UM ESTUDO SOBRE O USO DA GEOMETRIA POLIGONAL DA CADEIA PRINCIPAL COMO UMA ASSINATURA ESTRUTURAL MAIS CONSERVADA QUE O EMPACOTAMENTO DE RESÍDUOS

JOÃO ARTHUR FERREIRA GADELHA CAMPELO

UM ESTUDO SOBRE O USO DA GEOMETRIA POLIGONAL DA CADEIA PRINCIPAL COMO UMA ASSINATURA ESTRUTURAL MAIS CONSERVADA QUE O EMPACOTAMENTO DE RESÍDUOS

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

Orientador: Prof. Dra. Raquel Cardoso de Melo Minardi
Coorientador: Prof. Dra. Sabrina de Azevedo Silveira

Belo Horizonte

Julho de 2017

Ferreira Gadelha Campelo, João Arthur

Um estudo sobre o uso da geometria poligonal da cadeia principal como uma assinatura estrutural mais conservada que o empacotamento de resíduos / João Arthur Ferreira Gadelha Campelo. — Belo Horizonte, 2017
xxvi, 77 f. : il. ; 29cm

Tese (doutorado) — Universidade Federal de Minas Gerais
Orientador: Prof. Dra. Raquel Cardoso de Melo Minardi

1. Computação — Teses. 2. Bioinformática — Teses.
I. Orientador. II. Título.

[Folha de Aprovação]

Quando a secretaria do Curso fornecer esta folha,
ela deve ser digitalizada e armazenada no disco em formato gráfico.

Se você estiver usando o `pdflatex`,
armazene o arquivo preferencialmente em formato PNG
(o formato JPEG é pior neste caso).

Se você estiver usando o `latex` (não o `pdflatex`),
terá que converter o arquivo gráfico para o formato EPS.

Em seguida, acrescente a opção `approval={`*nome do arquivo*`}`
ao comando `\ppgccufmg`.

Se a imagem da folha de aprovação precisar ser ajustada, use:
`approval=[`*ajuste*`][`*escala*`]{`*nome do arquivo*`}`
onde *ajuste* Ãľ uma distância para deslocar a imagem para baixo
e *escala* é um fator de escala para a imagem. Por exemplo:
`approval=[-2cm][0.9]{`*nome do arquivo*`}`
desloca a imagem 2cm para cima e a escala em 90%.

Agradecimentos

Ao Todo, por nos ser, e, a nós, por sermo-lo. À Existência, ao Mistério e ao Universo pela magia da curiosidade, constatação, observação, investigação, pensamento e descoberta.

Aos meus pais, Douglas e Maria Ligia, pela Vida e por tudo que fizeram e fazem por mim. Sou eternamente grato pelo amor, carinho, felicidade e educação que sempre me deram, em todas as fases de minha vida. Vocês são os meus pilares moral e intelectual. Digo o mesmo do meu irmão Douglas. Amor de irmão será sempre a Amizade primeira.

À Luiza Melgaço, Amor de minha vida e minha Esposa Espiritual, pelo Amor Maior. Ao seu apoio incondicional, inquebrantável, paciente e cúmplice. Desculpe-me por todos os momentos de ausência e cansaço. Serei eternamente grato por esse infinito suporte. Nunca me esquecerei dos olhinhos que brilhavam a cada despedida. A caminhada de mãos dadas continua. Sempre.

Aos meus sogros, Celço e Lourdes, por todos os almoços oferecidos em finais de semana de trabalho! Por todas as orações e pedidos cheios de fé em prol dessa empreitada que, à luz da Razão, sempre se apresentou impossível. Por todos os cafezinhos que me impediam de dormir em cima dos livros!

Aos amigos, todos, pela torcida e compreensão de toda ausência.

À minha orientadora, professora Raquel Cardoso de Melo-Minardi, pelo convite para fazer o doutorado em Bioinformática, pela sugestão do tema, orientação em meu trabalho e pela oportunidade ofertada a mim de trabalhar no LBS.

Ao meu orientador (*in memorian*), professor Marcelo Matos Santoro, por toda sua inspiração e conhecimento. Sua intuição e criatividade originaram este trabalho. Exemplo de profissional e ser humano.

