Bioinformatics, 31(6), 2015, 864–870
doi: 10.1093/bioinformatics/btu746
Advance Access Publication Date: 10 November 2014
Original Paper

OXFORD

Structural bioinformatics

GASS: identifying enzyme active sites with genetic algorithms

Sandro C. Izidoro[1,*], Raquel C. de Melo-Minardi[2,3] and Gisele L. Pappa[2,3]

[1]Advanced Campus at Itabira, Universidade Federal de Itajubá, Itajubá, MG 35903-087, Brazil and [2]Department of Computer Science and [3]Department of Biochemistry and Immunology, Universidade Federal de Minas Gerais, Belo Horizonte, MG 31270-901, Brazil

*To whom correspondence should be addressed.
Associate Editor: Burkhard Rost

Received on July 1, 2014; revised on October 14, 2014; accepted on November 5, 2014

Abstract

Motivation: Currently, 25% of proteins annotated in Pfam have their function unknown. One way of predicting proteins function is by looking at their active site, which has two main parts: the catalytic site and the substrate binding site. The active site is more conserved than the other residues of the protein and can be a rich source of information for protein function prediction. This article presents a new heuristic method, named genetic active site search (GASS), which searches for given active site 3D templates in unknown proteins. The method can perform non-exact amino acid matches (conservative mutations), is able to find amino acids in different chains and does not impose any restrictions on the active site size.
Results: GASS results were compared with those catalogued in the catalytic site atlas (CSA) in four different datasets and compared with two other methods: amino acid pattern search for substructures and motif and catalytic site identification. The results show GASS can correctly identify >90% of the templates searched. Experiments were also run using data from the substrate binding sites prediction competition CASP 10, and GASS is ranked fourth among the 18 methods considered.
Availability and implementation: Source code and datasets (dcc.ufmg.br/ ∼glpappa/gass).
Contact: sandroizidoro@unifei.edu.br
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Over 14 000 protein families are currently annotated in Pfam (http://pfam.sanger.ac.uk/), from which more than 3500 still have their functions unknown (Finn *et al.*, 2014). Experimental tests are expensive and time consuming, and, in their absence, studies have shown that protein function can be successfully inferred based on the sequence or structure similarity between the hypothetical protein and proteins of known function (Zvelebil and Baum, 2008).

One way of predicting protein function is by searching for enzyme binding sites. According to Marhaman and Thornton (2008), enzyme binding sites are regions on the surface of an enzyme specially modelled to interact with other molecules. There are different types of binding sites, the most important being the active site. The active site is divided into two or three parts, which include the catalytic site and the substrate binding site. The former is usually a set of two to six residues that perform the catalytic reaction while the latter recognizes the molecule upon which the enzyme acts. For the sake of simplicity, as finding a binding, active or catalytic site refers to the same computational problem, this article always refers to active sites, which is a broader term and encompasses both catalytic and substrate binding sites.

Due to their importance to enzyme function, active sites amino acids are more conserved during evolution than the sequence as a whole. Consequently, they can be a rich source of information for function prediction (FN; Cassarino *et al.*, 2014; Torrance and Thornton, 2009). Several methods have been proposed to infer

protein function based on active site similarity (Jacobson *et al.*, 2014), and they can be grouped into three major categories: sequence-based, structure-based and hybrid. When using sequence, multiple alignments of various organisms have been widely used to verify conservation of residues that may be structurally or functionally important, including Henschel *et al.* (2007), Goldenberg *et al.* (2009), Torrance and Thornton (2009) and Lopez *et al.* (2011). Structure-based methods developed with the same purpose can be found in Wallace *et al.* (1997), Barker and Thornton (2003), Kristensen *et al.* (2008) and Brylinski and Skolnick (2008). Hybrid methods, in contrast, take advantage of different types of information, including surface accessibility (Huang and Schroeder, 2006), physiochemical properties (Andersson *et al.*, 2010) or homology modelling (Wass *et al.*, 2010).

This article is particularly interested in methods for searching active sites based on structural data. In general, structure-based methods proposed to identify active sites in proteins are based on graphs, where nodes represent atoms in the amino acid side chain and neighbour atoms are connected with edges, weighted by their distances. In this context, Stark and Russell (2003) proposed a simple graph-search method based on a depth-first search called patterns in non-homologous tertiary structures, which finds all possible residue patterns (considering all template atoms) common to the template of the target protein.

Nadzirin *et al.* (2012), in contrast, proposed amino acid pattern search for substructures and motifs (ASSAM), which models the problem as a sub-graph isomorphism problem. ASSAM searches for maximum common sub-graphs to find similar structures between the template active site and the enzyme. The graph represents the amino acids in the side chain, and each node consists of two pseudo-atoms. Distances among different structures are calculated using root-mean-squared deviation (RMSD).

Lightstone *et al.* (2013), in turn, introduced catalytic site identification (CatSId). The algorithm performs a protein-to-template matching using a sub-graph search method and a library of catalytic residue templates from catalytic site atlas (CSA; Porter *et al.*, 2004)—a database of catalytic sites in enzymes of known 3D structure. These results are refined using a logistic scoring procedure to re-score the matches found in the first phase and use information such as binding site predictions and others physical descriptors to improve the structure matching previously obtained.

