MeGASS: Multi-Objective Genetic Active Site Search

Sandro Izidoro
Computer Science Dep.
UNIFEI
Itabira, Brazil
sandroizidoro@gmail.com

Anisio M. Lacerda
Computer Science Dep.
CEFET-MG
Belo Horizonte, Brazil
anisio@decom.cefetmg.br

Gisele L. Pappa
Computer Science Dep.
UFMG
Belo Horizonte, Brazil
glpappa@dcc.ufmg.br

ABSTRACT

Active sites are regions in the enzyme surface designed to interact with other molecules. Given their importance to enzyme function, active site amino acids are more conserved during evolution than the whole sequence, and can be a useful source of information for function prediction. For this reason, great effort has been put into identifying active sites in proteins. The majority of methods for this purpose uses an active site template of a protein of known function to search for similar structures into proteins of unknown function. In this direction, we recently proposed GASS (Genetic Active Site Search), a method based on an evolutionary algorithm to search for active sites in proteins. Although the method obtained very accurate results, its main strength and weakness are related to using only the spatial distance from the template to the protein to evaluate candidate sites. In this direction, this paper proposes MeGASS, a multi-objective version of GASS that also considers the depth of the residues when looking for active sites. This is important, as active sites are known for being closer to the protein surface to allow interactions with ligands. Results showed the depth attribute improves over the results of GASS, and its role into the method is worth further investigation.

Categories and Subject Descriptors

J.3 [**Computer Applications**]: Life and medical sciences; I.2.8 [**Computing Methodologies**]: Artificial Intelligence—*problem solving, control methods, search*

1. INTRODUCTION

Binding sites are regions in the enzyme surface designed to interact with other molecules [16]. There are many types of binding sites, including allosteric, receptors, active, among others. Here we are interested in active sites. Active sites are commonly divided into two parts: catalytic site and substrate binding site. The first is usually a set of 2 to 6 amino acids responsible for catalytic reactions. The latter recognizes the molecules with which the enzyme works, and can be larger than the catalytic site, reaching up to 20 amino acids.

Given their importance to enzyme function, active site amino acids are more conserved during evolution than the whole sequence. Consequently, they can be a useful source of information for function prediction [21, 3], besides being a key element to the process of drug discovery. Hence, great attention has been given to active site identification methods.

Most of the methods for active site identification proposed in the literature represent enzymes as graphs, where each node corresponds to an amino acid of the side chain and is represented by one or more atoms, and each edge a connection between neighbour atoms. Then, classical and more sophisticated methods for graph search, including depth search and geometric hashing [22], are used to identify active sites based on known active site templates.

However, due to the problem of graph isomorphism, these methods usually impose some restrictions to make the search space more tractable. They include setting a maximum acceptable distance between two neighbour atoms [20] or restricting the maximum size of the active site template [17]. In order to tackled these problems, we recently proposed GASS (Genetic Active Site Search) [10], which does not impose any restrictions such as those aforementioned and, above all, can precisely identify the chain where the residues of the active site are located. Difficulties in correctly identifying the chain where the active site residues are located is one of the main drawbacks of the current methods, as showed in [10].

GASS is an evolutionary algorithm designed to find active sites in proteins. It receives as input a template and an unknown protein, and searches for the active site template in the protein residues. Although GASS presented very accurate results in active site identification, it still has room for improvements. Among these improvements are its evaluation function.

The original version of GASS evaluates individuals according to their spatial distance to the active site template, without accounting for any other properties of the proteins. This can be seem as an advantage of GASS, as it does not need a lot of information from the protein. At the same time, it may be considered a drawback, as it allows individuals with certain characteristics not common to active sites to be considered as such. One of this properties relates to the position of the active site regarding the protein surface. Active sites usually have at least one of their amino acids

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

GECCO '15, July 11 - 15, 2015, Madrid, Spain

© 2015 ACM. ISBN 978-1-4503-3488-4/15/07…$15.00

DOI: http://dx.doi.org/10.1145/2739482.2768436

close to the protein surface, facilitating the contact with potential ligands [16].

In this direction, this paper introduces MeGASS, a multi-objective version of GASS that uses alternative ways of evaluating the potential of a candidate active site. In particular, we investigate the advantages of using active sites depth in relation to the protein surface as an additional information. As previously mentioned, this is relevant because at least one residue of the active site is usually closer to the protein surface to facilitate interaction. We present two ways of incorporating depth to the current fitness, including a multi-objective lexicographic approach [6].