À minha orientadora, professora Sabrina Azevedo Silveira, pela enorme paciência em efetuar as correções dos erros dos meus textos (e por rir deles!).

Ao amigo e professor Marcos Augusto dos Santos por ter me apresentado à Bioinformática.

A todos os colegas do Laboratório de Bioinformática e Sistemas (LBS) que contribuiram de alguma forma para a conclusão desta tese: À Valdete Maria Gonçalves de Almeida, pelo carinho e atenção com que fui recebido no LBS e pela inspiração do seu trabalho. Ao Carlos Silveira, por sua capacidade intelectual, conhecimento e paciência para discutir comigo sobre questões teóricas. Ao Pedro Magalhães, Laerte Mateus e Wellisson Rodrigo, pela parceria nas disciplinas do curso, pelas risadas e zoeira no laboratório!

"O mais importante não é a decisão que você toma,
mas o grau de comprometimento que você tem."
(Bo Bartlett)

Resumo

Mesmo que a função não possa ser diretamente inferida puramente a partir do padrão de enovelamento específico adotado por uma determinada proteína, os dados estruturais podem ser usados para detectar proteínas com funções semelhantes. Neste contexto, uma estratégia possível é a definição de assinaturas estruturais, que são conjuntos de características capazes de identificar inequivocamente um tipo de enovelamento proteico.

O objetivo do corrente trabalho é encontrar possíveis padrões estruturais conservados em proteínas de mesma família. Para demonstrar que um modelo que utiliza as informações posicionais em nível atômico é conservado e discriminativo entre famílias, construiu-se classificadores estruturais. A partir deste estudo, foi possível discriminar os átomos do backbone como a melhor assinatura estrutural estudada.

Inserindo pontos intermediários fictícios entre os carbonos alfa, obtêm-se uma poligonal geométrica, similar à poligonal geométrica obtida pela inclusão dos átomos da cadeia principal. Dessa forma, consegue-se um efeito similar à poligonal geométrica da cadeia principal sem a influência do posicionamento dos átomos C e N e, indiretamente, sem a influência dos ângulos phi e psi. O corrente estudo sugere que não são as posições desses átomos os fatores determinantes no incremento da precisão da classificação, mas o fortalecimento do caráter geométrico linear da cadeia principal.

Portanto, a principal contribuição do presente trabalho é a investigação do uso da disposição geométrica poligonal da cadeia principal (ou dos próprios C_α, caso sejam adicionados pontos intermediários fictícios entre esses) como uma assinatura estrutural discriminativa.

Palavras-chave: Bioinformática, Assinatura Estrutural, Classificador Estrutural.

Resumo Estendido

Mesmo que a função não possa ser diretamente inferida puramente a partir do padrão de enovelamento específico adotado por uma determinada proteína, os dados estruturais podem ser usados para detectar proteínas com funções semelhantes cujas sequências divergiram durante a evolução. Neste contexto, uma estratégia possível é a definição de assinaturas estruturais, que são conjuntos de características capazes de identificar inequivocamente um tipo de enovelamento proteico e a natureza das interações que podem estabelecer com outras proteínas e ligantes.

O objetivo do corrente trabalho é encontrar possíveis padrões estruturais conservados em proteínas de mesma família. Almeja-se verificar se tais padrões são capazes de gerar assinaturas estruturais suficientemente discriminativas entre famílias. Almeja-se, também, encontrar ou propor modelos e algoritmos que possibilitem assinalar esses padrões de conservação.

Neste trabalho, investigou-se o uso de alguns conjuntos atômicos como assinaturas estruturais. Esta proposta tem o diferencial de trabalhar em uma granularidade refinada, através do cálculo de polaridade em nível atômico, ao invés de resíduos. Identificamos várias situações em que átomos hidrofóbicos conservam sua posição, apesar de não existirem resíduos hidrofóbicos conservados. O uso de informações fisico-químicas de mais fina granularidade pode ser importante, pois se sabe, atualmente, que em alguns casos a relação entre estrutura e função pode ser degenerada.

