered by fitness score of the matched residues (modified RMSD)

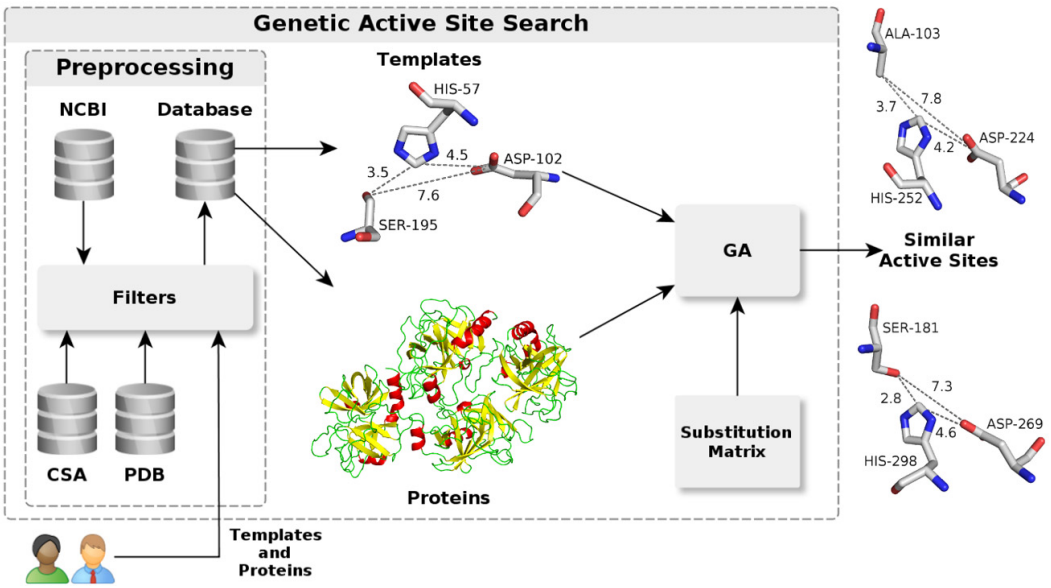

Figure 1. GASS method: data is extracted from PDB, CSA and NCBI and pre-processed. GASS performs a heuristic search to find matching active sites in the proteins of interest or, given an active site template it searches for it in the database of protein structures using a genetic algorithm. Matching active sites are then returned to the user.

of each solution (Figure 2C1), followed by a list of residues found by GASS-WEB (Figure 2C2), the template (PDB ID) (Figure 2C3), residues list of the matched template (Figure 2C4), EC number, Uniprot accession code and resolution of the matched template (Figure 2C5, C6 and C7). The columns C2 and C3 in Figure 2 have a small icon to visualize the predicted active site and the template using GLMol plugin.

In the case of searching for binding sites, a column displaying the ligand in the matched template is also added to the results. The output is also available for download as a CSV file. Reporting a ranking of candidate active sites allows users to evaluate the results closely, and easily identify potential false positives, given the complementary supporting information shown for each solution found.

RESULTS

The GASS method was previously tested, and proved to be very effective in a number of experiments. Based on the CSA annotation, it was able to correctly identify more than 90% of the catalytic sites catalogued. In specific enzymes sets as the Nitric Oxide Synthase (125 enzymes) and Trypsin-like (1085 enzymes), GASS-WEB found more than 90% of the active site correctly within the fifth place in the ranking (9). This property is very desirable since it facilitates the user's visual inspection. It also managed to achieve a Matthew correlation coefficient (MCC) of 0.63 using the Critical Assessment of protein Structure Prediction (CASP 10) dataset. In our analysis, GASS was ranking fourth among 18 methods (9).

To further evaluate the method's performance, we carried out a test involving a dataset composed of 6 subsets (DS1:DS6) randomly chosen from CSA literature entries. Each DS has a template and a specific protein family (distinct EC number). All proteins were extracted from PDB and their active sites validated in CSA. Analysing the results of GASS-WEB, the average accuracy for all the subsets was 99%.

Figure 3 (blue) shows a cumulative match score curve (CMS) for the experiment. This curve shows the relation between the number of correct catalytic sites found according to CSA and their position in the GASS-WEB ranking. As observed, the best catalytic site candidates appear mostly at the top five positions of the ranking. The accuracy value of each subset as well as a ROC graph are in the Supplementary Data.

