

SANDRO CARVALHO IZIDORO

**ALGORITMOS GENÉTICOS PARA
IDENTIFICAÇÃO DE SÍTIOS ATIVOS EM
ENZIMAS**

Belo Horizonte
02 de março de 2014

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

**ALGORITMOS GENÉTICOS PARA
IDENTIFICAÇÃO DE SÍTIOS ATIVOS EM
ENZIMAS**

Tese apresentada ao Curso de Pós-Graduação em Bioinformática da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Doutor em Bioinformática.

SANDRO CARVALHO IZIDORO

Belo Horizonte
02 de março de 2014



UNIVERSIDADE FEDERAL DE MINAS GERAIS

Algoritmos genéticos para identificação de sítios ativos em
enzimas

SANDRO CARVALHO IZIDORO

Ph. D. GISELE LOBO PAPPÀ – Orientadora
Universidade Federal de Minas Gerais

Dra. RAQUEL CARDOSO DE MELO-MINARDI – Co-orientadora
Universidade Federal de Minas Gerais

Belo Horizonte, 02 de março de 2014

Resumo Estendido

Mais de 14 mil famílias de proteínas estão anotadas no Pfam (*Protein Families Database*), das quais cerca de 3.500 ainda têm suas funções desconhecidas. Testes experimentais são caros e demorados e, na sua ausência, estudos têm demonstrado que a função de uma proteína pode ser inferida com sucesso baseando-se similaridade da sequência ou da estrutura de uma proteína hipotética e proteínas de função conhecida.

Uma maneira de prever a função de uma proteína é através da busca dos sítios de ligação (*binding sites*). Sítios de ligação são regiões na superfície de uma enzima especialmente modeladas para interagir com outras moléculas. Devido à sua importância para a função da enzima, os aminoácidos do sítio ativo são mais conservados durante a evolução do que a sequência como um todo. Consequentemente, eles podem ser uma rica fonte de informações para a predição de função.

Diversos métodos já foram propostos para identificar sítios ativos com base em *templates*. Porém, eles apresentam algumas limitações. Grande parte desses métodos não é capaz de lidar com mutações conservativas, onde enzimas com a mesma função podem variar em termos da composição dos aminoácidos do sítio ativo. Além disso, muitos deles não são capazes de identificar a cadeia ao qual um resíduo pertence ou restringem a busca em termos de número de resíduos no *template* ou distâncias máximas entre o *template* e o sítio candidato.

O principal objetivo desta tese é propor um novo método para a busca de sítios ativos baseados em *templates* utilizando algoritmos genéticos com base em dados estruturais. Para isso foi proposto o *Genetic Active Site Search* (GASS), um algoritmo genético modelado para utilizar informações estruturais de um sítio ativo *template* na busca de enzimas com sítios ativos similares. O método pode encontrar sítios ativos com resíduos em cadeias diferentes e é capaz de lidar com mutações conservativas, além de não impor quaisquer restrições quanto ao número de resíduos no sítio ativo e a distância entre eles.

Os resultados do GASS foram comparados com os sítios catalíticos anotados no *Catalytic Site Atlas* (CSA) utilizando quatro diferentes conjuntos de dados. Quando comparado com outros métodos de busca de sítios catalíticos, os resultados mostraram que o GASS pôde identificar corretamente mais de 90% dos sítios pesquisados.

Experimentos também foram realizados utilizando os dados de sítios de ligação da competição CASP 10 e, quando comparado com os 17 métodos participantes, o GASS apareceu em quarto lugar, embora não tenha sido inicialmente desenvolvido com este propósito.

Abstract

Currently, 25% of proteins annotated in the *Protein Families Database* (Pfam) have their function unknown. Experimental tests are expensive and time-consuming, and research has shown that the function of a protein can be successfully inferred based on the sequence or structure similarity of a hypothetical function and other functions of known function.

A way of predicting the function of a protein is to consider its binding sites. Binding sites are regions in the surface of an enzyme designed to interact with other molecules. Due to its importance to enzyme function, the residues in the active site are more conserved than the sequence as a whole, providing important information for function prediction. Hence, active sites are a rich source of information for protein function prediction.

Many methods have been previously proposed to identify active sites based on similarity. However, they do present some limitations, such as not being capable of dealing with conservative mutations (which occur when enzymes with the same function differ in terms of active site residues composition), having difficulties in assigning the active site to a chain or restricting the number of residues in the template.

The main goal of this thesis is to propose a new method for searching for active sites similar using genetic algorithms based on protein structural data, namely Genetic Active Site Search (GASS). The method is based on a genetic algorithm, modeled to use structural information from an active site template in the search for enzymes with similar active sites. The method can find active sites with residues in different chains and is able to handle conservative mutations, apart from not imposing any restrictions on the number of residues in the active site and the distance between them.

GASS results were compared with catalytic sites noted in the Catalytic Site Atlas (CSA) using four different data sets. When compared to other search methods of catalytic sites, the results showed that GASS identified correctly over 90% of the surveyed sites.

Experiments were also performed using data of binding sites from the competition CASP 10, and when compared with the 17 participants methods, GASS appeared in fourth, regardless of not being initially developed with this purpose.

Enquanto o sonho não vem

*Quando o vento para ou muda a direção
Em meio a dias negros, total escuridão
Se a única saída parece o desespero
Não deixe que seus sonhos morram no travesseiro*

*Com um ombro amigo não pode mais contar
Doença do egoísmo só faz contaminar
Na dura batalha de enfrentar o ego alheio
Seja persistente, nunca caia de joelhos*

*Melhor morrer na praia em meio a água e o sol
Do que num quarto escuro, escondido no lençol
Se a única saída parece o desespero
Não deixe que seus sonhos morram no travesseiro*

*Enquanto o sonho não vem
Enquanto o sono não vem*

Sandro Carvalho Izidoro

Agradecimentos

Aos meus pais, Crizo e Alete, por tudo o que fizeram e fazem por mim. Sou eternamente grato a eles pela educação que me deram, pelo apoio incondicional em todas as fases da minha vida e pelo amor sincero e verdadeiro que sempre nos uniu. Digo o mesmo das minhas irmãs Simone e Amanda, e é claro, da pequena Liz. O apoio que sempre me oferecem e a convivência com elas faz tudo ficar mais leve.

A minha esposa Edelma pelo amor, doçura e paciência durante o período do doutorado. Pelas suas aulas particulares de física, química, difração de raios-X, e por tantas outras. Pelo seu olhar encantador e acolhedor, que sempre funcionou como combustível quando mais precisei.

A minha orientadora, professora Gisele Lobo Pappa, por sua confiança, dedicação e amizade. Sou muito grato por suas aulas de Computação Natural e pelas correções em todos os textos que produzi durante o doutorado ("*É só Table 1, não the Table 1, Sandro!*"). Agradeço pela visita em Paris e pela reunião no Genoscope (você pegou o RER D até Évry!). Sem dúvida alguma eu posso dizer que tive uma orientadora de verdade!

A minha co-orientadora, professora Raquel Cardoso de Melo-Minardi, pelo convite para fazer o doutorado em Bioinformática, pela sugestão do tema, pela confiança em meu trabalho e pela oportunidade ofertada a mim de poder estagiar no Genoscope (França).

A todos os colegas do Laboratório de Bioinformática e Sistemas (LBS) que contribuíram de alguma forma para a conclusão deste trabalho. Ao meu grande amigo Douglas E. V. Pires, que por diversas vezes interrompeu o próprio trabalho para discutir comigo sobre o meu trabalho. Sua disponibilidade em ajudar e sua capacidade intelectual são admiráveis. Sou muito grato por todas as suas sugestões, críticas e pelo entusiasmo que tem quando conversamos sobre bioinformática. Muito obrigado pela parceria no GASS WEB.

À Valdete Maria Gonçalves de Almeida e Sabrina Azevedo Silveira pelo carinho com que fui recebido no LBS. Pela atenção e paciência que sempre demonstraram ao me ouvir falando sobre sítios ativos. É claro, pelas aulas de Pymol da Valdete.

À Elisa Boari de Lima pela parceria durante as aulas de EGTP, pela ajuda com o inglês, e pelo apoio nas primeiras semanas em Paris. Obrigado também pelo apoio no Genoscope e pelas discussões durante a elaboração do artigo.

Grato também à Nilma Rodrigues Alves e Moema Monteiro Batista pela parceria nas disciplinas do curso.

À Sheila Santana pela simpatia e profissionalismo que sempre demonstrou frente a secretaria do curso.

Ao meu amigo Daniel B. Roche pelas sugestões e críticas durante a elaboração do artigo. Pelas animadas conversas durante o almoço e pelas boas risadas entre os vários copos de café na cozinha do Genoscope.

Ao François Marie Artiguenave pela confiança em me receber no Genoscope, por sua disponibilidade, atenção e amizade.

É um agradecimento muito especial ao professor Marcelo Matos Santoro (*in memoriam*), exemplo de professor, orientador e ser humano. Apesar do pouco tempo de convívio, jamais esquecerei suas aulas de Bioquímica, suas dicas de música e suas *perguntinhas inocentes*.

Sumário

1	Introdução	1
1.1	Previsão de Função de Enzimas	2
1.2	Algoritmos Genéticos	4
1.3	Objetivos	5
1.4	Organização do Texto	6
2	Revisão da Literatura	7
2.1	Busca de sítios ativos e predição de função	7
2.1.1	Busca de sítios catalíticos	8
2.1.2	Busca de sítios de ligação	11
2.2	Predição de função	13
2.3	Algoritmos genéticos	13
3	Metodologia	16
3.1	Pré-processamento	17
3.2	Modelagem do algoritmo genético	18
3.2.1	Representação do indivíduo e inicialização da população	18
3.2.2	Função de avaliação (<i>fitness</i>)	19
3.2.3	Seleção e operadores genéticos	21
3.2.4	Sítios ativos candidatos	23
3.2.5	Parâmetros	23
3.2.6	Espaço de busca e análise de complexidade	24
4	Resultados e Discussões	25
4.1	Conjuntos de dados	25
4.2	Métricas de avaliação	26
4.3	Definição do átomo de referência para a representação do indivíduo	27
4.4	Parâmetros do GASS	29
4.5	Busca de sítios catalíticos similares utilizando o GASS	29
4.5.1	O GASS é capaz de encontrar sítios ativos em uma família de enzimas?	30
4.5.2	É possível classificar funcionalmente famílias de enzimas com o GASS?	34
4.5.3	Qual o resultado do GASS ao avaliar a evolução convergente?	36

4.5.4	Como o GASS lida com grandes conjuntos de dados?	39
4.5.5	Como o GASS se compara a outros métodos <i>estado da arte</i> para busca de sítios catalíticos?	47
4.5.6	Análise preliminar do impacto da acessibilidade e profundidade na <i>fitness</i>	55
4.6	GASS na busca de sítios de ligação	59
4.7	GASS-WEB	63
4.7.1	GASS-WEB utilizando <i>templates</i> LIT do CSA	63
4.7.2	GASS-WEB utilizando a base NCBI-VAST	66
5	Conclusões e Trabalhos Futuros	67
5.1	Direções de trabalhos futuros	68
5.1.1	Aprimoramento da <i>fitness</i>	68
5.1.2	Matriz de substituição	69
5.1.3	Extensão para sítios de ligação	69
5.1.4	Aprimoramentos no CSA	70
5.1.5	DUFs	70
	Referências Bibliográficas	71
	Informações adicionais	78
	CASP 10	78
	Artigo Publicado	84
	Comprovantes	92
	X-Meeting 2011	93
	3ª Escola Luso-Brasileira de Computação Evolutiva 2012	94
	X-Meeting 2012	95

Lista de Figuras

1.1	Funcionamento de um algoritmo genético.	4
2.1	Tela principal do PINTS.	9
2.2	Tela principal do ASSAM.	10
2.3	Funcionamento de um algoritmo genético.	13
3.1	Sítio ativo de função conhecida (A) - Proteína de função desconhecida (B). . .	16
3.2	Metodologia proposta para a busca de sítios ativos baseada em estrutura. . . .	17
3.3	Representação de um candidato a sítio ativo - (a) Indivíduo do GASS - (b) Sítio catalítico da enzima 3NOS com as distâncias (em Angstroms) entre os LHAs de cada resíduo.	19
3.4	Sítios ativos das enzimas 3NOS, 3NOD, 3NSE e 2BHJ, e seus respectivos valores de <i>fitness</i> utilizando a 3NOS como <i>template</i>	20
3.5	Representação dos operadores de cruzamento e mutação.	22
3.6	Exemplo de uma matriz de substituição para a enzima 1MUC.	22
4.1	<i>Fitness</i> médio e o desvio padrão das três referências para o CD 4: (a) AC - (b) SCC - (c) LHA.	28
4.2	Sítio catalítico da enzima 3NOS.	31
4.3	Sítio catalítico 3NOS com ligante (a) - Sítio catalítico 2BHJ com ligante (b). .	32
4.4	Valor médio de <i>fitness</i> da família NOS utilizando o template 3NOS.	33
4.5	Comportamento do GASS para a família NOS.	33
4.6	<i>Fitness</i> médio da família NOS em relação ao <i>template</i> 3NOS.	35
4.7	Curva ROC - Família NOS utilizando o <i>template</i> 3NOS.	36
4.8	Estrutura e sítio catalítico das enzimas 1ACB e 1R0R.	37
4.9	Sequência das enzimas 1ACB e 1R0R.	37
4.10	Comportamento do GASS para a enzima 1ACB.	38
4.11	Comportamento do GASS para as enzimas 1R0R e 1TEC.	38
4.12	<i>Fitness</i> médio utilizando <i>template</i> 1A0J - CD 3.	41
4.13	CMS do <i>template</i> 1A0J e do valor médio dos 9 <i>templates</i> (CD 3).	41
4.14	Enzima 1ARC: (a) Resíduos encontrados pelo GASS (vermelho) - (b) Localização dos resíduos na superfície (vermelho).	42

4.15	Alinhamento estrutural: (a) Enzima 1A0J (cinza) e 1ARC (vermelho). (b) Sítio catalítico da 1A0J (amarelo) e resíduos encontrados pelo GASS (azul).	42
4.16	Enzima 2GCT: (a) Resíduos encontrados pelo GASS (Cadeia B em vermelho - Cadeia C em cinza). (b) Localização dos resíduos na superfície da enzima (HIS e ASP em vermelho - SER em preto).	43
4.17	Enzima 2KAI: (a) Resíduos encontrados pelo GASS - HIS 57 (vermelho) na cadeia A (em vermelho claro), ASP 102 e SER 195 (amarelo) na cadeia B (cinza escuro). (b) Localização dos resíduos na superfície da enzima (HIS em vermelho escuro - ASP em amarelo).	44
4.18	(a) Distribuição dos valores de <i>fitness</i> - CD 4. (b) Distribuição dos valores de <i>fitness</i> para os sítios catalíticos encontrados corretamente conforme o CSA.	45
4.19	Enzima 3E4D: (a) Resíduos encontrados pelo GASS. (b) Localização dos resíduos na superfície da enzima (HIS em vermelho escuro - ASP e SER em amarelo).	46
4.20	Enzima 1HIA: (a) Resíduos encontrados pelo GASS (HIS 57 A em vermelho - ASP 102 B e SER 195 B em laranja). (b) Localização dos resíduos na superfície da enzima (vermelho).	46
4.21	Resultado parcial ASSAM - <i>Template</i> 1ACB.	50
4.22	CMS do CD 5.	53
4.23	Resultados do GASS x sítios anotados no CSA: enzima 1UK7 com sítio catalítico (em vermelho) (a) utilizada como entrada LIT pelo CSA para e enzima 1L7A (sítio catalítico em vermelho - CSA versão 2.2.12 - e SER 181 em verde). <i>Template</i> 1UK7 (c), parte do novo sítio catalítico informado pela nova versão do CSA (d), e sítio catalítico informado pelo CSA (versão 2.2.12) (e).	54
4.24	Grupos participantes do CASP 10 (categoria FN) classificados em ordem decrescente pelo valor médio de MCC juntamente com GASS. Preditores humanos são mostrados em cinza, preditores baseados em servidores em azul e o GASS em laranja.	63
4.25	Página principal do GASS-WEB.	64
4.26	Página GASS-WEB CSA <i>Templates</i> .	64
4.27	Página de resultados do GASS-WEB CSA <i>Templates</i> .	65
4.28	Página GASS-WEB CSA <i>Templates</i> .	65
1	X-Meeting 2011.	93
2	3ª Escola Luso-Brasileira de Computação Evolutiva 2012.	94
3	X-Meeting 2012.	95

Lista de Tabelas

4.1	Resultados da comparação entre as referências utilizadas para representar um indivíduo no GASS - carbono α (AC), centróide da cadeia lateral (SCC) e último átomo mais pesado da cadeia lateral (LHA).	29
4.2	Parâmetros do GASS, definidos em testes preliminares para cada CD.	30
4.3	Resultados GASS e CSA: para cada CD é apresentado o número de enzimas e templates, o número de sítios catalíticos anotados no CSA (<i>padrão-ouro</i>) e o número de sítios catalíticos encontrados corretamente pelo GASS, o percentual de sítios ativos encontrados em relação aos anotados no CSA e o tamanho do ranking utilizado pelo GASS.	31
4.4	Distribuição dos valores de <i>fitness</i> das enzimas da família NOS.	32
4.5	Distribuição dos valores de <i>fitness</i> das enzimas aleatórias.	34
4.6	Enzimas <i>Trypsin-like</i> e <i>Subtilisin-like</i> e seus respectivos sítios catalíticos.	37
4.7	<i>Templates</i> utilizados pelo GASS - CD 3.	39
4.8	Resultados por <i>templates</i> para o CD 3 considerando apenas uma execução do GASS, com os sítios catalíticos encontrados considerando diferentes tamanhos de ranking.	40
4.9	Distribuição dos valores de <i>fitness</i> do CD 4.	44
4.10	Posição dos resíduos dos sítios catalíticos - GASS x PINTS.	48
4.11	Disparidade nos resultados - GASS x ASSAM.	50
4.12	Resultado parcial do GASS - Template 1ACB.	51
4.13	Experimento GASS x CatSid - CD 5.	52
4.14	Enzimas e <i>templates</i> utilizados no teste com informações de acessibilidade.	56
4.15	Resultado das estratégias com acessibilidade.	57
4.16	Resultado da estratégia de acessibilidade binária.	58
4.17	Resultado das estratégias com profundidade.	59
4.18	Targets with biologically relevant ligands used in CASP 10.	60
4.19	Definição dos resíduos dos sítios de ligação alvo. Fonte: Cassarino et al. (2014).	61
4.20	Matriz de confusão cumulativa para todos os grupos no CASP 10, incluindo o GASS.	62
4.21	Resultados do GASS e dos métodos participantes do CASP 10, ordenados pelo valor médio de MCC, calculado sobre todos os alvos avaliados.	62

1	Pontuação para cada método executado no CASP 10, incluindo o GASS. TP:Verdadeiro Positivo, FP:Falso Positivo, FN:Falso Negativo, TN:Verdadeiro Negativo, MCC:Matthews Correlation Coefficient, BDT:Binding-site Distance Test.	78
2	Teste Wilcoxon signed-rank de todos os grupos no CASP 10, incluindo o GASS.	83

Capítulo 1

Introdução

A bioinformática pode ser definida como a ciência que aborda os problemas relacionados com o armazenamento, recuperação e análise de informações sobre sequência, função e estrutura biológica (Altman, 1998). Inicialmente essas informações eram basicamente de sequência. Porém, o rápido aumento no número de estruturas macromoleculares tridimensionais disponíveis em bases de dados, tais como o PDB (Berman et al., 2000), fez surgir uma nova subárea da bioinformática: a bioinformática estrutural.

A bioinformática estrutural também trata da representação, armazenamento, recuperação, análise e apresentação de informações estruturais, mas em escalas atômicas. A bioinformática estrutural possui duas grandes metas: a criação de métodos de uso geral para a manipulação de informações sobre macromoléculas biológicas e a aplicação destes métodos para resolver problemas da biologia gerando novos conhecimentos (Altman e Dugan, 2009). O tema desta tese pretende contribuir para que essas duas metas sejam alcançadas.

Como em toda área de pesquisa, a bioinformática estrutural também possui os seus desafios, e entre eles estão (Altman e Dugan, 2009):

- **Dados estruturais não são lineares:** Essa característica demanda novos algoritmos, uma vez que os algoritmos tradicionais que lidam com sequências não são apropriados neste contexto. No contexto estrutural, o que importam são as forças de atração e repulsão entre átomos e essas interações formam uma rede muito mais complexa que uma sequência simples. Dessa forma, algoritmos que trabalham com dados estruturais são normalmente mais caros computacionalmente que os que trabalham com sequência, e comumente trabalham com aproximações.
- **Espaço de busca contínuo:** As estruturas são representadas por coordenadas cartesianas dos átomos, que são variáveis contínuas. Assim, os algoritmos precisam lidar com um espaço de busca infinito.
- **Conexão fundamental entre a física e a estrutura molecular:** Embora esta afirmação pareça óbvia e trivial, certos tipos de aproximações, como por exemplo,

representações reduzidas de aminoácidos (e.g. pseudo-átomos) podem repercutir diretamente na identificação das interações moleculares.

- **Dados ruidosos e imperfeitos:** Apesar do avanço nas técnicas de elucidação das estruturas moleculares com altíssima resolução, ainda existem problemas relacionados a flexibilidade, dinâmica e ruídos experimentais.
- **Mais informação de sequência que estrutura:** Estruturas das proteínas são mais conservadas do que sua sequência, mas menos disponíveis. As sequências podem acumular mutações ao longo do tempo, dificultando a identificação de semelhanças, enquanto que as estruturas podem permanecer essencialmente idênticas.

Esses desafios aparecem ao tratar de diversos problemas que consideram a estrutura das proteínas, tais como visualização, classificação, simulação e previsão. Quanto à previsão, esta pode ser relacionada a estrutura ou a função de uma enzima, sendo esta última o enfoque deste trabalho. Essa tese propõe um novo arcabouço para previsão de função de enzimas baseado em duas fases principais: (i) a busca do sítio ativo de uma enzima utilizando algoritmos genéticos, e (ii) a inferência da função da enzima de acordo com o sítio e outros atributos relevantes, como detalhado a seguir.

1.1 Previsão de Função de Enzimas

As proteínas são produtos dos genes e de agentes estruturais e funcionais dos organismos vivos. No entanto, mais de 20% das famílias de proteínas ainda têm função desconhecida, e uma estimativa aponta que haverá mais famílias de proteína de função desconhecida do que as de funções conhecidas (Roberts, 2004).

Mesmo os organismos mais intensivamente estudados ainda possuem muitas proteínas que não foram caracterizadas, como é o caso da *Saccharomyces cerevisiae*¹, que ainda tem 17% das suas proteínas de funções desconhecidas. Geralmente, as proteínas de função desconhecida são anotadas como proteínas hipotéticas, e o conhecimento sobre elas pode ser ampliado pela descoberta do papel biológico que elas desempenham (Watson et al., 2005).

A função de uma proteína depende de sua estrutura. Proteínas que compartilham de uma mesma estrutura podem ter funções similares. Na falta de dados experimentais, a função de uma proteína pode ser inferida comparando sua estrutura com a estrutura de uma proteína de função conhecida (Zvelebil e Baum, 2008). Contudo, esse método falha nos casos em que uma proteína tem um tipo de enovelamento muito comum e conservado para famílias funcionalmente diversas.

Neste trabalho, o foco é um grupo determinado de proteínas conhecido como enzimas. Enzimas são proteínas que catalizam reações biológicas que, sem sua presença, dificilmente

¹Espécie de levedura utilizada na produção de pão e cerveja.

ocorreriam. Elas convertem substâncias, chamadas substratos, em outras, os produtos. Elas são extremamente específicas em relação aos seus substratos, e esta especificidade é conferida por um arranjo de aminoácidos que geralmente está localizado em cavidades ou bolsas estruturais da enzima, os quais são responsáveis pelo reconhecimento molecular (Dundas et al., 2006). Esse conjunto de resíduos que está diretamente envolvido na reação de catálise é chamado de sítio catalítico ou sítio ativo.

Devido à sua importância, os aminoácidos do sítio ativo foram mais conservados durante a evolução do que as sequências como um todo, e podem ser utilizados com sucesso na predição de função de enzimas. Eles são especialmente úteis quando as sequências são muito diferentes ou quando a inferência da função com base unicamente na homologia da sequência não é possível.

Alguns métodos têm sido descritos na literatura para inferir a função de uma enzima com base na similaridade do seu sítio ativo com outras enzimas de estrutura e função conhecidas (Zvelebil e Baum, 2008; Torrance e Thornton, 2009). Esses métodos trabalham normalmente em duas fases: primeiro, o sítio ativo é identificado na enzima para a qual se pretende inferir a função. Depois, com o sítio ativo localizado, a função da enzima pode ser inferida. Essas duas etapas podem ser realizadas utilizando tanto a sequência de aminoácidos quanto as coordenadas da estrutura tridimensional.

Quando trabalha-se com a sequência, métodos de alinhamento múltiplo são utilizados para encontrar os sítios ativos. Já ao lidar com a estrutura da enzima, o problema de encontrar os sítios ativos torna-se mais difícil, mas ao mesmo tempo os resultados tendem a ser mais precisos. De fato, a informação estrutural é muito mais conservada que a de sequência, ou seja, sequências diferentes podem se enovelar em estruturas similares.

Um fator importante a ser considerado quando se estuda sítios ativos em enzimas é a redundância no alfabeto dos aminoácidos, ou seja, aminoácidos diferentes podem desempenhar funções semelhantes. Por exemplo, Arginina e Lisina são os aminoácidos com carga formal positiva enquanto Aspartato e Glutamato são negativos. Da mesma forma, há um conjunto de aminoácidos capazes de doar ou aceitar pontes de hidrogênio ou de estabelecer interações hidrofóbicas. Deste modo, as proteínas podem sofrer mutações conservativas em sítios ativos com diferentes aminoácidos, mas mantendo função idêntica. Em outras palavras, a evolução promove mutações conservativas, e enzimas com a mesma função podem variar em termos da composição dos aminoácidos do sítio ativo. Isso é um importante dificultador tanto para os métodos baseados em sequência quanto em estrutura.

Quando se trata de estrutura há ainda a variabilidade inerente a conformação tridimensional. O problema advém do fato de que as enzimas são moléculas flexíveis, e as ligações covalentes entre os aminoácidos tem muitos graus de liberdade. Isto faz com que a conformação de sítios ativos similares sejam ligeiramente diferentes. Assim, qualquer método que se baseie na estrutura e geometria do sítio ativo deve ser robusto o suficiente para lidar com essas variações. Dessa forma, algoritmos para busca de sítios ativos simi-

lares devem ser capazes de fazer um casamento não exato. Neste contexto acredita-se que os algoritmos genéticos sejam apropriados para lidar com os desafios desse problema, pois são mecanismos de busca global, capazes de lidar com incerteza e dados ruidosos (Back et al., 1997; Unger, 2004).

1.2 Algoritmos Genéticos

Algoritmos Genéticos (AGs) são algoritmos de busca e otimização global baseados em mecanismos da seleção natural e sobrevivência do indivíduo mais adaptado, e foram desenvolvidos por John Holland e sua equipe na Universidade de Michigan (Goldberg, 1989). Assim como a Programação Genética, Programação Evolucionária e Estratégias Evolucionárias, os AGs fazem parte de uma classe de técnicas denominada Algoritmos Evolucionários, que por sua vez pertencem a área de estudo chamada Computação Evolucionária (Brownlee, 2011). Nesse trabalho, esses métodos serão utilizados para buscar sítios ativos em enzimas.

AGs são compostos por procedimentos iterativos que evoluem uma população de indivíduos, onde cada indivíduo representa uma solução candidata para o problema em questão. A cada iteração, denominada *geração*, os melhores indivíduos são selecionados com base em uma função de aptidão (*fitness*). Operadores genéticos (cruzamento e mutação) são aplicados aos indivíduos selecionados, visando produzir novos indivíduos a partir do material genético de seus pais. Esse processo é repetido até que uma condição de parada seja satisfeita. A Figura 2.3 ilustra o funcionamento de um AG.

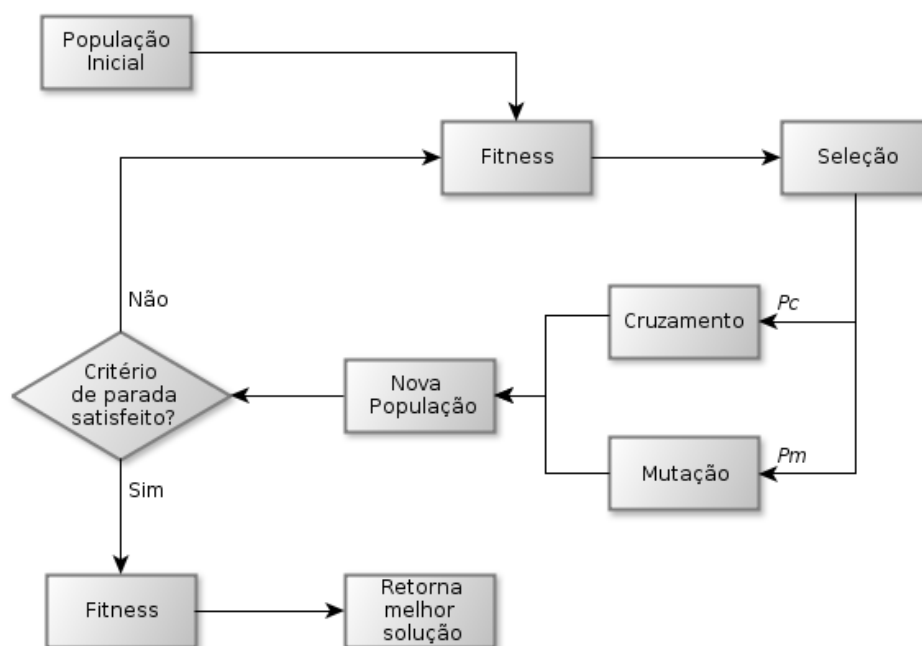


Figura 1.1: Funcionamento de um algoritmo genético.

Inspirado nos processos evolutivos, os AGs são métodos não-determinísticos, onde a geração da população inicial é feita de forma aleatória e a aplicação dos operadores genéticos (cruzamento e mutação) depende de valores de probabilidades definidos previamente pelo usuário. Isso faz com que cada execução possa gerar um resultado diferente. Os AGs também possuem um paralelismo implícito, uma vez que trabalha com uma população de soluções. Assim, eles efetuam uma busca paralela em diferentes áreas do espaço de soluções. Além disso, eles realizam uma busca global e lidam com incerteza e ruídos nos dados de entrada (Back et al., 1997).

1.3 Objetivos

O objetivo desse trabalho é propor um novo arcabouço para previsão de função de enzimas baseado em duas fases principais: (i) a busca do sítio ativo de uma enzima utilizando algoritmos genéticos, e (ii) a inferência da função da enzima de acordo com o sítio e outros atributos relevantes, com base em dados estruturais do PDB.

Os objetivos específicos são:

- Modelar e implementar um algoritmo genético para a busca de sítios ativos similares;
- Implementar scripts para filtrar e organizar dados do PDB, CSA e outras bases de dados que forem necessárias neste projeto;
- Construir conjuntos de dados de estruturas de enzimas com dados no PDB para montagem do repositório do algoritmo genético;
- Analisar e comparar os resultados obtidos a partir do algoritmo genético com os resultados de outras técnicas já implementadas verificando a eficácia e eficiência da modelagem proposta;
- Construir um conjunto de dados de sítios ativos de função conhecida que possa ser usado em experimentos de inferência de função;
- Coletar um conjunto de estruturas de proteínas de funções desconhecidas para que se possa inferir sua função;
- Utilizar o AG para realizar buscas de sítios conhecidos em proteínas de função desconhecida;
- Propor e validar modelos de inferência de função baseados, principalmente, em informação de similaridade de sítios e outras propriedades físicoquímicas de sítios e suas regiões vizinhas.

1.4 Organização do Texto

No Capítulo 2, Revisão da Literatura, foi feito um levantamento bibliográfico de trabalhos correlatos, discutindo predição de função de enzimas e algoritmos genéticos. O Capítulo 3, Metodologia, descreve a estrutura da metodologia proposta e a modelagem do algoritmo genético. Os resultados são apresentados e discutidos no Capítulo 4. O Capítulo 5 é referente à conclusão e é seguido pelo cronograma de trabalho para os próximos semestres no Capítulo 6. No Apêndice estão os comprovantes de apresentações de trabalhos e publicações.

Capítulo 2

Revisão da Literatura

Neste capítulo são apresentados trabalhos relacionados a busca de sítios ativos para predição de função de proteínas, e uma revisão das metodologias baseadas em algoritmos genéticos para bioinformática.

2.1 Busca de sítios ativos e predição de função

Como já mencionado no Capítulo 1, a predição de função de uma proteína pode ser realizada com base na similaridade do seu sítio ativo com outras enzimas de estrutura e função conhecidas. Para a busca de sítios ativos similares, dois tipos de informação molecular podem ser utilizados:

- Sequência de aminoácidos;
- Coordenadas de estrutura tridimensional.

Como os sítios ativos tendem a ser muito conservados ao longo da evolução, muitos trabalhos utilizam de múltiplos alinhamentos de sequência de diferentes organismos para detectar essas conservações (Casari et al., 1995; Armon et al., 2001; Landgraf et al., 2001; Pupko et al., 2002; Sol et al., 2003; Pazos et al., 2006; Henschel et al., 2007; Bernardes et al., 2008; Najmanovich et al., 2008; Goldenberg et al., 2009; Torrance e Thornton, 2009; Lopez et al., 2011).

No entanto, estes métodos possuem um inconveniente: se as sequências forem muito diferentes, as mesmas podem não ser alinhadas com sucesso. Sequências similares indicam funções similares, porém, existem casos de proteínas com um alto grau de identidade de sequência mas realizando funções diferentes (Whisstock e Lesk, 2003). Nestes casos, os métodos de trabalho baseados em estrutura são uma alternativa. Embora saiba-se que a estrutura é mais conservada e relacionada com a função do que a sequência, ainda existem poucos métodos baseados em estruturas quando comparados aos de sequência, tais como Lichtarge et al. (1996); Wallace et al. (1997); Madabushi et al. (2002); Barker e Thornton (2003); Kristensen et al. (2006, 2008); Brylinski e Skolnick (2008); Ward et al. (2009). O

motivo para essa discrepância entre a quantidade de métodos que utilizam sequência e os que utilizam estrutura pode estar no fato de que o número disponível de sequências é muito maior do que o número disponível de estruturas tridimensionais. Além disso, existem as dificuldades inerentes de se tratar o problema de busca em um espaço tridimensional.