Com o intuito de demonstrar que um modelo que utiliza as informações de polaridade atômica é conservado e discriminativo entre famílias, construiu-se classificadores estruturais. Alguns dos conjuntos atômicos utilizados foram formados por átomos polares e outros conjuntos formados por átomos apolares das estruturas classificadas. Surpreendentemente, os átomos hidrofílicos mostraram-se como uma assinatura estrutural muito mais discriminativa entre famílias do que os átomos hidrofóbicos. Sabe-se que a cadeia principal é polar, o que nos levou ao questionamento sobre a influência dessa cadeia na melhoria de precisão da classificação da assinatura polar. Na tentativa de elucidar esse questionamento, discriminou-se os átomos do backbone como a melhor

assinatura estrutural estudada.

Por este resultado, decidiu-se, então, verificar se os átomos C e N (átomos esses que compõem a cadeia principal) são indispensáveis para explicar a melhoria da capacidade discriminativa dos classificadores, uma vez que o uso de tais átomos poderia, inclusive, capturar indiretamente os ângulos ϕ e ψ. Inserindo pontos intermediários fictícios entre os C_α, obtem-se uma poligonal geométrica artificial, similar à poligonal geométrica obtida pela inclusão dos átomos C e N. Dessa forma, conseguimos o efeito geométrico da poligonal construída pelo posicionamento dos átomos da cadeia principal porém, sem a influência do posicionamento dos átomos C e N e, indiretamente, sem a influência dos ângulos ϕ e ψ. O corrente estudo sugere que não são as posições desses átomos os fatores determinantes no incremento da precisão da classificação, mas antes o fortalecimento do caráter poligonal da cadeia principal.

Portanto, a principal contribuição do presente trabalho é a investigação do uso da disposição geométrica poligonal da cadeia principal (ou dos próprios C_α, caso sejam adicionados pontos intermediários fictícios entre esses) como uma assinatura estrutural discriminativa.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

Um dos principais problemas em bioinformática é a predição de funções de proteínas. Graças aos avanços obtidos no sequencimento genômico e na resolução de estruturas, a quantidade de proteínas sequenciadas e com estruturas determinadas cresce rapidamente. Por outro lado, o processo experimental de anotação da função dessas estruturas é caro e demorado, sendo um limitador prático para o uso dessas informações biológicas [Brown et al., 2006]. Devido a essa limitação, inúmeras estruturas determinadas permanecem com funções desconhecidas. Segundo estimativa apresentada por Parasuram [Parasuram et al., 2010], 9.400 estruturas depositadas à época da publicação daquele artigo (aproximadamente 13% do total de estruturas), no Protein Data Bank (PDB [Berman et al., 2000]), possuíam funções desconhecidas, hipotéticas ou putativas (e esse número tem aumentado).

Neste contexto, o uso de métodos, paradigmas e modelos automatizados para anotação funcional torna-se importante. O repertório de técnicas computacionais disponível para essa finalidade (anotação funcional) é vasto [Lee et al., 2007, Nisius et al., 2012, Rentzsch & Orengo, 2009, Sael et al., 2012, Sleator & Walsh, 2010, Xin & Radivojac, 2011], dividido, principalmente, em métodos baseados em sequência e em métodos baseados em estruturas. Graças à grande quantidade de dados de sequências disponíveis (UniProt [Consortium et al., 2014]), as técnicas de predição de função baseadas em sequências predominaram na última década [Volkamer et al., 2013]. Porém, estudos de caso publicados mostram proteínas com baixa identidade de sequência, mas possuidoras de funções relacionadas [Rost, 2002, Almonacid et al., 2010], bem como o contrário: proteínas com alta similaridade de sequência, mas que diferem em suas funções enzimáticas [Galperin et al., 1998]. Além disso, uma vez que se sabe que as estruturas das proteínas são mais conservadas que suas sequências [Illergård et al., 2009], as técnicas baseadas em estruturas vêm ganhando importân-

cia [Volkamer et al., 2013].