The original version of GASS returns a ranking of candidate active sites to the user. So far, the results showed that the addition of depth has a small positive impact in the results found in the top 10 positions of the ranking, but overall MeGASS brings the correct active sites closer to the top of the ranking, indicating further investigation on this matter is relevant.

2. RELATED WORK

In the literature, the similarity search of active sites can be performed using information about the sequence or the structure of the proteins. Many of the works proposed in the literature are based on multiple sequence alignment of different organisms to detect active site conservation [21, 15]. These methods were the first proposed because initially the availability of sequences was much greater than the availability of structures. However, these methods have serious drawbacks, as the sequences might be similar and perform completely different functions [23] or be so different they cannot be successfully aligned.

Structure-based methods are an alternative to solve the problems aforementioned [2]. However, these methods have to deal with the inherent problems of searching in a 3 dimensional space. Most of methods for identifying active sites based on structure represent the proteins as graphs, where each node represents an atom in the side chain and edges are connections between neighbour atoms, weighted by their distances. Given this representation, methods based on simple depth-first search or graph isomorphism [17] have been proposed.

Here we describe two methods that are close related to ours, given their objectives. The first is ASSAM (Amino acid pattern Search for Substructures And Motifs) [17], which searches for maximum common sub-graphs to find similar structures between the template active site and the protein. Each graph node corresponds to an amino acid in the side chain, which in turn is represented by two pseudo-atoms. ASSAM calculates spatial differences between the protein and template using the Root Mean Squared Deviation (RMSD).

The second method is named Catalytic Site Identification (CatSId) [14]. It performs a protein-to-template matching using a sub-graph search method and a library of catalytic residue templates from CSA (Catalytic Site Atlas) [19] – a database of catalytic sites in enzymes of known 3D structure. These results are then refined using a logistic scoring procedure to re-score the matches found in the first phase, using information such as binding site predictions and others physical descriptors to improve the structure matching previously obtained.

The main problem with the methods previously described is that it is very difficult to compare them with GASS and MeGASS, as they work with their own templates and protein datasets, which were not made publicly available. However, we can highlight that the main differences from the proposed method to them are the heuristics used to search for active sites, the representation of the atoms, the metric used to calculate the distance from the templates and the fact that MeGASS does not impose any restrictions to the search space, as discussed in the next sections.

Finally, although evolutionary algorithms are popular methods in the context of bioinformatics [8, 18], to the best of our knowledge GASS is the first method to solve the problem of active site identification, although other related problems, such as multiple graph alignment for molecule structural analysis [7] and protein structure prediction [5], were previously investigated.

3. AN OVERVIEW OF GASS

GASS was created to solve the following problem. Given a set of N amino acids that compose the active site A_1 of a protein p_A of known function, and a second hypothetical protein p_B of unknown function and sequence size M. The problem is to search for a match of A_1 in p_B. The naive solution to this problem is to enumerate all possible arrangements of N amino acids considering all M available in p_B and select those with most similar amino acid conformation and relative position to p_A. However, this solution becomes intractable as N grows. Hence, GASS appears as an efficient alternative approach.

Figure 1 illustrates GASS, which receives as input a protein and one active site template. The given protein is identified by its PDB (Protein Data Bank) [1] identification number, and the catalytic site can be manually created or extracted from CSA (Catalytic Site Atlas)[19]. PDB is a well-known repository of 3D structures of proteins and nucleic acids, while CSA stores catalytic sites of enzymes of known 3D structure.

Apart from the protein and active site template, GASS also receives as input a substitution matrix, which represents amino acids conservative mutations. Conservative mutations occur when when one type of amino acid is replaced by another with similar biochemical properties, and they represent one of the difficulties of active site matching algorithms. This is because different amino acids may perform the same type of functions in active sites.

Notice that the method can be used under two different scenarios: (i) to find a given template in one or more proteins; (ii) to find sets of templates in one or more given proteins. Below we describe the main components of GASS, available as a webservice[1].