As próximas seções descrevem com mais detalhes alguns trabalhos com objetivos similares ao proposto nesta tese. Elas foram divididas em métodos para busca de sítio catalítico e sítio de ligação, respectivamente. Como mencionado anteriormente, no geral, os métodos se propõem a resolver apenas uma destas tarefas, embora computacionalmente o problema possa ser tratado da mesma forma.

2.1.1 Busca de sítios catalíticos

Em geral, os métodos baseados em estrutura para identificar sítios ativos nas proteínas são baseados em grafos, onde os nós representam os átomos dos aminoácidos da cadeia lateral e as arestas são as conexões entre átomos vizinhos, ponderadas por suas distâncias.

Um dos trabalhos que utiliza esta representação é Artymiuk et al. (1994), onde o problema de identificação de padrões tridimensionais de aminoácidos de cadeias laterais em estruturas de proteínas é modelado como um problema de isomorfismo de subgrafo. Dois grafos são ditos isomorfos se existe uma correspondência exata entre seus nós, ou seja, os dois grafos são idênticos. Um subgrafo é um subconjunto dos nós de um grafo. O isomorfismo de subgrafo existe quando os subgrafos de dois grafos são idênticos.

Wallace et al. (1997), por sua vez, apresentam o *Template Search and Superposition* (TESS), uma técnica que deriva de modelos tridimensionais a partir de estruturas do *Protein Data Bank* (PDB) com base em *hashing* geométrico. O método compila um conjunto de tabelas hash contendo informações geométricas sobre os átomos das estruturas PDB. Para cada resíduo referencial, todos os átomos a partir de uma distância limiar previamente definida são dispostos em um *grid* e suas distâncias calculadas. A busca é feita utilizando um *template* contra as informações armazenadas na tabela. Eles também compilaram um banco de dados de modelos tridimensionais de sítios ativos chamado PROCAT, que foi substituído pelo *Catalytic Site Atlas* (CSA) (Bartlett et al., 2002).

O PDB¹ é um repositório que mantém os dados estruturais de macromoléculas biológicas. Os dados depositados no PDB são extraídos utilizando várias técnicas, tais como Raio-X (determinação da estrutura cristalina), ressonância magnética nuclear (NMR) e microscopia eletrônica por criogenia (Berman et al., 2000).

O CSA² é um banco de dados que documenta sítios ativos de enzimas e resíduos catalíticos de estruturas tridimensionais com dois tipos de entradas: um conjunto original, anotado manualmente e derivado de literatura primária (LIT), e um conjunto encontrado por alinhamento utilizando PSI-BLAST (Porter et al., 2004). Os mesmos autores também

¹<http://www.pdb.org/>

²<http://www.ebi.ac.uk/thornton-srv/databases/CSA/>

estão envolvidos na proposta do JESS (Barker e Thornton, 2003), um método de busca de padrões tridimensionais baseado em restrições.

Laskowski et al. (2005b) apresenta o ProFunc, um servidor Web para predição de função que disponibiliza vários métodos baseados em sequência e estrutura. As coordenadas da estrutura pesquisada são enviadas ao servidor no formato PDB, e o ProFunc faz então uso de vários métodos baseados em sequência (Blast, UniProt, PROSITE, Pfam) e vários outros baseados em estrutura (Secondary Structure Matching, JESS e CSA) para realizar a predição.

Já Najmanovich et al. (2008) propôs um método baseado em correspondência de subgrafos para a detecção de similaridades atômicas tridimensionais introduzindo algumas simplificações, permitindo estender a sua aplicabilidade à análise de modelos dos átomos de sítios de ligação. O isomorfismo de subgrafo é realizado em duas etapas. Na primeira, são utilizados apenas os átomos de carbono α dos resíduos de interesse na busca de subgrafos. Na segunda etapa, todos os átomos que não os de hidrogênio são utilizados junto com um limiar de distância obtido com os subgrafos encontrados na primeira etapa. Essa associação faz diminuir o tamanho do espaço de busca.

A seguir são descritos os trabalhos de Stark e Russell (2003) (PINTS), Nadzirin et al. (2012) (ASSAM) e Lightstone et al. (2013) (CatSIId), que serão utilizados para comparação e validação da metodologia proposta.

Figura 2.1: Tela principal do PINTS.

Os autores do *Patterns In Non-homologous Tertiary Structures* (PINTS)³ (Figura 2.1) utilizam um método de busca em profundidade para encontrar todos os padrões possíveis de resíduos comuns aos dois conjuntos de coordenadas (as do sítio catalítico modelo e as dos aminoácidos da proteína alvo), e, em seguida, para cada padrão encontrado, é calculado o *Root Mean Deviation Square* (RMSD) entre os átomos correspondentes. Porém, o método não considera aminoácidos que estejam a mais de 12 Å de distância entre si como possíveis combinações. Além disso, alguns aminoácidos não são considerados na busca. Segundo Russell (1998), o carbono é não reativo nas proteínas e aminoácidos com cadeias laterais contendo apenas átomos de carbono e hidrogênio, e portanto, ALA, PHE, GLY, ILE, LEU, PRO, VAL são ignorados.

ASSAM
Amino acid pattern Search for Substructures And Motifs

SPRITE ASSAM NASSAM About GRASS Contact us HOME

Amino acid 3D pattern searching in protein structures
The ASSAM program searches for patterns of amino acid side chains in 3D space within PDB structures. The searches are based on distances between the amino side chains which are represented as pseudoatom vectors. This search is independent of the amino acid sequence order.

Citation for referencing ASSAM:
Nurul Nadzirin, Eleanor Gardiner, Peter Willett, Peter J. Artymiuk, Mohd Firdaus-Raih. 2012. SPRITE and ASSAM: web servers for side chain 3D-motif searching in protein structures. *Nucleic Acids Res.* 2012 Jul;40(Web Server issue):W380-6. Epub 2012 May 9. [HTML](#), [PDF](#).

Notes:

1. Your query should be a 3D amino acid pattern which will be used to search for other similar amino acid arrangements in a database of PDB structures.
2. Please ensure that the query is a PDB formatted file.
3. The query file cannot contain more than 12 amino acid residues.
4. An error will be returned for queries exceeding 12 residues including cases where 12 or less residues have multiple positions (ie. from NMR).
5. Queries submitted may take a few minutes to complete, please leave your browser on the page until the results are returned or use the link to retrieve the results (an ASSAM search typically runs for 1-5 minutes depending on the server load).
[see the [Help](#) page]

If your input is a protein structure for searching against a pattern or motif database, please use the [SPRITE](#) program.

Upload file to submit for ASSAM search:

Nenhum arquivo selecionado

Database to compare against:

* PDB input files will not be used for purposes other than the search and will be deleted after 2 days to a week upon search completion.

Precomputed examples:
View precomputed results of ASSAM search of:
[4cha_motif.pdb](#), a 3-residue motif of 4cha, an alpha-chymotrypsin. [\[View\]](#)
[1a0s_motif.pdb](#), a 4-residue motif from 1a0s, the sucrose-specific porin ScrY from *Salmonella typhimurium*. [\[View\]](#)

Demonstration run:
Click submit to do a demonstration ASSAM run for 4cha_motif. [\[Submit\]](#)

Figura 2.2: Tela principal do ASSAM.

No *Amino acid pattern Search for Substructures And Motifs* (ASSAM)⁴ (Figura 2.2), a estrutura da proteína é representada como um grafo, onde os nós são os aminoácidos da cadeia lateral. De fato, cada nó consiste de dois pseudo-átomos que são usados pra formar um vetor, e cada vetor corresponde a um nó no grafo. A relação geométrica entre os pares de resíduos é definida em termos de distâncias calculadas entre os vetores correspondentes no grafo. A versão avaliada do ASSAM utiliza o método de Subgrafo Máximo Comum (*Maximal Common Subgraph* - MCS) para encontrar estruturas similares, efetuando a

³<http://www.russelllab.org/cgi-bin/tools/pints.pl>

⁴<http://mfrlab.org/grafss/assam/>

busca de um conjunto de aminoácidos em uma base de dados de estruturas do PDB. A base de dados é composta por um conjunto de aproximadamente 28.500 estruturas do PDB (via NCBI VAST⁵) não redundantes, mas o usuário também pode executar o ASSAM com uma versão não acurada de 57.500 estruturas do PDB. Para esse conjunto, estruturas repetidas, tais como mutantes, foram manualmente removidas, mas versões da mesma proteína com e sem ligante foram mantidas. O usuário do programa pode fornecer as coordenadas no formato PDB de até 12 aminoácidos como referência. Os resultados são apresentados em uma lista ordenada pelo valor do RMSD.

Catalytic Site Identification (CatSid), por sua vez, realiza uma correspondência proteína-*template* utilizando um método de pesquisa de sub-gráfico e uma biblioteca de *templates* de resíduos catalíticos do CSA. Os resultados são refinados usando um procedimento de pontuação logística para reclassificar os padrões encontrados na primeira fase, que utiliza informações, tais como predições de sítio de ligação e outros descritores físicos, para melhorar os resultados obtidos.

2.1.2 Busca de sítios de ligação

Ao contrário da área de predição de sítios catalíticos, a área de predição de sítios de ligação é mais bem estruturada, e possui diversos mecanismos para permitir uma melhor avaliação dos métodos sendo criados e comparação entre diversos métodos propostos. Entre estes dois mecanismos estão as competições *Continuous Automated Model Evaluation* (CAMEO)⁶ e o *Critical Assessment of protein Structure Prediction* (CASP)⁷.

CAMEO foi originalmente criado para monitorar o desempenho dos servidores de predição de estruturas de proteínas registrados no *Protein Model Portal* (PMP). Atualmente ele avalia continuamente a precisão e confiabilidade das predições dos servidores nas categorias de estrutura de proteínas 3D e resíduos do sítio de ligação. Dentro de cada categoria, os servidores cadastrados tem acesso a um conjunto de sequências para efetuar a predição em um determinado prazo. Ao final do processo, os resultados são comparados com os dados experimentais e então são divulgados ao público pelo site (Hass et al., 2013).

O objetivo do CASP é avaliar o atual estado da arte dos métodos de predição de estrutura e função, para identificar as limitações e apontar oportunidades para novos desenvolvimentos. O CASP contempla a predição de estruturas através das categorias *template-based* (TBM) e *free-dodelling* (FM), e a predição de função através da categoria *function prediction* (FN), que foi introduzida na sexta edição (CASP 6) (Cassarino et al., 2014).

A predição de função no CASP é baseada na busca de resíduos em contato com ligantes biologicamente relevantes. São disponibilizados *targets* e *templates*, e a avaliação da

⁵<http://www.ncbi.nlm.nih.gov/>

⁶<http://cameo3d.org/>

⁷<http://predictioncenter.org/>

qualidade das predições é realizada utilizando o *Matthews Correlation Coefficient* (MCC) (Matthews, 1972) e o *Binding-site Distance Test* (BDT) (Roche et al., 2010). Para verificar a significância estatística da avaliação dos resultados, é aplicado o teste *Wilcoxon signed-rank* (Wilcoxon, 1945) com os valores de MCC para cada predição. Os métodos participantes são divididos em dois grupos: servidor (predição gerada a partir de métodos computacionais) e humano (predição gerada com base na análise de dados de literatura, servidores e bases de dados).

Vários métodos foram propostos no CASP 10 para a categoria FN. Aqui são detalhados os três métodos que apresentaram melhores resultados na competição: FIRESTAR, SP-ALIGN e o CNIO. O FIRESTAR (Lopez et al., 2011), por exemplo, é um sistema especialista para a predição de sítios de ligação e sítios catalíticos. Ele baseia suas predições em transferência de homologia de resíduos funcionalmente importantes. Ele faz uso tanto de informações de sequência quanto de estrutura. Em seu protocolo, ele utiliza o PSI-BLAST para um alinhamento inicial da sequência alvo com uma base de dados de resíduos funcionalmente importantes de proteínas com estrutura conhecida (FireDB) (Lopez et al., 2007). Os alinhamentos são então analisados utilizando o método HHsearch, que é baseado nos modelos ocultos de Markov (Soding, 2005), e os resultados são usados para prever locais funcionais e os resíduos de ligação. Quando estruturas são fornecidas, elas podem ser alinhadas com os modelos do FireDB utilizando o método de alinhamento estrutural LGA (Zemla, 2003), que efetua um alinhamento global e local utilizando matrizes de distâncias formadas com as posições de seus carbonos α .

O servidor SP-ALIGN é uma atualização do FINDSITE (Brylinski e Skolnick, 2008) para a predição de sítios de ligação. Para uma proteína alvo, estruturas templates são selecionadas. As estruturas com ligantes identificados são então sobrepostas a estrutura alvo utilizando o algoritmo de alinhamento estrutural TM-align (Zhang e Skolnick, 2005). Em seguida, os centros de massa dos ligantes são agrupados de acordo com sua proximidade espacial, e o centro geométrico de cada agrupamento irá corresponder ao centro de um sítio de ligação. Finalmente, os sítios de ligação previstos são classificados de acordo com o número de templates que compartilham de um mesmo *pocket*.

O CNIO é um método baseado em conhecimento humano que efetua uma análise nas predições realizadas pelo FIRESTAR e 3DLigandSite, gerando sua própria predição (Cassarino et al., 2014).

3DLigandSite (Wass et al., 2010), por sua vez, alinha estruturas semelhantes, sobrepondo os ligantes *templates* com a estrutura de consulta. Para este alinhamento é utilizado o *Matching molecular models obtained from theory* (MAMMOTH) (Ortiz et al., 2002), que utiliza as posições dos carbonos α como referência. Os melhores alinhamentos são separados e o agrupamento com o maior número de ligantes é selecionado como a área geral do sítio de ligação. Os resíduos que estão próximos aos ligantes são utilizados para prever o sítio de ligação.

2.2 Predição de função

A predição de função da maioria dos métodos que utiliza estrutura tem como base as informações dos sítios ativos similares encontrados. Em alguns casos, os sítios similares são agrupados segundo suas distâncias e características químicas, definindo a predição final (Roy e Zhang, 2012). O alinhamento estrutura-*pocket* também é utilizado com os sítios similares para definir a predição final (Brylinski e Skolnick, 2008). Outra alternativa para a predição é a transferência do número do *Enzyme Commission* (EC) (NC-IUBMB, 1999) do template utilizado para a estrutura na qual foi encontrado o sítio ativo com a maior similaridade. Este método é utilizado, por exemplo, por alguns métodos participantes do CASP (Cassarino et al., 2014) e também será considerado neste trabalho.

2.3 Algoritmos genéticos

AGs (Eiben e Smith, 2003) são procedimentos iterativos que evoluem uma população de indivíduos, onde cada indivíduo representa uma solução candidata para o problema em questão. A cada iteração, denominada *geração*, os melhores indivíduos são selecionados com base em uma função de aptidão (*fitness*). Operadores genéticos (cruzamento e mutação) são aplicados aos indivíduos selecionados, visando produzir novos indivíduos a partir do material genético de seus pais. Esse processo é repetido até que uma condição de parada seja satisfeita. A Figura 2.3 ilustra o funcionamento de um AG.

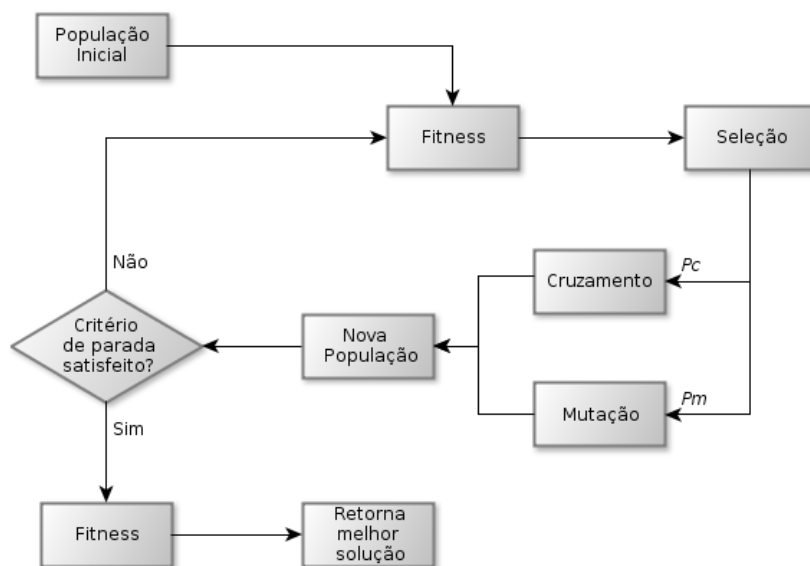


Figura 2.3: Funcionamento de um algoritmo genético.

Algoritmos Genéticos (AGs) têm sido utilizados com sucesso em uma variedade de domínios de aplicação, incluindo problemas de otimização combinatória e proteômica. Deaven e Ho (1995), por exemplo, propuseram um método usando AGs para determinar a estrutura de menor energia de um cluster atômico. Wild e Willett (1996), por sua vez,

fizeram uma pesquisa de similaridade em bases de dados de estruturas tridimensionais de produtos químicos, representados pelos seus campos potenciais eletrostáticos moleculares. Os autores utilizaram um AG para alinhar os campos moleculares, maximizando a sua sobreposição.

Jones et al. (1997) descreveram o *Genetic Optimisation for Ligand Docking* (GOLD) um método de docking automático de ligantes que utiliza um AG para explorar toda a gama de flexibilidade conformacional do ligante com a flexibilidade parcial da proteína.

No ano 2000, Szustakowski e Weng (2000) modelaram um AG para alinhar estruturas de proteínas. Dado que as estruturas das proteínas são mais conservadas no núcleo do que nos *loops* e alças (com exceção dos *loops* e alças envolvidos em sítios ativos), a estratégia utilizada foi a de alinhar os núcleos das proteínas, representados por seus elementos da estrutura secundária (SSE). Para essa tarefa, um AG foi utilizado para encontrar a menor diferença entre as matrizes de distância.

Quatro anos mais tarde, de Magalhães et al. (2004) utilizaram AGs para tratar do problema de docking para proteína-ligante. Nele, o AG trabalha com uma população de indivíduos, onde cada indivíduo é a posição do ligante com relação a proteína. Desta forma, a conformação do ligante é representada por um cromossomo constituído de genes de valores reais representando os graus de liberdade de orientação e conformação. A função de *fitness* foi baseada na interação total de energia entre a proteína e a molécula ligante. Os resultados mostraram que a distribuição da população inicial pode ser relevante para a performance do algoritmo.

Ainda em 2004, Unger (2004) apresentou um trabalho discutindo o uso e os resultados dos AGs em problemas de predição e alinhamento de estruturas de proteínas. Na predição de estruturas, são tratadas questões como representação dos indivíduos, função de *fitness* e operadores genéticos. Entre as questões discutidas está o problema de colisão entre os átomos quando um indivíduo é definido como sendo os valores dos ângulos ϕ e ψ ao longo da cadeia principal. Dessa forma, as aplicações que usam essa abordagem devem incluir, de alguma maneira, um procedimento para detectar essas colisões.

Em Fober et al. (2009) AGs foram usados na construção de alinhamentos de múltiplos grafos (MGA) para a análise estrutural de biomoléculas. No trabalho proposto, cada MGA corresponde a uma solução candidata (indivíduo).

Kernytsky e Rost (2009) utilizaram AGs para o problema de predição de função com uma abordagem diferente. O indivíduo do AG nessa abordagem é codificado a partir de informações de resíduo, estrutura secundária, acessibilidade do solvente, hélice transmembrana e conservação obtidos através de alinhamentos múltiplos de sequências. O AG seleciona os indivíduos mais aptos para a próxima geração a partir da avaliação feita por um algoritmo de aprendizagem baseado em redes neurais.

Kato et al. (2015) implementaram um AG para refinar parâmetros de campo de força com o objetivo de determinar a energia de RNA. Nesta abordagem, os nucleotídeos uracila, adenina, guanina ou citosina são utilizados como referências para os cálculos de mecânica

quântica, onde as energias diedro e electrostáticas são reparametrizadas.

No entanto, não foram encontradas quaisquer outras obras sobre o uso de AGs para busca de sítios ativos similares para predição de função de proteínas.

Capítulo 3

Metodologia

Esta tese propõe uma metodologia para a busca de sítios ativos utilizando dados estruturais de proteínas. A maior contribuição da metodologia proposta está no *Genetic Active Site Search* (GASS) (Izidoro et al., 2014), o AG proposto para realizar buscas baseadas em *templates*.

O problema de busca baseada em *templates* é definido da seguinte forma. Dado um conjunto de N aminoácidos que compõe o sítio ativo A de uma enzima de função conhecida (*template*), e uma proteína hipotética B com M aminoácidos de função desconhecida (Figura 3.1), o método procura o padrão A em B .

A Figura 3.2 ilustra a metodologia proposta para o problema definido acima. Proteínas e *templates* são selecionados pelo usuário para a etapa de pré-processamento. Nesta etapa é criado um repositório de proteínas com informações provenientes do PDB e do CSA, que serão acessados pelo GASS para criar sua população inicial, conforme detalhado nas próximas seções. Em seguida, o GASS executa uma busca heurística para encontrar os

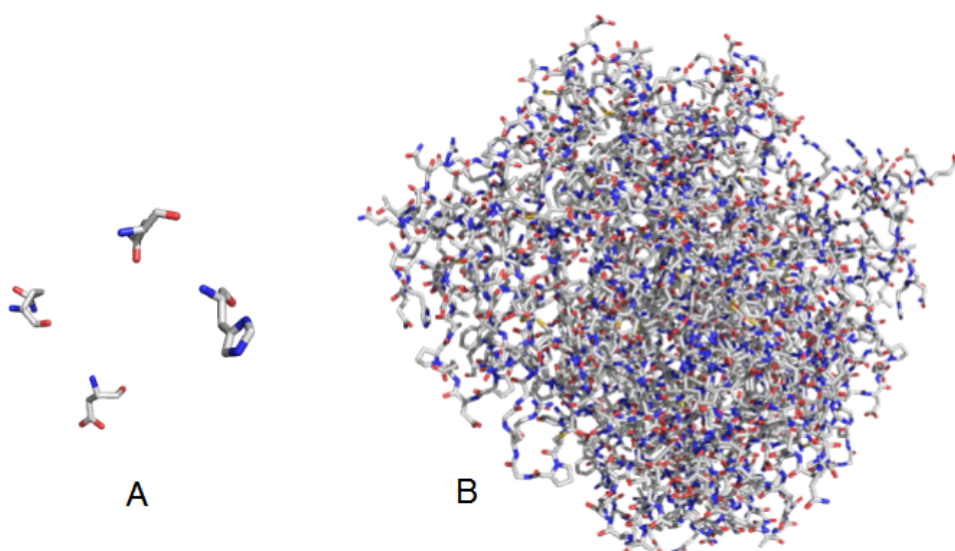


Figura 3.1: Sítio ativo de função conhecida (A) - Proteína de função desconhecida (B).

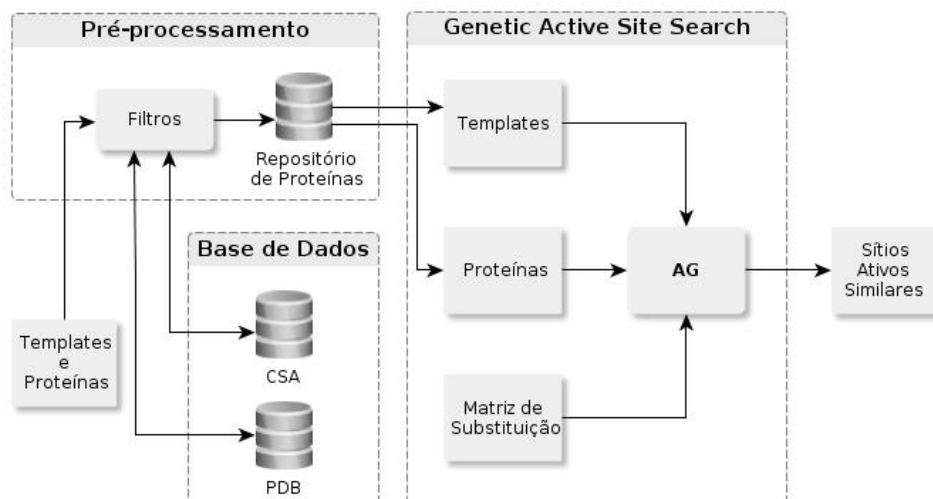


Figura 3.2: Metodologia proposta para a busca de sítios ativos baseada em estrutura.

sítios ativos correspondentes nas proteínas selecionadas, produzindo um ou mais sítios ativos candidatos. A fim de lidar com a mutação conservativa, o GASS também tem a opção de consultar uma matriz de substituição de resíduos.

O método pode ser explorado em dois cenários diferentes: para encontrar um *template* específico (sítio ativo conhecido) em uma ou mais proteínas ou, dado um conjunto de *templates*, para encontrá-los em uma ou mais proteínas. Todos esses passos serão detalhados a seguir.

O processo de obtenção de dados a partir do PDB e CSA é descrito na Seção 3.1. Encontrar a solução por enumeração de todas as possibilidades de arranjos de M aminoácidos de B e selecionar aqueles aminoácidos com maior similaridade conformacional se torna intratável a medida que o valor de M aumenta. Dessa forma, optou-se pelo uso de métodos heurísticos para realizar esta busca, e o método escolhido foi o algoritmo genético (AG), como detalhado na Seção 3.2.

3.1 Pré-processamento

O objetivo da etapa de pré-processamento, é criar um repositório de proteínas que será acessado pelo GASS para criar os sítios ativos candidatos. Esse repositório armazena informações sobre cada resíduo que pode compor um sítio ativo. Para cada resíduo é armazenado seu nome, o átomo de referência (para este trabalho, o último átomo mais pesado da cadeia lateral - ver Seções 3.2.1 e 4.3), o identificador da cadeia, e as coordenadas de referência do átomo na proteína (x , y e z).

A possibilidade de sítios ativos com resíduos em domínios (cadeias) diferentes também foi considerada nesta etapa. As informações são armazenadas conforme o número de cadeias que a proteína contém. Por exemplo, a proteína 2GCT (*gamma-chymotrypsin*) possui quatro cadeias (A, B, C e D), e assim, no repositório serão criadas cinco instâncias

diferentes: quatro delas com as informações dos resíduos separados por cadeia, e uma instância contendo informações de todas as cadeias. Dessa forma, um indivíduo do GASS pode conter resíduos de cadeias diferentes. No caso de uma proteína possuir apenas uma cadeia, somente uma instância será criada no repositório.

3.2 Modelagem do algoritmo genético

Como já mencionado, um AG trabalha com uma população de indivíduos, cada um representando uma solução para o problema em questão. A Seção 3.2.1 descreve como um indivíduo representará um sítio ativo, e como a população inicial do GASS é criada. O conjunto inicial de indivíduos deve ser então avaliado utilizando uma função *fitness* que diz o quão boa a solução que ele representa é para o problema em questão, como detalhado na Seção 3.2.2. Após a avaliação, os indivíduos são selecionados para passar por operações de cruzamento e mutação, gerando uma nova população, como detalhado na Seção 3.2.3. Esse processo se repete até que um critério de parada, como um número máximo de gerações, seja encontrado.

3.2.1 Representação do indivíduo e inicialização da população

A representação de um indivíduo é um ponto muito importante na modelagem de um AG, e depende muito do conhecimento disponível sobre o problema a ser resolvido. Para o problema em questão, um indivíduo representa um grupo de aminoácidos, o qual é um candidato a sítio ativo de uma enzima. O indivíduo é codificado como um vetor, onde cada posição recebe dados sobre um aminoácido, obtidos a partir do repositório de enzimas criado na fase de pré-processamento.

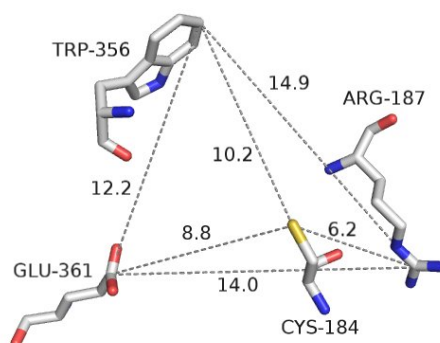
Neste ponto é importante ressaltar algumas informações relevantes sobre os aminoácidos. Cada aminoácido tem duas partes: uma cadeia principal e uma cadeia lateral. A cadeia principal não varia dentro do conjunto de 20 aminoácidos, e sua principal função é conectar diferentes aminoácidos para criar as proteínas. Portanto, é a cadeia lateral que confere especificidade ao aminoácido, tendo cada uma delas diferentes propriedades físico-químicas.

Como o sítio ativo tem importantes propriedades que o tornam único em sua representação, não são considerados os átomos da cadeia principal, apenas os da cadeia lateral. Mais especificamente, apenas a posição do último átomo mais pesado (último não-hidrogênio na cadeia lateral- *Last Heavy Atom* - LHA) de cada aminoácido é utilizado. Discussões sobre a escolha do LHA serão apresentadas na Seção 4.3. Contudo, cabe ressaltar aqui a versatilidade da metodologia, uma vez que outras informações de referência (por exemplo, o carbono α ou o centróide da cadeia lateral) poderiam ser utilizadas, bastando definir estas informações na estapa de pré-processamento.

Assim, para cada aminoácido que pode fazer parte de um sítio ativo, é armazenado o seu nome, o nome do átomo mais pesado na cadeia lateral e sua posição (x, y, z), a posição do aminoácido na sequência da enzima e sua cadeia. A Figura 3.3 mostra um exemplo de um indivíduo formado por 4 aminoácidos.

CYS SG A 184	ARG CZ A 187	TRP CH2 A 356	GLU CD A 361
17.125 8.914 23.94	14.206 4.532 27.145	13.702 14.611 16.21	21.359 16.429 25.582

(a)



(b)

Figura 3.3: Representação de um candidato a sítio ativo - (a) Indivíduo do GASS - (b) Sítio catalítico da enzima 3NOS com as distâncias (em Angstroms) entre os LHAs de cada resíduo.

A população inicial é gerada a partir do repositório de dados obtidos na etapa de pré-processamento. Cada indivíduo é formado por n aminoácidos que são aleatoriamente escolhidos do repositório, sempre respeitando seus tipos conforme o *template* dado, i.e., se a primeira posição requerida é um Glutamato, apenas aminoácidos desse tipo poderão ser selecionados para tal posição.

3.2.2 Função de avaliação (*fitness*)

Tendo a população inicial, o próximo passo do GASS é avaliar os indivíduos. Na metodologia implementada, a distância entre as coordenadas dos LHAs representadas por um vetor de coordenadas 3D é calculada para cada par de resíduos do *template* (\mathbf{v}), e as coordenadas de cada par de resíduos do sítio ativo candidato encontrado pelo GASS (\mathbf{w}), de acordo com a Equação 3.1, onde n é igual ao número de resíduos no *template* e no indivíduo. Note que a diferença entre a métrica na Equação 3.1 e o conhecido *Root Mean Square Deviation* (RMSD) está em não calcular o quadrado da média das distâncias dos resultados. Essa escolha se deve ao fato de que, como apresentado em Laskowski et al. (2005a), sítios ativos com pequenas diferenças entre as distâncias dos resíduos podem ter valores de RMSD semelhantes. Utilizando o valor absoluto das distâncias, evita-se este

problema. Note também que o intervalo do somatório equivale ao número máximo de combinações entre os pares de resíduos que compõe o indivíduo e o *template*.

$$Fit(\mathbf{v}, \mathbf{w}) = \sqrt{\sum_{i=1}^{(n^2-n)/2} \|v_i - w_i\|^2} \quad (3.1)$$

A Figura 3.4 apresenta os sítios ativos das enzimas 3NOS, 3NOD, 3NSE, 2BHJ, e as distâncias (em Angstroms) calculadas a partir da posição do último átomo mais pesado. O número que segue o nome do aminoácido representa a posição do mesmo na sequência da enzima. Por exemplo, GLU-361 representa um Glutamato na posição 361 da sequência de aminoácidos da enzima 3NOS. Ao lado de cada sítio ativo também aparece o valor da *fitness* calculado utilizando o sítio ativo da enzima 3NOS como *template*. Dessa forma, o cálculo da *fitness* da 3NOS utilizando ela mesma como *template* é zero. As demais enzimas apresentam valores diferentes de zero.

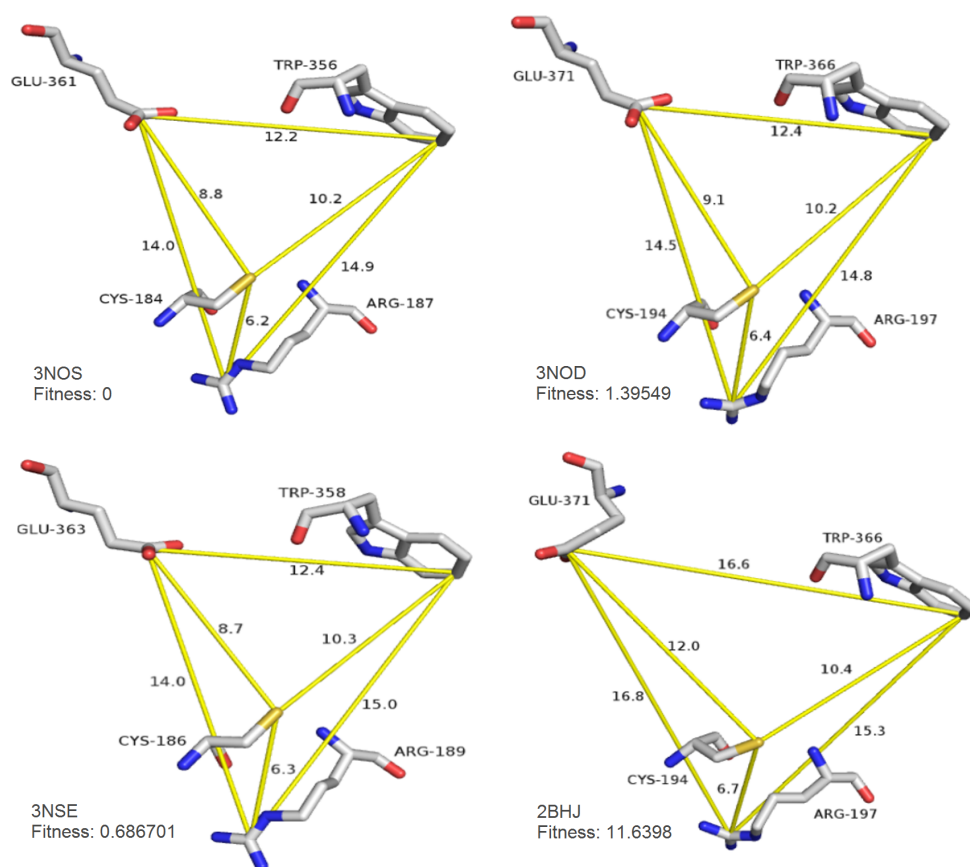


Figura 3.4: Sítios ativos das enzimas 3NOS, 3NOD, 3NSE e 2BHJ, e seus respectivos valores de *fitness* utilizando a 3NOS como *template*.