Mesmo que a função não possa ser diretamente inferida puramente a partir do enovelamento específico adotado por uma determinada proteína, os dados estruturais podem ser usados para detectar proteínas com funções semelhantes cujas sequências divergiram durante a evolução [Lee et al., 2007]. O enovelamento de proteínas é o processo pelo qual uma proteína assume sua conformação nativa, na qual é funcional ([Baker & Agard, 1994, Dill, 1999, Kauzmann, 1959, Privalov & Gill, 1988]) e é um processo de extrema complexidade ([Compiani & Capriotti, 2013, Dill et al., 2008, Schafer et al., 2014]). Seu estudo vem consumindo esforços de pesquisa há anos ([Dill & MacCallum, 2012, Dill et al., 2007]), tendo sido, inclusive, eleito em 2005, pela revista Science, como um dos 125 problemas mais importantes em aberto da ciência [Kennedy & Norman, 2005]. Apesar das muitas dificuldades que o problema traz consigo, avanços já foram obtidos ([Daggett & Fersht, 2003, Dill et al., 2007, Piana et al., 2014, Piana et al., 2013, Sharma et al., 2013]). Sabe-se, atualmente, que a estrutura conformacional nativa das proteínas é determinada, primariamente, por sua sequência de aminoácidos, conforme mostrado por Anfinsen [Anfinsen, 1972], e que são as ligações de hidrogênio e as interações hidrofóbicas os efeitos mais importantes que guiam o processo de enovelamento ([Baldwin, 1986, Baldwin, 2007, Baldwin, 2014, Baldwin & Rose, 2016, Bellissent-Funel et al., 2016, Chandler, 2005, Dill, 1990, Nick Pace et al., 2014, Pace et al., 2014, Pace et al., 1996, Perunov & England, 2014]).

A natureza emprega apenas alguns milhares de tipos de enovelamento para gerar todo o repertório de estruturas proteicas [Choi & Kim, 2006]. Chotia defende que todas as proteínas de todas as espécies podem ser representadas por cerca de 1.000 diferentes tipos de enovelamento [Chothia, 1992]. O SCOP *(Structural Classification of Proteins)* [Andreeva et al., 2004], utilizado no corrente trabalho, é o maior banco de dados manualmente curado de classificação de estruturas, baseado na similaridade dessas e em suas sequências de aminoácidos. O SCOP classifica as proteínas depositadas no PDB (*Protein Data Bank*) [Berman et al., 2000] em (1) *classes*, (2) *enovelamentos*, (3) *superfamílias* e (4) *famílias*. Para esse esquema de classificação, as proteínas de uma mesma família são as mais parecidas estruturalmente entre si.

Nas proteínas globulares, no estado enovelado, os resíduos apolares encontram-se, preferencialmente, mais protegidos do contato com o solvente, formando núcleos ([Cheung et al., 2002, Perunov & England, 2014, Rose et al., 1985, Zhou et al., 2004]). Acredita-se que proteínas estruturalmente similares, após respectivos processos de enovelamentos, adquiram núcleos semelhantes [Ding & Dokholyan, 2006]. Apesar do processo de enovelamento proteico ser fortemente relacionado às interações hidrofóbicas,

Larson e colaboradores [Larson et al., 2002] nos mostram que não há, necessariamente, conservação evolutiva preferencial nos resíduos que compõem o núcleo de enovelamento das proteínas. Ainda nessa linha, Bottini [Bottini et al., 2013] argumenta que definir a composição dos aminoácidos dos núcleos proteicos é fundamental para a compreensão do processo de enovelamento, uma vez que as diversas arquiteturas atingem suas estabilidades estruturais apenas na presença de redes de aminoácidos específicos. Soundararajan e colaboradores [Soundararajan et al., 2010] mostram que a rede de interação dos resíduos do núcleo não exposto ao solvente das proteínas conserva padrões de enovelamento, apesar de muitas vezes divergirem em suas sequências primárias. Utilizando mapas de contatos entre os resíduos que compõem o núcleo, um conjunto de teste de domínios, selecionados aleatoriamente, foi classificado com cerca de 97% de precisão, enquanto utilizando a totalidade dos resíduos das proteínas, o acerto do processo de classificação ficou em torno de 14%. A abordagem apresentada em [Soundararajan et al., 2010] consiste na obtenção de alinhamentos estruturais e na conservação posicional de resíduos hidrofóbicos. No nosso ponto de vista, os autores conseguem alta precisão de classificação por focarem nos resíduos extremamente conservados. Como aquele trabalho visava à classificação, o uso deste tipo de informação é apropriado.