Individual Representation: In GASS, each individual represents a candidate active site, and each gene an amino acid. As what differentiates amino acid is their side chain, each amino acid is represented by the last heavy atom (LHA) of its side chain [11], although experiments with the atom centroid and α-carbon were also performed. The size of the individual is dynamic, and usually assumes the size N of the active site template. Figure 2 shows an example of an individual. Each gene stores the name of the amino acid (considering the first gene in Figure 2, ALA), the name of the LHA (e.g. CB), its chain (e.g. A), position in the sequence

[1]http://gassweb.dcc.ufmg.br/

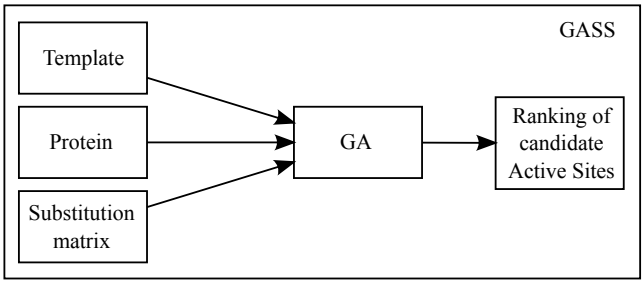

Figure 1: GASS framework.

(e.g. 103), and the 3D position of its LHA (10.551, 41.606, -5.105). The last field, 13.64, represents the depth of the amino acid in relation to the surface of the protein, which will be used as part of the evaluation process by MeGASS.

Population Initialization: The initialization of each individual takes into account the type of the amino acid in each position of the active site template. Hence, considering the example in Figure 2, the first position is always a ALA, extracted from a list of all positions of the protein where ALA is found, followed by an ASP and a HIS.

Fitness function: In the original version of GASS, the fitness function is based on the spatial distance between the individual and the active site template, defined according to Eq. 3, where n is the number of amino acids in the template, v is the candidate active site (individual) and w the active site template. As observed, i accounts for distances between all pairs of residues in the template. Notice that this equation differs from the well-known RMSD metric, as it does not average the squared distances of the results. As shown in [12], slightly different active sites may have very similar RMSD values when the square root is taken. By using their absolute distance values we avoid this problem.

$$Fit_{dist}(v,w) = \sqrt{\sum_{i=1}^{(n^2-n)/2} \|v_i - w_i\|^2} \quad (1)$$

Evolution process: After individuals are evaluated, they undergo a tournament selection and traditional one-point crossover. The mutation operator, apart from introducing variability to the population by replacing genes from the same type of amino acid, has another role: to deal with conservative mutations. If the amino acid selected to undergo mutation can suffer a conservative mutation (according to the substitution matrix described below), the mutated amino acid might be chosen to replace the original amino acid in the individual. The decision of which type of mutation should be applied is given by user defined probabilities, which might account for the number of possible available active site conservative mutations. At each generation, an elitist process saves the k best individuals, and automatically inserts them into the new population. At the end of the evolutionary process, GASS returns a ranking of k individuals, where the application user can analyse them and choose the best based not only on the spatial distance, but also on expert knowledge.

Substitution matrix: As previously mentioned, the mutation operator is responsible for dealing with conservative mutations, by replacing genes with different but compatible types of amino acids. These amino acids are defined according to a substitution matrix. The substitution matrix used in this work was borrowed from [14], where it was built using data from CSA. CSA entries may be of two types: those annotated as LIT (i.e., manually annotated and reported in the literature) or PSI-BLAST (i.e. annotated using the PSI-Blast tool for protein sequence search). The matrix is generated by comparing LIT and PSI-Blast entries with active sites with the same number of amino acids and Enzyme Commission (EC) number. For each active site template, a substitution matrix was built according to these comparisons. For more details on this approach, the user is referred to [14].

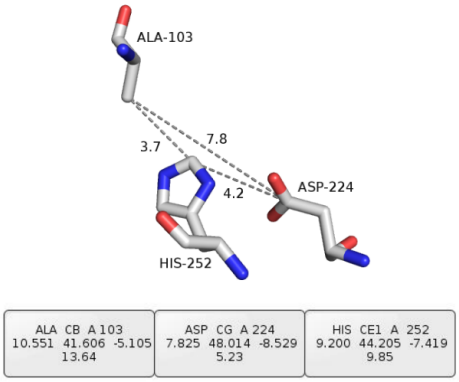

Figure 2: Example of an individual, which represents a candidate active site.