Many methods have also been proposed for substrate binding site in the context of the CASP competition (Cassarino *et al.*, 2014). SP-ALIGN (Brylinski and Skolnick, 2008), for instance, detects substrate binding sites by remote template identification and superimposition, structure-pocket alignment and binding site clustering guided by the template substrates. 3DLigandSite (Wass *et al.*, 2010), in contrast, aligns similar structures with the query, superimposing their bound ligands onto the query structure.

This article proposes a method to identify active sites using genetic algorithms (GAs) and information about the proteins 3D structure. GAs are widely used to solve combinatorial search problems and emulate the process of evolution and survival of the fittest (Goldberg, 1989). They have the advantage of performing a global search, being independent of application and tolerant to noise (Back *et al.*, 1997).

The proposed method can perform non-exact amino acid matches without restricting the number of amino acids in the template and finds catalytic residues and binding residues in different protein chains. Its global heuristic search is used to prune the search space, and only information about the 3D structure is required. Having the active sites identified in a second phase, protein function can be inferred using methods based on a similarity threshold or more sophisticated techniques, such as logistic regression (Lightstone *et al.*, 2013).

2 Materials and methods

This section introduces the principles of genetic active site search (GASS), details the evaluation strategy and describes the datasets used in the experimental evaluation.

2.1 Genetic active site search

The problem of identifying active sites in proteins can be defined as follows. 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 M amino acids in p_B and select those with most similar amino acid conformation and relative position to p_B. However, this solution becomes intractable as M grows. Hence, an alternative solution to this problem is to explore heuristic methods to perform this search, and here we investigate GAs.

Figure 1 illustrates the framework proposed to search similar active sites, named GASS. GASS receives as inputs the proteins and templates selected by the user and starts a preprocessing step. Note that the method can be explored in two different scenarios: to find a specific template (i.e. known active site) in one or more proteins or, given a set of templates, to find them in one or more proteins. The preprocessing step finds the selected proteins and active sites templates in protein data bank (PDB; Berman *et al.*, 2000) and CSA and returns, for each amino acid, its name, chain, reference atom and coordinates (x, y and z). This information is stored in a repository of proteins, and accessed by GASS to create its initial population, as detailed in the next sections. GASS then performs a heuristic search to find matching active sites in the selected proteins, and outputs one or more candidate active sites. In order to deal with conservative mutation, GASS also has the option of consulting a substitution matrix.

GASS is a method based on Darwin's theory of evolution and survival of the fittest. It evolves a population of individuals, where each individual represents a solution to the problem at hand. In this article, each solution corresponds to a candidate active site. These solutions are evaluated according to a fitness function, which assesses how good the individual is to solve the problem (e.g. we can

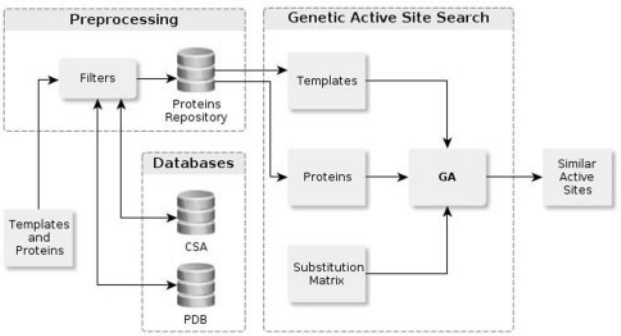

Fig. 1. Proposed methodology for active site matching: data is extracted from PDB and/or CSA, preprocessed and a search template is matched against a set of selected enzymes using GA

use the distance from the template to the candidate active site as a fitness measure). Individuals are selected to undergo crossover and mutation operations according to probabilities defined by the user (p_c and p_m). This process goes on until a stop criterion, which is usually based on a maximum number of generations, is met. This process is illustrated in the Supplementary Figure S1.

2.1.1 Individual representation and population initialization

For the problem of active site matching, an individual represents a group of amino acids, which is a candidate active site for an enzyme. The individual is encoded as a vector, where each position represents an amino acid. Figure 2a shows the catalytic site of enzyme synthesis of human endothelial nitric oxide arginine substrate (3NOS) and its GASS representation. Looking at the first amino acid in Figure 2b, note that GASS stores its name (CYS), chain (A), position in the sequence (184), the last heavy atom (LHA) of the side chain (SG) and its coordinates (17.125, 8.914, 23.94).

The choice of LHA as the reference atom was made after comparisons with two other references: α-carbon (AC) and side chain centroid (SCC; for more details, see Supplementary Table S1 and Fig. S2). Results showed that the performance of the method with different references varies slightly from one dataset to another. We chose LHA because it does not increase preprocessing computational cost (as SCC does) and uses information about the side chain instead of the backbone [and catalytic residues are more frequent in the side chain than in the main chain (Bartlett *et al.*, 2002)].