In another experiment, GASS-WEB found correctly the catalytic site of the enzyme 2GCT (Gamma-Chymotrypsin A), according to CSA at the first position of the ranking. Annotated by homology, the enzyme has HIS 57 and ASP 102 in chain B, and SER 195 in chain C, showing that finding inter-domain catalytic sites are not a limitation for the method. Figure 2C shows the result of this experiment. Further case studies can be found in Supplementary Data.

In a complementary experiment we evaluated the fitness distribution obtained by GASS-WEB on a large-scale search, with NCBI-VAST database (23 318 proteins) and the template enzyme 1ACB ((bovine alpha-chymotrypsin-eglin C complex). The number of catalytic sites found according to CSA was 270 (1.16% of all structures), and of these structures, 138 (51.11%) presented lower fitness than 1 Å, and 126 (46.67%) presented fitness between 1Å and 5Å. Thus, 97.78% of the sites found according to the CSA have a fitness value $\leq$5 Å. This indicates that, according to the fitness function (Supplementary Data), the sites found have a high degree of structural similarity with the template. Figure 3 (red) shows a CMS for the experiment.

Analysing the 200 similar active sites reported by GASS-WEB using the template 1ACB, the number of catalytic sites found according to CSA was 31. It is important to empha-

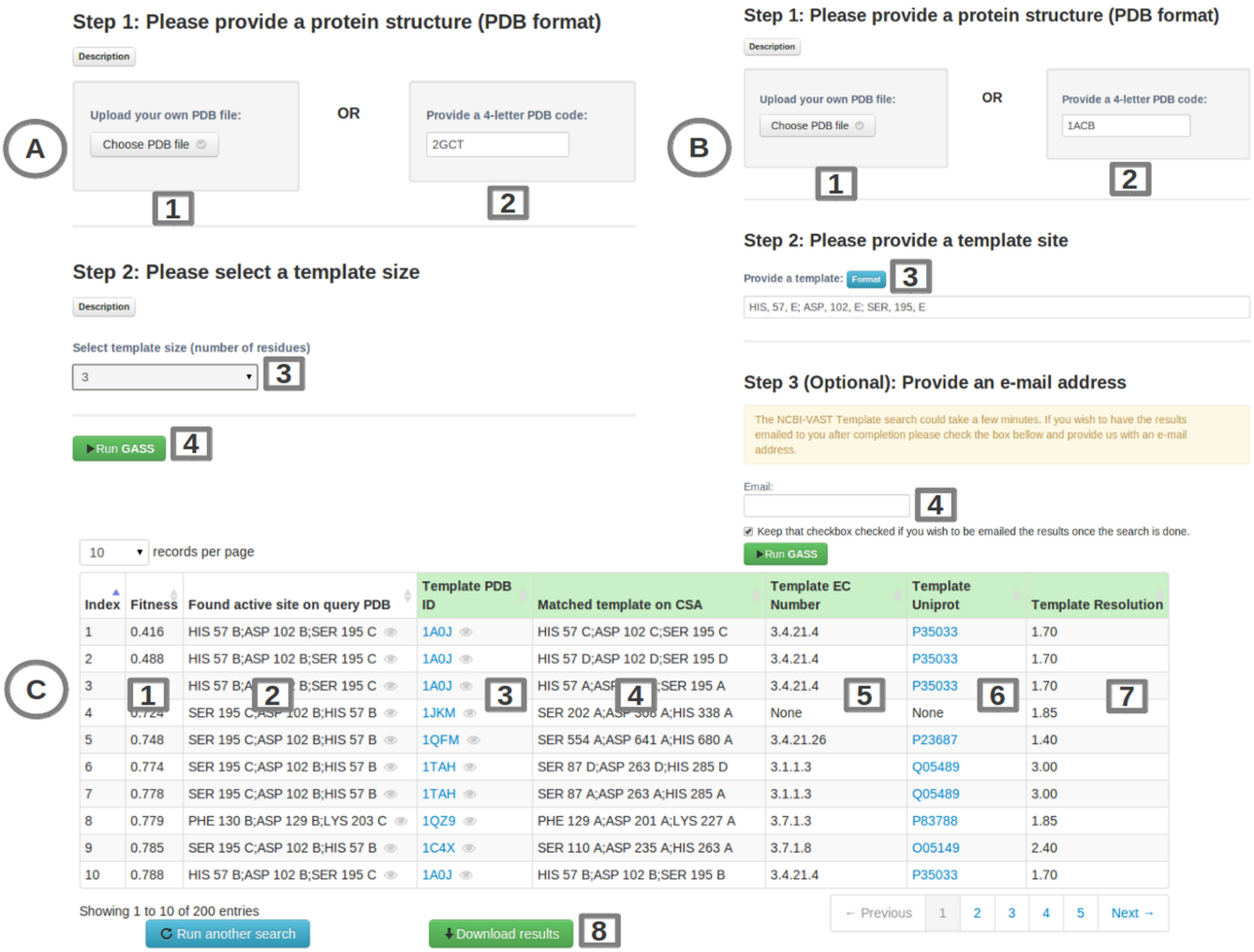

Figure 2. (**A**) Protein search for catalytic or binding sites requires the protein PDB file (Step 1) and the template size (Step 2). (**B**) NCBI-VAST Database search requires the PDB file (Step 1) and a template (Step 2), and has an optional field for email allowing the user to be emailed once the search finishes (Step 3). (**C**) Protein search for catalytic sites results page.

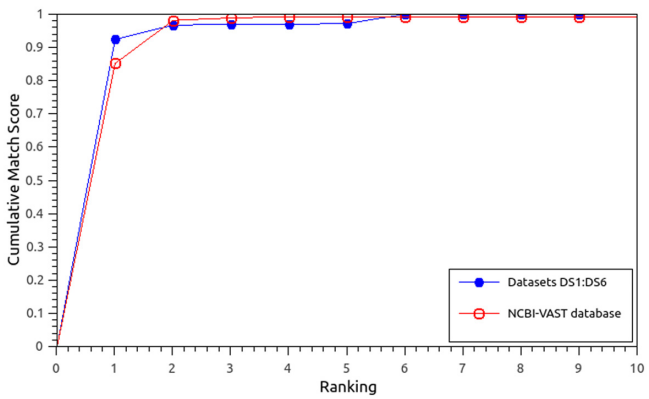

Figure 3. CMS score of catalytic sites found correctly according to the CSA—Datasets DS1:DS6 (blue). CMS score of catalytic sites found correctly according to the CSA—NCBI-VAST database experiment (red).

size that among the results reported by GASS-WEB there are correct active sites that are not annotated in the CSA (for more examples please see Supplementary Data).