4. MeGASS: MULTI-OBJECTIVE GASS

As previously mentioned, the original version of GASS evaluates candidate active sites simply based on their spatial distance from the template. This has advantages (as we do not need a lot of information about proteins properties) but at the same time does not account for other active site characteristics. Here we address one of this relevant characteristics: how close to the protein surface are the amino acids in the active site.

We used the depth of the active site [4] to measure this distance, although the accessibility [13] metric was also considered in initial experiments. The accessibility shows whether the residues of the active site are in the surface of the protein, and hence can recognize or react with other structures. The depth, in contrast, tells how buried the amino acids are in the protein, i.e., how far from the surface. Preliminary experiments showed the accessibility presented great variations in its results. This is because it does not distinguish between atoms just below the protein surface and those in the core of the protein [4]. Although atoms coordinates obtained from crystallography are a good approximation of their positions, atoms just below the surface might have contact with a ligand or solvent, and this is not reflected by the accessibility measure.

Hence, the depth, i.e., how close to the surface of the protein a residual is, was the measure incorporated to the fitness function. The depth of an active site was calculated

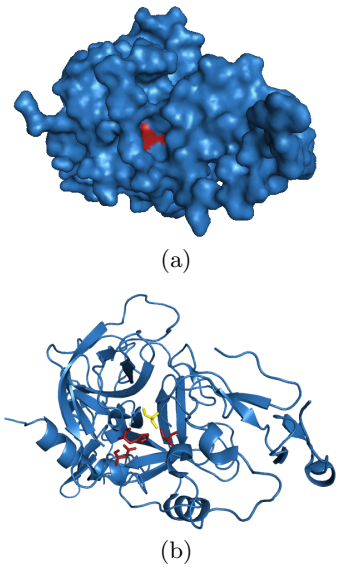

Figure 3: Example of an enzyme with its active site (in the surface) highlighted in red (a). In (b), the catalytic site residues are shown in red. In yellow, a candidate serine, disregarded as it was further from the surface than the serine in red.

using the software Depth[2]. Depth calculates the depth of an atom using its distance from the water molecule closer to the protein surface. It generates depth information for each atom on the side and main chain, and for each residue.

Figure 3 brings an example of a protein and its active site. In (a), we observe the active site in the surface of the enzyme. In (b) we see the catalytic triad. The residues in red represent the real active site. However, a candidate active site might consider the Serine in yellow as the correct active site residue instead of the one in red. By using the depth of the residues, the correct Serine would be preferred over the incorrect (yellow) one, as it is closer to the surface.

Given this motivation, the depth measure was incorporated to GASS in two different ways: (i) by adding it as a new term to the fitness function, considering two objectives simultaneously; (ii) by using a multi-objective lexicographic approach [6]. The latter was proposed because the spatial distance from the active site template and depth may be conflicting objectives, as GASS is able to find amino acid very close to the template but buried in the protein. In the lexicographic approach, a pre-defined ordering is established between the competing objectives, as one can be considered more relevant than the other. The next sections discuss these approaches.

4.1 Unique multi-objective fitness approach

This approach is the simplest to incorporate depth to the fitness function. Our first idea was to use the raw values of depth. However, it is known that the depth of active sites with similar function is also similar [4]. Hence, in the same way that we did with the spatial distance between the amino acids, we computed the depth distance from the template to the candidate active sites, as described in Eq. 2, where v_d and w_d represent the depth of each residual of the candidate and template active sites, respectively.

$$Fit_{depth}(v_d, w_d) = \sqrt{\sum_{i=1}^{(n^2-n)/2} \|v_{d_i} - w_{d_i}\|^2} \quad (2)$$

Hence, the total fitness function is defined by Eq. 3.

$$Fit_{MO}(v, w, v_d, w_d) = Fit_{dist}(v, w) + Fit_{depth}(v_d, w_d) \quad (3)$$

4.2 Lexicographic Multi-objective Approach

In contrast with traditional multi-objective optimization using the Pareto approach [6], where objectives are considered equally important, the lexicographic approach avoids the use of weight factors by explicitly incorporating priorities for the objectives of interest. Hence, it requires a specialist of the domain to establish a priority for each objective. After that, two solutions are compared with respect to the most important objective. If the result is a tie, the algorithm continues and compares the solutions according to the second most important objective. This tie-breaking process is repeated until no objectives are left to be accounted for.