The initial population is generated from the protein repository. Each individual corresponds to n amino acids that are randomly chosen from the repository, always respecting the types of the amino acids from the template and its size (i.e. if the first position in the template is a glutamate, only glutamate may be selected for that position and the size of the individual is equal to the size of the template). In this way, it is possible to have individuals with amino acids from different chains. As explained later, conservative mutations are handled by the *mutation* operator.

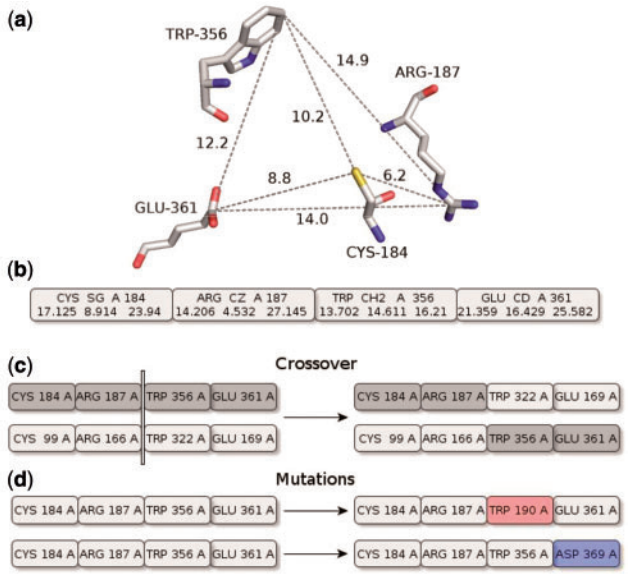

Fig. 2. Representation of the individual and genetic operators: (**a**) catalytic site of the enzyme 3NOS and distances (in Å) between the LHAs. (**b**) Representation of 3NOS as a GASS individual. (**c**) One-point *Crossover* recombines segments of individuals. (**d**) *Mutation* changes the same (TRP-356 replaced by TRP-190) or different types (GLU-361 was changed by ASP-369) of amino acids

2.1.2 Fitness function and selection

GASS individuals are evaluated by calculating the distance between the coordinates of the LHA of the template, represented by a vector of its 3D coordinates (**v**) and the candidate active site found by GASS (**w**), according to Equation (1). Note that the difference between the metric in Equation (1) and the well-known RMSD is that we do not average the squared distances of the results. This is because, as shown in Laskowski *et al.* (2005), slightly different active sites may have similar RMSD values. By using their absolute value distances we try to avoid this problem.

$$\mathrm{Fit}(\mathbf{v},\mathbf{w}) = \sqrt{\sum_{i=1}^{n} \|v_i - w_i\|^2} \quad (1)$$

The individual's evaluation is followed by the selection phase. This phase is crucial for the evolution of the population, as it gives a greater chance of survival to the best individuals, according to their fitness function. There are several selection methods in the literature (Back *et al.*, 1997). We used tournament selection, where a subset of k individuals is randomly selected from the population, and the one with best fitness value is chosen to undergo crossover and mutation operations.

2.1.3 Genetic operators

After selection, two genetic operators are used to generate a new population: standard one-point crossover and single-point mutation. Figure 2 illustrates both methods. Two individuals are required for crossover and one for mutation. In crossover, a random position in the individual is selected, and the amino acids before that point in the first parent merged with the amino acids after that point in the second parent (Fig. 2c). These new individuals are then added to the new population.

In the case of the single-point mutation, only the point chosen is replaced by either (i) a random amino acid of the same type from the selected enzyme (TRP 356 by TRP 190 in Fig. 2d) or (ii) a different type of amino acid indicated by the substitution matrix in the same enzyme (GLU 361 by ASP 369 in Fig. 2d). The substitution matrix was borrowed from Lightstone *et al.* (2013) and indicates possible conservative mutations in active sites annotated in CSA.

2.1.4 Candidate active sites

GAs have one characteristic that differs them from other search methods: they explore the search space by searching different sets of solutions (individuals) in parallel. Hence, at the end of the evolution process, we end up with a set of candidate active sites as big as the population size. When the user needs a unique solution, the individual with the best fitness is returned. In some cases, however, it might be interesting for the user to analyse a set of solutions, and then use another set of criteria, perhaps more subjective, to choose the best.

For instance, it might be that the best solution—the one with smallest distance from the template, is buried in the protein, instead of being in a pocket. The specialist can immediately recognize the candidate active site is not a real one. In order to avoid situations like that, the method returns a ranking of the n best solutions found. In this way, a specialist can choose the most appropriate according to his background knowledge.

More details about GASS search space and computational complexity can be found in the Supplementary Material.