3.2.3 Seleção e operadores genéticos

Após a avaliação dos indivíduos vem a fase de seleção. Essa fase é crucial para a evolução da população, pois dá uma maior chance de sobrevivência aos melhores indivíduos, i.e., aqueles com melhor *fitness*. Existem diversos métodos de seleção na literatura (Eiben e Smith, 2003). Aqui foi utilizado a seleção por torneio, onde um subconjunto de k indivíduos é sorteado aleatoriamente da população, e o melhor indivíduo desse subconjunto de acordo com a *fitness* é selecionado. Este método tem a propriedade de não requerer conhecimento sobre a população, apenas utilizando uma relação de ordem para classificar k indivíduos. Outra característica positiva do torneio é a facilidade de controlar a pressão seletiva (convergência prematura) por meio do parâmetro k . A convergência prematura acontece quando, em poucas gerações, o AG estabiliza, ficando limitado a explorar apenas uma parte do espaço de busca. Este comportamento pode fazer com que o AG encontre apenas o melhor indivíduo *local*, e não o melhor indivíduo *global*. Desta forma, quanto maior o valor de k , maior será a pressão seletiva.

Uma vez feita a seleção, dois operadores genéticos são usados para gerar uma nova população: cruzamento de um ponto e mutação de um ponto. A Figura 3.5 ilustra ambos os métodos. Dois indivíduos são necessários para o cruzamento e um para mutação. No cruzamento, uma posição aleatória no indivíduo é selecionada, e os resíduos antes desse ponto no primeiro pai são concatenados com os resíduos depois desse ponto no segundo pai (Figura 3.5). Estes novos indivíduos são então adicionados à nova população.

No caso da mutação de um ponto, apenas o ponto escolhido é substituído por um resíduo aleatório, que pode ser do mesmo tipo a partir da enzima selecionada (TRP 356 trocado pelo TRP 190 - em vermelho na Figura 3.5), ou por um tipo diferente de resíduo (mutação conservativa), indicado pela matriz de substituição de resíduos da mesma enzima (GLU 361 trocado pelo ASP 369 - em azul na Figura 3.5).

A matriz de substituição de resíduos utilizada pelo GASS foi criada por Lightstone et al. (2013) e sua principal diferença com relação a outras matrizes de substituição existentes, como a Blosum62 (Henikoff e Henikoff, 1992) ou a MIQS (Yamada e Tomii, 2014), é que ela é específica para cada *template* LIT do CSA.

Como mencionado na Seção 2.1, o CSA possui como entradas um conjunto anotado manualmente e derivado de literatura primária (LIT), e um conjunto encontrado por alinhamento com as entradas LIT utilizando PSI-BLAST. Para criar a matriz de substituição, as entradas LIT são comparadas com as respectivas entradas PSI-BLAST, seguindo dois critérios Lightstone et al. (2013):

1. O número de resíduos na entrada PSI-BLAST deve corresponder ao mesmo número de resíduos na entrada LIT;
2. O número do *Enzyme Commission* (EC) deve ser igual entre as duas entradas (LIT e PSI-BLAST).

O EC é um sistema numérico e hierárquico de classificação de enzimas. Existem 6 principais categorias: (1) Oxidoredutases, (2) Transferases, (3) Hidrolases, (4) Liases, (5) Isomerases e (6) Ligases. O número EC é formado por 4 números (#.#.#.#) que classificam as enzimas, fornecendo detalhes sobre suas reações enzimáticas ((NC-IUBMB, 1999)).

As substituições observadas são anotadas para cada entrada LIT, gerando uma lista de possíveis substituições. Nesta lista, por exemplo, a anotação *2BMI, ASP, GLU* indica que os sítios catalíticos anotados através do PSI-BLAST utilizando a enzima LIT 2BMI podem ter o resíduo ASP substituído por um resíduo GLU ou vice-versa. A partir desta lista, uma matriz binária de substituição é montada, onde a substituição de um resíduo por outro é permitida se a linha de interseção entre os resíduos possui o valor um (1), e não permitida quando o valor é zero (0). Como as substituições são equivalentes, a matriz é simétrica. Boa parte das substituições observadas são pontuais e capazes de conservar a função catalítica, como no exemplo da enzima 2BMI ($ASP \rightleftharpoons GLU$). No entanto, também foram observadas substituições menos óbvias, indicando substituições mais específicas para algumas famílias de enzimas. Ao todo, foram observadas 567 substituições possíveis entre as entradas LIT. Assim, cada entrada LIT terá sua matriz de substituição. A Figura 3.6 apresenta a matriz de substituição para a enzima LIT 1MUC, com as possíveis substituições: $HIS \rightleftharpoons LYS$, $ASN \rightleftharpoons ASP$, $ASP \rightleftharpoons GLU$.

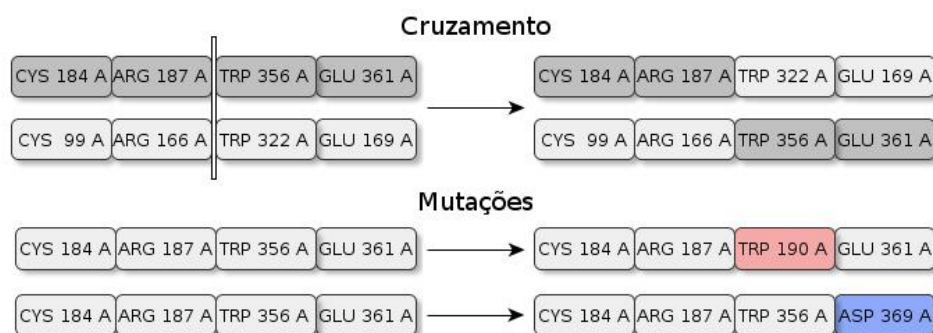


Figura 3.5: Representação dos operadores de cruzamento e mutação.

	Asn	Asp	Glu	His	Lys
Asn	1	0	0	0	0
Asp	1	1	0	0	0
Glu	0	1	1	0	0
His	0	0	0	1	0
Lys	0	0	0	1	1

Figura 3.6: Exemplo de uma matriz de substituição para a enzima 1MUC.

3.2.4 Sítios ativos candidatos

AGs exploram o espaço de busca, de maneira paralela, uma vez que cada indivíduo pode se concentrar em uma região diferente do espaço. Assim, ao final do processo de evolução, pode-se obter um conjunto de sítios ativos candidatos tão grande quanto o tamanho da população. Quando o usuário precisa de uma solução única, o indivíduo com a melhor aptidão é retornado. Em alguns casos, no entanto, pode ser interessante para o usuário analisar um conjunto de soluções, e, em seguida, usar um outro conjunto de critérios, talvez mais subjetivo, para escolher o melhor indivíduo.

Por exemplo, pode ser que a melhor solução (indivíduo com a menor distância a partir do *template*) esteja enterrada na proteína, em vez de estar em um *pocket*. O especialista pode reconhecer imediatamente que este sítio ativo candidato não é um sítio ativo real. Para evitar situações como essa, o método retorna um *ranking* com as n melhores soluções encontradas. Desta forma, um especialista pode escolher o mais adequado de acordo com seu conhecimento do problema.

Note também que a versão atual do GASS considera a distância entre os resíduos como única forma de avaliação. Porém, conforme ilustrado na Seção 4.5.6, outros critérios também podem ser incorporados à função *fitness*.

3.2.5 Parâmetros

Um AG possui um conjunto de parâmetros que influencia diretamente o seu comportamento, e cada problema requer uma configuração particular a partir de testes e análises de resultados preliminares. Para isso, são utilizados valores padrões como ponto de partida até a obtenção dos valores finais para os parâmetros (Brownlee, 2011; Eiben et al., 1999; Haupt e Haupt, 2004). Esses parâmetros são:

- **Número de Gerações:** Determina o critério de parada de execução do algoritmo, e depende muito do problema sendo resolvido. Esse critério pode ser um número de gerações pré-definido, a detecção de convergência ou de um erro mínimo, ou até mesmo o tempo de processamento;
- **Tamanho da População:** Uma população pequena de indivíduos pode afetar negativamente o desempenho do AG, não favorecendo a cobertura do espaço de busca para o problema em questão. Uma população grande demais pode consumir mais tempo e recursos computacionais;
- **Probabilidade de Cruzamento:** Valor que determina a probabilidade de cruzamento dentro de uma população. Taxas altas têm a característica de introduzir novos elementos na população mais rapidamente. Entretanto, o AG pode perder *bons indivíduos* pois a maior parte da população será substituída. Com valores baixos, o AG pode se tornar muito lento. Geralmente, os valores *default* ficam entre 0,6 a 0,95;

- **Probabilidade de Mutação:** Determina a probabilidade que uma mutação ocorrerá. Responsável por manter a diversidade na população, os valores iniciais sugeridos ficam entre 0,001 e 0,01. Outra forma de calcular esse valor é através da fórmula $pm = 1/L$, onde L é o tamanho do indivíduo (Muhlenbein, 1992). Assim, por exemplo, com indivíduos de tamanho 4, tem-se o valor de 0,25;
- **Tamanho do torneio:** Com este valor pode-se controlar a pressão seletiva, e com isso evitar uma convergência prematura. O valor utilizado para o torneio (k) geralmente é configurado em dois (2), dependendo do tamanho da população.

Além dos cinco parâmetros citados acima, o GASS inclui também um sexto parâmetro que determina o número de soluções que serão exibidas ao final da execução:

- **Tamanho do Ranking:** Número das melhores soluções encontradas após a execução do AG. Pode variar de um (1) até o tamanho da população.

Os parâmetros do GASS foram ajustados de forma empírica.

3.2.6 Espaço de busca e análise de complexidade

O tamanho do espaço de busca do problema que está sendo abordado neste trabalho depende do tamanho M da proteína que está sendo pesquisada e o tamanho N do *template*. O número total de soluções que podem ser exploradas é dado, no pior dos casos, pela combinação de m resíduos em subconjuntos de tamanho N . Nota-se que esse espaço pode ser menor, uma vez que não são permitidas combinações de resíduos que não estejam presentes no *template*.

Existem muitos métodos sofisticados para calcular a ordem de complexidade de um AG, incluindo aqueles que se baseiam em processos de cadeia de Markov (Davis e Principe, 1993; SUZUKI, 1995). No entanto, uma boa estimativa pode ser efetuada tendo em conta a operação mais custosa do algoritmo: a função de *fitness*. No GASS, a função de *fitness* é a distância entre o sítio ativo candidato e o *template*, e considera distâncias entre pares de amino ácidos. Isso leva a $(N^2 - N)/2$ subtrações, de acordo com a Equação 3.1. Esse valor deve ser multiplicado pelo número de indivíduos i e gerações g utilizados, que são controlados pelo usuário e definidos de acordo com os recursos computacionais disponíveis. Assim, o algoritmo é executado em $O(N^2ig)$, onde N é o tamanho do *template*.

Capítulo 4

Resultados e Discussões

Este capítulo discute os resultados obtidos ao se avaliar o GASS na busca de sítios ativos baseados em *templates*. Serão apresentados detalhes sobre os conjuntos de dados utilizados nos experimentos (Seção 4.1), da definição do LHA como átomo de referência no GASS (Seção 4.3), e dos parâmetros utilizados nos experimentos (Seção 4.4). As Seções 4.5 e 4.6 apresentam os resultados do GASS na busca de sítios catalíticos e sítios de ligação, respectivamente. A Seção 4.7 apresenta o GASS-WEB, um servidor web com dois recursos disponíveis: (i) busca de sítios catalíticos similares utilizando *templates* LIT do CSA com base em um arquivo PDB fornecido pelo usuário; (ii) busca de sítios ativos similares na base NCBI-VAST utilizando *template* fornecido pelo usuário.

4.1 Conjuntos de dados

Com o objetivo de testar e validar a metodologia implementada para a identificação de sítios ativos, seis conjuntos de dados foram utilizados nos experimentos. Todos os dados foram extraídos do PDB e CSA, e pré-processados (veja Seção 3.1). Os conjuntos foram selecionados visando tentar responder às seguintes perguntas:

- O GASS é capaz de encontrar sítios ativos em uma família de enzimas?
- É possível classificar funcionalmente famílias de enzimas com o GASS?
- Qual o resultado do GASS ao avaliar a evolução convergente?
- Como o GASS lida com grandes conjuntos de dados?
- Como o GASS se compara a outros métodos do estado da arte?

Os conjuntos de dados (CD) utilizados foram os seguintes:

- **CD 1:** 125 enzimas da família sintase de óxido nítrico (SON) (EC: 1.14.13.39) com sítios catalíticos anotados no CSA. Este conjunto também foi testado com outras

126 enzimas, escolhidas aleatoriamente a partir do PDB, com números EC diferentes de 1.-.-.

- **CD 2:** 9 enzimas *Serine Protease* (4 *Trypsin-like* e 5 *Subtilisin-like*), de acordo com (de Almeida, 2011).
- **CD 3:** 1.085 enzimas *Trypsin-like* escolhidas aleatoriamente a partir do PDB, utilizando a classificação SCOP¹ (superfamília 1A0J).
- **CD 4:** 24.437 enzimas extraídas a partir do banco de dados NCBI-VAST não-redundante (p-value 10e-80), conforme relatado em Nadzirin et al. (2012), e um conjunto contendo 100 enzimas escolhidas a partir do PDB com base nos resultados do ASSAM.
- **CD 5:** 61 enzimas e 1.800 *templates* selecionados do CSA, como feito pelo CatSIId (Lightstone et al., 2013).
- **CD 6:** 13 enzimas alvo e 25 *templates* de sítio de ligação para cada enzima, de acordo com o CASP 10, na categoria *Function Prediction (FN)* (Cassarino et al., 2014).

4.2 Métricas de avaliação

Devido as características dos CD e dos métodos utilizados para comparação com o GASS, as métricas de avaliação consideradas durante a validação diferiram de um CD para outro. Como as perguntas relacionadas a cada CD mediam aspectos diferentes do GASS, em alguns casos métricas utilizadas para predição foram mais apropriadas, enquanto em outros casos métricas de ranking foram necessárias.

Para avaliação dos resultados obtidos com o CD 1, foi utilizado o gráfico *Receiver Operating Characteristic* (ROC), que é uma ferramenta para avaliação de algoritmos de aprendizado e predição. A criação deste gráfico é baseada nos valores da matriz de confusão, relacionados como verdadeiro positivo (TP), verdadeiro negativo (TN), falso positivo (FP) e falso negativo (FN). O gráfico ROC é um gráfico bidimensional, onde a taxa de TP (Equação 4.1) é representada no eixo de Y e a taxa de FP (Equação 4.2) é representada no eixo de X, mas que também pode ser representado pelo valor da área abaixo da curva ROC (*Area Under the ROC Curve* - AUC) (Hand, 2009; Fawcett, 2006).

A curva *Cumulative Match Score* (CMS) (Bolle et al., 2005) foi utilizada nos experimentos com os conjuntos CD 3 e CD 5. Esta curva mostra a relação entre o número de sítios catalíticos encontrados corretamente conforme o CSA e sua posição no ranking do GASS. Note que essa métrica foi utilizada apenas nos conjuntos de dados onde o tamanho do ranking foi maior que 1.

¹<http://scop.berkeley.edu/>

$$Taxa\ TP = \frac{Positivos\ classificados\ corretamente}{Total\ de\ positivos} \quad (4.1)$$

$$Taxa\ FP = \frac{Negativos\ classificados\ incorretamente}{Total\ de\ negativos} \quad (4.2)$$

No experimento utilizando o CD 6, comparados com outros métodos da competição CASP, os resíduos do sítio de ligação foram classificados como verdadeiros positivos (TP: resíduos do sítio de ligação preditos corretamente), verdadeiros negativos (TN: predição correta para resíduos não vinculados ao sítio), falsos negativos (FN: resíduos do sítio de ligação incorretamente não preditos), falsos positivos (FP: resíduos não vinculados ao sítio de ligação incorretamente predito) (Cassarino et al., 2014). Assim, a avaliação da qualidade das predições foi realizada utilizando o *Matthew Correlation Coefficient* (MCC) (Equação 4.3), conforme recomendado pelos organizadores da competição. O valor de MCC varia de +1 (predição perfeita) a -1 (predição inversa), onde um MCC de valor 0 corresponde a uma predição aleatória.

O *Binding-site Distance Test* (BDT) (Roche et al., 2010) também foi utilizado no CD 6. Uma vez que o MCC não leva em consideração a proximidade dos resíduos de ligação preditos com os reais, o BDT foi proposto como uma avaliação alternativa, efetuando uma pontuação contínua (entre 0 e 1) relativa às distâncias entre estes resíduos. Desta forma, sítios de ligação, mesmo que preditos incorretamente, poderão ter uma pontuação mais elevada por estarem mais próximos do sítio real, do que aqueles mais distantes.

$$MCC = \frac{(TP.TN - FP.FN)}{\sqrt{(TP + FP).(TP + FN).(TN + FP).(TN + FN)}} \quad (4.3)$$

4.3 Definição do átomo de referência para a representação do indivíduo

Como mencionado no Capítulo 3, o indivíduo representa um sítio ativo a partir de um átomo de referência. Entre as opções de representação tem-se: carbono α (AC), o centróide da cadeia lateral (SCC), e o LHA (Zhang e Skolnick, 2005; Nadzirin et al., 2012). Para o experimento foram utilizados subconjuntos de enzimas de três conjuntos de dados (CD 1, CD 3 e CD 4). A razão da utilização de subconjuntos está no fato de que, em muitas estruturas, não havia informações suficientes para o cálculo do SCC. Para cada experiência, o GASS foi executado 30 vezes. Isso é necessário para obter significância estatística dos resultados, uma vez que o método é não-determinista e cada execução

pode retornar um resultado diferente. A Tabela 4.1 apresenta o número de enzimas que foram utilizadas em cada experimento e o número de enzimas cujos sítios ativos foram encontrados corretamente nas 30 execuções para cada tipo de referência (AC, SCC e LHA).

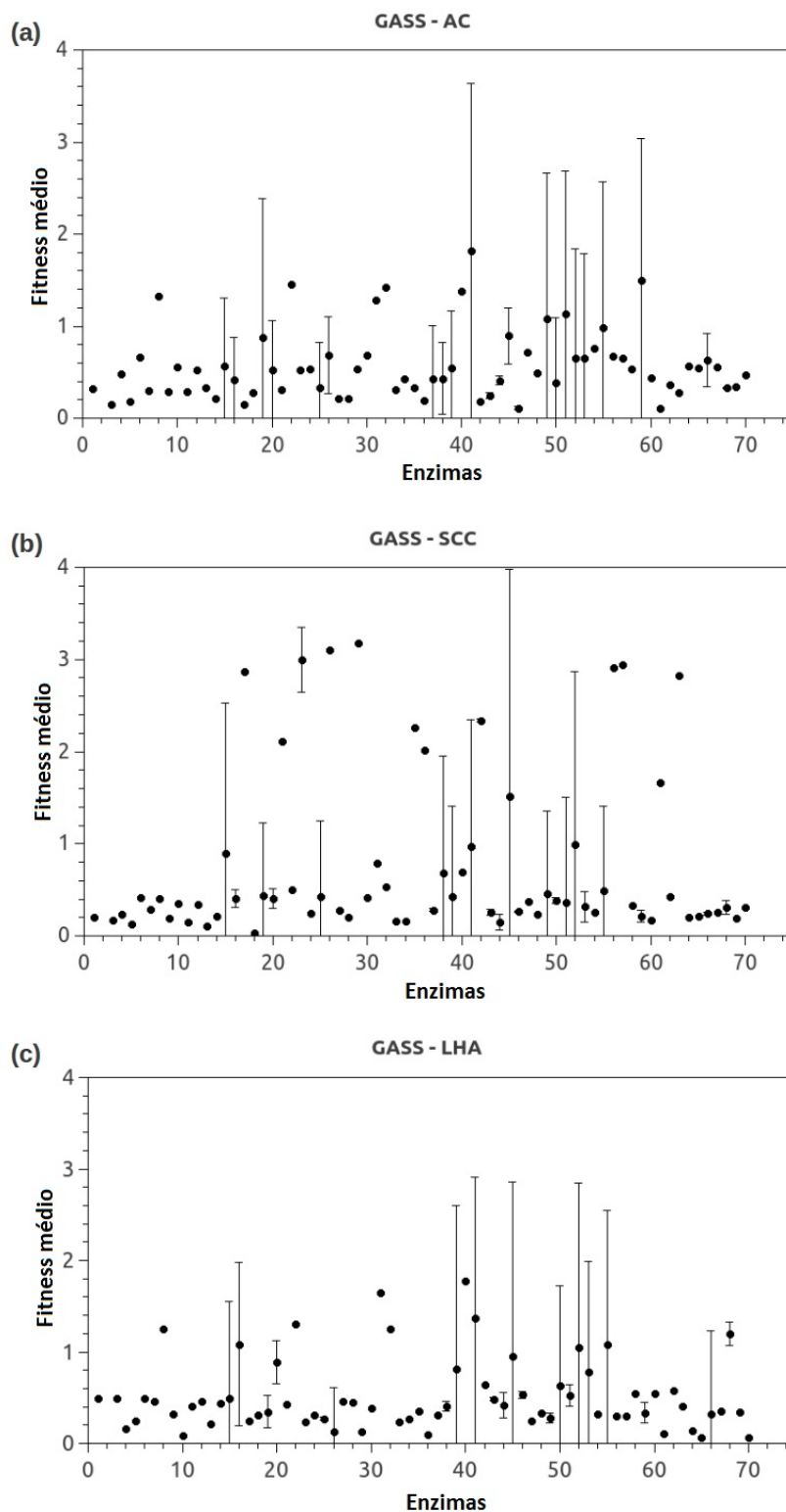


Figura 4.1: *Fitness* médio e o desvio padrão das três referências para o CD 4: (a) AC - (b) SCC - (c) LHA.

Pode-se perceber que não há um vencedor entre as três referências. Os resultados mostraram que o desempenho do método com diferentes referências varia de um conjunto de dados para outro. No CD 4, por exemplo, tanto o LHA quanto o SCC apresentam desvio padrão total da função de *fitness* de 0.22, e o AC apresenta o valor 0.25. Assim, o LHA foi escolhido por não aumentar o custo computacional do pré-processamento (como acontece com o SCC) e por utilizar informações sobre a cadeia lateral em vez da cadeia principal (Bartlett et al., 2002; Kato e Nagano, 2010). A Figura 4.1 mostra o valor de *fitness* médio e o desvio padrão das três referências para o CD 4.

Tabela 4.1: Resultados da comparação entre as referências utilizadas para representar um indivíduo no GASS - carbono α (AC), centróide da cadeia lateral (SCC) e último átomo mais pesado da cadeia lateral (LHA).

Dados	Enzimas	Referência	Corretos
CD 1	126	SCC	123(98%)
		AC	121(96%)
		LHA	121(96%)
CD 3	184	SCC	150(82%)
		AC	156(85%)
		LHA	152(83%)
CD 4	70	SCC	46(66%)
		AC	47(67%)
		LHA	48(69%)

4.4 Parâmetros do GASS

Os parâmetros do GASS foram definidos em testes preliminares para cada CD. Para cada experiência, o GASS foi executado 30 vezes. A Tabela 4.2 apresenta o valor dos parâmetros utilizados em cada CD. É importante notar que a mutação que utiliza a matriz de substituição (Mutaç o B) s o foi utilizada nos conjuntos CD 3 e CD 5 pois somente nestes casos havia informa oes de muta o na matriz. Todos os testes foram executados considerando a varia o de um par metro por vez, e os resultados avaliados em termos de *fitness* m dio e tempo de converg ncia.

4.5 Busca de s tios catal ticos similares utilizando o GASS

Nesta se o os resultados gerados pelo GASS s o validados de acordo com os s tios catal ticos das enzimas catalogadas no CSA (vers o 2.2.12). O CSA pode ser considerado o banco de dados dispon vel mais completo sobre s tios catal ticos, e por isso pode ser usado como um *padr o-ouro* para compara o dos resultados dos m todos de identifica o de

Tabela 4.2: Parâmetros do GASS, definidos em testes preliminares para cada CD.

Parâmetros	CD 1	CD 2	CD 3	CD 4	CD 5	CD 6
População	150	400	400	300	300	300
Gerações	100	100	100	100	100	100
Cruzamento	90%	90%	90%	90%	90%	90%
Mutação A ¹	20%	20%	20%	30%	30%	30%
Mutação B ²	-	-	10%	-	10%	-
Ranking	1	1	10	1	10	10
Torneio	2	2	2	2	2	2

¹Simple mutação que substitui um aminoácido por outro do mesmo tipo.

²Mutação que altera um aminoácido por outro com base em uma matriz de substituição.

sítios catalíticos. No CSA, um resíduo é definido catalítico se preencher qualquer um dos seguintes critérios (Furnham et al., 2013):

- Estiver envolvido diretamente no mecanismo catalítico;
- Alterar o pKA do resíduo ou molécula de água diretamente envolvida no mecanismo catalítico;
- Representar uma estabilização de um estado de transição ou intermediário;
- For responsável pela ativação de um substrato.

A Tabela 4.3 apresenta um resumo dos resultados obtidos pelo GASS em comparação com os sítios catalíticos anotados no CSA, detalhados nas próximas seções. As diferenças entre o número de enzimas pesquisadas e aquelas com sítios catalíticos anotados no CSA acontecem por duas razões: (i) algumas vezes, uma enzima possui mais do que um sítio catalítico e (ii) nem todas as enzimas têm seus sítios catalíticos catalogados no CSA.

As próximas seções foram baseadas nas questões formuladas na Seção 4.1, visando avaliar os resultados do GASS ao tratar a busca de sítios catalíticos similares.

4.5.1 O GASS é capaz de encontrar sítios ativos em uma família de enzimas?

A fim de responder a esta pergunta, o GASS foi utilizado para encontrar os sítios catalíticos em um conjunto de 125 enzimas da família NOS (CD 1). Duas abordagens diferentes foram empregadas nos experimentos executados: a utilização de um único *template* contra toda a família NOS, e a utilização de todas as enzimas como *templates* (*todos contra todos*).

Na primeira abordagem, utilizou-se o *template* 3NOS (*Endothelial nitric-oxide synthase*), uma vez que é a única entrada LIT no CSA entre todas as enzimas da família NOS. Como já mencionado, a preferência pela utilização de enzimas LIT como *templates*

Tabela 4.3: Resultados GASS e CSA: para cada CD é apresentado o número de enzimas e templates, o número de sítios catalíticos anotados no CSA (*padrão-ouro*) e o número de sítios catalíticos encontrados corretamente pelo GASS, o percentual de sítios ativos encontrados em relação aos anotados no CSA e o tamanho do ranking utilizado pelo GASS.

CD	Enzimas	Templates	Sítios Catalíticos		Match (%)	GASS Rank
			CSA	GASS		
1	125	1	248	248	100,00	1
	125	125	248	235	94,76	1
2	9	9	9	9	100,00	1
	9	9	9	9	100,00	1
3	1.085	9	1.085	899	82,85	1
	1.085	9	1.085	987	90,94	5
	1.085	9	1.085	1.015	93,52	10
4	100	1	79	79	100,00	1
	23.318	1	-	-	-	1
5	61	1.800	182	162	89,01	1
	61	1.800	182	165	90,65	5
	61	1.800	182	165	90,65	10

se dá pelo fato delas terem sido anotadas manualmente, baseadas em referências de literatura e experimentos. O *template* 3NOS é composto pelos resíduos Cys 184, Arg 187, Trp 356 e Glu 361. A Figura 4.2 mostra o sítio catalítico da enzima 3NOS com as distâncias (em Angstrom - Å) entre cada par de resíduos.

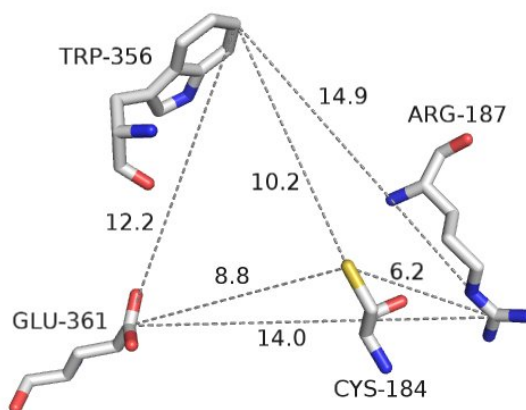


Figura 4.2: Sítio catalítico da enzima 3NOS.

Neste caso, o GASS encontrou corretamente todos os 248 sítios catalíticos anotados no CSA (versão 2.2.12). Observando os valores de *fitness* (definido na equação 3.1) de todos os sítios catalíticos candidatos (indivíduos) na população final, 85,50% apresentaram distâncias, a partir do *template*, menores ou iguais a 5 Å. Isto mostra que a maioria das enzimas da mesma família têm pequena variação de distâncias entre os sítios catalíticos, indicando que esta técnica tem um bom potencial a ser explorado na busca de sítios ativos

similares. A Tabela 4.4 apresenta um resumo da distribuição dos valores de *fitness* para este experimento.

Tabela 4.4: Distribuição dos valores de *fitness* das enzimas da família NOS.

Valores de <i>fitness</i>	Número de Enzimas	Porcentagem
$fitness \leq 5 \text{ \AA}$	106	85,50
$5 \text{ \AA} < fitness \leq 10 \text{ \AA}$	16	12,90
$fitness > 10 \text{ \AA}$	2	1,60
Total	124	100

No entanto, há exceções. Um exemplo é a enzima 2BHJ (*Murino Ino sintase com Coumarin Inibidor*), que apresentou um valor de *fitness* de 11,64 Å, o dobro do valor de aptidão encontrado para a maioria das enzimas da família NOS. Esta diferença ocorre por causa do ligante da enzima. Na 3NOS, o ligante HAR-512 (*N-omega-hidroxi-L-arginina*) ocupa um pequeno volume quando comparado ao ligante FC1-1499 (*Thiocoumarin*) na 2BHJ, como apresentado na Figura 4.3 (Ligantes aparecem em amarelo).

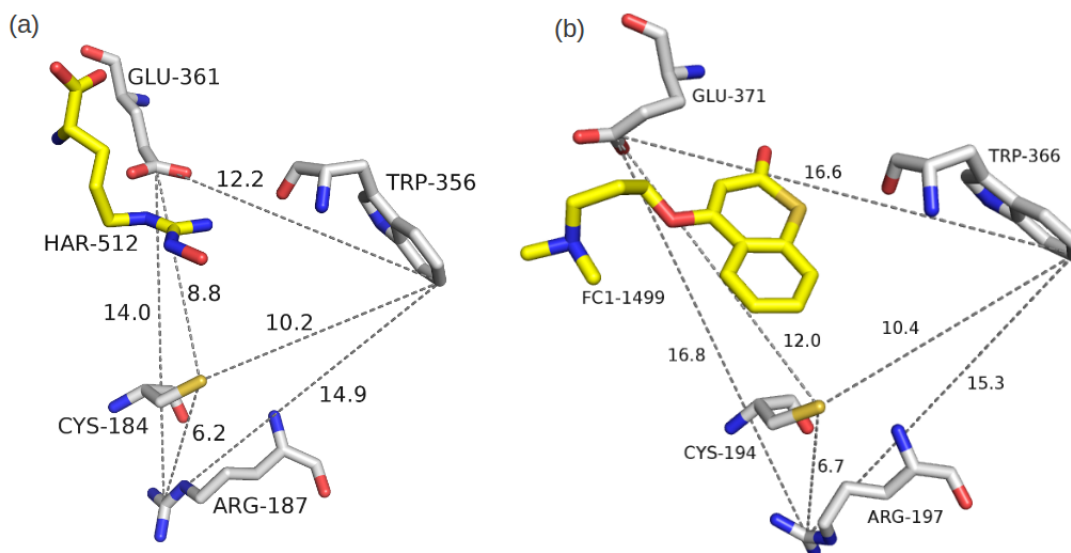


Figura 4.3: Sítio catalítico 3NOS com ligante (a) - Sítio catalítico 2BHJ com ligante (b).

Considerando 30 execuções diferentes do GASS para todas as enzimas com o *template* 3NOS, foi calculado o valor médio e desvio padrão da *fitness* para cada enzima. Apenas 3 enzimas apresentaram desvio padrão diferente de zero (1NOC, 1NOS e 2NOS - *Inducible nitric oxide synthase*). A enzima 1NOC apresentou *fitness* médio de 7,44 e desvio padrão de 2,94 (6 erros em 30 execuções); a enzima 1NOS apresentou *fitness* médio de 9,31 e desvio padrão 1,31 (4 erros em 30 execuções); e a 2NOS obteve *fitness* médio de 10,24 e desvio padrão de 1,61 (7 erros em 30 execuções). Isso mostra que os resultados encontrados têm uma variabilidade muito baixa entre as diferentes execuções do GASS. A baixa variabilidade é necessária para garantir robustez ao método. A Figura 4.4 apresenta o valor médio de *fitness* da família NOS após 30 execuções do GASS. Na figura, cada círculo

representa uma enzima, o eixo x corresponde ao indivíduo do GASS que representa o sítio catalítico de cada enzima e o eixo y corresponde ao valor de *fitness*.