Contudo, posto que nosso objetivo é compreender o papel dos efeitos interatômicos, não poderíamos utilizar uma abordagem em nível de resíduos, sob pena de importantes padrões serem perdidos. Nossa proposta tem o diferencial de trabalhar em uma granularidade mais refinada, através do cálculo de polaridade em nível atômico, ao invés de resíduos. Somos capazes de identificar várias situações em que átomos hidrofóbicos conservam sua posição, apesar de não existirem resíduos hidrofóbicos conservados. Por exemplo, um resíduo volumoso como o Triptofano poderia ser equivalente a mais de um resíduo hidrofóbico menor, como a Valina e a Alanina. Além disso, resíduos polares podem apresentar grandes porções hidrofóbicas, como é o caso da Lisina. O uso de informações fisico-químicas de mais fina granularidade pode ser importante, pois se sabe, atualmente, que em alguns casos a relação entre estrutura e função pode ser degenerada. Há muitos casos de proteínas com mesma função e estruturas diferentes e o inverso: funções diferentes para a mesma estrutura ([Gherardini et al., 2007, Russell, 1998, Bork et al., 1993]).

Buscamos por padrões posicionais em porções hidrofóbicas / hidrofílicas em proteínas globulares de mesma família. Nossa hipótese inicial era de que existiam padrões conservados no posicionamento dos átomos com polaridades similares nas proteínas globulares de uma mesma família. Nossas premissas foram inspiradas por um trabalho anterior de nosso grupo de pesquisa, denominado Hydropace

[Gonçalves-Almeida et al., 2012]. O Hydropace é um estudo que versa sobre inibição cruzada (fenômeno observado quando estruturas com topologias diferentes são inibidas por um mesmo inibidor), considerando a região intermolecular (interface de contato). Foram identificadas ilhas hidrofóbicas, que são compostas por aglomerados de átomos apolares. O estudo utiliza como abordagem uma análise de fina granularidade, considerando as redes de interações entre os átomos. As interações atômicas foram modeladas como grafos e, posteriormente, algoritmos para cálculo de componentes conexos e comunidades foram utilizados para a identificação das ilhas hidrofóbicas, chamadas no trabalho como *patches* hidrofóbicos.

Como resultado do citado trabalho (*Hydropace*), foi possível observar uma conservação da condição atômica que não era possível ser observada em nível de resíduo. A conservação de resíduos, na maioria das vezes, sequer é observada em proteínas de uma mesma família, mesmo essas possuindo estruturas similares. Entretanto, o método foi capaz de identificar essa conservação atômica em proteínas de famílias distintas, apesar de possuírem estruturas não homólogas. Trabalhando em nível atômico, foi possível abstrair a informação de resíduo e localizar regiões densamente conectadas, atraídas pelo efeito hidrofóbico. O estudo se concentrou em duas famílias proteicas, classificadas na base MEROPS [Rawlings et al., 2008] como serino proteases (tipo tripsina e tipo subtilisina). O Hydropace nos presta na indicação da estratégia de trabalhar em nível atômico, bem como na indicação da existência de ilhas hidrofóbicas em interfaces de contato, compostas por aglomerados de átomos apolares.

Para a investigação de nossa hipótese inicial de que existiam padrões conservados no posicionamento dos átomos com polaridades similares nas proteínas de uma mesma família, construímos algoritmos para detecção de ilhas hidrofóbicas nessas estruturas. Por meio do estudo manual de casos ilustrativos e da execução do algoritmo de detecção de ilhas para toda a base do *SCOP*, evidenciamos a existência desses grupamentos atômicos apolares. Verificamos, através desses mesmos casos ilustrativos, conservação na forma, volume e posicionamento dessas ilhas apolares, mesmo que, em algumas vezes, os átomos que as compõem fossem de elementos diferentes. Com tal resultado em mãos, acreditávamos que poderíamos utilizar essas ilhas hidrofóbicas como uma assinatura estrutural discrimativa. Assinaturas estruturais são conjuntos de características capazes de identificar, inequivocamente, um padrão de enovelamento proteico e a natureza das interações que podem estabelecer com outras proteínas e ligantes. Esses conjuntos de características são representações concisas de estruturas proteicas.