As well as the fitness value, ranking position can be very useful information, assisting the user in the inspection and validation of the results.

CONCLUSION

GASS-WEB is a free and a user-friendly web server created for searching similar active sites based on data from the PDB, CSA and NCBI-VAST. Based on these three different resources, GASS-WEB can use catalytic and binding sites templates to search similar sites in a protein. It also can use an specific active site template to search similar active sites in a protein database.

Without the limitations of the traditional methods (performing only exact matches, restricting the number of amino acids in the template or pruning the search space using *ad-hoc* procedures besides finding inter-domain active sites) the method showed to be effective in finding similar active sites in most datasets.

In our analysis, GASS-WEB managed to achieve a MCC of 0.63 on the Critical Assessment of protein Structure Prediction (CASP 10) dataset (ranking fourth among 18 methods) besides being able to correctly identify >90% of the catalogued active sites in CSA. In a dataset with six specific protein families (551 proteins and 6 templates), the GASS-WEB average accuracy was 99%.

On a large-scale search (NCBI-VAST), 97.78% of the sites found according to the CSA had a fitness value ≤5 Å and appeared mostly at the top five positions of the ranking. This implies that both fitness value and ranking position can help the user in the inspection and validation of the results.

During the experiments some sites were found according to the literature but are not included in the CSA (for more details please see Supplementary Data). This indicates that our method can be of great help in improving and increasing coverage of resources including the CSA and similar databases. We believe, therefore, that GASS-WEB could be an invaluable tool for assisting protein function prediction and active site annotation.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Dr Edelma Eleto da Silva helpful discussions and aid in developing the graphical abstract of this work.