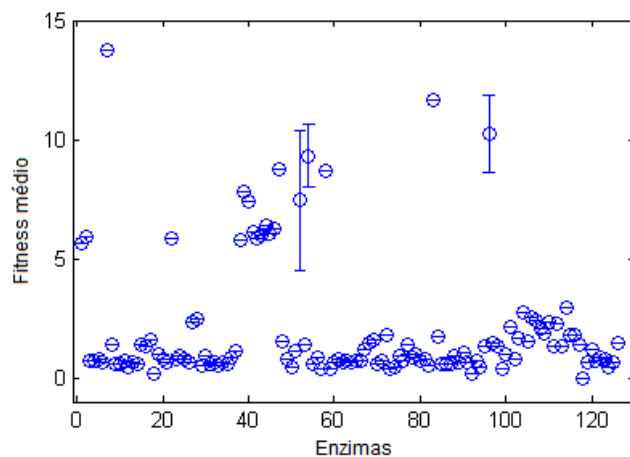


Figura 4.4: Valor médio de *fitness* da família NOS utilizando o template 3NOS.

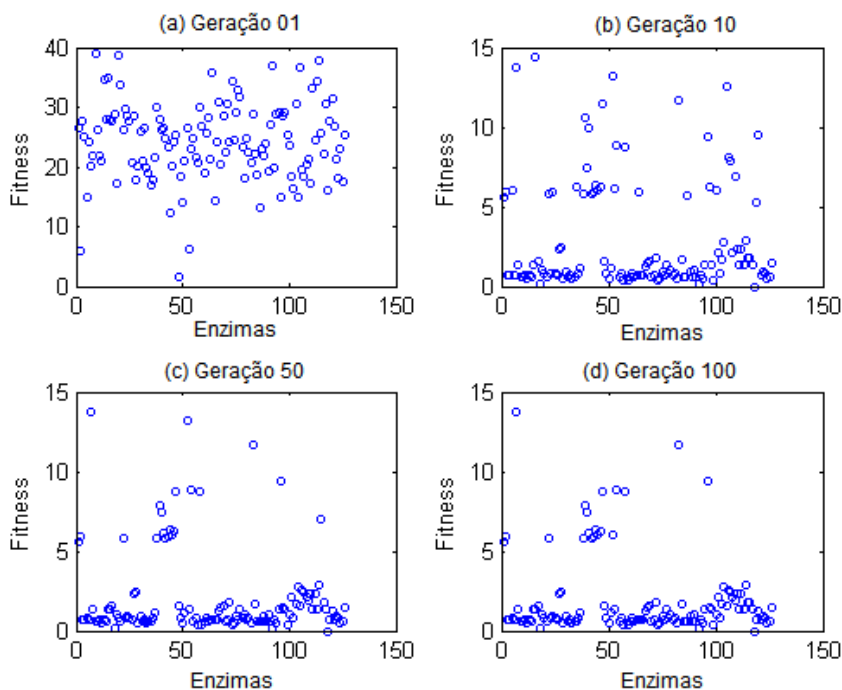


Figura 4.5: Comportamento do GASS para a família NOS.

A Figura 4.5 mostra a evolução do GASS para encontrar o sítio catalítico em 4 diferentes pontos: 1, 10, 50 e 100 gerações. A Figura 4.5 (a) apresenta a população resultante da primeira geração criada aleatoriamente a partir do repositório de proteínas gerado na fase de pré-processamento, em seguida tem-se a evolução do GASS após 10 (b), 50 (c) e 100 (d) gerações. Após 10 gerações, pode-se notar uma separação no espaço das enzimas, com muitas enzimas com sítios ativos muito parecidos e próximos de zero no eixo y . Com

50 gerações (Figura 4.5 (c)), os valores de *fitness* dos indivíduos começam a convergir, existindo pouca variação até o gráfico (d).

Em uma segunda abordagem de teste com a família NOS, foram utilizadas as outras 124 enzimas da família NOS anotadas no CSA (PSI-BLAST) como *templates* (teste *todos contra todos*). Para o conjunto de 125 enzimas, existem 248 sítios catalíticos anotados no CSA. Em média, para cada um dos 125×30 experimentos realizados, o GASS encontrou 235 sítios catalíticos corretamente de acordo com a CSA (94,76 %).

Um exemplo da robustez do método está novamente na enzima 2BHJ (Figura 4.3). Mesmo possuindo os resíduos do seu sítio catalítico um pouco mais distantes entre si, devido ao volume do ligante, o GASS conseguiu, ao utilizá-la como *template*, uma média de 236 sítios catalíticos corretos conforme o CSA.

Apesar da família NOS ser muito conservada em relação ao seu sítio catalítico, os resultados demonstram que o método é capaz de encontrar sítios catalíticos similares baseado na distância de um *template* e uma proteína alvo.

4.5.2 É possível classificar funcionalmente famílias de enzimas com o GASS?

Como explicado anteriormente, o GASS retorna como resultado um ranking dos sítios catalíticos mais semelhantes ao template. Por isso, mesmo quando uma enzima não possui o sítio desejado, o conjunto mais semelhante de resíduos é retornado.

Nesse experimento foi utilizado um conjunto com 251 enzimas, onde 125 são da família NOS (CD 1 - EC: 1.14.13.39) e 126 escolhidas aleatoriamente a partir do PDB com números EC diferentes de 1.-.-.-. Como *template* foi utilizado a enzima 3NOS.

A análise foi feita considerando os valores de *fitness*, esperando que eles sejam mais elevados para as enzimas do conjunto aleatório do que aqueles que pertencem à Família NOS. Dessa forma, um limiar de distância pode ser definido a partir dos valores de *fitness*, e utilizado para classificar famílias de enzimas.

Assim, calculou-se a média e o desvio padrão da *fitness* para cada enzima nesse novo conjunto. A Tabela 4.5 apresenta o número absoluto e a percentagem dos valores de *fitness* encontrados em diferentes intervalos. Observa-se que 80,95% dos valores de *fitness* são maiores do que 10 Å para as enzimas do conjunto aleatório.

Tabela 4.5: Distribuição dos valores de *fitness* das enzimas aleatórias.

Valores de <i>fitness</i>	Número de Enzimas	Percentagem
$fitness \leq 5 \text{ \AA}$	1	0,79
$5 \text{ \AA} < fitness \leq 10 \text{ \AA}$	23	18,25
$fitness > 10 \text{ \AA}$	102	80,95
Total	126	100

No entanto, existem ainda algumas enzimas com valores entre 5 Å e 10 Å (18,25% contra 12,90% nas enzimas da mesma família - Tabela 4.4). Porém, é visível que enquanto 0,79% tem distância menor igual a 5 Å, na Tabela 4.4, esse número é de 85,50% para enzimas da mesma família. Isto pode sugerir que as enzimas de diferentes famílias tendem a ter sítios catalíticos diferentes, e o GASS foi capaz de identificar isso.

A Figura 4.6 mostra os resultados do GASS considerando as distâncias das enzimas da família NOS e as enzimas aleatórias em relação ao *template*. Como esperado, as enzimas dentro da mesma família estão mais próximas do *template* do que as enzimas aleatórias.

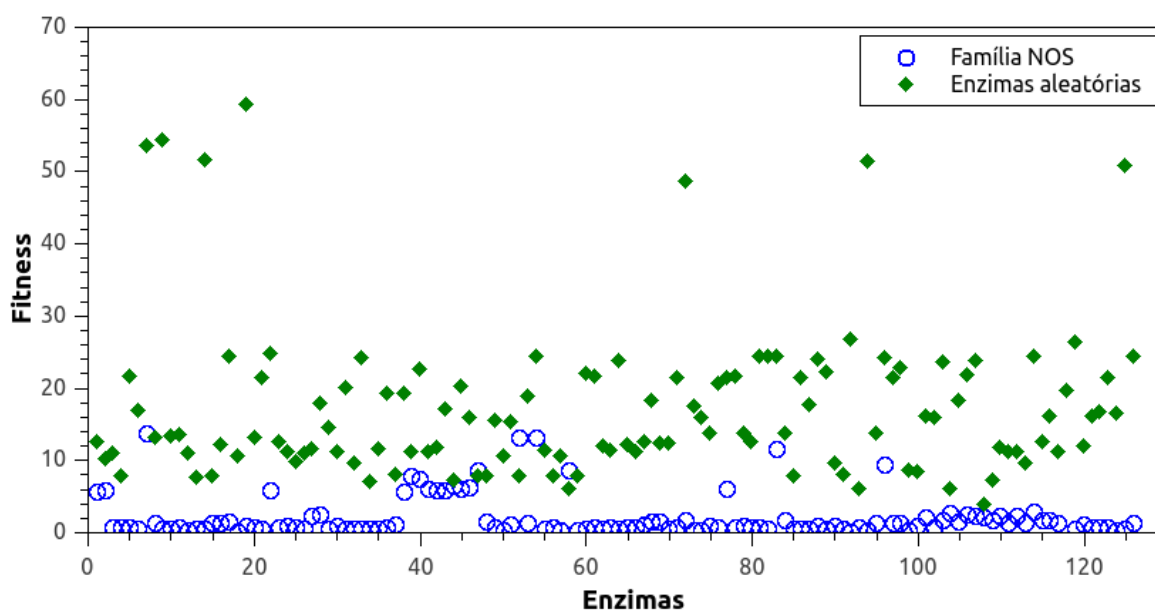


Figura 4.6: *Fitness* médio da família NOS em relação ao *template* 3NOS.

A Figura 4.7 mostra uma curva ROC indicando a capacidade do GASS de associar corretamente enzimas de uma família com base em um simples limiar de distância. A área abaixo da curva (AUC), considerando o limiar de distância é 0,97.

Os resultados mostram que o método é robusto e que a escolha da função *fitness*, a qual é baseada na distância entre o sítio catalítico real e o encontrado, pode ser usada junto com um limiar de distância para determinar se duas enzimas pertencem a uma mesma família e, conseqüentemente, compartilham de uma mesma função.

Deve-se ressaltar mais uma vez que a família NOS é muito conservada em relação ao seu sítio catalítico, e por isso outros testes foram realizados com outros grupos de enzimas com o objetivo de validar a função *fitness* ou de propor alguma melhoria na mesma, conforme discutido nas próximas seções.

Os primeiros resultados deste experimento foram divulgados em forma de pôster no X-meeting (2011) e na 3ª Escola Luso-Brasileira de Computação Evolutiva (2012). Os comprovantes estão no Anexo 5.1.5 (Figuras 1 e 2).

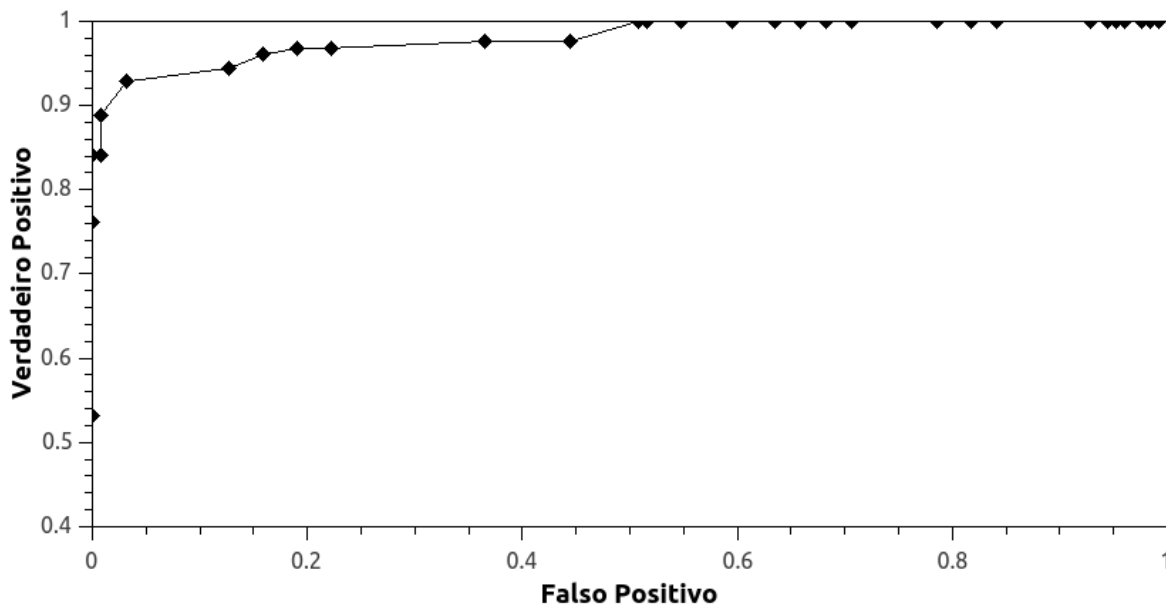


Figura 4.7: Curva ROC - Família NOS utilizando o *template* 3NOS.

4.5.3 Qual o resultado do GASS ao avaliar a evolução convergente?

Evolução convergente é um fenômeno curioso observado na natureza que ocorre quando diferentes enzimas se adaptam e adquirem funções similares (Zvelebil e Baum, 2008). Por exemplo, duas famílias de enzimas não homologas (*Serine Protease*), a *Trypsin-like* (*mammalian enzymes*) e a *Subtilisin-like* (*bacterial enzymes*) não compartilham uma estrutura tridimensional similar e suas sequências têm uma notável diferença na composição, mas elas compartilham da mesma tríade catalítica e tem o mesmo mecanismo de ação. Essas enzimas (provavelmente) não tiveram uma enzima ancestral comum, e assim evoluíram independente uma da outra.

Dois conjuntos de enzimas pertencentes a família Serine Protease (*Trypsin-like* e *Subtilisin-like*) foram analisadas (CD 2). A Tabela 4.6 apresenta as enzimas dos dois conjuntos e seus respectivos sítios catalíticos, conforme o CSA. Os dois conjuntos de enzimas possuem estruturas tridimensionais diferentes e sua identidade de sequência é menor que 20%, mas ambos os conjuntos têm a tríade catalítica (SER-ASP-HIS) no seu sítio catalítico. A Figura 4.8 apresenta a estrutura de duas enzimas, uma *Trypsin-Like* (1ACB) e uma *Subtilisin-like* (1R0R), e seus respectivos sítios catalíticos. É visível a diferença estrutural de ambas as enzimas. Já a Figura 4.9 mostra a sequência de ambas as enzimas, onde também pode-se notar a diferença.

Os experimentos realizados foram divididos em dois grupos. Um grupo utilizou as enzimas *Trypsin-like* como *template* para o GASS, e o outro grupo utilizou as enzimas *Subtilisin-like*. Para cada enzima de referência, o GASS foi executado 30 vezes para todas as enzimas do conjunto de dados (*Trypsin-like* e *Subtilisin-like*).

Tabela 4.6: Enzimas *Trypsin-like* e *Subtilisin-like* e seus respectivos sítios catalíticos.

PDB ID	Família	Sítios catalíticos
1ACB	Trypsin-like	HIS 57 - ASP 102 - SER 195
1PPF	Trypsin-like	HIS 57 - ASP 102 - SER 195
1CHO	Trypsin-like	HIS 57 - ASP 102 - SER 195
3SGB	Trypsin-like	HIS 57 - ASP 102 - SER 195
1TEC	Subtilisin-like	HIS 71 - ASP 38 - SER 225
1CSE	Subtilisin-like	HIS 64 - ASP 32 - SER 221
1MEE	Subtilisin-like	HIS 64 - ASP 32 - SER 221
1SBN	Subtilisin-like	HIS 64 - ASP 32 - SER 221
1R0R	Subtilisin-like	HIS 64 - ASP 32 - SER 221

Em um primeiro teste, foram utilizadas as enzimas *Trypsin-like* (1ACB, 1PPF, 1CHO e 3SGB) como *templates* para o GASS. Quanto aos parâmetros do GASS, o tamanho da população foi alterado em relação aos primeiros experimentos, e o novo valor está na Tabela 4.2.

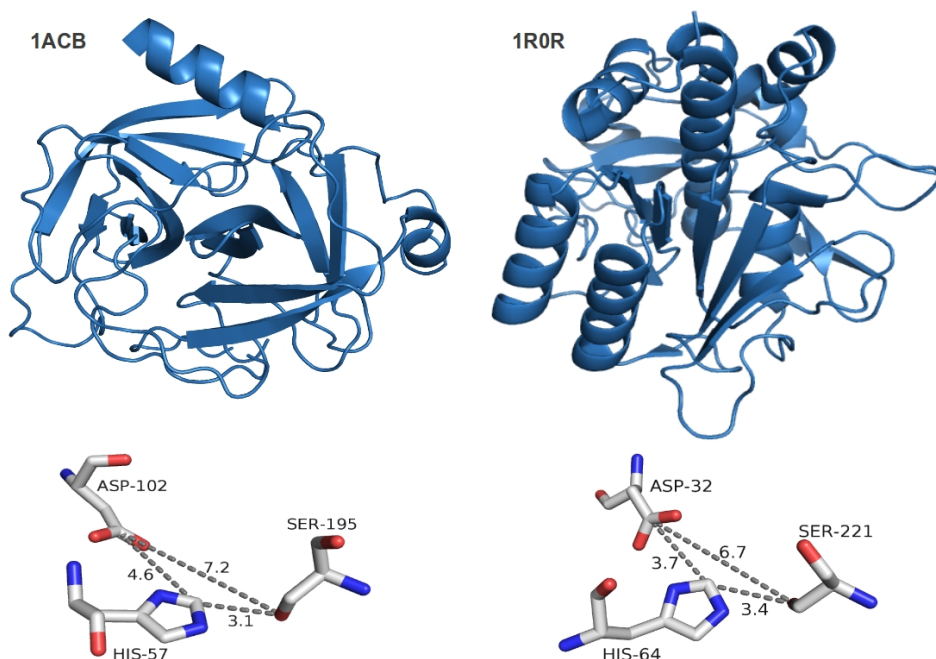


Figura 4.8: Estrutura e sítio catalítico das enzimas 1ACB e 1R0R.

CLUSTAL O(1.1.0) multiple sequence alignment

```

1ACB: E | PDBID | CHAIN | SEQUENCE  -----CGVPATIQ-----PVL SGL--SRIVNGEE-----AVPGSNPWQVSLQDKTGFHFCGGSLLINENWVVTAAHC-
1R0R: E | PDBID | CHAIN | SEQUENCE  AQTVPYGIPLTKADKVAQGFKGANWVKVAVLDTGIQASHFDLNVVGGASFVAGEAYNTDGNHGHTHVAGTVAALDNTTGVLVGAPSVSLYAVKVLNSSLGSGSYGIVSGIENATTNGMDV
      *:*:*
1ACB: E | PDBID | CHAIN | SEQUENCE  -----GVTT-----SDVVAAGEFDQGSSEKIQKLIKIAKVFKNKYNSLTIINNDITLLKLTSTAASFQTVSAVCLPSASDDFAAGTTCVTTGWGLTRYTNA-----NTPD
1R0R: E | PDBID | CHAIN | SEQUENCE  INMSLGGASGSTMKQAVDNAYARGVVVVAAGNSGNSG-STNTIG-----YPAKYDSVIIVGAVD--SNSNRASFSSVGAEL-----EVMAPGAG-VYSTYPTINTYATLNGTSMASPH
      *:*
1ACB: E | PDBID | CHAIN | SEQUENCE  RLQQA-----LPLLSN-----TNCKYWGTKIKDAMI CAGASGVSSCMGDSGGPLVCKKNGAWTLVGVSWGSSTCSTSTPGVYARVTLVWVQQLAAN
1R0R: E | PDBID | CHAIN | SEQUENCE  VAGAAAILLSKHPNLSASQVRNRLSSSTATYLGSSFYGKGLINVEAAAQ-----
      *:*

```

Figura 4.9: Sequência das enzimas 1ACB e 1R0R.

Em todos os testes o GASS encontrou corretamente o sítio catalítico esperado, inclusive para as enzimas *Subtilisin-like*.

A Figura 4.10 apresenta o comportamento do GASS utilizando a enzima 1ACB como *template*. A Figura 4.10 (a) mostra a primeira geração do GASS e a Figura 4.10 (b) a geração de número 100. O eixo *horizontal* corresponde ao indivíduo que representa o sítio catalítico de cada enzima e o eixo *vertical* corresponde ao valor de *fitness* (cada círculo representa uma enzima da Tabela 4.6). Após 100 gerações, pode-se perceber que os valores de *fitness* convergiram para próximo de zero.

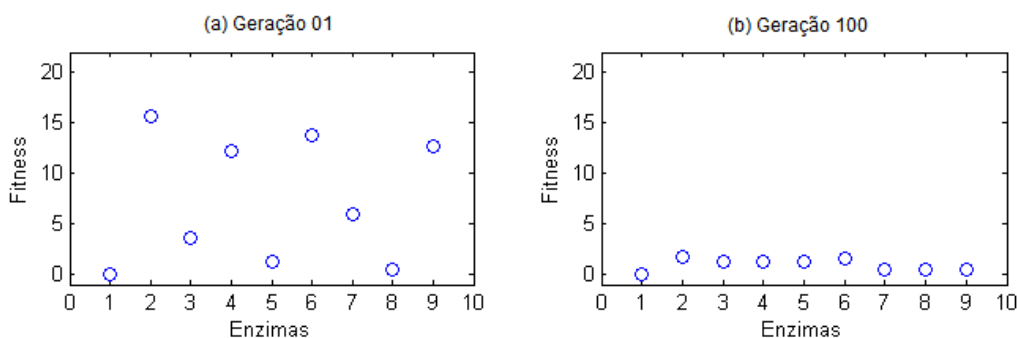


Figura 4.10: Comportamento do GASS para a enzima 1ACB.

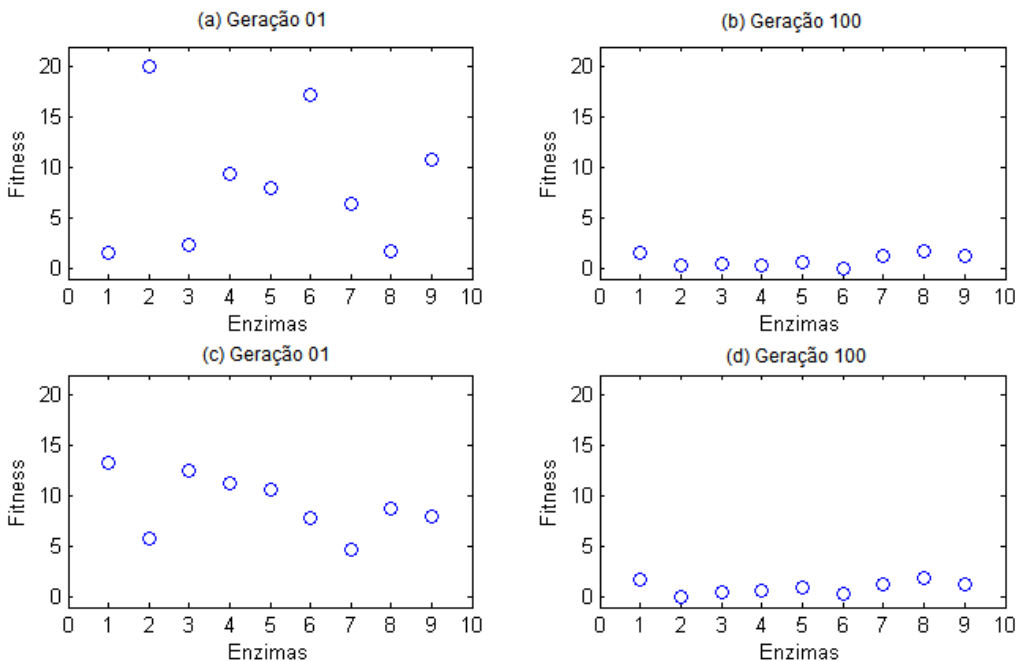


Figura 4.11: Comportamento do GASS para as enzimas 1R0R e 1TEC.

No segundo teste foram utilizadas as enzimas *Subtilisin-like* (1TEC, 1CSE, 1MEE, 1SBN, e 1R0R) como *templates* para o GASS. Para todos os templates o GASS encontrou corretamente o sítio catalítico esperado, inclusive para as enzimas *Trypsin-like*.

A Figura 4.11 apresenta o comportamento do GASS utilizando as enzimas 1R0R e 1TEC como *templates*. As Figuras 4.11(a) e 4.11(b) mostram a geração 1 e a geração 100 para a enzima 1R0R como *template*, e as Figuras 4.11(c) e 4.11(d) mostram a geração 1 e a geração 100 para a enzima 1TEC. Após 100 gerações, percebe-se a convergência dos valores de *fitness* para próximo de zero.

Mesmo em enzimas com estruturas diferentes, o GASS conseguiu encontrar os sítios catalíticos com precisão, uma vez que ele está interessado apenas nos resíduos do sítio catalítico. No entanto, outros testes com grupos maiores e mais diversificados de enzimas foram necessários para validar esta estratégia e para sugerir melhorias.

Os resultados deste experimento foram divulgados em forma de pôster no X-meeting (2012). O comprovante está no Anexo 5.1.5 (Figura 3).

4.5.4 Como o GASS lida com grandes conjuntos de dados?

Testes anteriores foram realizados com conjuntos pequenos e muito controlados de enzimas, especificamente escolhidos para testar algumas propriedades do GASS. Esta seção compreende experimentos com o CD 3, composto por 1.085 enzimas *Trypsin-like* escolhidas aleatoriamente a partir do PDB, e o CD 4, composto por 23.318 enzimas a partir do banco de dados NCBI-VAST não-redundante (p-value 10e-80).

4.5.4.1 Enzimas *Trypsin-like*

Para este experimento foram utilizadas enzimas *Trypsin-like* por possuírem um sítio catalítico bem conhecido: HIS-ASP-SER. Foram utilizadas nove (9) enzimas LIT (CSA) como *templates*, uma vez que, como já mencionado, as enzimas LIT são mais confiáveis pelo fato de terem sido anotadas manualmente, com base na literatura e experimentos. A Tabela 4.7 apresenta os 9 *templates* utilizados pelo GASS neste experimento.

Tabela 4.7: *Templates* utilizados pelo GASS - CD 3.

Enzimas	Sítios Catalíticos
1A0J	HIS 57 A - ASP 102 A - SER 195 A
1CA0	HIS 57 B - ASP 102 B - SER 195 C
1DDJ	HIS 603 A - ASP 646 A - ALA 741 A
1DS2	HIS 57 E - ASP 102 E - SER 195 E
1HJA	HIS 57 B - ASP 102 B - SER 195 C
1N8O	HIS 57 B - ASP 102 B - SER 195 C
1RTF	HIS 57 B - ASP 102 B - SER 195 B
1SSX	HIS 57 A - ASP 102 A - SER 195 A
2LPR	HIS 57 A - ASP 102 A - SER 195 A

Depois de executar os nove *templates* contra 1.085 enzimas, o GASS encontrou, em média, 899 sítios catalíticos conforme o CSA (82,85%) na primeira posição do ranking. Aumentando o tamanho do ranking para 5, obteve-se 987 sítios catalíticos corretamente

identificados (90,94%). Quando o tamanho do ranking foi de 10, o número de sítios catalíticos foi 1.015 (93,52%).

Uma análise mais detalhada por *template* é apresentada na Tabela 4.8. Perceba que apenas o template 1DDJ não obteve sucesso com ranking de tamanho 1, encontrando apenas 8,57% dos sítios catalíticos conforme o CSA. O valor aumenta significativamente quando o ranking é aumentado para 5 (69,31%) e 10 (87,83%). Isso acontece porque a enzima 1DDJ tem uma mutação conservativa (ALA no lugar da SER), o que aumenta o espaço de busca do GASS, podendo gerar indivíduos com melhores valores de *fitness* em relação ao *template* do que o sítio real.

Tabela 4.8: Resultados por *templates* para o CD 3 considerando apenas uma execução do GASS, com os sítios catalíticos encontrados considerando diferentes tamanhos de ranking.

<i>Templates</i>	Ranking (sítios catalíticos encontrados (%))		
	1	5	10
1A0J	1.002 (92,35)	1.020 (94,01)	1.027 (94,65)
1CA0	1.003 (92,44)	1.020 (94,01)	1.025 (94,47)
1DDJ	93 (8,57)	752 (69,31)	953 (87,83)
1DS2	1.002 (92,35)	1.019 (93,92)	1.026 (94,56)
1HJA	1.003 (92,44)	1.016 (93,64)	1.024 (94,38)
1N80	998 (91,98)	1.010 (93,09)	1.029 (94,84)
1RTF	1.002 (92,35)	1.015 (93,55)	1.023 (94,29)
1SSX	1.004 (92,53)	1.018 (93,82)	1.024 (94,38)
2LPR	995 (91,71)	1.016 (93,64)	1.025 (94,47)

Considerando 30 diferentes execuções do GASS para todas as enzimas, calculou-se a média e o desvio padrão da *fitness*. Por exemplo, para o conjunto de 1.085 enzimas, 67 dos sítios catalíticos encontrados (indivíduos) tiveram um desvio padrão diferente de zero. A Figura 4.12 apresenta o valor de *fitness* médio após 30 execuções do GASS utilizando o *template* 1A0J.

A Figura 4.13 apresenta uma curva CMS para o *template* (1A0J - melhor valor médio de *fitness*) e para o valor médio de *fitness*, considerando todos os 9 *templates*. A escolha desta métrica de avaliação (CMS) se deve ao fato de que nesse teste não há um conjunto com valores *negativos*, como na Seção 4.5.2, para a geração de uma curva ROC. Analisando as curvas desse teste, observa-se que os melhores candidatos a sítio catalítico aparecem principalmente entre os 5 primeiros do ranking.

Uma análise dos sítios catalíticos não encontrados pelo GASS também foi realizada, e identificou-se duas razões:

1. O sítio catalítico não está anotado no CSA;
2. O sítio catalítico está anotado no CSA mas foi encontrado usando o PSI-BLAST, e aparece dividido em cadeias diferentes;

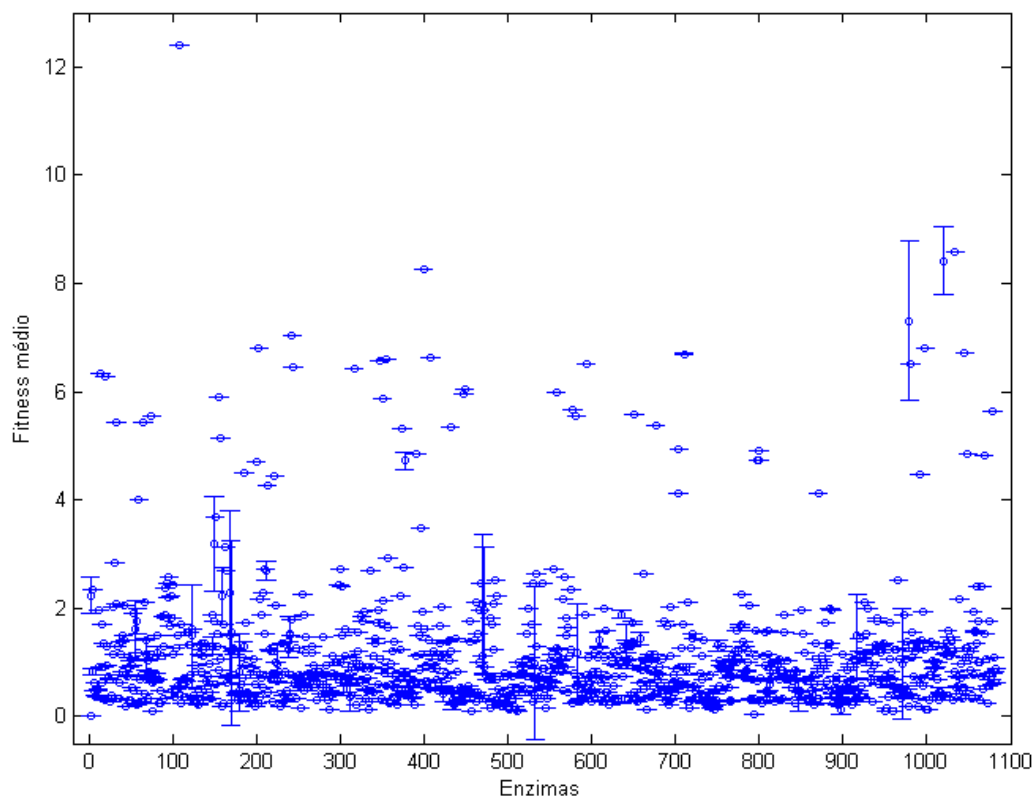


Figura 4.12: *Fitness* médio utilizando *template* 1A0J - CD 3.

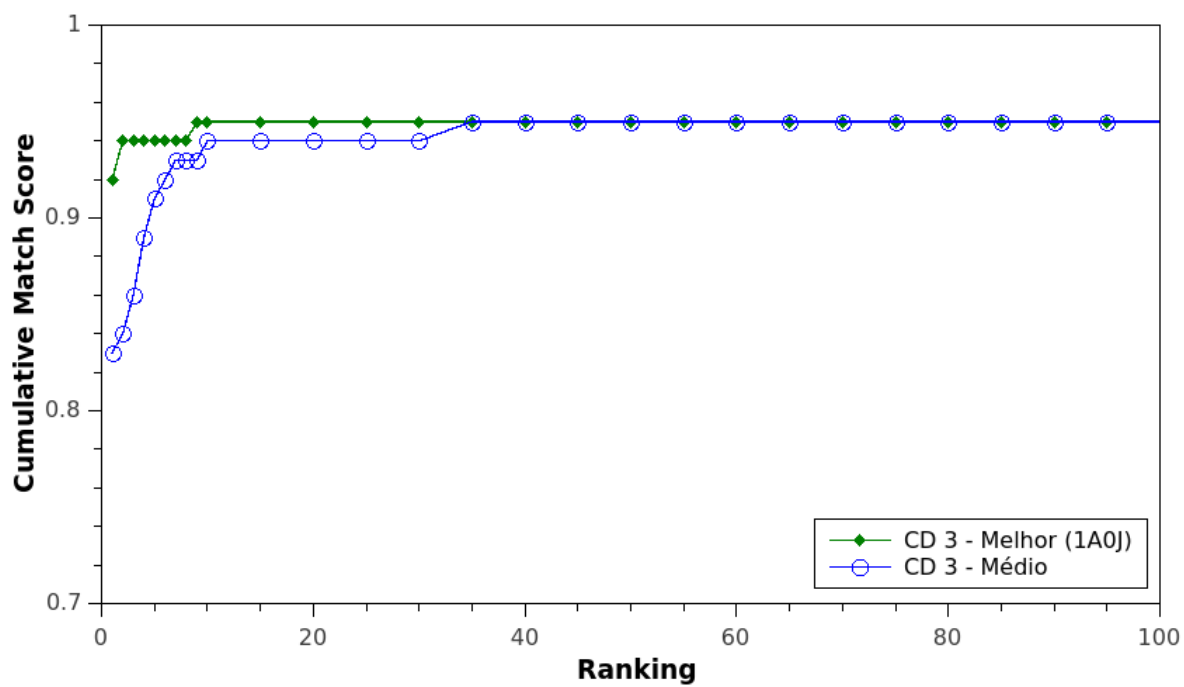


Figura 4.13: CMS do *template* 1A0J e do valor médio dos 9 *templates* (CD 3).