2.2 Evaluation strategies

Two evaluation strategies were proposed to evaluate GASS. In the first, results generated by GASS are validated according to the enzymes catalytic sites catalogued in CSA (Bartlett *et al.*, 2002). CSA is a database of 3D structures of catalytic residues that stores two types of entries: originally annotated and manually derived from the primary literature (LIT) and automatically identified sites found by alignment using PSI-BLAST. In CSA, a residue is defined as catalytic if it fulfils any one of the following criteria: (i) It is direct involved in the catalytic mechanism; (ii) It alters the pKA of another residue or water molecule directly involved in the catalytic mechanism; (iii) It represents a stabilization of a transition state or intermediate and (iv) It is responsible for the activation of a substrate (Furnham *et al.*, 2014). As far as we are concerned, CSA is the most complete database that can be used as a gold standard for comparing the results of catalytic sites identification methods. The second evaluation strategy uses a set of templates of substrate binding sites considered in the 2012 CASP competition, namely CASP 10, and compares the results obtained by GASS with those of the 17 competing methods.

For each experiment, GASS was executed 30 times. This is necessary to obtain statistical significance in the results, once the method is non-deterministic and each execution may return a different result. Preliminary tests were performed to configure a set of parameters required by GASS to find active sites. As all GAs, GASS requires the definition of the following parameters: number of generations, population size, probability of crossover and mutation, tournament size and, in our case, candidate ranking size. The values of these parameters are listed in the Supplementary Table S2).

2.3 Datasets

Five datasets were used in the experiments, and they were selected to answer the following questions: (i) Can GASS find catalytic sites within a family? (ii) Can GASS help functionally classify enzyme families? (iii) How does GASS handle less-controlled datasets? (iv) How does GASS compare with other state-of-the-art methods? They are the following:

DS 1: 125 enzymes from the nitric oxide synthase (NOS) family (EC:1.14.13.39) with catalytic sites annotated in CSA. This group was also tested with other 126 enzymes, randomly chosen from PDB, and with EC numbers different from EC 1.-.-.-.

DS 2: 1085 enzymes *Trypsin-like* randomly chosen from PDB using SCOP (http://scop.berkeley.edu/) classification (superfamily 1A0J).

DS 3: 24,437 enzymes from the database NCBI VAST non-redundant (*P*-value 10e-80), as reported in Nadzirin *et al.* (2012), and one set containing 100 enzymes chosen from PDB based on the results of ASSAM.

DS 4: 61 enzymes and 1800 templates selected from CSA, as done in CatSId (Lightstone *et al.*, 2013).

DS 5: 13 target enzymes and 25 binding site templates for each enzyme, according to CASP 10 FN category (Cassarino *et al.*, 2014).

3 Results

This section discusses the results obtained when evaluating GASS using the two strategies previously described. A summary of the results obtained by GASS when comparing its results with CSA templates is presented in Table 1. The differences among the number of enzymes searched and those with known catalytic sites in CSA happen for two reasons: (i) sometimes an enzyme has more than one catalytic site; (ii) not all enzymes have their catalytic sites catalogued in CSA. The results are detailed in the following sections.

Table 1. GASS and CSA results

DS	Enzymes	Templates	Catalytic sites		Match	GASS
			CSA	GASS	(%)	Rank
1	125	1	248	248	100.00	1
	125	125	248	235	94.49	1
2	1085	9	1085	899	82.85	1
	1085	9	1085	987	90.94	5
	1085	9	1085	1015	93.52	10
3	100	1	79	79	100.00	1
	24 437	1	–	–	–	1
4	61	1800	182	162	89.01	1
	61	1800	182	165	90.65	5
	61	1800	182	165	90.65	10

For each DS we show the number of enzymes and templates, the number of catalytic sites annotated in CSA (gold standard) and the number of catalytic sites found correctly by GASS, the percentage number of catalytic sites found in relation to those annotated in CSA and the ranking size used by GASS.

3.1 Can GASS find catalytic sites within a family?

In order to answer this question, GASS was used to find one catalytic site in the set of 125 NOS family enzymes (DS 1). Note that the quality of the results of GASS highly depends on the quality of the templates. Hence, catalytic sites annotated as literature are more appropriate for this type of search. First, we tested 3NOS as a template, as it is the only CSA LIT entry among all NOS enzymes.

In this case, GASS correctly found all 248 catalytic sites (CSA—version 2.2.12). Observing the values of fitness [defined in Equation (1)] of all candidate catalytic sites (individuals) in the final population, we noticed that 84.13% presented distances from the template ≤ 5 Å. This shows that the majority of enzymes within the same family have small distances variation between their catalytic sites.

However, there are exceptions. An example is the enzyme murine ino synthase with coumarin inhibitor (2BHJ), which presented a fitness value of 11.64 Å, which is twice the value of fitness found for most enzymes in the NOS family. This difference occurs because of the enzyme's ligand. In 3NOS, the ligand HAR-512 (*N*-omega-hydroxy-*L*-arginine) occupies a small volume when compared with ligand FC1-1499 (thiocoumarin) in 2BHJ, as showed in the Supplementary Figure S4.

Considering 30 different runs of the GASS for all enzymes, we also calculated the mean and standard deviation of the fitness for each enzyme. Only 3 out of 125 catalytic sites found (individuals) had a standard deviation different from 0 (1NOC, 1NOS and 2NOS; see Supplementary Figure S3), which shows that the results found have a very low variability between different GASS runs. Low variability is necessary to guarantee a robust search method.