In our approach, being closer to the template is initially more important than being similar in terms of protein depth. Hence, for all individuals, the depth is only analysed when two individuals are similar regarding spatial distance from the template. These two objectives are taken into account during the tournament selection process.

We opt for using a lexicographic approach over a Pareto-based approach because we know distance is more important than depth. However, a Pareto approach where the decision process would account for the lowest distance with the benefit of a non-dominated depth is the subject of future work.

5. EXPERIMENTAL RESULTS

The tests reported in this section consider datasets of catalytic sites, although MeGASS can be also used for subtract binding site identification [10]. We start with catalytic sites because they are smaller and easier to deal with. Preliminary experiments are reported in two datasets:

NOS: 125 enzymes from the Nitric Oxide Synthase (NOS) family (EC:1.14.13.39) with 248 catalytic sites annotated in CSA.

TRP: 1,085 enzymes *Trypsin-like* randomly chosen from PDB using SCOP (http://scop.berkeley.edu/) classification (superfamily 1A0J). For these 1,085 enzymes we had 1,085 templates annotated in CSA.

For both datasets, we used as templates the enzymes appearing in CSA as LIT entries, i.e., those that were manually annotated and hence have a higher confidence of being correct than the active sites annotated via Psi-BLAST. For NOS, only the enzyme *human endothelial nitric oxide synthase with arginine substrate* (PDB id 3NOS) was obtained from the literature. For TRP, 9 LIT templates were found, and their PDB ids are: 1A0J, 1CA0, 1DDJ, 1DS2, 1HJA, 1N8O, 1RTF, 1SSX and 2LPR.

MeGASS was executed 30 times with the following parameters: 300 individuals evolved for 100 generations, with

[2]http://mspc.bii.a-star.edu.sg/tankp/intro.html

Table 1: Results obtained by GASS (in terms of number of correctly identified active sites) and the two versions MeGASS considering different ranking sizes.

Data	# Enzymes	# Lit Templ.	Rank	# of Active Sites Found			
				CSA	GASS	MeGASS-Depth	MeGASS-Lexic.
NOS	125	1	1	248	248	248	248
TRP	1,085	9	1	1,085	899	893	901
	1,085	9	5	1,085	987	973	986
	1,085	9	10	1,085	1,015	1,008	1,017

a crossover probability of 0.9, a mutation probability considering the same amino acid of 0.2 and a mutation probability considering a different amino acid (conservative mutation) of 0.1. The 5 best individuals are always conserved via elitism. Results were evaluated considering two different criteria: the number of active sites found in the first position of the ranking and the Cumulative Match Score Curve (CMS). This curve shows the relation between the number of correct active sites found according to CSA and their position in the ranking, and allows us to clearly see how MeGASS improves the ranks of the correct active sites when compared to GASS. From now on, the version of MeGASS with a unique multi-objective function is referred as MeGASS-Depth and the lexicographic version as MeGASS-Lexic.

Table 1 shows the results obtained for rankings of size 1, 5 and 10 with the original version of GASS and the two versions of MeGASS. The first thing to notice is how much room for improvement exists. For NOS, the results obtained by GASS were already the best possible to obtain, as the algorithm found all the 248 active sites available using the single LIT template catalogued in CSA. In this case, our intention was to show that the information coming from the depth attribute would not interfere into the results already obtained.

For TRP, in contrast, there is room for improvement, but note that the results of template matching for GASS were already high. From the 1,085 active sites, 82.85% appeared in the first position of the raking. The objective of MeGASS was to improve over the 17.15% left. Table 1 shows the number of active sites recovered considering rankings of 1, 5 and 10 active sites. Notice that, considering the top 10 ranked active sites, MeGASS-Depth obtained worse results than GASS while MeGASS-Lexic inserted two new active sites into the top 1 and top 10 rankings. However, looking at Figure 4, it is interesting to observe that, considering bigger rankings, MEGASS-Lex classifies 95% of active sites within the top-20 candidate solutions. GASS, in contrast, needs to consider almost 60 positions of the ranking to obtain these same results. It is important to observe that, in both cases, around 5% of the active sites were not found in the final ranking, regardless of the algorithm used. Suggestions to solve this problem are discussed as future work.