Um exemplo de enzima não anotada no CSA é a 1ARC (*Achromobacter Protease I*), onde o GASS identificou o sítio catalítico HIS 57, ASP 113 e SER 194, que está de acordo com Tsunasawa et al. (1989) (Figuras 4.14 e 4.15). Portanto, esse não pode ser considerado um erro, e sim uma nova enzima que pode ser inserida no CSA.

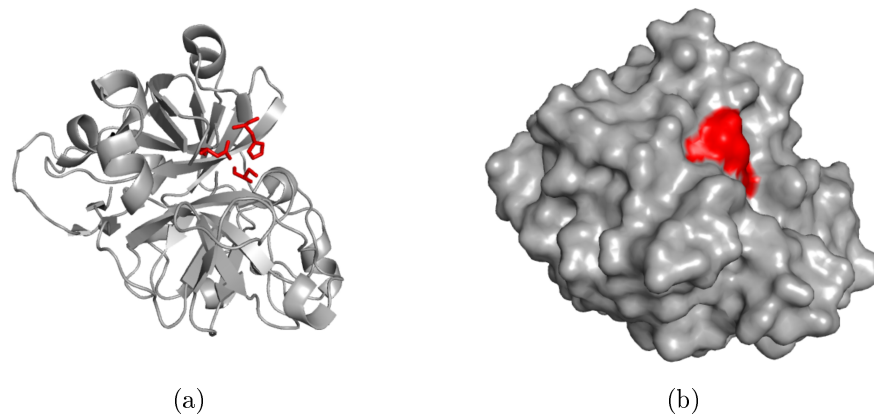


Figura 4.14: Enzima 1ARC: (a) Resíduos encontrados pelo GASS (vermelho) - (b) Localização dos resíduos na superfície (vermelho).

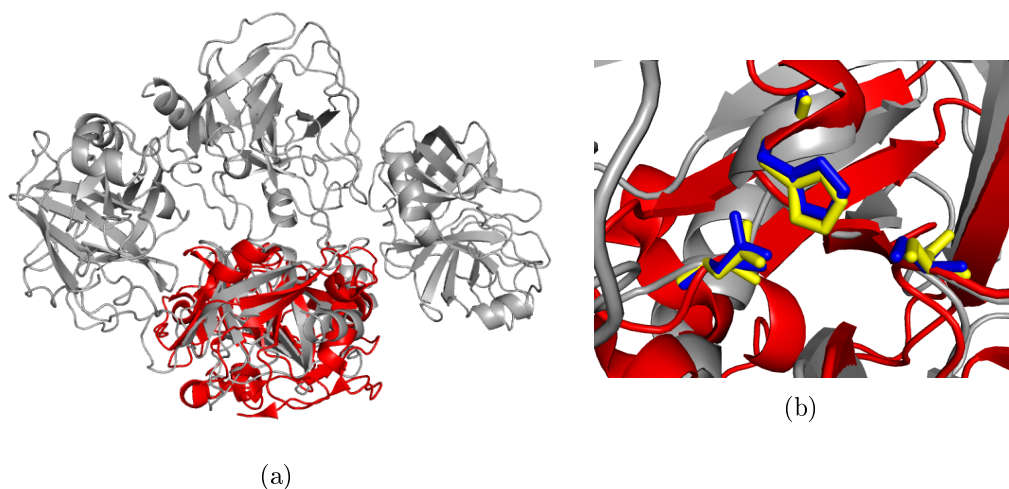


Figura 4.15: Alinhamento estrutural: (a) Enzima 1A0J (cinza) e 1ARC (vermelho). (b) Sítio catalítico da 1A0J (amarelo) e resíduos encontrados pelo GASS (azul).

Já o sítio catalítico da enzima 2GCT (*Gamma-Chymotrypsin A*) está armazenado no CSA como três sítios distintos (HIS 57 e ASP 102 (cadeia B), GLY 193 e 195 SER (cadeia C), SER 195 e GLY 196 (cadeia C)). O GASS encontrou os resíduos HIS 57 e ASP 102 (cadeia B) e SER 195 (cadeia C), como mostrado na Figura 4.16. Este exemplo mostra uma desvantagem de alinhamentos PSI-BLAST feitos pelo CSA, que, neste caso, dividiu o que poderia ser um único sítio em três sítios diferentes. Assim, embora isto possa parecer um erro do GASS, é um problema do CSA quando se lida com sítios catalíticos em cadeias diferentes. Um ponto importante é que o valor de *fitness* apresentado pelo GASS foi de

0,41612, o que significa que a diferença entre as distâncias do sítio ativo template (1A0J) e os aminoácidos encontrados é pequena.

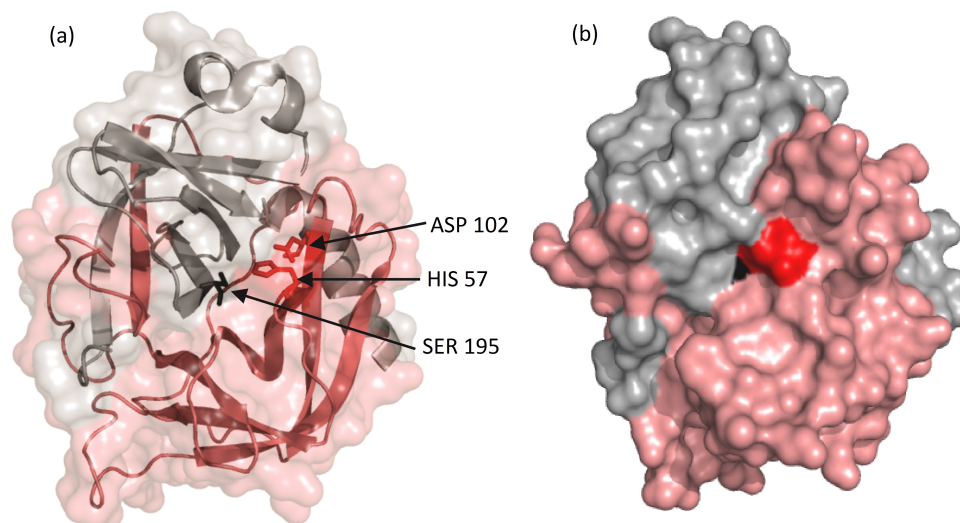


Figura 4.16: Enzima 2GCT: (a) Resíduos encontrados pelo GASS (Cadeia B em vermelho - Cadeia C em cinza). (b) Localização dos resíduos na superfície da enzima (HIS e ASP em vermelho - SER em preto).

Em outra situação, a enzima 2KAI possui duas seções de resíduos anotadas no CSA (versão 2.2.12), uma com SER 195 e GLY 196, e outra com GLY 193 e SER 195 (ambas na cadeia B). O GASS, por sua vez, encontra os resíduos HIS 57 na cadeia A, ASP 102 e SER 195 na cadeia B (Figura 4.17 (a)). No detalhe da superfície (Figura 4.17 (b)) pode-se perceber parte dos resíduos encontrados pelo GASS (HIS 57 em vermelho escuro - ASP 102 em amarelo).

Porém, em uma nova versão do CSA (Furnham et al., 2013), aparecem 21 sítios homologos para a enzima 2KAI. O resíduo HIS 57 (cadeia A) e os resíduos ASP 102 e SER 195 (cadeia B) sempre aparecem em separado (sítios distintos - cadeia A e cadeia B), reforçando mais uma vez o problema do uso do PSI-BLAST pelo CSA. Um ponto interessante está nas enzimas que o PSI-BLAST utilizou como LIT para encontrar os sítios homologos. Entre elas estão as enzimas 1A0J, 1CA0, 1DDJ, 1HJA, 1N8O e 1RTF, que também foram utilizadas pelo GASS como templates (Tabela 4.7).

Os resultados deste experimento mostraram que o GASS poderia melhorar as anotações dos sítios catalíticos do CSA quando estes são feitos através do uso do PSI-BLAST, principalmente quando se tratam de sítios com resíduos em cadeias diferentes.

4.5.4.2 Busca de sítios catalíticos utilizando dados do NCBI VAST

Este experimento foi feito com o mesmo conjunto de dados utilizado pelo ASSAM (CD 4). Os dados foram obtidos via NCBI VAST² de acordo com Nadzirin et al. (2012).

²<http://www.ncbi.nlm.nih.gov/Structure/VAST/nrpdb.html>

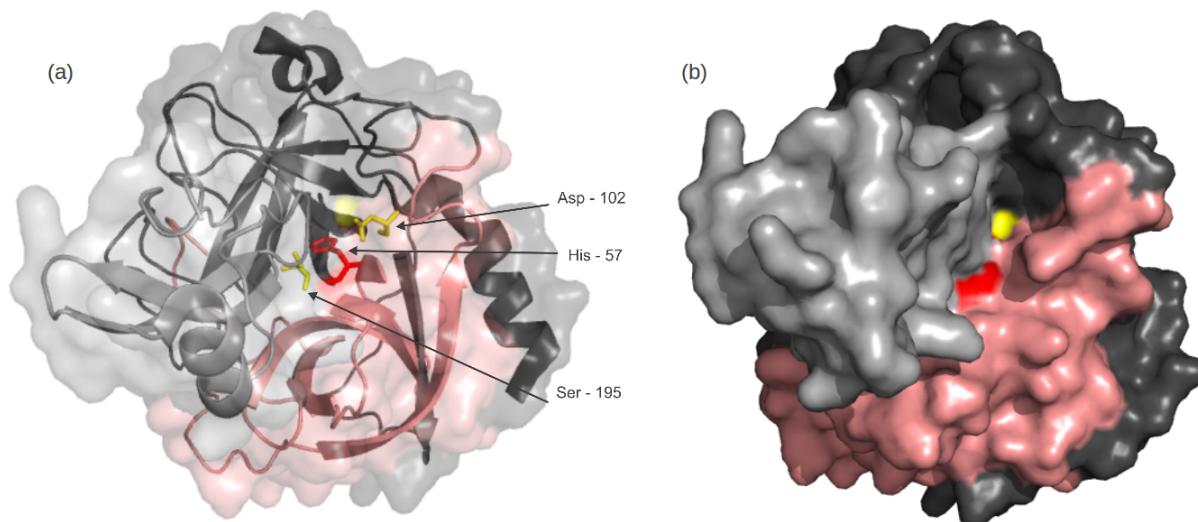


Figura 4.17: Enzima 2KAI: (a) Resíduos encontrados pelo GASS - HIS 57 (vermelho) na cadeia A (em vermelho claro), ASP 102 e SER 195 (amarelo) na cadeia B (cinza escuro). (b) Localização dos resíduos na superfície da enzima (HIS em vermelho escuro - ASP em amarelo).

Inicialmente o conjunto era formado por 28.554 estruturas não redundantes do PDB com P-value de 10×10^{-80} . Após filtros para retirada de arquivos inconsistentes, o conjunto final utilizado foi de 23.318 estruturas. Os parâmetros do GASS estão mencionados na Tabela 4.2. O *template* escolhido para este experimento não foi uma enzima LIT, mas a enzima 1ACB, que tem sua anotação no CSA por meio do PSI-BLAST utilizando as enzimas LIT 1A0J, 1DDJ, 1CA0, 1HJA, 1N8O e 1RTF. A intenção também foi verificar se, um *template* PSI-BLAST poderia obter resultados compatíveis a um *templates* LIT.

Dessa vez, o conjunto de dados é bem maior, envolvendo vários tipos de famílias de enzimas. Por esse motivo, houve uma distribuição maior entre os valores de *fitness*. A Tabela 4.9 apresenta os valores de *fitness* e as faixas em que eles aparecem. Das 23.318 estruturas, 5.430 estruturas (23,29%) tiveram *fitness* entre 1 Å e 5 Å, e 551 (2,36%) tiveram *fitness* menores iguais a 1 Å. Acima de 20 Å foram observadas 2.196 estruturas (9,42%) (Figura 4.18 (a)).

Tabela 4.9: Distribuição dos valores de *fitness* do CD 4.

Valores de <i>fitness</i>	Número de enzimas	Porcentagem
$fitness \leq 1$	551	2,36
$1 < fitness \leq 5$	5.430	23,29
$5 < fitness \leq 10$	8.689	37,26
$10 < fitness \leq 20$	6.452	27,67
$fitness > 20$	2.196	9,42
Total	23.318	100

O número de sítios catalíticos encontrados de acordo com o CSA foi de 353 (1,51% do total de estruturas), e dessas estruturas, 194 (54,96%) apresentaram *fitness* menor

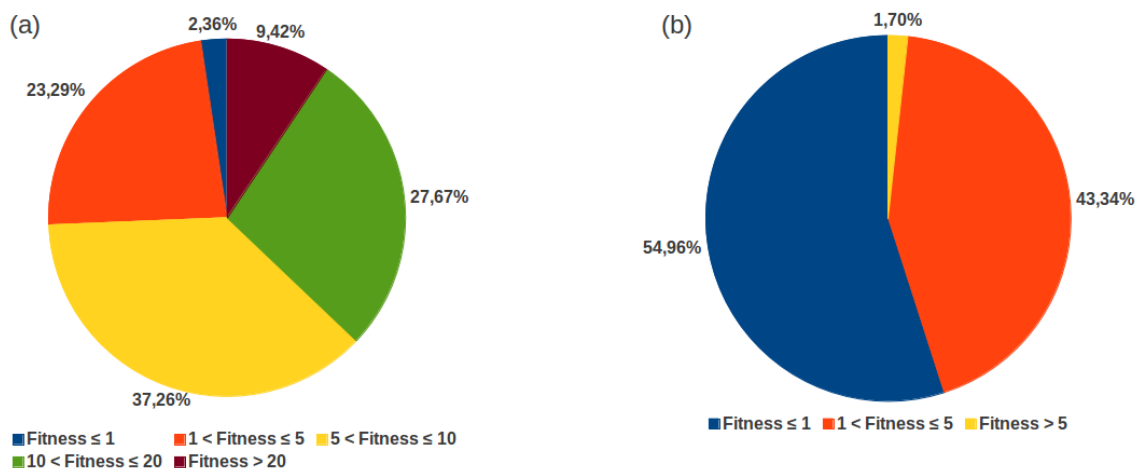


Figura 4.18: (a) Distribuição dos valores de *fitness* - CD 4. (b) Distribuição dos valores de *fitness* para os sítios catalíticos encontrados corretamente conforme o CSA.

que 1 Å, e 153 (43,34%) apresentaram *fitness* entre 1 Å e 5 Å. Assim, 98,30% dos sítios encontrados conforme o CSA possuem valor de *fitness* menor ou igual a 5 Å (Figura 4.18 (b)). Dessa forma, pode-se acreditar que na Figura 4.18 (a), 25% dos sítios encontrados podem estar corretos por terem o *fitness* menor ou igual a 5 Å. Contudo tal validação ainda não foi possível visto que o CSA não traz informações sobre essas enzimas.

Durante os testes anteriores não houve a necessidade em avaliar o desempenho do GASS com relação ao tempo de processamento, dado o tamanho pequeno dos conjuntos de dados utilizados. Já para o CD 4, esse tempo foi observado. Em um computador desktop Intel Core i5 com 4GB de memória RAM, o GASS precisou de 7 minutos para realizar a tarefa com o tamanho do ranking em 1, enquanto que sua implementação no servidor (ver Seção 4.7) rodou em 16 minutos. O ASSAM realiza a mesma tarefa em 25 minutos. Note que o ASSAM calcula o RMSD, no qual a *fitness* do GASS foi inspirada, que também tem tempo quadrático em relação ao número de resíduos. Porém, parte do processamento do ASSAM também é gasto no cálculo dos pseudo-átomos que representam um sítio ativo candidato.

Ao confrontar os resultados fornecidos pelo GASS com os dados cadastrados no CSA, pode-se perceber as mesmas três situações encontradas no CD 3. No caso da enzima 3E4D (*Hydrolase*), a versão 2.2.12 do CSA apresentava apenas os resíduos ASP 223 e HIS 256 em suas respectivas cadeias (A, B, C, D, E e F). Na nova versão do CSA (Furnham et al., 2013), a enzima 3E4D não aparece com seus resíduos catalogados. Já o GASS apresentou os resíduos SER 147, ASP 223 e HIS 256 na cadeia F. O *fitness* para esse conjunto de resíduos foi de 0.0534, o que implica uma alta similaridade com a enzima template. Os resíduos encontrados pelo GASS estão de acordo com van Straaten et al. (2009). A Figura 4.19 apresenta em vermelho os resíduos encontrados pelo GASS para a 3E4D (a) e sua posição na superfície da enzima (b).

Para a enzima 1HIA (*protease/inhibitor*), o CSA (versão 2.2.12) apresenta os resíduos

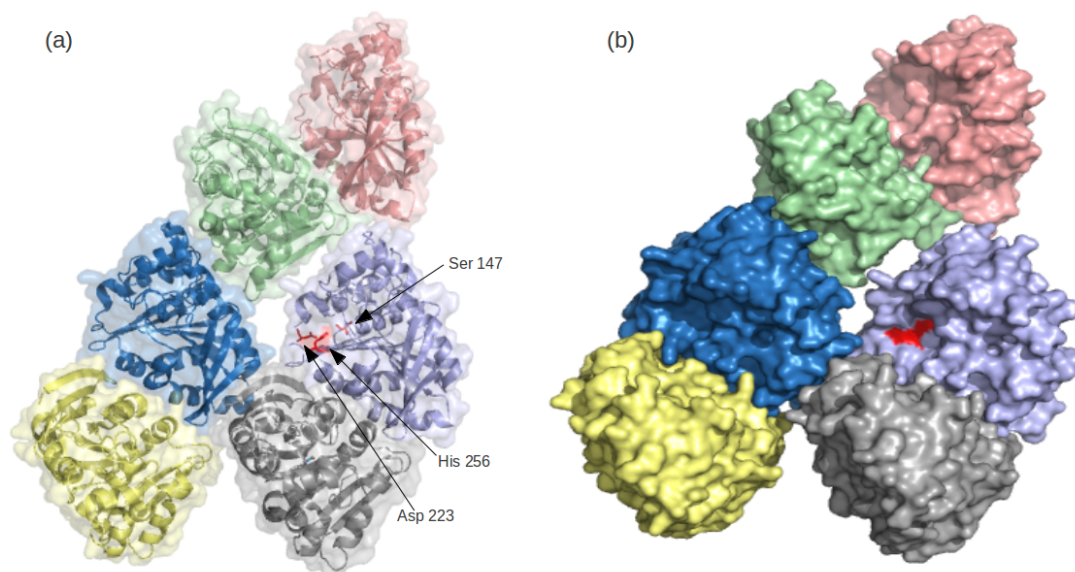


Figura 4.19: Enzima 3E4D: (a) Resíduos encontrados pelo GASS. (b) Localização dos resíduos na superfície da enzima (HIS em vermelho escuro - ASP e SER em amarelo).

SER 195 e GLY 196 para as cadeias Y e B, e GLY 193 e SER 195 para as cadeias Y e B. O GASS apresenta HIS 57 na cadeia A, ASP 102 e SER 195 na cadeia B. O valor de *fitness* para esse conjunto de resíduos foi de 0.104, sugerindo também uma alta similaridade com o sítio template. A Figura 4.20 apresenta em laranja e vermelho os resíduos encontrados pelo GASS para a 1HIA (a). No detalhe da superfície (b) pode-se perceber em vermelho parte dos resíduos encontrados pelo GASS.

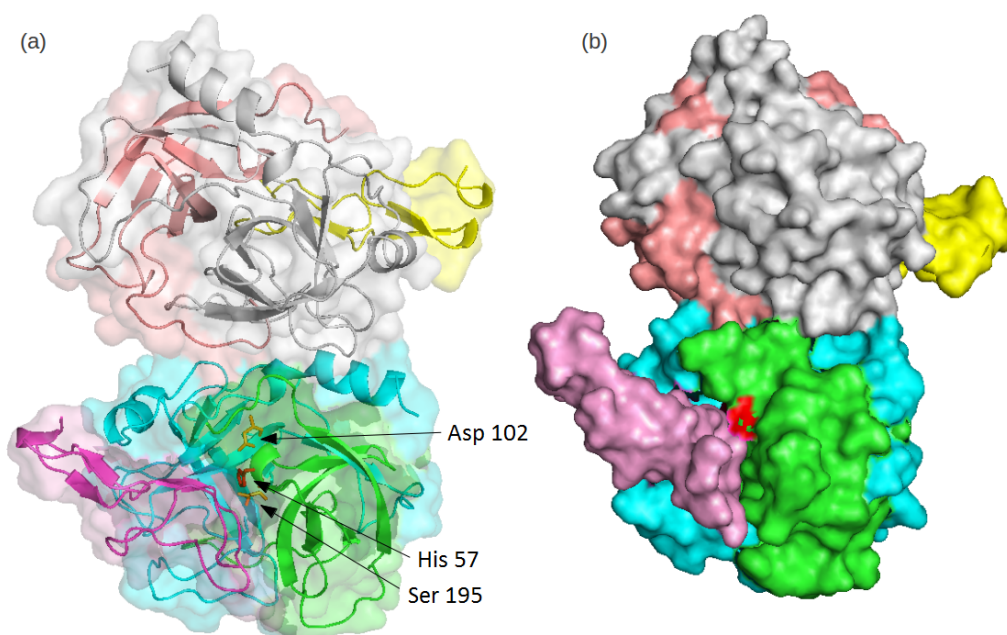


Figura 4.20: Enzima 1HIA: (a) Resíduos encontrados pelo GASS (HIS 57 A em vermelho - ASP 102 B e SER 195 B em laranja). (b) Localização dos resíduos na superfície da enzima (vermelho).

Na nova versão do CSA (Furnham et al., 2013), aparecem 42 sítios homólogos para a enzima 1HIA. O resíduo HIS 57 e os resíduos ASP 102 e SER 195 sempre aparecem em separado (em cadeias diferentes - sítios distintos). Como aconteceu com a enzima 2KAI, o CSA (via PSI-BLAST) utilizou as mesmas enzimas LIT que o GASS utilizou como templates para encontrar os sítios homólogos. Entre elas estão as enzimas 1A0J, 1CA0, 1DDJ, 1HJA, 1N8O e 1RTF (Tabela 4.7). Os resíduos encontrados pelo GASS estão de acordo com Mittl et al. (1997).

4.5.5 Como o GASS se compara a outros métodos *estado da arte* para busca de sítios catalíticos?

Esta seção compara o GASS com outros três métodos descritos na literatura para encontrar sítios catalíticos: PINTS, ASSAM e CatSIId, descritos na Seção 2.1.

4.5.5.1 GASS x PINTS

Para testar o PINTS foram utilizados os dados do CD 1. Inicialmente foi feito um teste de sanidade, onde o *template* da enzima 3NOS (CYS 184, ARG 187, TRP 356 e GLU 361) foi utilizado para encontrar ele mesmo na própria enzima. Neste teste, o GASS obteve êxito, enquanto com o PINTS, foram encontrados 3 possíveis conjuntos de resíduos com apenas 3 resíduos em cada conjunto. Como mencionado no Capítulo 2, o PINTS implementa o método de busca em profundidade, e dependendo das características do espaço de busca, a solução pode não ser encontrada (Russell e Norvig, 1995).

Deve-se lembrar que o sítio catalítico procurado possui quatro resíduos, e é verdade que o GASS tem uma vantagem sobre o PINTS, conhecendo de antemão o número de resíduos do sítio catalítico. Porém, o PINTS possui outras duas características que podem contribuir negativamente para sua performance. A primeira é que ele usa informação de conservação de sequência e, se a sequência não alinha perfeitamente com a família, o resultado pode ser influenciado. A segunda é que alguns resíduos que o autor considerou serem não reativos são ignorados pelo método (Russell, 1998).

Em um segundo teste, foram escolhidas três diferentes enzimas para serem utilizadas como templates (Tabela 4.10): 3NOS, 3NSE e 1K2T (*Nitric Oxide Synthase*). Foram utilizadas cada uma destas enzimas como template para as outras (*todos contra todos*). Em nenhuma das seis combinações o PINTS encontrou o sítio catalítico corretamente, embora em quatro testes tenha encontrado três dos quatro resíduos pertencentes ao sítio catalítico correto. Já o GASS encontrou corretamente todos os sítios esperados.

Em um outro momento, as três enzimas que tiveram desvio padrão diferente de zero no experimento do GASS com o CD 1 (1NOC, 1NOS e 2NOS) foram testadas com o PINTS, já que as mesmas geraram dificuldade para o GASS. Neste teste, o PINTS novamente não encontrou o sítio catalítico corretamente. O mau desempenho do PINTS também

Tabela 4.10: Posição dos resíduos dos sítios catalíticos - GASS x PINTS.

Resíduos	3NOS	3NSE	1K2T
CYS	184	186	415
ARG	187	189	418
TRP	356	358	587
GLU	361	363	592

pode ser explicado por sua restrição em relação a distância entre os resíduos, que é de no máximo 12 Å.

4.5.5.2 GASS x ASSAM

Esta seção compara os resultados do GASS e do ASSAM. É importante enfatizar as principais diferenças entre estes dois métodos. O ASSAM representa os resíduos de aminoácidos utilizando pseudo-átomos, enquanto o GASS utiliza o LHA. No que diz respeito a métrica usada para calcular as distâncias a partir do *template*, o ASSAM utiliza RMSD enquanto o GASS utiliza a Equação 3.1. Após a execução, o ASSAM apresenta os 100 sítios catalíticos mais semelhantes ao *template*, ordenados por RMSD, e os *templates* são limitados a 12 resíduos. O GASS não tem limite para o tamanho do *template*, e o limite para o número de sítios catalíticos similares que ele pode encontrar é o próprio tamanho da população.

Para comparar o GASS e o ASSAM, foi utilizado como *template* a estrutura 1A0S (*Salmonella typhimurium* sacarose porina ScrY), relatada em Nadzirin et al. (2012) e o CD 4 (23.318 estruturas - NCBI VAST). O GASS foi executado contra cada uma das enzimas do CD 4, e os resultados ordenados de acordo com os valores de *fitness*.

Entre os 100 resultados reportados pelo ASSAM estão as estruturas 1A0T e 1OH2, apresentadas em Nadzirin et al. (2012), e que são exemplos de *sacarose porina*. O GASS encontrou todos os três sítios catalíticos da estrutura 1A0T (cadeias R, P e Q) nas posições 1, 4 e 7 do ranking com 100 enzimas. Ao perceber a ausência da estrutura 1OH2 nos resultados, verificou-se que a estrutura não estava no CD 4.

Analisando mais de perto os resultados do ASSAM, foi constatado que apenas 23 dos 100 resultados apresentados estão no conjunto de dados original (NCBI VAST). Supõe-se que o ASSAM adicionou mais estruturas no conjunto de dados original utilizando o SPRITE, que é um programa de busca que mantém uma base de dados de resíduos biologicamente relevantes (Nadzirin et al., 2012). No entanto, uma vez que este procedimento não foi documentado, não pode ser replicado.

Como alternativa, foi simulado o que teria acontecido se o GASS tivesse acesso as 100 enzimas apresentadas como resposta pelo ASSAM, analisando especialmente se a ordem relativa das enzimas iria mudar, tendo em conta que os métodos utilizam diferentes métricas de distância. Este não é o ideal, mas é uma maneira de comparar os resultados do GASS com o ASSAM.

Estas 100 enzimas foram fornecidas como dados de entrada para o GASS e a estrutura 1ACB (*Alpha-Chymotrypsin*) utilizada como *template*. Nesse caso, os resultados obtidos pelo ASSAM e pelo GASS foram muito semelhantes. Ambos encontraram os mesmos 79 sítios catalíticos de acordo com CSA, e os 21 restantes foram descartados porque não estavam catalogados no CSA.

No entanto, em oito enzimas, o ASSAM não apresentou a cadeia dos resíduos do sítio catalítico corretamente. Este é o caso das enzimas 1AUJ e 6CHA. O GASS encontrou o sítio catalítico para a estrutura 1AUJ na cadeia A (de acordo com CSA), enquanto o ASSAM não informou a cadeia. Para a estrutura 6CHA, o GASS encontrou o sítio catalítico e as respectivas cadeias para cada resíduo (HIS-57 (B), ASP-102 (B), 195-SER (C)), enquanto o ASSAM localizou os resíduos na cadeia A. Contudo, o arquivo PDB da estrutura 6CHA tem apenas 9 resíduos de aminoácidos na cadeia A, e nenhum deles correspondem a HIS, ASP, ou SER, que são os resíduos que fazem parte do sítio catalítico da 6CHA. Este erro pode ter acontecido porque os resíduos do sítio catalítico estão em diferentes cadeias (B e C).

Outras diferenças interessantes foram observadas nos resultados. Dentro das 100 enzimas que o ASSAM apresenta como resposta, algumas são repetidas. Isso ocorre porque, para enzimas com mais de uma cadeia, onde cada cadeia possui um sítio catalítico, o ASSAM pode retornar um resultado para cada cadeia. Dessa forma, o ASSAM encontra sítios similares ao *template* dentro da mesma enzima, com pequena diferença nos valores de RMSD. Por exemplo, tanto o GASS quanto o ASSAM encontraram corretamente o sítio catalítico para a enzima 2BM2 (conforme CSA), mas o GASS encontrou o sítio na cadeia A e o ASSAM na cadeia D. O CSA apresenta 4 sítios ativos para essa enzima (cadeias A, B, C e D). A questão da cadeia não coincidir está no fato de ambos usarem métricas diferentes. Para o GASS, um sítio catalítico encontrado na cadeia A pode ter um valor de *fitness* melhor do que os demais sítios das outras cadeias na mesma enzima, e para o ASSAM, o melhor sítio pode estar em outra cadeia que não a cadeia A. A Tabela 4.11 apresenta as enzimas em que o GASS e o ASSAM encontraram disparidade nos resultados.

A Figura 4.21 apresenta parte do resultado do ASSAM. Os dados são apresentados em colunas com a descrição da enzima, os resíduos, o alinhamento dos resíduos com suas posições na sequência, a cadeia, e o valor de RMSD. A primeira proteína a aparecer é a utilizada como *template* e aparece com o valor de RMSD zero. As demais enzimas aparecem de acordo com o valor crescente de RMSD.

A Tabela 4.12 apresenta parte do resultado do GASS que utilizou as enzimas encontradas pelo ASSAM. O GASS apresenta o nome da enzima, os resíduos do sítio catalítico e suas respectivas posições na sequência, as posições espaciais (x , y e z) do átomo mais pesado, a cadeia e o valor de *fitness*. Percebe-se a proteína 1ACB com valor de *fitness* zero. A configuração do GASS está descrita na Tabela 4.2.

Em mais um teste com grupos pequenos de enzimas, o GASS demonstra ser eficaz e

Tabela 4.11: Disparidade nos resultados - GASS x ASSAM.

Enzimas(Cadeia)		Descrição
GASS	ASSAM	
4DGJ(A),4DGJ(A),3T62(A), 3CP7(A),3S7K(B).	4DGJ(D),4DGJ(C),3T62(B), 3CP7(B),3S7K(D).	Sítios encontrados em cadeias diferentes (Não constam no CSA).
2F9P(A),2F9N(C),2BM2(A), 3BEU(B),2R3Y(B),2F9N(D), 2F9N(A),2F9P(C),2F9P(A), 2F9F(A),2FWW(D),1TX6(C), 3RP2(B).	2F9P(D),2F9N(B),2BM2(D), 3BEU(A),2R3Y(A),2F9N(C), 2F9N(D),2F9P(A),2F9P(C), 2F9F(B),2FWW(B),1TX6(D), 3RP2(A).	Sítios encontrados em cadeias diferentes (Constam no CSA).
1BJV(A),1AUJ(A),2ULL(A), 2TIO(B),8GCH(FFG),2ALP(A), 1GMH(FFG),6CHA(BBC).	1BJV(-),1AUJ(-),2ULL(-), 2TIO(-),8GCH(-),2ALP(-), 1GMH(-),6CHA(A).	ASSAM não apresentou cadeia ou apresentou cadeia incorreta (Constam no CSA).
-	1FY8(E)	Sítios incorretos (Constam no CSA).



Results of ASSAM search and righthanded superposition for 1acb_motif

→ [Download text version of the ASSAM output](#)

Matches found in 1acb_motif (PDB ID)	Description	Residues	Residue Matches			Heteroatoms Notes in Database hit	RMSD
			Query	Database	Hits		
1acb PDB PDBsum	ALPHA-CHYMOTRYPSIN	H D S	E 57 E 102 E 195	matches matches matches	E 57 E 102 E 195	0.0 A from site CAT 0.0 A from site CAT 0.0 A from site CAT	0.00 A
4dgi PDB PDBsum	ENTEROPEPTIDASE CATALYTIC LIGHT CHAIN	H D S	E 57 E 102 E 195	matches matches matches	D 41 D 92 D 187	none none none	0.09 A
3vpk PDB PDBsum	CATIONIC TRYPSIN	H D S	E 57 E 102 E 195	matches matches matches	A 57 A 102 A 195	4.9 A from C10 GHS A 300 8.1 A from C10 GHS A 300 -1.4 A from C10 GHS A 300s	0.09 A
1pq7 PDB PDBsum	TRYPSIN	H D S	E 57 E 102 E 195	matches matches matches	A 56 A 99 A 195	4.2 A from N BARG 703 6.8 A from N BARG 703 2.2 A from O AARG 703	0.10 A
1jrt PDB PDBsum	TRYPSIN	H D S	E 57 E 102 E 195	matches matches matches	A 57 A 102 A 195	10.0 A from C ACE B 1A 12.6 A from C ACE B 1A 9.1 A from C ACE B 1A	0.10 A

Figura 4.21: Resultado parcial ASSAM - *Template* 1ACB.

robusto. Apesar de utilizar apenas as posições dos átomos mais pesados da cadeia lateral de cada resíduo, o GASS foi capaz de encontrar os sítios ativos da mesma forma que o ASSAM utilizando pseudo-átomos.

4.5.5.3 GASS x CatSid

Por último, o GASS foi comparado com o CatSid. O CatSid difere do GASS nos seguintes pontos:

Tabela 4.12: Resultado parcial do GASS - Template 1ACB.