In a second step, we also used each of the other 124 enzymes annotated in CSA using PSI-BLAST as templates for searching the remaining enzymes, including 3NOS for completeness (all against all). Considering 125 enzymes, we had 248 catalytic sites annotated in CSA. In average, for each of the 125×30 experiments performed, GASS found 235.31 catalytic sites correctly according to CSA (94.49%).

3.2 Can GASS help classifying families?

As previously explained, GASS always returns as a result a ranking of the most similar catalytic sites to the template. Hence, even when

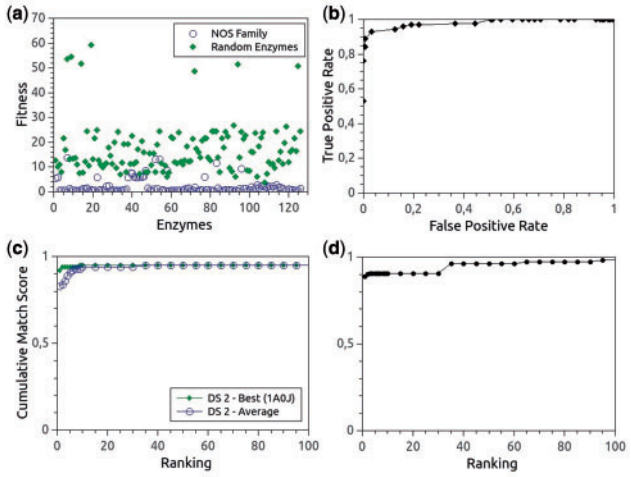

Fig. 3. Results for DS 1, DS 2 and DS 4: (**a**) fitness of NOS family and random enzymes—each symbol is an enzyme, *x* corresponds to a catalytic sites found and *y* to its fitness value (DS 1). (**b**) ROC curve considering DS 1. (**c**) CMS of the template 1A0J and average over all nine templates (DS 2). (**d**) CMS of the catalytic sites found by GASS (DS 4)

an enzyme does not have that particular site, the most similar set of amino acids is returned. This second experiment was run in a group of 251 enzymes, 125 from the NOS family (DS 1) and 126 randomly chosen from PDB using 3NOS as a template. This test shows how many false positives (enzymes that do not have the catalytic site but had it detected) the method generates given a distance threshold. Analysing the values of fitness of the identified catalytic sites, 80.95% presented values >10 Å for the 126 enzymes in the random set. Only one catalytic site (0.79%) presented template distances smaller than 5 Å. This may suggest that enzymes in different families tend to have very different catalytic sites, and GASS was able to identify that. Figure 3a shows the distance results of GASS considering enzymes from the NOS family and random enzymes. As expected, enzymes within the same family are closer to the template than random enzymes.

Figure 3b shows a receiver operating characteristic (ROC) curve (Hand, 2009), indicating the ability of catalytic sites distances from the family template to correctly assign the family of an enzyme based on a simple distance threshold. The area under the curve (AUC) considering the distance threshold is 0.97.

3.3 How does GASS handle less-controlled datasets?

Previous tests were performed with small and very controlled sets of enzymes, specifically chosen to test some properties of the algorithm. This section comprises experiments with DS 2, composed of 1085 *Trypsin-like* enzymes randomly chosen from PDB. The nine templates annotated as LIT in CSA and given as input to GASS are listed in the Supplementary Table S3.

After running the nine templates against 1085 enzymes, GASS found, in average, 899 catalytic sites annotated in CSA (82.85%) in the first position of the ranking. Increasing the ranking size to 5, we had 987 catalytic sites correctly identified (90.94%). When the size of the ranking was 10, the number of catalytic sites was 1015 (93.52%). A more detailed analysis per template can be found in the Supplementary Table S4.

Figure 3c shows a cumulative match score curve (CMS) for the most successful template (1A0J) and the average considering all nine templates of the catalytic sites found by GASS. This curve

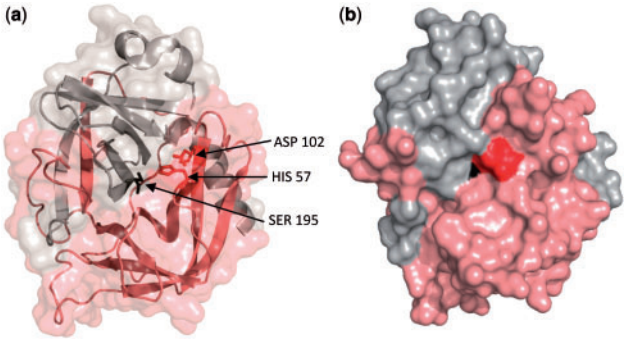

Fig. 4. Enzyme 2GCT: (**a**) amino acids found by GASS. (**b**) Location of the amino acids on the surface

shows the relation between the number of correct catalytic site found according to CSA and their position in the ranking. Analysing the curves, we observe that the best catalytic site candidates appear mostly in the top five positions of the ranking.