Dado um conjunto de proteínas estruturalmente similares, uma estratégia possível para a predição funcional é a definição de assinaturas estruturais. Acreditamos que seu uso na predição funcional de proteínas é um passo além dos métodos baseados apenas

na homologia de sequências. Por exemplo, Pires [Pires et al., 2011] investiga padrões de distâncias inter-resíduos em uma matriz de *cuttof* de distâncias para a classificação estrutural e predição de função, técnica que ele denominou *CSM* [Pires et al., 2011].

A *CSM* é o estado da arte em classificação de estruturas em larga escala e é totalmente independente de algoritmos de alinhamento estrutural. A técnica gera vetores de características, que representam padrões de distância entre resíduos de proteínas (cada resíduo é representado por um centroide). Em seu artigo original, os centroides utilizados para representar cada resíduo foram os C_α). Estes vetores de características são, então, utilizados como evidência para a classificação.

Pires [Pires et al., 2011] reporta que experiências foram conduzidas com outros centroides em vez do C_α, como o C_β ou o último átomo pesado (LHA) da cadeia lateral. O C_α obteve melhor desempenho em todos os experimentos, um fato que ele registrou como "exigindo uma investigação mais aprofundada". Almejamos elucidar o motivo dessa diferença de resultados reportada. No corrente trabalho, investigamos o uso de outros conjuntos atômicos como assinaturas estruturais. Para demonstrarmos que um modelo que utiliza as informações de polaridade atômica é conservado e discriminativo entre famílias, construímos alguns classificadores seguindo a estratégia da *CSM*. Alguns dos conjuntos atômicos utilizados foram formados por átomos polares e outros conjuntos formados por átomos apolares das estruturas classificadas.

Acreditávamos que poderíamos utilizar as ilhas hidrofóbicas (representadas pelos átomos polares que as compõem) como uma assinatura estrutural para diferenciação inter-famílias. Percebemos, porém, que apesar das estruturas de mesma família guardarem muita semelhança na forma, volume e posicionamento dessas ilhas apolares, não havia discriminação suficiente entre estruturas de famílias diferentes. Uma vez que os átomos apolares possuem a capacidade de se aglomerarem formando ilhas, provavelmente, famílias diferentes guardam alguma similaridade nas características dessas ilhas. Essa característica globular dificultaria uma tentativa de classificação seguindo apenas critérios estruturais, uma vez que famílias diferentes podem vir a guardar similaridades em suas assinaturas apolares.

Os átomos hidrofílicos mostraram-se como uma assinatura estrutural muito mais discriminativa entre famílias do que os átomos hidrofóbicos. Analisando diversos experimentos de classificação, constatamos que a assinatura polar das famílias é melhor diferencial de classificação que seus C_α e seus átomos apolares. Acreditamos que a diferença de precisão de classificação entre a assinatura apolar em relação à precisão de classificação da assinatura polar (em parte) deve-se à sua distribuição espacial mais globular, ao passo que a distribuição polar é mais poligonal (seguindo o backbone).

Sabe-se de antemão que a cadeia principal é polar, isso nos levou a questionar

a influência dessa cadeia na melhoria de precisão da classificação da assinatura polar. Na tentativa de elucidar esse questionamento, discriminou-se os átomos do backbone como a melhor assinatura estrutural estudada. Utilizando os átomos da cadeia principal como uma assinatura estrutural, fomos capazes de melhorar em até 10.3% a precisão da classificação de famílias para a base *SCOP* em relação à precisão obtida pela técnica *CSM* original, que utiliza somente o posicionamento dos carbonos alfa.

Uma vez que a classificação utilizando-se os átomos da cadeia principal obteve melhor desempenho em praticamente todas as métricas comparadas, verificou-se se os átomos C e N (átomos esses que compõem a cadeia principal) são indispensáveis para explicar a melhoria da capacidade discriminativa dos classificadores, eis que o uso de tais átomos poderia, inclusive, capturar indiretamente os ângulos ϕ e ψ. Uma hipótese era que, a adição de pontos geométricos promovida pelo uso dos demais átomos do backbone, reforçou a característica geométrica poligonal da cadeia principal, o que auxiliou a diferenciação inter-famílias pela *CSM*.