Enzima	Resíduos	Sequência	X, Y, Z	Cadeia	<i>Fitness</i>
1ACB	HIS	57	5,459 9,297 17,268	E	0
	ASP	102	1,655 8,402 19,615	E	
	SER	195	8,32 8,218 16,781	E	
4DGJ	HIS	41	41,822 80,175 84,048	A	0,360
	ASP	92	40,836 79,715 88,406	A	
	SER	187	40,444 79,657 81,4	A	
3VPK	HIS	57	21,815 -3,233 30,189	A	0,314
	ASP	102	24,68 -6,241 28,244	A	
	SER	195	19,164 -1,501 29,788	A	
1PQ7	HIS	56	7,56 -0,01 -2,923	A	0,125
	ASP	99	7,501 -4,341 -1,626	A	
	SER	195	5,701 2,407 -3,696	A	
1JRT	HIS	57	39,227 23,531 44,106	A	0,245
	ASP	102	36,188 20,297 43,654	A	
	SER	195	39,643 26,699 44,172	A	
3E8L	HIS	55	-12,448 6,629 -32,704	B	1,096
	ASP	99	-11,321 6,055 -38,025	B	
	SER	192	-15,341 7,076 -31,948	B	

- O *template* é representado através das coordenadas do α -carbono, cofatores e/ou ions;
- Ele realiza uma busca otimizada por isomorfismo de subgrafo na primeira fase do método, utilizando um limiar de 1,5 Å para podar subgrafos não promissores;
- Ele utiliza RMSD para medir as distâncias entre enzimas/*template*;
- A segunda fase realiza um procedimento de pontuação logística, que utiliza muito mais informações sobre as enzimas do que o GASS, incluindo descritores físico-químicos.

Para uma comparação justa entre os métodos, os mesmos *templates* e as mesmas enzimas utilizadas por CatSid em sua primeira fase devem ser consideradas. No entanto, como o CatSid não apresenta as informações das operações realizadas em sua primeira fase, foram utilizadas as enzimas e os *templates* considerados na segunda fase.

O CatSid utiliza 1.993 *templates* (LIT - CSA) para buscar os sítios catalíticos em 66 enzimas escolhidas aleatoriamente (CSA - Versão 2.2.12). O GASS utilizou 1.800 *templates* para buscar os sítios catalíticos em 61 enzimas. Esta diferença no número de *templates* e de enzimas é devido à falta de informação (posição do LHA da cadeia lateral) em alguns arquivos do PDB e o fato do GASS não utilizar aminoácidos não-padrão. Os dados utilizados neste experimento pertencem ao conjunto CD 5, mencionado na Tabela 4.3.

No total, dos 182 sítios catalíticos em 61 enzimas, o GASS encontrou 165 sítios corretamente. Os 17 sítios não encontrados pelo GASS pertencem a sete enzimas. A Tabela 4.13 apresenta todas as enzimas utilizadas no experimento .

Tabela 4.13: Experimento GASS x CatSid - CD 5.

Enzimas utilizadas no CD 5	1ADO, 1AJB, 1CIB, 1D4E, 1E2R, 1EP9, 1EUS, 1F3X, 1F49, 1FDV, 1G1Y, 1G87, 1GGF, 1I45, 1IB4, 1IU8, 1JOL, 1K3T, 1KAK, 1KG4, 1KHN, 1KVY, 1L7A, 1NTO, 1NWR, 1RRY, 1RU1, 1TSL, 1WDD, 1XPT, 1XV8, 1XWW, 1YJA, 2A3T, 2AYL, 2AYO, 2C0H, 2CBA, 2CNH, 2EWN, 2EZ9, 2FBP, 2FPT, 2NU8, 2NZE, 2O3Q, 2OTC, 2POV, 2PPY, 2QD4, 2QU9, 2VEG, 2WFP, 2WHR, 2ZJ3, 2ZYD, 3BBF, 3C52, 3CZN, 3DHE, 3EHB, 3FGD, 3GTD, 3IT1.
Enzimas com os sítios catalíticos não encontrados pelo GASS	1G1Y, 1L7A, 1NWR, 2EWN, 2EZ9, 2FBP, 2PPY

Foram identificadas duas situações em que os erros ocorreram. Em cinco enzimas o GASS encontra corretamente os sítios catalíticos, mas não dentro do *top 5* (ranking). Isso acontece pois todos os *templates* para estas enzimas possuem muitas substituições. Estas substituições fazem com que o GASS utilize mais tipos de resíduos na criação dos indivíduos. Isso aumenta o espaço de busca, e contribui para que o GASS gere indivíduos com melhores valores de *fitness* do que o do sítio real em comparação com o *template*.

A enzima 2PPY, por exemplo, cujo sítio é GLU 139 - GLY 147 (cadeia A), tem quatro possíveis substituições na matriz de substituição ($ALA \rightleftharpoons ASP$, $ASP \rightleftharpoons GLY$, $GLU \rightleftharpoons PHE$, $GLU \rightleftharpoons THR$). O GASS encontrou o sítio catalítico corretamente com valor *fitness* de 0,6374 na trigésima quinta posição do ranking. Dentro dos resultados top 5, os indivíduos tinham valores de *fitness* entre 0,0442 e 0,0884. Esta situação foi observada em outras quatro enzimas (1NWR, 2EWN, 2EZ9, 2FBP).

A Figura 4.22 mostra o CMS dos sítios catalíticos encontrados pelo GASS. Utilizando ranking de tamanho 5, o GASS encontrou 165 sítios corretamente, o que corresponde a 90,65% de acerto. Os outros sítios catalíticos foram encontrados a partir da posição 35, devido a um grande número de substituições possíveis.

Outro problema surge por causa de erros do CSA (enzimas 1L7A e 1G1Y). Na Figura 4.23(a), a enzima 1UK7 (sítio catalítico em vermelho) é a entrada LIT no CSA para a enzima 1L7A (b) (sítio catalítico em vermelho e Ser 181 em verde) (CSA - versão 2.2.12). Não há substituições de resíduos anotadas para o sítio catalítico da enzima 1UK7. Comparando o template (c) e o sítio informado pelo CSA (e), percebe-se que o resíduo ALA está muito distante dos outros resíduos. A nova versão do CSA revelou uma nova referência LIT para a enzima 1L7A (a enzima 1ODT) e um novo sítio catalítico (GLN 182 - ASP 269 - SER 181 - HIS 298 (cadeia A)). Ao definir uma possível substituição para

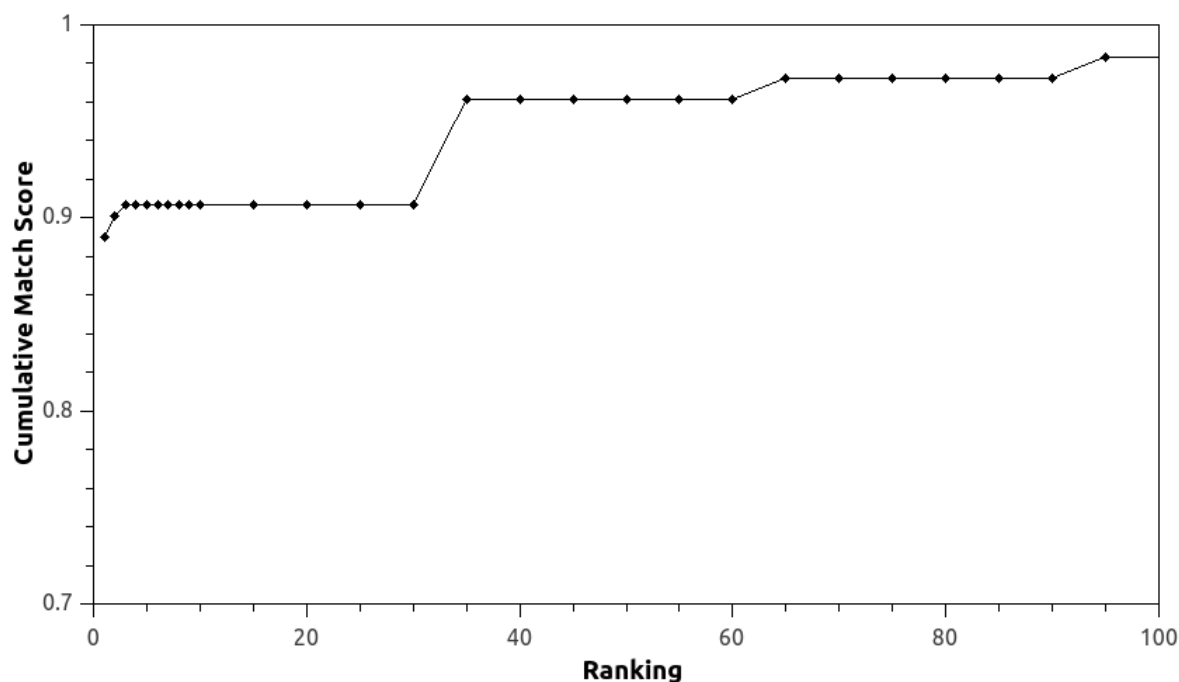


Figura 4.22: CMS do CD 5.

a 1UK7, substituindo ALA por SER (d), o GASS encontrou o sítio com as distâncias entre os resíduos muito semelhante ao *template* original (1UK7) e que faz parte do novo *template* (1ODT). Com o novo *template* (1ODT), o GASS encontra os sítios para a enzima 1L7A corretamente na primeira posição do ranking (*fitness* 1,5428 na cadeia A e 1,65276 em cadeia B). O mesmo aconteceu com a enzima 1G1Y.

Mesmo sem contar com todas as informações que o CatSIId utiliza em sua busca, o GASS conseguiu encontrar os sítios catalíticos corretos com uma taxa de acerto entre 89,01% com ranking de tamanho 1 e 90,65% com ranking de tamanho 5. Os sítios catalíticos encontrados com valores de *fitness* acima da posição 35 no ranking indicam a necessidade de melhorias, que podem ser efetuadas com uma reavaliação da matriz de substituição e com a utilização de outras informações na função de *fitness*.

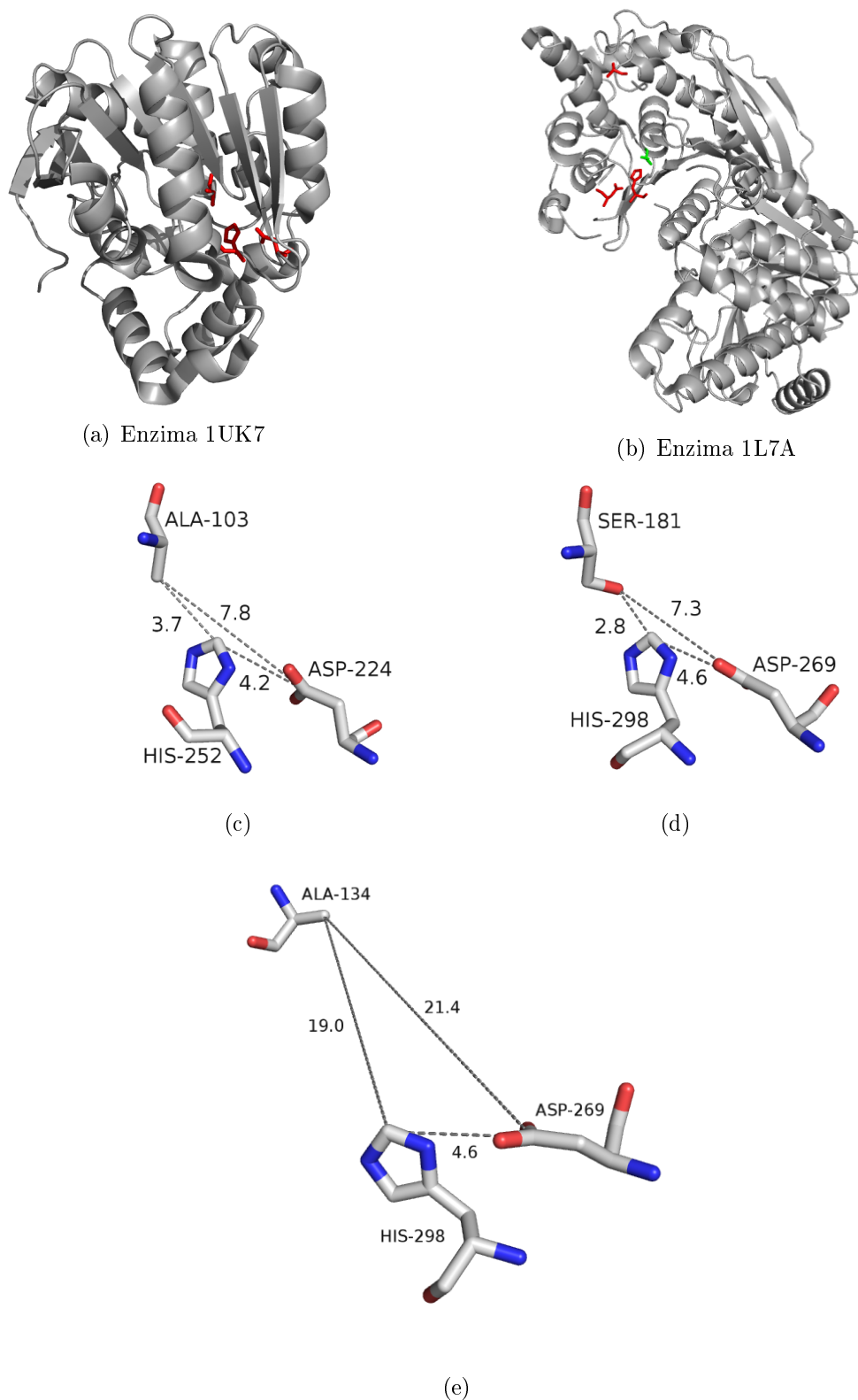


Figura 4.23: Resultados do GASS x sítios anotados no CSA: enzima 1UK7 com sítio catalítico (em vermelho) (a) utilizada como entrada LIT pelo CSA para e enzima 1L7A (sítio catalítico em vermelho - CSA versão 2.2.12 - e SER 181 em verde). *Template* 1UK7 (c), parte do novo sítio catalítico informado pela nova versão do CSA (d), e sítio catalítico informado pelo CSA (versão 2.2.12) (e).

4.5.6 Análise preliminar do impacto da acessibilidade e profundidade na *fitness*

Na Seção 3.2.4 foi apresentado o conceito de ranking utilizado pelo GASS. O objetivo do ranking era habilitar o GASS a apresentar as n melhores soluções encontradas, uma vez que a melhor solução (indivíduo com a menor distância a partir do *template*) poderia estar enterrada na proteína ao invés de estar na superfície ou em um *pocket*.

A motivação para essa mudança tornou-se mais visível quando a mutação conservativa passou a ser considerada (sítios catalíticos - CSA), e o espaço de busca tornou-se maior. Em alguns experimentos notamos um número muito maior de indivíduos com valores de distância muito inferiores a do sítio real em relação ao *template*.

Esse foi o caso, por exemplo, da enzima 2PPY (*Enoyl-CoA Hydratase*)(Seção 4.5.5.3). A enzima possui como sítio catalítico apenas dois resíduos: ASP 204 e GLU 196 (cadeia A), tendo 4 possíveis substituições. Esta situação acaba gerando muitos indivíduos com valores de *fitness* muito próximos ao do *template*. Os cinco melhores indivíduos do GASS para este caso apresentaram valores de *fitness* entre 0,044 e 0,088. Contudo, o sítio catalítico correto (conforme CSA) apresentou valor de *fitness* de 0,637, na trigésima quinta posição do ranking. Apesar da posição disprezível no ranking, o valor de *fitness* apresentado pelo GASS para o sítio correto é muito pequeno, o que demonstra a sensibilidade do método.

Uma alternativa para minimizar este problema é verificar se os resíduos estão enterrados na proteína ou acessíveis ao ligante. Esta informação pode ser adicionada a população do GASS e utilizada no cálculo da *fitness*, penalizando os indivíduos enterrados na proteína e favorecendo os indivíduos da superfície. Isto poderia ajudar o GASS a lidar com o aumento do espaço de busca quando mutações conservativas são consideradas.

As seções seguintes descrevem testes realizados com informações de acessibilidade e profundidade. O objetivo é verificar o impacto das duas abordagens na busca do GASS.

4.5.6.1 GASS utilizando informações de acessibilidade

O GASS foi testado utilizando informações de acessibilidade provenientes do software NACCESS³ (Hubbard e Thornton, 1993). Este software basicamente implementa o método proposto por Lee e Richards (1971), o qual calcula a acessibilidade atômica da superfície de uma proteína rolando uma sonda de tamanho específico ao redor da superfície de Van der Waals. Como resultado do NACCESS, são geradas informações de acessibilidade de cada átomo, das cadeias principal e lateral, e dos resíduos.

Para os testes foram utilizadas as enzimas 1NWR, 2EWN, 2FBP e 2PPY, com os respectivos *templates* (LIT - CSA) (CD 5), uma vez que as mesmas apresentaram dificuldade para o GASS devido ao tamanho do sítio catalítico (entre 2 e 3 resíduos) e ao

³<http://www.bioinf.manchester.ac.uk/naccess/>

número de substituições possíveis. Uma vez que os *templates* também podem apresentar elementos (átomos, resíduos ou cadeias) inacessíveis, a idéia é penalizar os indivíduos com valores de acessibilidade diferentes aos do *template*. Para isso, o valor de acessibilidade do *template* é adicionado ao seu valor de *fitness*, e o mesmo é feito para todos os indivíduos do GASS. Foram testadas quatro estratégias diferentes para incorporar os valores de acessibilidade ao valor de *fitness* (Equação 3.1):

- A** - Diferença entre os valores de acessibilidade do LHA do *template* e do indivíduo GASS;
- B** - Diferença entre os valores médios de acessibilidade da cadeia lateral do *template* e do indivíduo GASS;
- C** - Diferença entre os valores médios de acessibilidade dos resíduos do *template* e do indivíduo GASS;
- D** - A acessibilidade dos resíduos é incorporada de forma binária (1-não acessível/0-acessível);

Nas estratégias A, B e C, o valor da diferença entre as informações de acessibilidade utilizadas é adicionado ao valor final da *fitness*. Na estratégia D, os resíduos são tratados de forma binária (1-não acessível/0-acessível). Quando o valor de acessibilidade de um resíduo do *template* é diferente ao do indivíduo GASS, um valor de penalização é adicionado ao valor final da *fitness*. Testes foram realizados variando este valor de penalização entre 1 e 5 Å. A Tabela 4.14 apresenta as enzimas e os *templates* com os respectivos sítios catalíticos, e as possíveis substituições de resíduos que foram utilizadas nos testes (matriz de substituição).

Tabela 4.14: Enzimas e *templates* utilizados no teste com informações de acessibilidade.

Enzima	<i>Template</i>	Substituições
1NWR ASP 136 A;LEU 140 A	1ITX ASP 200 A;GLU 204 A	ASN,ASP GLU,LEU GLU,MET
2EWN ARG 118 A;LYS 183 A;ARG 317 A	1BIB ARG 118 A;LYS 183 A;ARG 317 A	-
2FBP ASP 68 A;ASP 74 A;GLU 98 A	1EYI ASP 68 A;ASP 74 A;GLU 98 A	ASP,GLU
2PPY GLU 139 A;GLY 147 A	1DCI GLU 196 A;ASP 204 A	ALA,ASP ASP,GLY GLU,PHE GLU,THR

Na Tabela 4.15 são apresentadas as posições do ranking onde o GASS encontrou o sítio catalítico correto para cada uma das enzimas avaliadas. As colunas de A a D correspondem as quatro estratégias utilizadas para incorporar a acessibilidade ao valor

da *fitness*. A coluna *Sem acessibilidade* apresenta as posições do ranking onde o GASS encontrou o sítio correto originalmente (sem a informação de acessibilidade). Em negrito estão os melhores valores de ranking encontrados para as enzimas.

Para cada teste foi considerado um ranking de tamanho máximo em 40. Este tamanho foi escolhido com base nos valores da figura 4.22, onde se percebe um aumento significativo no número de sítios corretos entre o ranking 30 e 40.

O que se esperava com o experimento era encontrar os sítios corretos com valores de ranking menores do que os valores originais (sem acessibilidade). Apesar da melhora significativa para as enzimas 2FBP e 2PPY, a estratégia A apresentou uma piora para a enzima 1NWR e não conseguiu achar o sítio dentro do ranking de 40 para a enzima 2EWN. Embora com a estratégia D todos os sítios tenham sido encontrados abaixo limite de 40, houve piora nos valores para as enzimas 1NWR e 2FBP.

Tabela 4.15: Resultado das estratégias com acessibilidade.

Enzima	Ranking				Sem acessibilidade
	A	B	C	D	
1NWR	37	31	31	35	28
2EWN	>40	>40	>40	22	25
2FBP	20	>40	>40	>40	36
2PPY	7	33	28	25	35

Uma explicação para o desempenho irregular das estratégias utilizando acessibilidade pode estar nos valores apresentados pelo NACCESS. Ao se utilizar a estratégia A, por exemplo, os valores do LHA dos resíduos ASP 68, ASP 74 e GLU 98 do *template* 1EYI foram, respectivamente, 7,688, 0,502 e 0,016. Já para a enzima 2FBP os valores foram 10,177, 0,392 e 0. Neste caso, os valores estão relativamente próximos, o que não acontece com o *template* 1BIB e a enzima 2EWN. Para os resíduos ARG 118, LYS 183 e ARG 317 do *template* 1BIB, os valores foram 8,588, 39,748 e 0,110, enquanto que para a enzima 2EWN os valores foram 0,988, 2,900 e 0,909. Essa diferença também foi observada nas outras estratégias testadas, indicando que a acessibilidade é menos conservada do que as distâncias entre os resíduos do sítio ativo.

Ainda sobre a estratégia que utiliza a acessibilidade de forma binária, outros testes foram feitos para verificar qual o melhor valor a ser utilizado para penalizar os indivíduos com acessibilidade diferente do *template* ou mesmo enterrados na proteína. Assim, cinco testes foram feitos variando a penalização entre 1 a 5 Å. A Tabela 4.16 apresenta as posições do ranking onde o GASS encontrou o sítio catalítico correto para cada uma das enzimas. O aumento na penalização não tem um impacto significativo na enzima 2EWN, e nas enzimas 1NWR e 2PPY o impacto chega a ser desfavorável, aumentando o valor do ranking.

Tabela 4.16: Resultado da estratégia de acessibilidade binária.

Enzima	Ranking				
	1	2	3	4	5
1NWR	35	37	>40	>40	>40
2EWN	22	23	22	22	22
2FBP	>40	>40	38	36	28
2PPY	25	28	30	28	28

4.5.6.2 GASS utilizando informações de profundidade

O GASS também foi testado utilizando informações de profundidade baseadas no software Depth⁴. Segundo Chakravarty e Varadarajan (1999), o parâmetro de acessibilidade não faz distinção entre os átomos logo abaixo da superfície e aqueles no núcleo da proteína. As coordenadas dos átomos obtidas por meio de cristalografia representam uma boa aproximação de suas posições na solução, mas flutuações podem ocorrer e muitos átomos pouco abaixo da superfície podem vir a ter contato com o solvente/ligante. Desta forma, o Depth efetua o cálculo da profundidade de um átomo em uma proteína utilizando a distância deste até uma molécula de água mais próxima na superfície. Como resultado do Depth, são geradas informações de profundidade de cada átomo, das cadeias principal e lateral, e dos resíduos.

As mesmas enzimas do teste de acessibilidade foram utilizadas neste teste (Tabela 4.14). Foram testadas duas estratégias diferentes para incorporar os valores de profundidade ao valor da *fitness* (Equação 3.1):

A - Diferença entre os valores de profundidade do LHA do *template* e do indivíduo GASS;

B - Diferença entre os valores de profundidade dos resíduos do *template* e do indivíduo GASS.

A profundidade da cadeia lateral não foi utilizada como estratégia devido a falta das informações do Depth para vários resíduos das enzimas testadas. Nas estratégias A e B, o valor da diferença entre as informações de profundidade utilizadas é adicionado ao valor final da *fitness*. A Tabela 4.17 apresenta os resultados para as duas estratégias (A e B). A coluna *Sem profundidade* apresenta a posição do ranking onde o GASS encontrou o sítio correto originalmente (sem a informação de profundidade). Em negrito estão os melhores valores de ranking encontrados para as enzimas.

Como no teste de acessibilidade, neste teste foi considerado um ranking de tamanho máximo em 40. Pelos resultados percebe-se que o valor do ranking para a enzima 2FBP piorou em relação ao GASS sem a informação de profundidade. Utilizando a estratégia A (LHA), a enzima 2PPY conseguiu uma melhora significativa no ranking, e com a utilização estratégia B (profundidade do resíduo), as enzimas 1NWR e 2PPY foram as

⁴<http://mspc.bii.a-star.edu.sg/tankp/intro.html>

Tabela 4.17: Resultado das estratégias com profundidade.

Enzima	Ranking		
	A	B	Sem profundidade
1NWR	39	6	28
2EWN	18	39	25
2FBP	>40	>40	36
2PPY	8	5	35

que conseguiram obter uma melhora significativa nos resultados, tendo os sítios corretos encontrados nas posições 6 e 5 do ranking respectivamente.

Levando em consideração todas as estratégias de acessibilidade e profundidade, e mesmo considerando que nenhuma estratégia obteve sucesso em todas as enzimas, a estratégia utilizando informações de profundidade dos resíduos foi a que obteve o melhor resultado, encontrando os sítios corretos em melhores posições do ranking. Além disso, é importante salientar também que as enzimas testadas representam casos complicados devido ao tamanho do sítio (entre 2 e 3 resíduos) e ao número de substituições possíveis.

Testes preliminares utilizando informações de profundidade com os dados do CD1 e CD2 foram realizados e os resultados obtidos foram semelhantes aos dos testes anteriores, sem afetar negativamente o valor de *fitness*. Apesar do resultado promissor, ainda é prematuro agregar informações de profundidade na função de *fitness* de maneira definitiva, sendo necessários mais testes envolvendo outros conjuntos de dados.

4.6 GASS na busca de sítios de ligação

Nesta seção, um conjunto de modelos de sítios de ligação de substratos considerados na competição CASP 10 é utilizado para avaliação do GASS. Como mencionado anteriormente, o objetivo do CASP é avaliar o atual estado da arte dos métodos de predição de estrutura e função, para identificar as limitações e apontar oportunidades para novos desenvolvimentos. A predição de função no CASP é baseada na busca de resíduos em contato com ligantes biologicamente relevantes.

O GASS foi comparado com 17 métodos apresentados na categoria *function prediction* (FN). O conjunto de dados utilizado (CD 6) foi o mesmo proposto no CASP 10, com 13 enzimas e 25 estruturas *templates* para cada enzima. Cada método participante deve utilizar apenas as 25 estruturas para montar o seu conjunto de *templates* de sítio de ligação e buscar em cada uma das 13 enzimas o seu sítio de ligação (alvo).

Segundo Cassarino et al. (2014), um sítio de ligação é definido por todos os resíduos da proteína (estrutura alvo) que possuem pelo menos um átomo (não hidrogênio) dentro de uma certa distância ($d_{i,j}$) dos átomos de ligantes biologicamente relevantes. Na Equação 4.4, $d_{i,j}$ é a distância entre um átomo do resíduo (i) e um átomo do ligante (j), r_i e r_j são os raios de Van der Waals dos átomos envolvidos, e c é uma distância de tolerância

de 0.5 Å.

$$d_{i,j} \leq r_i + r_j + c \quad (4.4)$$

Para a realização do experimento foram definidos os *templates* que o GASS utilizaria para cada sítio de ligação alvo. Os dados foram obtidos a partir das estruturas fornecidas pelo CASP 10 (Tabelas 4.18 e 4.19) de acordo com Cassarino et al. (2014).

A Tabela 4.18 apresenta todos os alvos do experimento CASP 10 (categoria FN), indicando o nome do alvo, o PDB ID (quando existe), o ligante e o tipo. A Tabela 4.19 apresenta todos os alvos do experimento CASP 10.

Tabela 4.18: Targets with biologically relevant ligands used in CASP 10.

Target	PDB	Ligand	Type
T0652	4HG0	AMP	Non-metal
T0657	2LUL	ZN	Metal
T0659	4ESN	ZN	Metal
T0675	2LV2	ZN	Metal
T0686	4HQO	MG	Metal
T0696	-	NA	Metal
T0697	-	LLP	Non-metal
T0706	-	MG	Metal
T0720	4IC1	MN/SF4	Metal
T0721	4FK1	FAD	Non-metal
T0726	4FGM	ZN	Metal
T0737	3TD7	FAD	Non-metal
T0744	2YMV	FNR	Non-metal

O GASS foi executado com a mesma configuração de parâmetros utilizada nos testes contra o ASSAM e CastId (Tabela 4.2), sem o uso da matriz de substituição pois a mesma é específica para sítios catalíticos anotados no CSA.

Para cada alvo, os resultados obtidos foram ordenados de acordo com o seu valor de *fitness*. A predição do GASS foi realizada com base no número do EC da estrutura *template* com o sítio de menor valor de *fitness* encontrado. Também foram calculados, para cada alvo, o MCC, o MCC Z-scores e o BDT (Tabela 1 - Apêndice 5.1.5), bem como a matriz de confusão cumulativa (Tabela 4.20).

Na Tabela 4.21 são apresentados os resultados de todos os métodos participantes do CASP 10 na categoria FN juntamente com o GASS. Os resultados foram ordenados pelo valor médio de MCC sobre todos os alvos avaliados. Para a significância estatística, o teste *Wilcoxon Signed-rank* foi utilizado (Tabela 2 - Apêndice 5.1.5).

Para duas das 13 enzimas alvo (T0659 e T0721), não se conseguiu identificar o ligante nas estruturas *template*, e, portanto, a comparação com o GASS utiliza 11 das 13 enzimas. Obedecendo as regras do CASP 10, os dois alvos que não foram identificados foram considerados como tendo valor de MCC zero. É importante ressaltar ainda que o GASS foi

Tabela 4.19: Definição dos resíduos dos sítios de ligação alvo. Fonte: Cassarino et al. (2014).

Enzimas Alvo	Sítios de ligação (número do resíduo)
T0652	74, 79, 80, 99, 100, 101, 102, 103, 104, 165, 180, 182, 183
T0657	121, 132, 133, 143
T0659	43, 48, 63
T0675	21, 24, 37, 42, 49, 52, 65, 70
T0686	28, 30, 103
T0696	18, 69, 104
T0697	91, 150, 151, 152, 190, 243, 245, 247, 272, 274, 301, 303, 304, 351
T0706	25, 27, 99, 101, 129, 130
T0720	32, 34, 35, 62, 99, 113, 114, 115, 182, 188, 191, 194, 197, 200
T0721	10, 12, 13, 14, 33, 34, 35, 36, 37, 38, 39, 42, 45, 46, 60, 78, 79, 80, 109, 110, 111, 114, 126, 136, 235, 237, 268, 269, 277, 278, 281
T0726	273, 277, 307
T0737	37, 40, 41, 42, 44, 45, 49, 78, 83, 114, 117, 118, 120, 121, 123, 124, 128, 130, 135, 138, 174, 237
T0744	22, 23, 24, 26, 58, 61, 120, 121, 122, 124, 196, 214, 216, 270, 271, 272, 273, 314, 316

originalmente implementado para encontrar sítios catalíticos similares, que, geralmente, possuem menos resíduos que os sítios de ligação.

A Figura 4.24 mostra o desempenho geral do GASS e de todos os grupos participantes do CASP 10. Entre as 11 enzimas alvo consideradas, o GASS encontrou 5 sítios de ligação corretamente, e aparece em quarto lugar no ranking, com valor médio de MCC de 0,63. É importante observar que, dos três métodos com melhor desempenho que o GASS, um é validado por especialistas (humano), enquanto o GASS é completamente automático. Assim, o GASS é o terceiro entre os métodos automáticos (servidor).

De acordo com o CASP 10, as enzimas alvo T0657 e T0659 eram as mais difíceis, com os métodos preditores obtendo as notas mais baixas de MCC para elas. A enzima alvo T0659 não foi considerada nos testes do GASS, uma vez que o seu ligante não foi localizado nos *templates* disponibilizados. Para a enzima alvo T0657, o GASS encontrou o sítio de ligação corretamente.

O GASS poderia ter obtido um desempenho melhor se tivesse conseguido identificar o ligante nas estruturas T0659 e T0721, podendo assim configurar os seus *templates*. Como dito anteriormente, o GASS foi originalmente implementado para encontrar sítios catalíticos, e este experimento é uma extensão da sua aplicação. Assim, uma otimização dos parâmetros do GASS para sítios de ligação também poderia contribuir positivamente para os resultados.

Tabela 4.20: Matriz de confusão cumulativa para todos os grupos no CASP 10, incluindo o GASS.

N	Grupos	TP	FP	FN	TN	TPR	ACC
1	FIRESTAR	106	38	39	3040	0.731	0.976
2	3DLIGANDSITE	78	39	21	2298	0.788	0.975
3	CNIO	117	54	25	2953	0.824	0.975
4	MCGUFFIN	90	27	55	3051	0.621	0.975
5	GASS	50	29	44	2718	0.532	0.974
6	INTFOLD2	95	35	50	3043	0.655	0.974
7	SEOK	92	35	50	2972	0.648	0.973
8	COFACTOR_ HUMAN	103	51	39	2956	0.725	0.971
9	SEOK-SERVER	85	35	57	2972	0.599	0.971
10	SP-ALIGN	110	61	35	3017	0.759	0.970
11	HHPREDA	102	63	35	2948	0.745	0.969
12	3DLIGANDSITE2	90	53	35	2592	0.720	0.968
13	COFACTOR	99	70	46	3008	0.683	0.964
14	FNGUSHAK	93	73	39	2786	0.705	0.963
15	BINDING_ KIHARA	42	29	103	3049	0.290	0.959
16	ATOME2_ CBS	79	69	63	2938	0.556	0.958
17	CONPRED-UCL	82	83	63	2995	0.566	0.955
18	CHUO-BINDING-SITES	117	467	28	2611	0.807	0.846

Tabela 4.21: Resultados do GASS e dos métodos participantes do CASP 10, ordenados pelo valor médio de MCC, calculado sobre todos os alvos avaliados.