An analysis of the catalytic sites not found by GASS was also performed, and we identified three reasons for that: (i) the catalytic site is not in CSA, (ii) the catalytic site is in CSA but was found using PSI-BLAST, and appears divided into different chains; (iii) conservative mutations not reflected in the substitution matrix borrowed from Lightstone *et al.* (2013) occurred. One example of an enzyme not annotated in CSA is 1ARC, where GASS identified the catalytic site HIS 57, ASP 113 and SER 194, which is in agreement with Tsunasawa *et al.* (1989; Supplementary Figs. S5 and S6).

Concerning situation (ii), the catalytic site of 2GCT is stored in CSA as three distinct sites [HIS 57 and ASP 102 (chain B), GLY 193 and SER 195 (chain C), SER 195 and GLY 196 (chain C)]. GASS found amino acids HIS 57 and ASP 102 (chain B) and SER 195 (chain C), as shown in Figure 4. This example shows a drawback of PSI-BLAST alignments made by CSA, which in this case split what might be a single site into three different ones. Hence, although this may seem like a GASS error, it is actually a problem of CSA when dealing with catalytic sites in different chains.

3.4 How do GASS results compare with those obtained by other state-of-the-art methods?

This section compares GASS against two others recently reported in the literature to find catalytic sites: ASSAM (Nadzirin *et al.*, 2012) and CatSId (Lightstone *et al.*, 2013). It also compares the results of GASS with 17 methods submitted to the CASP 10 competition regarding substrate-binding site templates.

3.4.1 GASS ×ASSAM

First, it is important to emphasize the main differences between ASSAM and GASS. ASSAM represents amino acids using pseudo-atoms, while GASS uses LHA. Concerning the metric used to calculate the distances from the template, ASSAM uses RMSD while GASS uses Equation (1). After search, ASSAM reports the 100 most similar active sites to the template, ordered by RMSD, and reference templates are limited to 12 amino acids. GASS has no limit for the latter, and can return as many active sites as its population site.

In order to compare GASS with ASSAM, we used as template the structure 1A0S (*Salmonella typhimurium* sucrose specific porin ScrY), reported in Nadzirin *et al.* (2012) and DS3. GASS was

run against each of the enzymes in DS 3, and the results ordered according to their fitness values.

Among the 100 results reported by ASSAM are the structures 1A0T and 1OH2, discussed in Nadzirin *et al.* (2012), which are examples of specific porin sucrose. GASS found all the three catalytic sites of 1A0T structure (chains R, P and Q) at positions 1, 4 and 7 of a ranking with 100 enzymes. Realizing the absence of the structure 1OH2, we noticed it was not in DS3. Checking more closely the results of ASSAM, only 23 out of 100 results returned are in the original dataset. We believe ASSAM added more structures to the original database using SPRITE (Nadzirin *et al.*, 2012). However, once this procedure was not documented, it could not be mimicked.

As an alternative, we simulated what would have happened if GASS had access to the 100 enzymes output by ASSAM, analysing specially if the relative order of the enzymes would change, given the methods use different template distance metrics. This is not ideal, but it is one way to compare our results. These 100 enzymes were given as input data to GASS and the structure 1ACB (bovine alpha-chymotrypsin-eglin C complex) used as template. In this case, the results obtained by ASSAM and GASS were very similar. Both found the same 79 catalytic sites in accordance with CSA, and the remaining 21 were discarded because they were not catalogued in CSA. However, for some enzymes, ASSAM omitted or incorrectly reported the chain of the catalytic site amino acids. This is the case of enzymes 1AUJ and 6CHA. GASS found the catalytic site for 1AUJ in chain A (in agreement with CSA) while ASSAM did not report the chain. For 6CHA, GASS found the catalytic site and the respective chain of each residue (HIS-57 (B), ASP-102 (B), 195-SER (C)), while ASSAM located the site in chain A. However, the PDB file for 6CHA has only nine amino acids in chain A, and none of them correspond to HIS, ASP, or SER, which are the amino acids of the catalytic site of 6CHA. This error may have happened because the amino acids of the catalytic site are in different chains (B and C).

3.4.2 GASS × CatSId

CatSId differs from GASS in the following: (i) it represents a template using the coordinates of the α-carbon, cofactor and/or ion; (ii) the optimized sub-graph isomorphism search performed in the first phase of the method uses a threshold (1.5 Å) to prune non-promising sub-graphs; (iii) it uses RMSD to measure enzymes/template distances; (iv) its second phase performs a logistic scoring procedure, which uses much more information about the enzymes than GASS, including physicochemical descriptors.

For a fair comparison between the methods, the same templates and the same enzymes used by CatSId in its first phase should be considered. However, as CatSId does not report it, we used the enzymes and templates considered in the second phase. CatSId used 1993 templates (LIT—CSA) to search catalytic sites in 66 randomly chosen enzymes (CSA—version 2.2.12). GASS used 1800 templates to search catalytic sites in 61 enzymes. This difference in the number of templates and enzymes is due to the lack of information (position of the LHA of the side chain) in some PDB files and the fact that we did not use non-standard amino acids. DS 4 enzymes are listed in Supplementary Table S5.