A deeper analysis of how often MeGASS-Lexic actually uses the second objective, namely depth, showed that, during the evolution, in average 81% of the times only the spatial distance is considered, while in the other 18.95% the depth criterion is taken into account. Figure 5 shows how the fitness of the individuals, in this case represented by the spatial distance from an individual to a template, evolves over time. Note the wide range of values, varying from numbers smaller than one to almost 200.

As already mentioned, one of the main problems with the

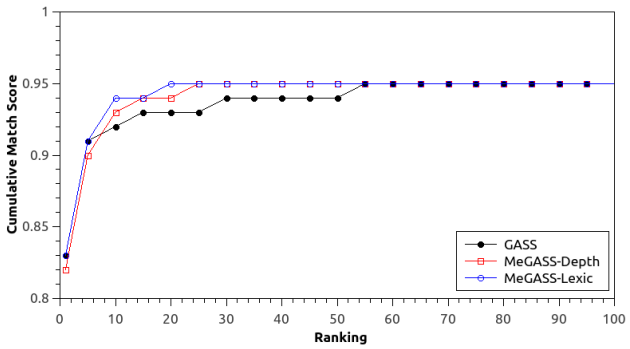

Figure 4: Results of CMS for dataset TRP.

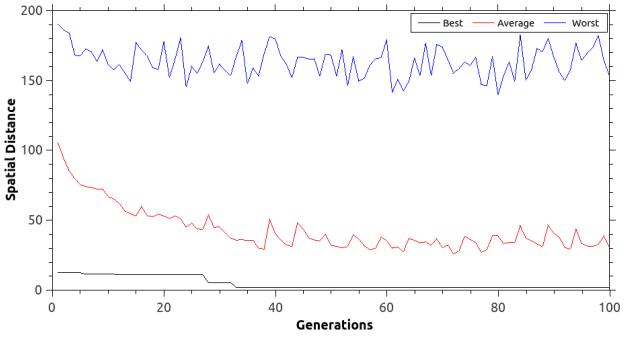

Figure 5: Fitness evolution over time for MeGASS-Lexic

experimental analysis of GASS is that we cannot directly compare it with other methods proposed in the literature. However, indirect comparisons with small sets of proteins used by them has already shown they are at least as good as and most time better than ASSAM and CatSid [10].

6. CONCLUSIONS AND FUTURE WORK

This paper introduced MeGASS, a first multi-objective version of GASS for active site search. The main differences of MeGASS over GASS is the use of the depth of the residual into the protein to improve the active site identification process. MeGASS was tested with two classical approaches for multi-objective optimization: (i) considering different objectives (spatial distance and depth distance) in the same fitness function or (ii) using a lexicographic approach, as the spatial distance is crucial in the identification process.

Preliminary results showed that MeGASS-Lexic brings the active sites to rank positions that are closer to the top than those obtained by GASS. However, as GASS already

obtains very accurate results ($\approx 87\%$), improving over them is not a straightforward task. We intend to further investigate the role of depth in the fitness function. Another approach would be, instead of rerunning the algorithms with an additional objective to be optimized, to propose a method for rescoring the active sites ranks according to depth or any other protein properties, such as whether they are in a pocket or not.

Apart from the fitness function, we also would like to make MeGASS population a bit more dynamic, so that it could search templates of different sizes simultaneously (nowadays it is done on batch). For example, we could have different sub-populations where, apart from looking for the complete template, we would also search for subsets of residues. This is particularly useful when dealing with substrate binding sites, which are bigger and might make the search space more difficult to explore.

Other directions not related to the evolutionary computation components include studying a more sophisticated substitution matrix, perhaps more generic and based on other classical ones, such as Blosum62 and MIQS [9]. We believe this would increase the chances of bringing active sites currently left out of the ranking (not found by the algorithm) to the final population, as they might have gone through conservative mutations not reflected in the current matrix. Finally, investing on using the identified active sites for protein function prediction is a relevant task.

7. ACKNOWLEDGMENTS

This work was partially supported by the following Brazilian Research Support Agencies: CNPq, FAPEMIG, and CAPES.

8. REFERENCES

[1] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, and P. Bourne. The protein data bank. *Nucleic acids research*, 28(1):235, 2000.