N	Método	MCC
1	FIRESTAR	0,715
2	SP-ALIGN	0,707
3	CNIO	0,673
4	GASS	0,630
5	COFACTOR_ HUMAN	0,629
6	SEOK	0,601
7	MCGUFFIN	0,593
8	SEOK-SERVER	0,579
9	HHPREDA	0,576
10	INTFOLD2	0,569
11	COFACTOR	0,564
12	FNGUSHAK	0,486
13	CONPRED-UCL	0,457
14	3DLIGANDSITE2	0,405
15	BINDING_ KIHARA	0,403
16	3DLIGANDSITE	0,379
17	CHUO-BINDING-SITES	0,331
18	ATOME2_ CBS	0,301

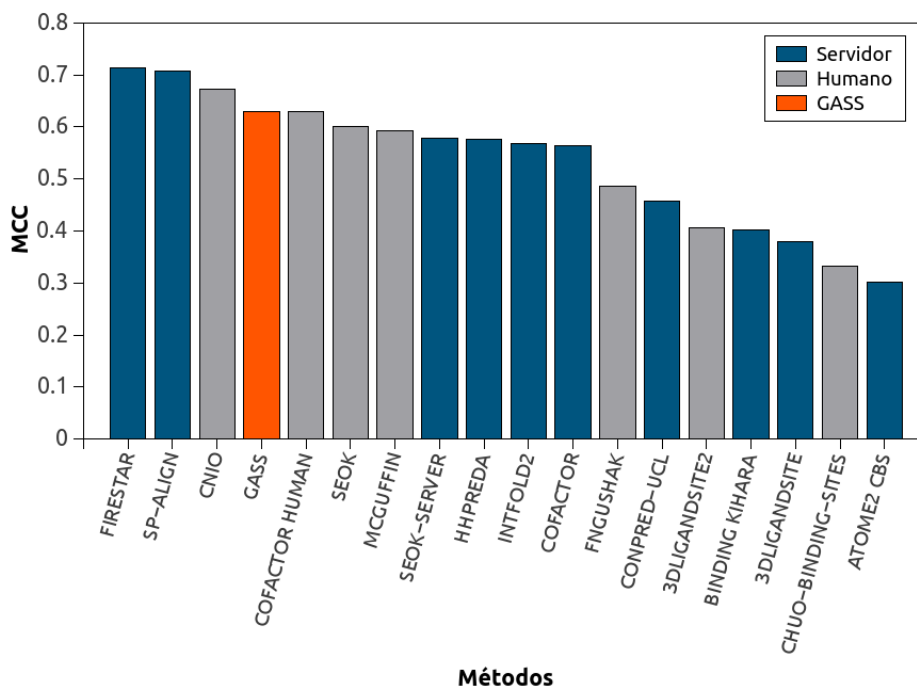


Figura 4.24: Grupos participantes do CASP 10 (categoria FN) classificados em ordem decrescente pelo valor médio de MCC juntamente com GASS. Preditores humanos são mostrados em cinza, preditores baseados em servidores em azul e o GASS em laranja.

4.7 GASS-WEB

Com o intuito de disponibilizar o GASS para a comunidade científica, foi implementado o GASS-WEB, um servidor web com dois recursos disponíveis:

- Busca de sítios catalíticos similares utilizando *templates* LIT do CSA;
- Busca de sítios ativos similares na base NCBI-VAST.

A Figura 4.25 apresenta a página principal do GASS-WEB e as seções seguintes descrevem os recursos disponíveis. O GASS-WEB está disponível no endereço <http://gassweb.dcc.ufmg.br/>.

4.7.1 GASS-WEB utilizando *templates* LIT do CSA

A Figura 4.26 apresenta a opção CSA Templates. Como primeiro passo, o usuário deve fornecer um arquivo PDB ou indicar o nome da proteína desejada, utilizando a nomenclatura PDB (PDB ID). Depois, é necessário que o usuário forneça o número de resíduos que serão considerados na busca, e então executar o GASS.

Os resultados são apresentados em uma tabela (Figura 4.27) ordenados pelo valor de fitness. O GASS-WEB informa o nome da enzima e o *template*, além do EC Number, Uniprot e resolução da estrutura. O usuário ainda pode escolher quantos resultados visualizar (10, 25, 50, 100) e fazer o download deste.

GASS
Genetic Active Site Search

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

Sandro C. Izidoro*, Douglas E. V. Pires*, Raquel C. de Melo-Minardi, Gisele L. Pappa

Abstract

Structure-guided methods have been proposed over the years to infer protein function based on active site similarity. Given an active site template, these methods use different mathematical modelling and searching procedures to match the template to a given set of proteins. Many of the current available methods present, however, limitations such as performing only exact matches on template residues (not accounting for conservative changes), pruning the search space using ad-hoc procedures, besides finding inter-domain active sites. In order to tackle these problems, we have recently proposed GASS (Genetic Active Site Search), a search method based on genetic algorithms that aims to cope with the aforementioned issues. Here we propose a user-friendly web server implementing the method's capabilities, called GASS-WEB.

GASS-WEB can be used under two different scenarios: (a) given a protein of interest, to try to match a set of specific templates (i.e., known active sites), or (b) 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. Based on the Catalytic Site Atlas (CSA) annotation, it was able to correctly identify >90% of the catalogued active sites. It also managed to achieve a MCC of 0.63 on the CASP 10 data set (ranking fourth among 18 methods).

Available Resources

- CSA templates**: Performs active site search using literature-derived and PSI-BLAST templates from CSA, given a PDB file.
- NCBI-VAST database**: Performs active site search using a user-provided template on the NCBI-VAST database.

Genetic Active Site Search

Preprocessing: Filters, Proteins Repository, Databases (Templates and Proteins, CSA, PDB). Genetic Active Site Search: Templates, Proteins, GA, Substitution Matrix. Similar Active Sites.

Figura 4.25: Página principal do GASS-WEB.

GASS
Genetic Active Site Search

Run example

Disclaimer

No PDB files will be retained on the system after being uploaded by the user.

Step 1: Please provide a protein structure (PDB format)

Description

Upload your own PDB file: 1ARC.pdb OR Provide a 4-letter PDB code:

Step 2: Please select a template size

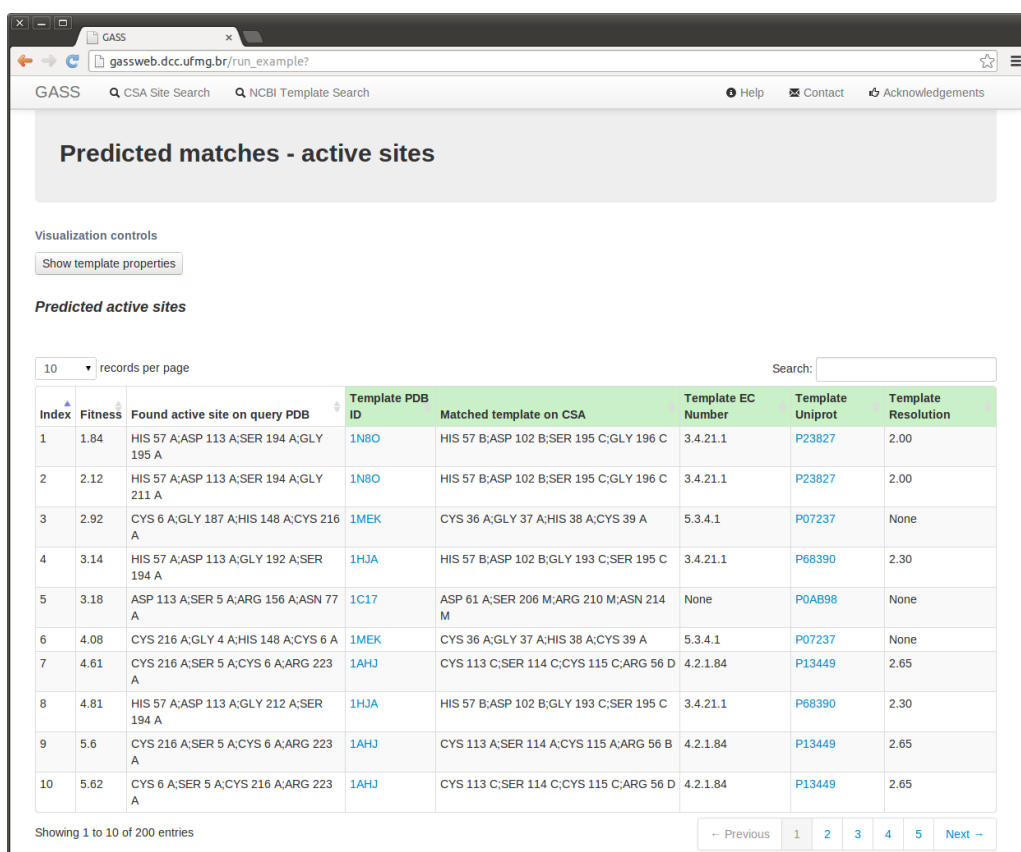
Description

Select template size (number of residues)

UFMG

Best viewed using Chrome on 1280x1024 resolution and above

Figura 4.26: Página GASS-WEB CSA Templates.



Predicted matches - active sites

Visualization controls

Predicted active sites

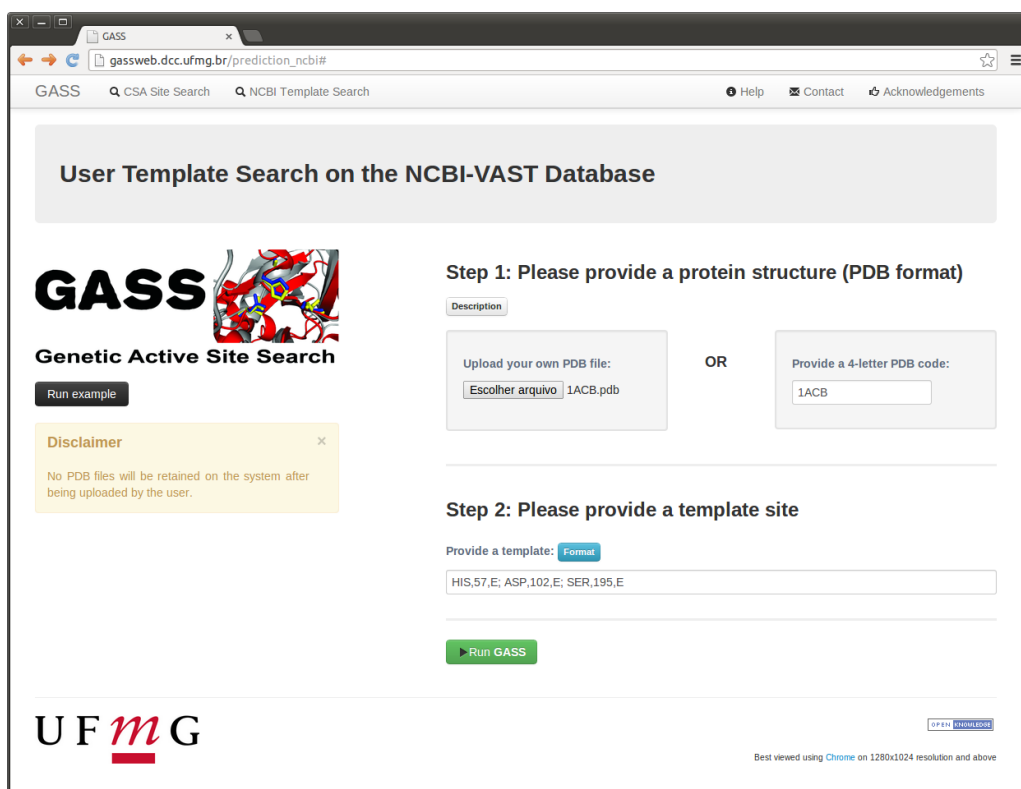
10 records per page Search:

Index	Fitness	Found active site on query PDB	Template PDB ID	Matched template on CSA	Template EC Number	Template Uniprot	Template Resolution
1	1.84	HIS 57 A;ASP 113 A;SER 194 A;GLY 195 A	1N8O	HIS 57 B;ASP 102 B;SER 195 C;GLY 196 C	3.4.21.1	P23827	2.00
2	2.12	HIS 57 A;ASP 113 A;SER 194 A;GLY 211 A	1N8O	HIS 57 B;ASP 102 B;SER 195 C;GLY 196 C	3.4.21.1	P23827	2.00
3	2.92	CYS 6 A;GLY 187 A;HIS 148 A;CYS 216 A	1MEK	CYS 36 A;GLY 37 A;HIS 38 A;CYS 39 A	5.3.4.1	P07237	None
4	3.14	HIS 57 A;ASP 113 A;GLY 192 A;SER 194 A	1HJA	HIS 57 B;ASP 102 B;GLY 193 C;SER 195 C	3.4.21.1	P68390	2.30
5	3.18	ASP 113 A;SER 5 A;ARG 156 A;ASN 77 A	1C17	ASP 61 A;SER 206 M;ARG 210 M;ASN 214 M	None	P0AB98	None
6	4.08	CYS 216 A;GLY 4 A;HIS 148 A;CYS 6 A	1MEK	CYS 36 A;GLY 37 A;HIS 38 A;CYS 39 A	5.3.4.1	P07237	None
7	4.61	CYS 216 A;SER 5 A;CYS 6 A;ARG 223 A	1AHJ	CYS 113 C;SER 114 C;CYS 115 C;ARG 56 D	4.2.1.84	P13449	2.65
8	4.81	HIS 57 A;ASP 113 A;GLY 212 A;SER 194 A	1HJA	HIS 57 B;ASP 102 B;GLY 193 C;SER 195 C	3.4.21.1	P68390	2.30
9	5.6	CYS 216 A;SER 5 A;CYS 6 A;ARG 223 A	1AHJ	CYS 113 A;SER 114 A;CYS 115 A;ARG 56 B	4.2.1.84	P13449	2.65
10	5.62	CYS 6 A;SER 5 A;CYS 216 A;ARG 223 A	1AHJ	CYS 113 C;SER 114 C;CYS 115 C;ARG 56 D	4.2.1.84	P13449	2.65

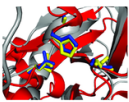
Showing 1 to 10 of 200 entries

← Previous 1 2 3 4 5 Next →

Figura 4.27: Página de resultados do GASS-WEB CSA *Templates*.



User Template Search on the NCBI-VAST Database

GASS  **Genetic Active Site Search**

Disclaimer

No PDB files will be retained on the system after being uploaded by the user.

Step 1: Please provide a protein structure (PDB format)

Description

Upload your own PDB file: 1ACB.pdb


OR

Provide a 4-letter PDB code:

Step 2: Please provide a template site

Provide a template:

HIS,57,E; ASP,102,E; SER,195,E

UFMG 

Best viewed using Chrome on 1280x1024 resolution and above

Figura 4.28: Página GASS-WEB CSA *Templates*.

4.7.2 GASS-WEB utilizando a base NCBI-VAST

A Figura 4.28 apresenta a opção NCBI-VAST database. Para essa opção, o usuário deve fornecer um arquivo PDB ou indicar o nome da proteína desejada, utilizando a nomenclatura PDB (PDB ID). Depois, é necessário fornecer o *template* do sítio ativo para efetuar a busca na base NCBI-VAST. O *template* é fornecido através de uma linha onde as informações dos resíduos (nome, posição na sequência e cadeia) são separados por vírgula, e cada resíduo é separado por um ponto-e-vírgula.

Os resultados são apresentados da mesma forma que foi utilizada pelo CSA *Templates* (Figura 4.27). A ordenação dos resultados, bem como as informações apresentadas seguem o mesmo padrão do CSA *Templates*.

Capítulo 5

Conclusões e Trabalhos Futuros

Neste trabalho foi proposto o *Genetic Active Site Search* (GASS), um novo método de busca de sítios ativos similares baseado em algoritmos genéticos. O método trata das limitações de métodos anteriores, tais como a identificação de sítios ativos em cadeias diferentes, mutações conservativas e restrições quanto ao tamanho dos sítios que podem ser buscados e a distância entre os resíduos.

Ao lidar com uma família específica de enzimas, o GASS conseguiu encontrar todos os sítios catalíticos conforme o CSA. Mesmo quando o sítio procurado possuía uma grande diferença na distância em relação ao *template*, como no caso da enzima 2BHJ e o *template* 3NOS, o GASS demonstrou ser robusto o bastante para encontrar o sítio corretamente.

No experimento utilizado a família NOS (CD 1 - EC: 1.14.13.39) e 126 enzimas escolhidas aleatoriamente a partir do PDB (números EC diferentes de 1.-.-), 85,50% das enzimas da NOS apresentaram distâncias em relação ao *template* menores que 5 Å. Este valor foi de apenas 0,79% nas enzimas escolhidas aleatoriamente. Para este experimento, o GASS apresentou o valor de AUC em 0,97, demonstrando que o método tem potencial para determinar se duas enzimas pertencem a uma mesma família e, conseqüentemente, compartilhar de uma mesma função.

O método também mostrou não ser afetado pela evolução convergente (CD 2). Em grandes conjuntos de dados, o GASS também apresentou bons resultados. No CD 3, composto por 1.085 enzimas *Trypsin-like*, a média de sítios catalíticos encontrados corretamente conforme o CSA foi de 82,85% utilizando somente a primeira posição do ranking. Com um ranking de tamanho 5, esse valor subiu para 90,94%. Mesmo utilizando um sítio *template* com mutação (1DDJ - ALA no lugar da SER), o GASS encontrou 69,31% de sítios conforme o CSA, utilizando um ranking de tamanho 5. Esse valor subiu para 87,83% quando o ranking foi de tamanho 10. Como dito anteriormente, o desempenho do GASS está diretamente ligado a qualidade dos *templates*. O uso da matriz de substituição pode auxiliar o GASS a encontrar sítios similares com mutações conservativas em alguns resíduos. No entanto, dependendo do número de mutações, isso pode aumentar significativamente o espaço de busca, afetando seu desempenho.

Já no CD 4, composto por 23.318 proteínas extraídas do banco de dados NCBI-VAST, foram encontrados 353 sítios conforme o CSA (1,51% do total das estruturas). No entanto, outros sítios encontrados pelo GASS já foram citados na literatura mas não estão catalogados no CSA. Além disso, existe o problema dos sítios com resíduos em diferentes cadeias. O CSA, através do PSI-BLAST, não consegue tratar bem essas situações. Dessa forma, o GASS poderia ser utilizado para ampliar os sítios anotados no CSA e corrigir possíveis erros de anotação.

O GASS também foi comparado com outros métodos para encontrar sítios catalíticos. Na comparação com o PINTS, o GASS se mostrou superior. Em relação ao ASSAM, o GASS mostrou um desempenho equivalente, sendo mais preciso ao tratar sítios catalíticos com resíduos em cadeias diferentes. Nos experimentos, o GASS conseguiu apresentar todas as cadeias corretamente (de acordo com o CSA), enquanto o ASSAM omitiu a cadeia ou informou de maneira incorreta. Por último, no experimento com o CatSid, a comparação direta não foi possível. No entanto, o GASS apresentou uma taxa de acerto de 90,65%, utilizando o ranking de tamanho 5. Além disso, o GASS conseguiu esse resultado sem utilizar todas as informações que o CatSid utiliza, incluindo descritores físico-químicos.

Mesmo não sendo originalmente projetado para lidar com sítios de ligação, o GASS foi testado contra os 17 métodos participantes do CASP 10, na categoria Function Prediction (FN). No experimento, o GASS aparece em quarto lugar geral, com valor médio de MCC de 0,63. Se comparado apenas aos métodos automáticos, o GASS aparece em terceiro lugar. Mesmo com esse bom resultado, ficou claro a necessidade de uma matriz de substituição para os *templates* utilizados, bem como um melhor ajuste nos parâmetros do GASS para sítios de ligação.

Com base nos bons resultados do GASS, foi implementado e disponibilizado o GASS-WEB, um servidor web contemplando tanto a busca de sítios catalíticos similares utilizando *templates* LIT do CSA quanto a busca de sítios ativos similares na base NCBI-VAST.

5.1 Direções de trabalhos futuros

Após análise dos resultados, percebeu-se que algumas melhorias poderiam aprimorar e expandir a metodologia proposta. Entre elas estão a inclusão de mais informações e melhorias na função de *fitness*, introduzindo novas formas de se tratar mutações conservativas, e a ampliação do GASS para lidar com sítios de ligação.

5.1.1 Aprimoramento da *fitness*

A utilização de outras informações na função de *fitness*, que não só a posição do último átomo mais pesado da cadeia lateral, pode aprimorar os resultados. Informações sobre a profundidade dos resíduos foram utilizadas em testes preliminares e os resultados

se mostraram promissores, indicando a necessidade de novos testes para confirmar os resultados obtidos. Outras propriedades físico-químicas, tais como eletronegatividade e índice de hidropatia também deverão ser analisadas e testadas visando tornar a função de *fitness* mais robusta e precisa do que a função atual.

Estas novas informações podem ser incorporadas na função de *fitness* de diversas formas. Uma delas é através da utilização de um método de otimização multiobjetivo, como os baseados em fronteiras de Pareto, ou métodos lexicográficos, onde os objetivos podem ser ordenados de acordo com sua ordem de importância (Coello et al., 2007).

5.1.2 Matriz de substituição

A utilização de uma matriz de substituição para sítios catalíticos fez com que o GASS conseguisse melhores resultados ao lidar com mutações conservativas. No entanto, esta matriz de substituição precisa ser revista e testada novamente, uma vez que, ao longo dos testes, vários erros de anotação foram encontrados no CSA. Desta forma, pretende-se fazer um estudo mais detalhado sobre matrizes de substituição e mutações conservativas. A matriz de substituição já utilizada pelo GASS será analisada e reformulada conforme a nova versão do CSA, visando evitar trocas desnecessárias e incorretas. Além disso, outras matrizes de substituição como a Blosum62 (Henikoff e Henikoff, 1992) e a MIQS (Yamada e Tomii, 2014) serão testadas e, conforme os resultados, incluídas no GASS. A utilização de novas maneiras de se tratar a mutação conservativa podem repercutir também em alterações na função de *fitness*. Todas as alterações que surtirem efeito positivo serão gradualmente adicionadas ao GASS-WEB.

5.1.3 Extensão para sítios de ligação

Outro ponto a ser explorado é a ampliação do GASS para tratar sítios de ligação. Apesar do bom desempenho do GASS no CASP 10, ficou evidente durante os testes a necessidade de se ajustar melhor os seus parâmetros para um melhor desempenho juntos aos sítios de ligação. Novos testes deverão ser realizados para ajustar esses parâmetros e posteriormente anexar mais essa funcionalidade (sítios de ligação) ao GASS-WEB. Nesta direção, a combinação do GASS com métodos de busca local pode trazer benefícios na exploração do espaço de busca maior.

Além disso, um novo estudo será realizado com o objetivo de construir uma biblioteca de *templates* para sítios de ligação. Assim, como acontece com os *templates* do CSA, esta biblioteca seria incorporada pelo GASS-WEB, juntamente com as novas configurações do GASS para sítios de ligação.

5.1.4 Aprimoramentos no CSA

Ao longo deste trabalho, foram identificados vários sítios catalíticos faltantes no CSA e disponíveis na literatura, e vários erros de anotação providos pelo PSI-BLAST. Uma grande contribuição que este trabalho pode oferecer é em relação à disponibilidade de uma base mais ampla e consistente. Para isso, os desenvolvedores do CSA foram contactados para um possível trabalho em conjunto.

5.1.5 DUFs

Um outro direcionamento para este trabalho está em abordar os *Domains of unknown function* (DUFs), que são um grande conjunto de famílias de proteínas sem função conhecida anotadas no Pfam (Bateman et al., 2010). Normalmente a anotação destes domínios dependem de diversas análises experimentais. Pode-se alinhar estruturas DUFs e seus pockets para encontrar resíduos conservados. Estes resíduos podem ser utilizados como *templates* no GASS para buscar estruturas similares no PDB.

Referências Bibliográficas

- Altman, R. B. (1998). A curriculum for bioinformatics: the time is ripe. *Bioinformatics*, 14:549–550.
- Altman, R. B. e Dugan, J. M. (2009). *Structural Bioinformatics*. Wiley-Blackwell, 2 edição.
- Armon, A.; Graur, D. e Ben-Tal, N. (2001). ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J. Mol. Biol.*, 307:447–463.
- Artymiuk, P. J.; Poirrette, A. R.; Grindley, H. M.; Rice, D. W. e Willett, P. (1994). A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures. *J. Mol. Biol.*, 243:327–344.
- Back, T.; Fogel, D. B. e Michalewicz, Z. (1997). *Handbook of Evolutionary Computation*. Oxford University Press.
- Barker, J. A. e Thornton, J. M. (2003). An algorithm for constraint-based structural template matching: application to 3d templates with statistical analysis. *Bioinformatics*, 19(13):1644–1649.
- Bartlett, G. J.; Porter, C. T.; Borkakoti, N. e Thornton, J. M. (2002). Analysis of catalytic residues in enzyme active sites. *J. Mol. Biol.*, 324:105–121.
- Bateman, A.; Coghill, P. e Finn, R. D. (2010). DUFs: families in search of function. *Acta Crystallographica Section F-Structural Biology and Crystallization Communications*, 66(10):1148–1152.
- Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissing, H.; Shindyalov, I. N. e Bourne, P. E. (2000). The protein data bank. *Nucleic Acids Res*, 28:235–242.
- Bernardes, J. S.; Fernandez, J. H. e Vasconcelos, A. T. R. (2008). Structural descriptor database: a new tool for sequence-based functional site prediction. *BMC bioinformatics*.

- Bolle, R.; Connell, J.; Pankanti, S.; Ratha, N. e Senior, A. (2005). The relation between the ROC curve and the CMC. In Martin, DC, editor, *Fourth IEEE Workshop on Automatic Identification Advanced Technologies, Proceedings*, pp. 15–20. IEEE; Univ Buffalo, Ctr Unified Biometr & Sensors; Ultra-Scan; CUBRC.
- Brownlee, J. (2011). *Clever Algorithms: Nature-Inspired Programming Recipes*. LuLu.
- Brylinski, M. e Skolnick, J. (2008). A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *Proceedings of the National Academy of Sciences of the United States of America*, 105(1):129–134.
- Casari, G.; Sander, C. e Valencia, A. (1995). A method to predict functional residues in proteins. *Nat. Struct. Biol.*, 2:171–178.
- Cassarino, T. G.; Bordoli, L. e Schwede, T. (2014). Assessment of ligand binding site predictions in CASP 10. *Proteins*, 82((Suppl 2)):154–163.
- Chakravarty, S. e Varadarajan, R. (1999). Residue depth: a novel parameter for the analysis of protein structure and stability. *Structure with Folding & Design*, 7(7):723–732.
- Coello, C. A. C.; Lamont, G. B. e Veldhuizen, D. A. V. (2007). *Evolutionary Algorithms for Solving Multi-Objective Problems*. Genetic and Evolutionary Computation. Springer Science+Business Media, LLC, 2nd edição.
- Davis, T. E. e Principe, J. C. (1993). A Markov Chain Framework for the Simple Genetic Algorithm. *Evolutionary Computation*, 1(3):269–288.
- de Almeida, V. M. G. (2011). *Hydropace: Uma Metodologia Para Análise de Inibição Cruzada em Serino Proteases Através de Centroides de Regiões Hidrofóbicas*. PhD thesis, UFMG.
- de Magalhães, C. S.; Barbosa, H. J. C. e Dardenne, L. E. (2004). A genetic algorithm for the ligand-protein docking problem. *Genetics and Molecular Biology*, 27:605–610.
- Deaven, D. M. e Ho, K. M. (1995). Molecular Geometry Optimization with a Genetic Algorithm. *Phys. Rev. Lett.*, 75:288–291.
- Dundas, J.; Ouyang, Z.; Tseng, J.; Binkowski, A.; Turpaz, Y. e Liang, J. (2006). CASTp: computer atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic Acids Res.*, 34:116–118.
- Eiben, A. E.; Hinterding, R. e Michalewicz, Z. (1999). Parameter Control in Evolutionary Algorithms. *IEEE Transactions on Evolutionary Computation*, 3(2):124–141.

- Eiben, A. E. e Smith, J. E. (2003). *Introduction to Evolutionary Computing*. Springer Verlag.
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874.
- Fober, T.; Mernberger, M.; Klebe, G. e Hüllermeier, E. (2009). Evolutionary construction of multiple graph alignments for the structural analysis of biomolecules. *Bioinformatics*, 25(16):i2110–i2117.
- Furnham, N.; Holliday, G. L.; de Beer, T. A. P.; Jacobsen, J. O. B.; Pearson, W. R. e Thornton, J. M. (2013). The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes. *Nucleic Acids Research*, pp. 1–5.
- Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison - Wesley.
- Goldenberg, O.; Erez, E.; Nimrod, G. e Ben-Tal, N. (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. *Machine Learning*, 77(1):103–123.
- Hass, J.; Roth, S.; Arnold, K.; Kiefer, F.; Schmidt, T.; Bordoli, L. e Schwede, T. (2013). The Protein Model Portal—a Comprehensive Resource for Protein Structure and Model Information. *Database*, 2013.
- Haupt, R. L. e Haupt, S. E. (2004). *Practical Genetic Algorithms*. Wiley-Interscience, 2 edição.
- Henikoff, S. e Henikoff, J. G. (1992). Amino-acid Substitution Matrices from Protein Blocks. *Proceedings of The National Academy of Sciences of The United States of America*, 89(22):10915–10919.
- Henschel, A.; Winter, C.; Kim, W. K. e Schroeder, M. (2007). Using structural motif descriptors for sequence-based binding site prediction. *BMC bioinformatics*, 8.
- Hubbard, S. J. e Thornton, J. M. (1993). *NACCESS computer program*. Department of Biochemistry and Molecular Biology. University College of London, UK.
- Izidoro, S. C.; de Melo-Minardi, R. C. e Pappa, G. L. (2014). GASS: identifying enzyme active sites with genetic algorithms. *Bioinformatics*.
- Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R. e Taylor, R. (1997). Development and validation of a genetical gorithm for flexible docking. *Journal of Molecular Biology*, 267:727–748.

- Kato, R. B.; Silva, F. T.; Pappa, G. L. e Belchior, J. C. (2015). Genetic algorithms coupled with quantum mechanics for refinement of force fields for RNA simulation: a case study of glycosidic torsions in the canonical ribonucleosides. *Phys. Chem. Chem. Phys.*, 17:2703–2714.
- Kato, T. e Nagano, N. (2010). Metric learning for enzyme active-site search. *Bioinformatics*, 26(21):2698–2704.
- Kernytsky, A. e Rost, B. (2009). Using genetic algorithms to select most predictive protein features. *Proteins-Structure Function and Bioinformatics*, 75(1):75–88.
- Kristensen, D. M.; Chen, B. Y.; Fofanov, V. Y.; Ward, R. M.; Lisewske, A. M.; Kimmel, M.; Kavradi, L. E. e Lichtarge, O. (2006). Recurrent use of evolutionary importance for functional annotation of proteins base on local structural similarity. *Protein Sci.*, 15(6):1530–1536.
- Kristensen, D. M.; Ward, R. M.; Lisewske, A. M.; Erdin, S.; Chen, B. Y.; Fofanov, V. Y.; Kimmel, M.; Kavradi, L. E. e Lichtarge, O. (2008). Prediction of enzyme function based on 3D templates of evolutionary important amino acids. *BMC Bioinformatics*, 9(17):1–7.
- Landgraf, R.; Xenarios, I. e Eisenberg, A. (2001). Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J. Mol. Biol.*, 307(5):1487–1502.
- Laskowski, R. A.; Watson, J. D. e Thornton, J. (2005a). Protein function prediction using local 3D templates. *Journal of Molecular Biology*, 351:614–626.
- Laskowski, R. A.; Watson, J. D. e Thornton, J. M. (2005b). ProFunc: a server for predicting proteins function form 3d structure. *Nucleic Acids Res.*, 33:W89–W93.
- Lee, B. e Richards, F. M. (1971). Interpretation of Protein Structures - Estimation of Static Accessibility. *Journal of Molecular Biology*, 55(3):379–&.
- Lichtarge, O.; Bourne, H. R. e Cohen, F. E. (1996). An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.*, 257(2):342–358.
- Lightstone, F. C.; Wong, S. E.; Kirshner, D. A. e Nilmeier, J. P. (2013). Rapid Catalytic Template Searching as an Enzyme Function Prediction Procedure. *PLoS ONE*, 8(5):1–17.
- Lopez, G.; Maietta, P.; Rodriguez, J. M.; Valencia, A. e Tress, M. L. (2011). firestar-advances in the prediction of functionally important residues. *Nucleic Acids Research*, 39(2):W235–W241.

- Lopez, G.; Valencia, A. e Tress, M. (2007). FireDB - a database of functionally important residues from proteins of known structure. *Nucleic Acids Research*, 35(SI):D219–D223.
- Madabushi, S.; Yao, H.; Marsh, M.; Kristensen, D. M.; Philippi, A.; Sowa, M. E. e Lichtarge, O. (2002). Structural cluster of evolutionary trace residues are statistically significant and common in proteins. *J. Mol. Biol.*, 316(1):139–154.
- Matthews, B. W. (1972). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et biophysica acta*, 405(2):442–451.
- Mittl, P.; DiMarco, S.; Fendrich, G.; Pohlig, G.; Heim, J.; Sommerhoff, C.; Fritz, H.; Priestle, J. e Grutter, M. (1997). A new structural class of serine protease inhibitors revealed by the structure of the hirustasin-kallikrein complex (vol 5, pg 253, 1997). *Structure*, 5(4):585.
- Muhlenbein, H. (1992). How Genetic Algorithms Really Work. 1. Mutation and Hillclimbing. In Manner, R and Manderick, B, editor, *Parallel Problem Solving from Nature*, 2, pp. 15–25. European Commiss, Esprit Basic Res; Natl Fund Sci Res; Parsytec; Free Univ Brussels, Res Council; SIEMENS Nixdorf Belgium. 2nd Conf on Parallel Problem Solving From Nature (PPSN-92), Free Univ Brussels, Brussels, Belgium, Sep 28-30, 1992.
- Nadzirin, N.; Gardiner, E. J.; Willett, P.; Artymiuk, P. J. e Firdaus-Raih, M. (2012). SPRITE and ASSAM: web servers for side chain 3D-motif searching in protein structures. *Nucleic Acids Res.*, 40:W380–W386.
- Najmanovich, R.; Kurbatova, N. e Thornton, J. (2008). Detection of 3D atomic similarities and their use in the discrimination of small molecule protein-binding sites. *Bioinformatics*, 26(16):i105–i111.
- NC-IUBMB (1999). Nomenclature committee of the international union of biochemistry and molecular biology (NC-IUBMB), Enzyme Supplement 5 (1999). *Eur J Biochem*, 264(2):610–50.
- Ortiz, A.; Strauss, C. e Olmea, O. (2002). MAMMOTH (Matching molecular models obtained from theory): An automated method for model comparison. *Protein Science*, 11(11):2606–2621.
- Pazos, F.; Rausell, A. e Valencia, A. (2006). Phylogeny-independent detection of functional residues. *Bioinformatics*, 22(12):1440–1448.
- Porter, C. T.; Bartlett, G. J. e Thornton, J. M. (2004). The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res.*, 32:D129–D133.