In total, there were 182 catalytic sites found in 61 enzymes. GASS found 165 catalytic sites correctly. The 17 sites not found belong to seven enzymes (see Supplementary Table S5). We identified two situations where errors occurred. In five of them GASS finds the catalytic sites, but not within the top five rank. This might happen because all templates for these enzymes have substitutions. This increases even further the search space and makes GASS generates many individuals with better fitness values than those in the real site in comparison with the template. Examples of this case are available in the Supplementary Figure S7. Another problem emerges because of CSA errors, such as for enzymes 1L7A and 1G1Y (Supplementary Fig. S7).

Figure 3d shows the CMS of the active sites found by GASS. Using ranking size 5, GASS found 165 sites correctly, which corresponds to an accuracy of 90.65%. The other catalytic sites were found from the 35th position on due to the large number of possible substitutions.

3.4.3 GASS × CASP 10 methods

We also compared GASS to the 17 methods submitted to the FN category of the CASP 10 competition. The dataset used has 13 target enzymes and 25 binding sites templates for each target. First, we defined GASS templates for each target using the residues from templates provided by CASP 10 according to Cassarino *et al.* (2014). For 2 out of the 13 targets we could not identify the ligand in the templates structures, and hence the comparison used 11 out of the 13 targets.

GASS was run with the same parameter configuration used against ASSAM and CastId, without the substitution matrix. For each target, the results obtained were ordered according to their fitness value, and the function of the top-ranking individual used as GASS final prediction. On the basis of this result, we calculated the Matthew correlation coefficient (MCC), MCC Z-scores and binding-site distance test (BDT; Supplementary Table S7) as well as the cumulative confusion matrices, and for statistical significance the Wilcoxon signed-rank test was used (Supplementary Tables S8 and S9) (Matthews, 1975; Roche *et al.*, 2010). These results were compared with those obtained by 17 other methods proposed in CASP 10. As CASP 10 guidelines, the two targets with unidentified ligands (T0659 and T0721) were considered as having MCC zero.

Figure 5 shows the overall performance of GASS and all participating groups from CASP 10. The groups were ranked according to the average value of their MCCs normalized on all prediction targets (see Supplementary Table S6). Among the 11 targets considered, GASS found five binding sites correctly, and appears fourth in the ranking, with average MCC value of 0.63. Note that, from the three methods with better performance than GASS one is validated by

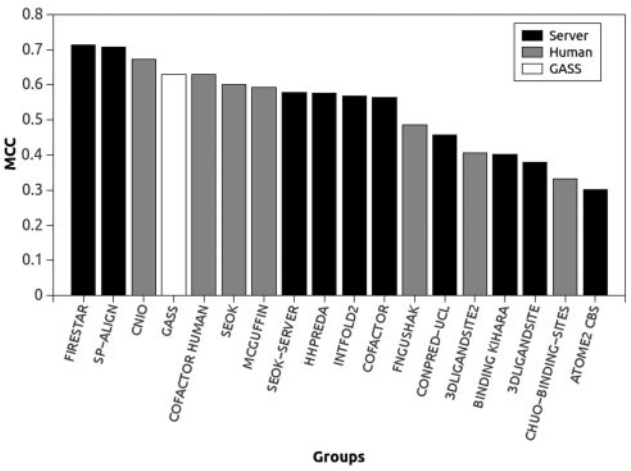

Fig. 5. Groups from CASP 10 (FP category) ranked in decreasing order by average MCC together with GASS. Human predictors are shown in gray, server predictors in black and GASS in white

human experts, while GASS is completely automatic. GASS is third among the automatic methods.

According to CASP 10, T0657 and T0659 were the most challenging targets, as predictors obtained the lowest MCC scores for them. T0659 was not considered in GASS tests, as its ligand was not found. For target T0657, GASS found the binding site correctly. Note that GASS results can be further improved by using a more complete substitution matrix (recall that the one used here was created by analysing the CSA conservative mutations) and a fine-tuned parameter optimization for GASS.

4 Conclusion

This work proposed GASS, a method for active site search based on GAs. The method receives as input one or more active sites templates, and looks for them in one or more proteins, returning a ranking of solutions as big as its population size. It also takes into account conservative mutations during the search by using a substitution matrix. The method can also find active sites in different protein chains, and its results can be further improved by using additional attributes to describe the sites. GASS has no predefined criteria to search the candidate solution space, such as CatSId, nor uses a limit for the size of the active site, as ASSAM does. Results show the method is effective in finding catalytic sites already catalogued in CSA, with accuracy rates above 90% in most datasets. Besides, when considering the dataset used in the FN task in CASP 10, when compared with the other 17 methods, GASS is ranked fourth according to values of MCC. It can be a powerful tool to improve the current catalytic sites and add new ones to CSA.

As future work, we intend to enhance the presented techniques so it verifies the accessibility and location of the amino acids (pockets) found by GASS. The current solution can be further extended to consider physicochemical attributes during GASS search. Additional tests with other substitution matrices will also be performed (Yamada and Tomii, 2014). Finally, we plan to make GASS available from a web server after the aforementioned issues are addressed.