FUNDING

Medical Research Council (MRC) Newton Fund RCUK-CONFAP Grant; Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) [APQ-00828-15 to D.E.V.P.]; Centro de Pesquisas René Rachou (CPqRR/FIOCRUZ Minas), Brazil (to D.E.V.P.); CAPES; CNPq; Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) (to S.C.I., G.L.P., J.P.A.M.). Funding for open access charge: MRC Newton Fund RCUK-CONFAP Grant; FAPEMIG; Centro de Pesquisas René Rachou (CPqRR/FIOCRUZ Minas), Brazil; CAPES; CNPq.

Conflict of interest statement. None declared.

REFERENCES

1. Kahraman,A. and Thornton,J. M.(2008) Methods to Characterize the Structures of Enzyme Binding Sites. In: Schwede,T and Peitsch,M (eds). *Computational Structural Biology: Methods and Applications*. World Scientific Publishing, London, pp. 189–221.
2. Torrance,J.W. and Thornton,J.M. (2008) Structure-based Prediction of Enzymes and Their Active Sites. In: Bujnicki,JM (ed). *Prediction of Protein Structures, Functions, and Interactions*. John Wiley & Sons, Ltd, pp. 187–209.
3. Roche,D.B., Brackenridge,D.A. and McGuffin,L.J. (2015) Proteins and Their Interacting Partners: an introduction to protein-ligand binding site prediction methods. *Int. J. Mol. Sci.*, **16**, 29829–29842.
4. Stark,A. and Russell,R.B. (2003) Annotation in three dimensions. PINTS: patterns in non-homologous tertiary structures. *Nucleic Acids Res.*, **31**, 3341–3344.
5. Brylinski,M. and Skolnick,J. (2008) A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 129–134.
6. Wass,M.N., Kelley,L.A. and Sternberg,M.J. (2010) 3DLigandSite: predicting ligand-binding sites using similar structures. *Nucleic Acids Res.*, **38**, W469–W473.
7. Nadzirin,N., Gardiner,E.J., Willett,P., Artymiuk,P.J. and Firdaus-Raih,M. (2012) SPRITE and ASSAM: web servers for side chain 3D-motif searching in protein structures. *Nucleic Acids Res.*, **40**, W380–W386.
8. Nilmeier,J.P., Kirshner,D.A., Wong,S.E. and Lightstone,F.C. (2013) Rapid catalytic template searching as an enzyme function prediction procedure. *PLoS One*, **8**, e62535.
9. Izidoro,S.C., de Melo-Minardi,R.C. and Pappa,G.L. (2015) GASS: identifying enzyme active sites with genetic algorithms. *Bioinformatics (Oxford, England)*, **31**, 864–870.
10. Izidoro,S.C., Lacerda,A.M. and Pappa,G.L. (2015) Megass: multiobjective genetic active site search. In: *Genetic and Evolutionary ComputationConference (GECCO 2015 - Madrid - Spain)*. ACM, NY, pp. 905–910.
11. Furnham,N., Holliday,G.L., de Beer,T.A., Jacobsen,J.O., Pearson,W.R. and Thornton,J.M. (2014) The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes. *Nucleic Acids Res.*, **42**, D485–D489.
12. Madej,T., Lanczycki,C.J., Zhang,D., Thiessen,P.A., Geer,R.C., Marchler-Bauer,A. and Bryant,S.H. (2014) MMDB and VAST+: tracking structural similarities between macromolecular complexes. *Nucleic Acids Res.*, **42**, D297–D303.
13. Baeck,T., Fogel,D.B. and Michalewicz,Z. (1997) *Handbook of Evolutionary Computation*. Oxford University Press, Bristol.
14. Goldberg,D.E. (1989) *Genetic Algorithmsin Search, Optimization and Machine Learning*. Addison-Wesley, Boston.
15. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
16. Grinberg,M. (2014) *Flask Web Development: Developing Web Applications with Python*. O'Reilly Media, Sebastopol.
17. Spurlock,J. (2013) *Bootstrap: Responsive Web Development*. O'Reilly Media, Sebastopol.