- Pupko, T.; Bell, R. E.; Mayrose, Y.; Glaser, F. e Ben-Tal, N. (2002). Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants with their homologues. *Bioinformatics*, 18(Supplement 1):S71–S77.
- Roberts, R. J. (2004). Identifying Protein Function - A Call for Community Action. *PLoS Biology*, 2:293–294.
- Roche, D. B.; Tetchner, S. J. e McGuffin, L. J. (2010). The binding site distance test score: a robust method for the assessment of predicted protein binding sites. *Bioinformatics*, 26(22):2920–2921.
- Roy, A. e Zhang, Y. (2012). Recognizing Protein-Ligand Binding Sites by Global Structural Alignment and Local Geometry Refinement. *Structure*, 20(6):987–997.
- Russell, R. B. (1998). Detection of Protein Three-dimensional Side-chain Patterns: New Examples of Convergent Evolution. *J.Mol. Biol.*, 279:1211–1227.
- Russell, S. J. e Norvig, P. (1995). *Artificial Intelligence: a modern approach*. Prentice-Hall.
- Soding, J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics*, 21(9):2144.
- Sol, A. D.; Pazos, F. e Valencia, A. (2003). Automatic methods for predicting functionally important residues. *J. Mol. Biol.*, 326(4):1289–1302.
- Stark, A. e Russell, R. B. (2003). Annotation in three dimensions. PINTS: Patterns in Non-homologous Tertiary Structures. *Nucleic Acids Res.*, 31(13):3341–3344.
- SUZUKI, J. (1995). A Markov-chain Analysis on Simple Genetic Algorithms. *IEEE Transactions on Systems Man and Cybernetics*, 25(4):655–659.
- Szustakowski, J. D. e Weng, Z. (2000). Protein Structure Alignment Using a Genetic Algorithm. *Proteins: Structure, Function, and Genetics*, 38:428–440.
- Torrance, J. W. e Thornton, J. M. (2009). *Structure-based Prediction of Enzymes and Their Active Sites*. Wiley.
- Tsunasawa, S.; Masaki, T.; Hirose, M.; Soejima, M. e Sakiyama, F. (1989). The Primary Structure and Structural Characteristics of *Achromobacter lyticus* Protease I, a Lysine-specific Serine Protease. *The Journal of Biological Chemistry*, 264(7):3832–3839.
- Unger, R. (2004). The Genetic Algorithm Approach to Protein Structure Prediction. *Structure and Bonding*, 110:153–175.

- van Straaten, K. E.; Gonzalez, C. F.; Valladares, R. B.; Xu, X.; Savchenko, A. V. e Sanders, D. A. R. (2009). The structure of a putative S-formylglutathione hydrolase from *Agrobacterium tumefaciens*. *Protein Science*, 18(10):2196–2202.
- Wallace, A. C.; Borkakoti, N. e Thornton, J. M. (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.
- Ward, R. M.; Venner, E.; Daines, B.; Murray, S.; Erdin, S.; Kistensen, D. M. e Lichtarge, O. (2009). Evolutionary trace annotation server: Automated enzyme function prediction in protein structures using 3D templates. *Bioinformatics*, 25(11):1426–1427.
- Wass, M. N.; Kelley, L. A. e Sternberg, M. J. E. (2010). 3DLigandSite: predicting ligand-binding sites using similar structures. *Nucleic Acids Research*, 38(2):W469–W473.
- Watson, J. D.; Laskowski, R. A. e Thornton, J. M. (2005). Predicting protein function from sequence and structural data. *Current Opinion in Structural Biology*, 15:275–284.
- Whisstock, J. C. e Lesk, A. M. (2003). Prediction of protein function from protein sequence and structure. *Q. Rev. Biophys*, 36:307–340.
- Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6):80–83.
- Wild, D. J. e Willett, P. (1996). Similarity Searching in Files of Three-Dimensional Chemical Structures. Alignment of Molecular Electrostatic Potential Fields with a Genetic Algorithm. *J. Chem. Inf. Comput. Sci.*, 36 (2):159–167.
- Yamada, K. e Tomii, K. (2014). Revisiting amino acid substitution matrices for identifying distantly related proteins. *Bioinformatics*, 30(3):317–325.
- Zemla, A. (2003). LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Research*, 31(13):3370–3374.
- Zhang, Y. e Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, 33(7):2302–2309.
- Zvelebil, M. e Baum, J. O. (2008). *Understanding Bioinformatics*. Garland Science.

Informações adicionais

CASP 10

Tabela 1: Pontuação para cada método executado no CASP 10, incluindo o GASS. TP:Verdadeiro Positivo, FP:Falso Positivo, FN:Falso Negativo, TN:Verdadeiro Negativo, MCC:Matthews Correlation Coefficient, BDT:Binding-site Distance Test.

Alvo	N	Grupo	TP	FP	FN	TN	MCC	Z-score	BDT score
T0652	1	SEOK-SERVER	11	0	2	219	0.916	1.414	0.910
	2	SEOK	11	0	2	219	0.916	1.414	0.910
	3	INTFOLD2	12	3	1	216	0.850	0.714	0.830
	4	MCGUFFIN	12	3	1	216	0.850	0.714	0.830
	5	ATOME2_CBS	13	5	0	214	0.840	0.602	0.720
	6	FIRESTAR	12	4	1	215	0.821	0.400	0.770
	7	3DLIGANDSITE2	12	4	1	215	0.821	0.400	0.770
	8	HHPREDA	12	4	1	215	0.821	0.400	0.770
	9	SP-ALIGN	11	3	2	216	0.804	0.215	0.840
	10	COFACTOR_HUMAN	11	4	2	215	0.774	-0.104	0.790
	11	COFACTOR	11	4	2	215	0.774	-0.104	0.790
	12	CNIO	12	8	1	211	0.726	-0.618	0.620
	13	CONPRED-UCL	11	7	2	212	0.700	-0.902	0.650
	14	GASS	8	2	5	217	0.687	-1.046	0.662
	15	CHUO-BINDING-SITES	9	4	4	215	0.674	-1.180	0.810
	16	BINDING_KIHARA	8	6	5	213	0.568	-2.320	0.650
T0657	1	BINDING_KIHARA	4	0	0	150	1	1.949	1
	2	GASS	4	0	0	150	1	1.949	1
	3	SP-ALIGN	4	1	0	149	0.891	1.668	0.800
	4	CNIO	4	14	0	136	0.449	0.524	0.220
	5	COFACTOR_HUMAN	4	15	0	135	0.435	0.489	0.210
	6	COFACTOR	4	15	0	135	0.435	0.489	0.210
	7	FIRESTAR	4	16	0	134	0.423	0.457	0.200
	8	CHUO-BINDING-SITES	4	44	0	106	0.243	-0.008	0.080
	9	SEOK-SERVER	0	7	4	143	-0.036	-0.728	0.020
	10	SEOK	0	8	4	142	-0.038	-0.735	0.020
	11	3DLIGANDSITE2	0	9	4	141	-0.041	-0.741	0.010
	12	3DLIGANDSITE	0	9	4	141	-0.041	-0.741	0.010
	13	INTFOLD2	0	9	4	141	-0.041	-0.741	0.010
	14	MCGUFFIN	0	9	4	141	-0.041	-0.741	0.010
	15	FNGUSHAK	0	13	4	137	-0.050	-0.764	0.010
	16	ATOME2_CBS	0	14	4	136	-0.052	-0.769	0.010
	17	HHPREDA	0	15	4	135	-0.054	-0.774	0.010
	18	CONPRED-UCL	0	17	4	133	-0.058	-0.784	0.010
T0659	1	SP-ALIGN	3	3	0	68	0.692	2.054	0.5
	2	HHPREDA	2	1	1	70	0.653	1.909	0.73
	3	FNGUSHAK	1	6	2	65	0.168	0.127	0.18
	4	CONPRED-UCL	0	0	3	71	0.000	-0.489	0

Continua na página seguinte

Tabela 1 – Continuação da página anterior

Alvo	N	Grupo	TP	FP	FN	TN	MCC	Z-score	BDT score
	5	FIRESTAR	0	0	3	71	0.000	-0.489	0
	6	BINDING_ KIHARA	0	0	3	71	0.000	-0.489	0
	7	INTFOLD2	0	0	3	71	0.000	-0.489	0
	8	MCGUFFIN	0	0	3	71	0.000	-0.489	0
	9	CHUO-BINDING-SITES	0	0	3	71	0.000	-0.489	0
	10	COFACTOR	0	4	3	67	-0.049	-0.669	0.06
	11	GASS	0	0	0	0	0.000	-0.489	0
T0675	1	COFACTOR_ HUMAN	8	0	0	67	1	1.044	1
	2	SEOK	8	0	0	67	1	1.044	1
	3	CNIO	8	0	0	67	1	1.044	1
	4	GASS	8	0	0	67	1	1.044	1
	5	BINDING_ KIHARA	8	1	0	66	0.936	0.771	0.890
	6	FIRESTAR	7	0	1	67	0.929	0.740	0.900
	7	SEOK-SERVER	6	1	2	66	0.780	0.108	0.820
	8	INTFOLD2	4	0	4	67	0.687	-0.289	0.510
	9	MCGUFFIN	4	0	4	67	0.687	-0.289	0.510
	10	SP-ALIGN	4	0	4	67	0.687	-0.289	0.510
	11	CONPRED-UCL	6	4	2	63	0.627	-0.545	0.660
	12	FNGUSHAK	3	1	5	66	0.495	-1.108	0.430
	13	COFACTOR	5	6	3	61	0.467	-1.225	0.500
	14	CHUO-BINDING-SITES	7	29	1	38	0.273	-2.051	0.210
	15	ATOME2_ CBS	0	0	8	67	0.000	-3.215	0.000
T0686	1	GASS	3	0	0	251	1	1.421	1
	2	CONPRED-UCL	3	1	0	250	0.864	0.912	0.750
	3	3DLIGANDSITE	3	1	0	250	0.864	0.912	0.750
	4	MCGUFFIN	3	1	0	250	0.864	0.912	0.750
	5	FIRESTAR	3	2	0	249	0.772	0.565	0.600
	6	COFACTOR_ HUMAN	3	2	0	249	0.772	0.565	0.600
	7	INTFOLD2	3	2	0	249	0.772	0.565	0.600
	8	SEOK	3	2	0	249	0.772	0.565	0.600
	9	3DLIGANDSITE2	2	1	1	250	0.663	0.157	0.750
	10	SP-ALIGN	2	1	1	250	0.663	0.157	0.750
	11	HHPREDA	3	4	0	247	0.649	0.107	0.430
	12	SEOK-SERVER	2	2	1	249	0.572	-0.185	0.560
	13	CNIO	2	2	1	249	0.572	-0.185	0.560
	14	FNGUSHAK	3	6	0	245	0.570	-0.189	0.330
	15	COFACTOR	3	14	0	237	0.408	-0.797	0.180
	16	BINDING_ KIHARA	1	5	2	246	0.223	-1.492	0.240
	17	CHUO-BINDING-SITES	3	67	0	184	0.177	-1.663	0.040
	18	ATOME2_ CBS	0	0	3	251	0.000	-2.327	0.000
T0696	1	FIRESTAR	3	0	0	97	1	1.410	1
	2	BINDING_ KIHARA	3	0	0	97	1	1.410	1
	3	GASS	3	0	0	97	1	1.410	1
	4	CNIO	3	1	0	96	0.862	0.937	0.750
	5	SP-ALIGN	2	0	1	97	0.812	0.769	0.700
	6	FNGUSHAK	3	3	0	94	0.696	0.371	0.500
	7	COFACTOR_ HUMAN	2	1	1	96	0.656	0.235	0.700
	8	COFACTOR	2	1	1	96	0.656	0.235	0.700
	9	SEOK-SERVER	2	1	1	96	0.656	0.235	0.700
	10	CONPRED-UCL	3	6	0	91	0.559	-0.097	0.330
	11	HHPREDA	3	11	0	86	0.436	-0.519	0.210
	12	SEOK	2	5	1	92	0.411	-0.603	0.310
	13	3DLIGANDSITE	1	2	2	95	0.313	-0.940	0.450
	14	INTFOLD2	1	2	2	95	0.313	-0.940	0.420
	15	MCGUFFIN	1	2	2	95	0.313	-0.940	0.420
	14	CHUO-BINDING-SITES	3	22	0	75	0.305	-0.967	0.120
	16	ATOME2_ CBS	0	0	3	97	0.000	-2.009	0.000
T0697	1	COFACTOR_ HUMAN	13	1	1	447	0.926	0.796	0.930

Continua na página seguinte

Tabela 1 – *Continuação da página anterior*

Alvo	N	Grupo	TP	FP	FN	TN	MCC	Z-score	BDT score
	2	COFACTOR	13	1	1	447	0.926	0.796	0.930
	3	ATOME2_CBS	12	1	2	447	0.886	0.597	0.890
	4	SEOK-SERVER	11	0	3	448	0.883	0.584	0.830
	5	INTFOLD2	12	2	2	446	0.853	0.432	0.870
	6	MCGUFFIN	12	2	2	446	0.853	0.432	0.870
	7	HHPREDA	11	1	3	447	0.844	0.391	0.820
	8	SEOK	10	0	4	448	0.841	0.376	0.780
	9	FNGUSHAK	13	4	1	444	0.837	0.356	0.770
	10	FIRESTAR	12	3	2	445	0.823	0.283	0.820
	11	3DLIGANDSITE2	12	3	2	445	0.823	0.283	0.830
	12	3DLIGANDSITE	12	3	2	445	0.823	0.283	0.820
	13	SP-ALIGN	12	4	2	444	0.795	0.148	0.780
	14	CNIO	13	6	1	442	0.790	0.122	0.690
	15	GASS	6	5	8	444	0.469	-1.066	0.442
	16	CHUO-BINDING-SITES	14	44	0	404	0.467	-1.476	0.240
	17	BINDING_KIHARA	2	6	12	442	0.170	-2.941	0.260
	18	CONPRED-UCL	1	2	13	446	0.143	-3.075	0.170
T0706	1	HHPREDA	5	1	1	197	0.828	0.831	0.880
	2	COFACTOR	4	0	2	198	0.812	0.771	0.770
	3	SP-ALIGN	4	0	2	198	0.812	0.771	0.770
	4	CONPRED-UCL	4	1	2	197	0.723	0.433	0.770
	5	FIRESTAR	4	1	2	197	0.723	0.433	0.770
	6	SEOK-SERVER	4	1	2	197	0.723	0.433	0.770
	7	SEOK	4	1	2	197	0.723	0.433	0.770
	8	CNIO	4	1	2	197	0.723	0.433	0.770
	9	COFACTOR_HUMAN	5	3	1	195	0.712	0.393	0.670
	10	GASS	3	0	3	198	0.702	0.354	0.500
	11	INTFOLD2	4	2	2	196	0.657	0.183	0.770
	12	MCGUFFIN	4	2	2	196	0.657	0.183	0.770
	13	FNGUSHAK	3	1	3	197	0.603	-0.019	0.570
	14	CHUO-BINDING-SITES	5	23	1	175	0.352	-0.966	0.190
	15	ATOME2_CBS	0	0	6	198	0.000	-2.294	0.000
	16	BINDING_KIHARA	0	3	6	195	-0.021	-2.374	0.020
T0720	1	HHPREDA	9	1	7	185	0.694	1.835	0.620
	2	CNIO	8	0	8	186	0.692	1.829	0.630
	3	SP-ALIGN	5	0	11	186	0.543	1.096	0.430
	4	FIRESTAR	4	0	12	186	0.485	0.808	0.270
	5	GASS	6	4	8	184	0.477	0.770	0.550
	6	SEOK	4	1	12	185	0.425	0.515	0.270
	7	MCGUFFIN	3	3	13	183	0.273	-0.235	0.240
	8	BINDING_KIHARA	2	1	14	185	0.267	-0.262	0.260
	9	FNGUSHAK	4	7	12	179	0.253	-0.333	0.290
	10	CHUO-BINDING-SITES	5	13	11	173	0.230	-0.445	0.310
	11	INTFOLD2	3	5	13	181	0.222	-0.482	0.250
	12	3DLIGANDSITE2	2	2	14	184	0.221	-0.486	0.310
	13	COFACTOR_HUMAN	2	2	14	184	0.221	-0.486	0.190
	14	COFACTOR	2	2	14	184	0.221	-0.486	0.190
	15	SEOK-SERVER	2	2	14	184	0.221	-0.486	0.180
	16	CONPRED-UCL	0	0	16	186	0	-1.575	0.000
	17	ATOME2_CBS	0	0	16	186	0.000	-1.575	0.000
T0721	1	HHPREDA	29	5	2	263	0.880	1.430	0.860
	2	INTFOLD2	25	4	6	264	0.815	0.883	0.850
	3	CNIO	27	8	4	260	0.798	0.733	0.790
	4	FNGUSHAK	26	7	5	261	0.791	0.674	0.810
	5	SP-ALIGN	28	11	3	257	0.780	0.589	0.740
	6	COFACTOR_HUMAN	23	5	8	263	0.757	0.391	0.810
	7	COFACTOR	23	5	8	263	0.757	0.391	0.810
	8	MCGUFFIN	21	3	10	265	0.747	0.312	0.750

Continua na página seguinte

Tabela 1 - *Continuação da página anterior*

Alvo	N	Grupo	TP	FP	FN	TN	MCC	Z-score	BDT score
	9	3DLIGANDSITE	29	16	2	252	0.747	0.305	0.650
	10	FIRESTAR	23	7	8	261	0.726	0.134	0.810
	11	ATOME2_CBS	22	8	9	260	0.690	-0.173	0.780
	12	SEOK	22	8	9	260	0.690	-0.173	0.790
	13	3DLIGANDSITE2	26	16	5	252	0.683	-0.226	0.660
	14	SEOK-SERVER	21	7	10	261	0.681	-0.243	0.770
	15	CONPRED-UCL	20	11	11	257	0.604	-0.893	0.710
	16	CHUO-BINDING-SITES	31	46	0	222	0.577	-1.117	0.400
	17	BINDING_KIHARA	6	2	25	266	0.352	-3.017	0.310
	18	GASS	0	0	0	0	0	0.000	0
T0726	1	FIRESTAR	3	0	0	584	1	1.355	1
	2	SEOK-SERVER	3	0	0	584	1	1.355	1
	3	MCGUFFIN	3	0	0	584	1	1.355	1
	4	GASS	3	0	0	584	1	1.355	1
	5	INTFOLD2	3	2	0	582	0.773	0.571	0.600
	6	SEOK	3	2	0	582	0.773	0.571	0.600
	7	3DLIGANDSITE	3	3	0	581	0.705	0.336	0.500
	8	CONPRED-UCL	3	5	0	579	0.610	0.006	0.380
	9	3DLIGANDSITE2	3	5	0	579	0.610	0.006	0.380
	10	COFACTOR_HUMAN	3	7	0	577	0.544	-0.220	0.300
	11	COFACTOR	3	7	0	577	0.544	-0.220	0.300
	12	HHPREDA	3	8	0	576	0.519	-0.309	0.270
	13	CNIO	3	8	0	576	0.519	-0.309	0.270
	14	FNGUSHAK	3	10	0	574	0.476	-0.456	0.230
	15	SP-ALIGN	3	18	0	566	0.372	-0.816	0.140
	16	ATOME2_CBS	3	21	0	563	0.347	-0.902	0.130
	17	CHUO-BINDING-SITES	3	97	0	487	0.158	-1.556	0.030
	18	BINDING_KIHARA	0	3	3	581	-0.005	-2.120	0.040
T0737	1	3DLIGANDSITE	20	2	2	229	0.900	1.266	0.920
	2	MCGUFFIN	18	1	4	230	0.870	1.023	0.860
	3	INTFOLD2	18	2	4	229	0.845	0.825	0.860
	4	CNIO	18	2	4	229	0.845	0.825	0.860
	5	SEOK-SERVER	16	1	6	230	0.814	0.571	0.790
	6	SEOK	16	1	6	230	0.814	0.571	0.790
	7	3DLIGANDSITE2	17	4	5	227	0.772	0.234	0.830
	8	COFACTOR_HUMAN	17	4	5	227	0.772	0.234	0.830
	9	COFACTOR	17	4	5	227	0.772	0.234	0.830
	10	FIRESTAR	16	3	6	228	0.764	0.171	0.770
	11	ATOME2_CBS	20	10	2	221	0.755	0.098	0.680
	12	HHPREDA	16	4	6	227	0.741	-0.007	0.800
	13	FNGUSHAK	18	9	4	222	0.711	-0.250	0.710
	14	SP-ALIGN	16	7	6	224	0.683	-0.474	0.770
	15	CONPRED-UCL	18	14	4	217	0.642	-0.801	0.580
	16	BINDING_KIHARA	5	0	17	231	0.460	-2.259	0.370
	17	CHUO-BINDING-SITES	19	40	3	191	0.460	-2.259	0.340
	18	GASS	9	13	8	223	0.421	-2.121	0.564
T0744	1	FIRESTAR	15	2	4	306	0.825	1.636	0.830
	2	CNIO	15	4	4	304	0.776	1.326	0.830
	3	FNGUSHAK	16	6	3	302	0.768	1.273	0.750
	4	3DLIGANDSITE2	16	9	3	299	0.716	0.937	0.660
	5	SP-ALIGN	16	13	3	295	0.658	0.571	0.570
	6	INTFOLD2	10	2	9	306	0.647	0.498	0.580
	7	MCGUFFIN	9	1	10	307	0.639	0.450	0.550
	8	3DLIGANDSITE	10	3	9	305	0.619	0.318	0.590
	9	COFACTOR_HUMAN	12	7	7	301	0.609	0.256	0.710
	10	COFACTOR	12	7	7	301	0.609	0.256	0.710
	11	CONPRED-UCL	13	15	6	293	0.531	-0.239	0.510
	12	SEOK	9	7	10	301	0.489	-0.508	0.540

Continua na página seguinte

Tabela 1 - *Continuação da página anterior*

Alvo	N	Grupo	TP	FP	FN	TN	MCC	Z-score	BDT score
13		HHPREDA	9	8	10	300	0.472	-0.618	0.580
14		ATOME2_ CBS	9	10	10	298	0.441	-0.814	0.540
15		GASS	7	5	12	303	0.438	-0.833	0.410
16		CHUO-BINDING-SITES	14	38	5	270	0.392	-1.125	0.290
17		SEOK-SERVER	7	13	12	295	0.318	-1.597	0.430
18		BINDING_ KIHARA	3	2	16	306	0.289	-1.787	0.320

Tabela 2: Teste Wilcoxon signed-rank de todos os grupos no CASP 10, incluindo o GASS.

		FN119	FN326	FN475	FN208	FN473	FN285	FN261	FN430	FN273	FN227	FN082	GASS
Firestar	FN119	NA	0.59	0.56	0.31	0.22	0.17	0.07	0.33	0.07	0.07	0.01	0.32
SP-ALIGN	FN326	0.59	NA	0.79	0.54	0.54	0.5	0.27	0.74	0.5	0.08	0.05	0.54
CNIO	FN475	0.56	0.79	NA	0.45	0.31	0.33	0.22	0.67	0.27	0.09	0.03	0.83
COFACTOR_ human	FN208	0.31	0.54	0.45	NA	0.76	0.91	0.31	0.55	0.62	0.36	0.13	1
Seok	FN473	0.22	0.54	0.31	0.76	NA	0.84	0.91	0.95	0.48	0.74	0.24	0.97
McGuffin	FN285	0.17	0.5	0.33	0.91	0.84	NA	0.93	0.74	0.29	0.59	0.27	0.96
Seok-server	FN261	0.07	0.27	0.22	0.31	0.91	0.93	NA	0.95	0.67	0.82	0.45	0.83
HHPredA	FN430	0.33	0.74	0.67	0.55	0.95	0.74	0.95	NA	0.81	0.95	0.31	0.89
IntFold2	FN273	0.07	0.5	0.27	0.62	0.48	0.29	0.67	0.81	NA	0.64	0.31	0.68
COFACTOR	FN227	0.07	0.08	0.09	0.36	0.74	0.59	0.82	0.95	0.64	NA	0.64	0.64
FNGUSHAK	FN082	0.01	0.05	0.03	0.13	0.24	0.27	0.45	0.31	0.31	0.64	NA	0.34
GASS		0.32	0.54	0.83	1	0.97	0.96	0.83	0.89	0.68	0.64	0.34	NA

Artigo Publicado

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}

¹Advanced Campus at Itabira, Universidade Federal de Itajubá, Itajubá, MG 35903-087, Brazil and ²Department of Computer Science and ³Department 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), physicochemical 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

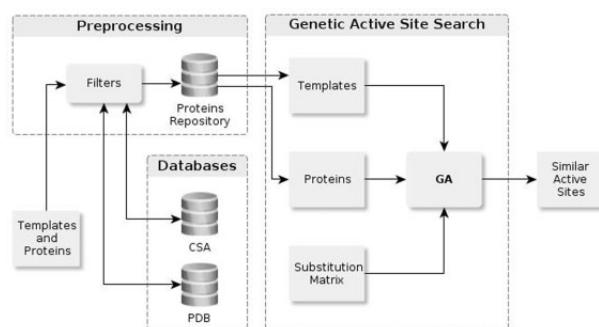


Fig. 1. Proposed methodology for active site matching: data is extracted from PDB and/or CSA, preprocessed and a search template is matched against a set of selected enzymes using GA

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.

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 (\mathbf{v}) and the candidate active site found by GASS (\mathbf{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.

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

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](#).

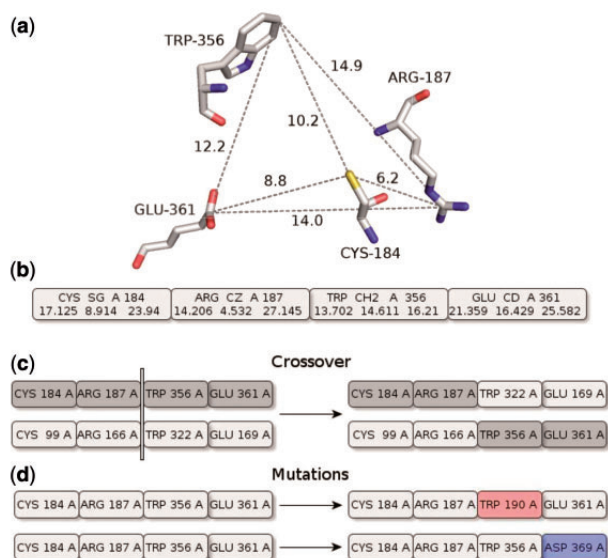


Fig. 2. Representation of the individual and genetic operators: (a) catalytic site of the enzyme 3NOS and distances (in Å) between the LHAs. (b) Representation of 3NOS as a GASS individual. (c) One-point *Crossover* recombines segments of individuals. (d) *Mutation* changes the same (TRP-356 replaced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

2.2 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

Table 1. GASS and CSA results

DS	Enzymes	Templates	Catalytic sites		Match (%)	GASS Rank
			CSA	GASS		
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
	24437	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.

catalytic site; (ii) not all enzymes have their catalytic sites catalogued in CSA. The results are detailed in the following sections.

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 $\leq 5 \text{ \AA}$. 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 \AA , 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

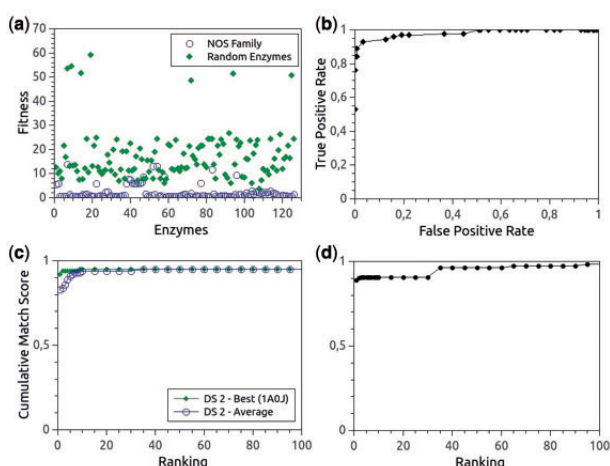


Fig. 3. Results for DS 1, DS 2 and DS 4: (a) fitness of NOS family and random enzymes—each symbol is an enzyme, x corresponds to a catalytic sites found and y to its fitness value (DS 1). (b) ROC curve considering DS 1. (c) CMS of the template 1A0J and average over all nine templates (DS 2). (d) CMS 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 $>10 \text{ \AA}$ for the 126 enzymes in the random set. Only one catalytic site (0.79%) presented template distances smaller than 5 \AA . 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

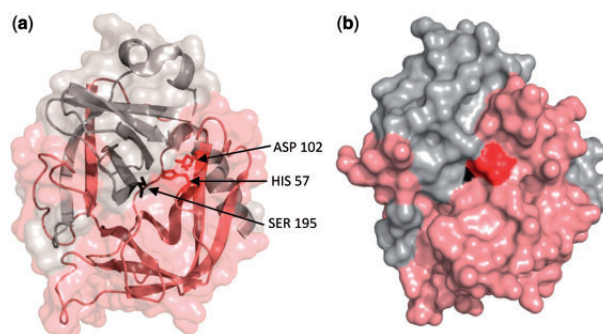


Fig. 4. Enzyme 2GCT: (a) amino acids found by GASS. (b) Location of the amino acids on the surface

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 \times 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 size.

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

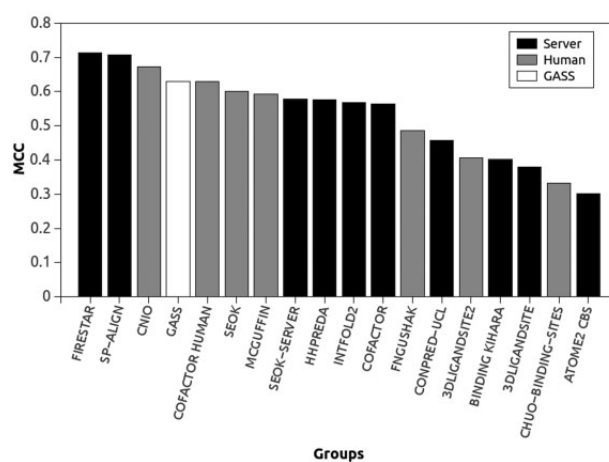


Fig. 5. Groups from CASP 10 (FP category) ranked in decreasing order by average MCC together with GASS. Human predictors are shown in gray, server predictors in black and GASS in white

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.

Comprovantes

X-Meeting 2011



Figura 1: X-Meeting 2011.

3ª Escola Luso-Brasileira de Computação Evolutiva 2012



Figura 2: 3ª Escola Luso-Brasileira de Computação Evolutiva 2012.

X-Meeting 2012

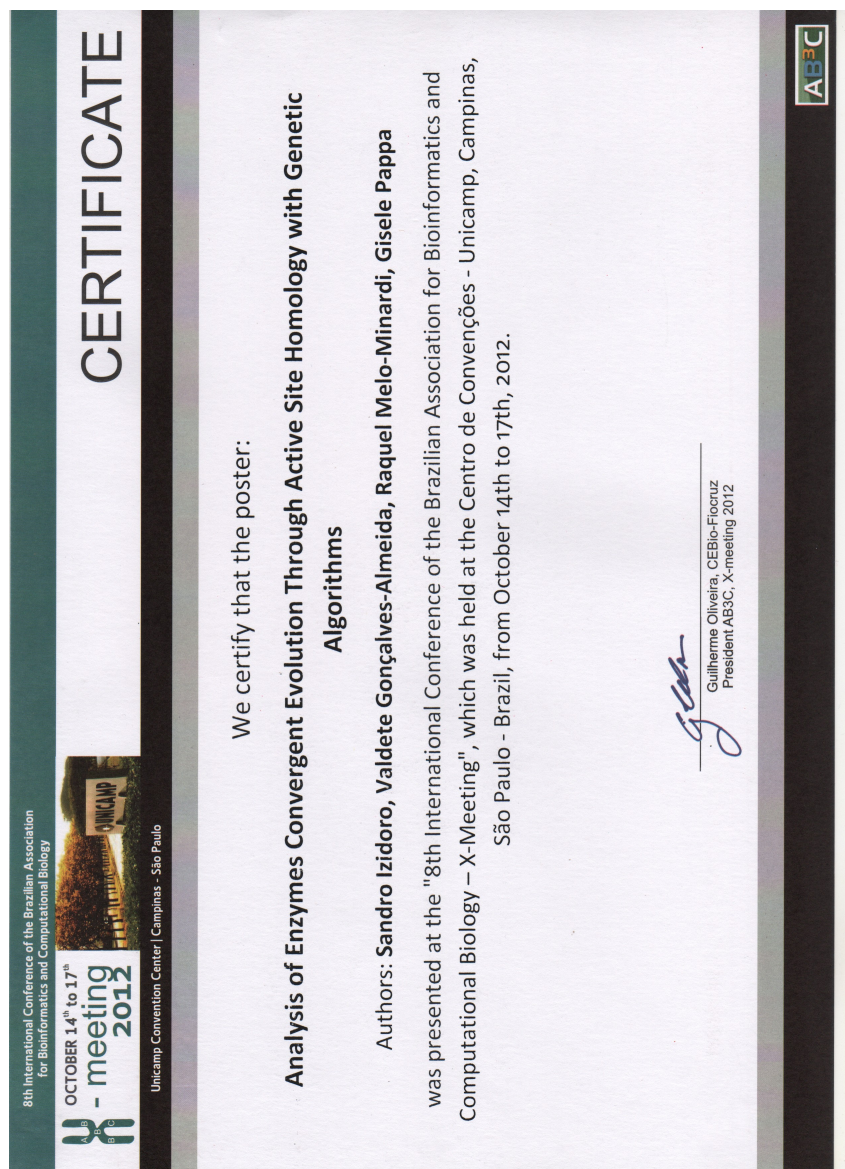


Figura 3: X-Meeting 2012.