[2] M. Brylinski and J. Skolnick. A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *Proc. of the National Academy of Sciences of the USA*, 105(1):129–134, 2008.

[3] T. G. Cassarino, L. Bordoli, and T. Schwede. Assessment of ligand binding site predictions in CASP 10. *Proteins*, 82((Suppl 2)):154–163, 2014.

[4] S. Chakravarty and R. Varadarajan. Residue depth: a novel parameter for the analysis of protein structure and stability. *Structure with Folding & Design*, 7(7):723–732, Jul 15 1999.

[5] F. L. Custodio, H. J. Barbosa, and L. E. Dardenne. A multiple minima genetic algorithm for protein structure prediction. *Applied Soft Computing*, 15(0):88 – 99, 2014.

[6] K. Deb and D. Kalyanmoy. *Multi-Objective Optimization Using Evolutionary Algorithms*. John Wiley & Sons, Inc., New York, NY, USA, 2001.

[7] T. Fober, M. Mernberger, G. Klebe, and E. Hüllermeier. Evolutionary construction of multiple graph alignments for the structural analysis of biomolecules. *Bioinformatics*, 25(16):i2110–i2117, 2009.

[8] G. B. Fogel and D. W. Corne, editors. *Evolutionary Computation in Bioinformatics*. Morgan Kaufmann, 2002.

[9] S. Henikoff and J. G. Henikoff. Amino-acid Substitution Matrices from Protein Blocks. *Proc. of The National Academy of Sciences of The USA*, 89(22):10915–10919, Nov 15 1992.

[10] S. C. Izidoro, R. C. de Melo-Minardi, and G. L. Pappa. GASS: identifying enzyme active sites with genetic algorithms. *Bioinformatics*, 2014.

[11] T. Kato and N. Nagano. Metric learning for enzyme active-site search. *Bioinformatics*, 26(21):2698–2704, 2010.

[12] R. A. Laskowski, J. D. Watson, and J. M. Thornton. Protein function prediction using local 3D templates. *Journal of Molecular Biology*, 351:614–626, 2005.

[13] B. Lee and F. M. Richards. Interpretation of Protein Structures - Estimation of Static Accessibility. *Journal of Molecular Biology*, 55(3):379–&, 1971.

[14] F. C. Lightstone, S. E. Wong, D. A. Kirshner, and J. P. Nilmeier. Rapid Catalytic Template Searching as an Enzyme Function Prediction Procedure. *PLoS ONE*, 8(5):1–17, may 2013.

[15] G. Lopez, P. Maietta, J. M. Rodriguez, A. Valencia, and M. L. Tress. firestar-advances in the prediction of functionally important residues. *Nucleic Acids Research*, 39(2):W235–W241, Jul 2011.

[16] A. Marhaman and J. M. Thornton. Methods to Characterize the Structure of Enzyme Binding Sites. In T. Schwede and M. Peitsch, editors, *Computational Structural Biology - Methods and Applications*, pages 189–221. World Scientific Publishing, 2008.

[17] N. Nadzirin, E. J. Gardiner, P. Willett, P. J. Artymiuk, and M. Firdaus-Raih. SPRITE and ASSAM: web servers for side chain 3D-motif searching in protein structures. *Nucleic Acids Res.*, 40:W380–W386, May 2012.

[18] S. Pal, S. Bandyopadhyay, and S. Ray. Evolutionary computation in bioinformatics: a review. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 36(5):601–615, 2006.

[19] C. T. Porter, G. J. Bartlett, and J. M. Thornton. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res.*, 32:D129–D133, 2004.

[20] A. Stark and R. B. Russell. Annotation in three dimensions. PINTS: Patterns in Non-homologous Tertiary Structures. *Nucleic Acids Res.*, 31(13):3341–3344, 2003.

[21] J. W. Torrance and J. M. Thornton. *Structure-based Prediction of Enzymes and Their Active Sites*. Wiley, 2009.

[22] A. C. Wallace, N. Borkakoti, and J. M. Thornton. Tess: A geometric hashing algorithm for deriving 3d coordinate templates for searching structural databases application to enzyme active sites. *Protein Sci.*, 6:2308–2323, 1997.

[23] J. C. Whisstock and A. M. Lesk. Prediction of protein function from protein sequence and structure. *Q. Rev. Biophys*, 36:307–340, 2003.