Acknowledgements

Thanks to Daniel B. Roche and Douglas E. V. Pires for discussions and suggestions, and François Marie Artiguenave and Genoscope staff (CEA, France).

Funding

This work was supported by CAPES (BIOCOMPUTACIONAL process number 23038004007/2014-82, PVE process number 403076/2012-9), CNPq, FAPEMIG and all Brazilian funding agencies.

Conflict of interest: none declared.

References

Andersson,C.D. *et al.* (2010). Mapping of ligand-binding cavities in proteins. *Proteins*, **78**, 1408–1422.

Back,T. *et al.* (1997). *Handbook of Evolutionary Computation.* Oxford University Press, Bristol, UK.

Barker,J.A. and Thornton,J.M. (2003). An algorithm for constraint-based structural template matching: application to 3d templates with statistical analysis. *Bioinformatics*, **19**, 1644–1649.

Bartlett,G.J. *et al.* (2002). Analysis of catalytic residues in enzyme active sites. *J. Mol. Biol.*, **324**, 105–121.

Berman,H. *et al.* (2000). The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.

Brylinski,M. and Skolnick,J. (2008). A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *Proc. Natl Acad. Sci. USA*, **105**, 129–134.

Cassarino,T.G. *et al.* (2014). Assessment of ligand binding site predictions in CASP 10. *Proteins*, **82**, 154–163.

Finn,R.D. *et al.* (2014). Pfam: the protein families database. *Nucleic Acids Res.*, **42**, D222–D230.

Furnham,N. *et al.* (2014). The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes. *Nucleic Acids Res.*, **42**, D485–D489.

Goldberg,D.E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison - Wesley, Boston, MA.

Goldenberg,O. *et al.* (2009). The ConSurf-DB: pre-calculated evolutionary conservation profiles of protein structures. *Nucleic Acids Res.*, **37**, D323–D327.

Hand,D. (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Mach. Learn.*, **77**, 103–123.

Henschel,A. *et al.* (2007). Using structural motif descriptors for sequence-based binding site prediction. *BMC Bioinformatics*, **8**, 12.

Huang,B. and Schroeder,M. (2006). LIGSITE(csc): predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Struct. Biol.*, **6**, 19.

Jacobson,M.P. *et al.* (2014). Leveraging structure for enzyme function prediction: methods, opportunities, and challenges. *Trends Biochem. Sci.*, **39**, 363–371.

Kristensen,D.M. *et al.* (2008). Prediction of enzyme function based on 3D templates of evolutionary important amino acids. *BMC Bioinformatics*, **9**, 1–7.

Laskowski,R.A. *et al.* (2005). Protein function prediction using local 3D templates. *J. Mol. Biol.*, **351**, 614–626.

Lightstone,F.C. *et al.* (2013). Rapid catalytic template searching as an enzyme function prediction procedure. *PLoS One*, **8**, 1–17.

Lopez,G. *et al.* (2011). Firestar-advances in the prediction of functionally important residues. *Nucleic Acids Res.*, **39**, W235–W241.

Matthews,B.W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.

Marhaman,A. and Thornton,J.M. (2008). Methods to characterize the structure of enzyme binding sites. In: T., Schwede and M., Peitsch (eds.) *Computational Structural Biology: Methods and Applications,* Chapter 8, pp. 189–221. World Scientific Publishing, London, UK.

Nadzirin,N. *et al.* (2012). SPRITE and ASSAM: web servers for side chain 3D-motif searching in protein structures. *Nucleic Acids Res.*, **40**, W380–W386.

Porter,C.T. *et al.* (2004). The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res.*, **32**, D129–D133.

Roche,D.B. *et al.* (2010). The binding site distance test score: a robust method for the assessment of predicted protein binding sites. *Bioinformatics*, **26**, 2920–2921.

Stark,A. and Russell,R.B. (2003). Annotation in three dimensions. PINTS: patterns in non-homologous tertiary structures. *Nucleic Acids Res.*, **31**, 3341–3344.

Torrance,J.W. and Thornton,J.M. (2009). *Structure-Based Prediction of Enzymes and Their Active Sites*. Wiley, Chichester, UK.

Tsunasawa,S. *et al.* (1989). The primary structure and structural characteristics of Achromobacter lyticus Protease I, a Lysine-specific Serine Protease. *J. Biol. Chem.*, **264**, 3832–3839.

Wallace,A.C. *et al.* (1997 Tess: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases application to enzyme active sites. *Protein Sci.*, **6**, 2308–2323.

Wass,M.N. *et al.* (2010). 3DLigandSite: predicting ligand-binding sites using similar structures. *Nucleic Acids Res.*, **38**, W469–W473.

Yamada,K. and Tomii,K. (2014). Revisiting amino acid substitution matrices for identifying distantly related proteins. *Bioinformatics*, **30**, 317–325.

Zvelebil,M. and Baum,J.O. (2008). *Understanding Bioinformatics*. Garland Science, New York, USA.