Verificou-se, então, se os posicionamentos dos átomos que compõem a cadeia principal são o principal fator de melhoria discriminativa entre famílias ou se bastaria a disposição geométrica poligonal, característica da cadeia principal. Inserindo pontos intermediários fictícios entre os C_α, obtêm-se uma poligonal "artificial", similar à poligonal obtida pela inclusão dos átomos C e N. Dessa forma, conseguimos o efeito da geométrico da poligonal da cadeia principal sem a influência do posicionamento dos átomos C e N e, indiretamente, sem a influência dos ângulos ϕ e ψ.

Concluíu-se, assim, que a melhoria da precisão de classificação, ao se utilizar um modelo com os átomos da cadeia principal, em relação ao uso exclusivo dos C_α dessa mesma cadeia, deve-se, principalmente, pela evidenciação do caráter geométrico poligonal da cadeia. O corrente estudo sugere que não são as posições dos átomos C e N (o que captura de maneira indireta os ângulos ϕ e ψ) os fatores determinantes no incremento da precisão da classificação, mas, sim, o fortalecimento do caráter geométrico poligonal da cadeia principal. Concluíu-se, também, que a assinatura da geometria poligonal da cadeia principal das proteínas de uma mesma família é melhor diferencial de classificação que a assinatura que utiliza somente seus carbonos alfas. Ou seja, o caráter sequencial da cadeia polipeptídica e sua disposição geométrica poligonal são mais relevantes que apenas o empacotamento dos resíduos, como usado na *CSM* original.

Não encontrou-se classificadores estruturais que explorassem a característica geométrica poligonal da cadeia principal como uma informação relevante [Zhou, 1998, Kedarisetti et al., 2006, Bu et al., 1999, Sahu & Panda, 2010, Ding et al., 2007, Liu & Jia, 2010, Kurgan et al., 2008, Chen et al., 2008b, Chen et al., 2008a, Dehzangi et al., 2013, Liu et al., 2010, Deschavanne & Tuffery, 2008,

Ding et al., 2012, Zheng et al., 2010, Dai et al., 2013, Levy et al., 2006, Røgen & Fain, 2003, Lee et al., 2007, Binkowski & Joachimiak, 2008].

Portanto, as contribuições do presente trabalho são a indicação da existência de ilhas hidrofóbicas nas estruturas globulares de mesma família e a investigação do uso da disposição geométrica poligonal da cadeia principal (ou dos próprios C_α, caso sejam adicionados pontos intermediários fictícios entre esses) como uma assinatura estrutural discriminativa. Comparou-se a capacidade discriminativa das assinaturas dos átomos polares e da poligonal da cadeia principal para as tarefas de classificação e de predição funcional com a assinatura original promovida pelo uso exclusivo dos C_α, constatando-se melhorias.

1.1 Objetivos

O objetivo do corrente trabalho é a verificação da existência de possíveis padrões estruturais conservados em proteínas de mesma família. Almeja-se verificar se tais padrões são capazes de gerar assinaturas estruturais suficientemente discriminativas entre famílias bem como encontrar ou propor modelos e algoritmos que possibilitem detectar esses padrões de conservação.

1.1.1 Objetivos específicos

Os objetivos específicos são:

- Verificar se o uso de informações com granularidade mais refinada, em nível atômico, inclusive de polaridade, agrega conhecimentos relevantes para a busca de padrões conservados.

- Verificar se os padrões conservados encontrados podem ser usados como possíveis assinaturas estruturais.

- Comparar os graus de conservação entre famílias de cada uma das assinaturas encontradas.

- Investigar a relevância das informações geométrica e de polaridade atômica dessas assinaturas estruturais.

1.2 Organização do texto

O Capítulo 2, Metodologia, apresenta na subseção Conhecimentos Prévios uma discussão teórica e conceitual necessária para o bom entendimento do trabalho. Na subseção Dados (2.2.1), as bases de dados utilizadas nos experimentos. Na subseção Técnica (2.2.2), a estrutura da metodologia proposta, indicando todos os passos e motivos experimentais. Os resultados são apresentados e discutidos no Capítulo 3. O Capítulo 4 versa sobre a Conclusão e os Trabalhos Futuros.

Capítulo 2

Metodologia

Esta tese propõe uma metodologia para encontrar modelos e construir algoritmos que possibilitem assinalar padrões de conservação em estruturas de mesma família proteica. A maior contribuição está na indicação da existência de ilhas hidrofóbicas nas estruturas globulares de mesma família e a investigação do uso da disposição geométrica poligonal da cadeia principal como uma assinatura estrutural viável. Comparou-se a capacidade discriminativa das assinaturas dos átomos polares e da geometria poligonal da cadeia principal para as tarefas de classificação e de predição funcional com a assinatura original promovida pelo uso exclusivo dos C_α, constatando-se melhorias.

A metodologia desse trabalho dividiu-se em **dois grupos de experimentos** principais: o **primeiro grupo de experimentos** objetiva conferir, em exemplos de famílias SCOP, a existência de grupamentos atômicos apolares nas famílias proteicas globulares. O **segundo**, de mais larga escala e automatizado, objetiva demonstrar que um modelo que utiliza informações de polaridade atômica ou da linearidade geométrica da cadeia principal é conservado e discriminativo entre famílias. Para tanto, efetuou-se diversos experimentos de classificação que utilizaram assinaturas estruturais formadas por conjuntos atômicos variados.

Para um melhor entendimento do trabalho apresentamos, primeiro, na sessão 2.1 e suas respectivas sub seções, os detalhes conceituais e de implementação.

2.1 Conhecimentos prévios

Para um melhor entendimento do corrente trabalho, são necessários alguns conhecimentos prévios, apresentados e discutidos a seguir. Primeiramente, serão apresentados alguns conceitos biológicos e, em seguida, algumas construções conceituais computacionais.

2.1.1 Conceitos biológicos

2.1.1.1 Polaridade atômica

A polaridade das ligações químicas altera o seu potencial de atraírem cargas elétricas. Normalmente, quando átomos de elementos distintos estão fazendo ligações covalentes, aquele mais eletronegativo exerce maior força de atração sobre a nuvem eletrônica, acarretando, desta forma, um compartilhamento desigual dos elétrons [Sherwood, 2005]. Esses compartilhamentos desiguais de elétrons nas ligações químicas são a causa da formação de dipolos, representados em cargas parciais positivas e negativas [Heinz & Suter, 2004]. As ligações podem chegar ao extremo de serem completamente polares e apolares. Uma ligação completamente apolar ocorre quando as eletronegatividades são idênticas. Uma ligação completamente polar é uma ligação iônica, ocorrendo quando a diferença de eletronegatividade é suficientemente grande para que um átomo retire completamente um elétron de outro.

A polaridade atômica é um conceito-chave para nossa análise, uma vez que ela é uma das informações fisico-químicas utilizadas para a caracterização das ilhas apolares (2.1.1.3). Entende-se, por este trabalho, que essas ilhas possuem conservação nos monômeros de mesma família, mesmo quando há divergência dos resíduos desses núcleos [Tseng & Liang, 2004, Larson et al., 2002]. Segundo Sobolev, são hidrofóbicos Cl, Br, I e todos os C que não estejam fazendo ligação covalente com O ou N. [Alexandre V. Fassio, 2017] adiciona ainda alguns carbonos aromáticos. Considerou-se, de maneira simplificada no corrente trabalho, como polares, todos os átomos que não fossem considerados hidrofóbicos. A tabela 2.1 lista esses átomos.

2.1.1.2 Cadeia principal

Os carbonos alfa (C_α) de resíduos de aminoácidos adjacentes são separados por três ligações covalentes, dispostas como $C_\alpha - C - N - C_\alpha$.

As ligações peptídicas $C - N$ não podem rotacionar livremente. A rotação é permitida sobre as ligações $N - C_\alpha$ e $C_\alpha - C$.

Residue	Atom	Criterion	
		[Sobolev et al., 1999]	[Alexandre V. Fassio, 2017]
ALA	CB	x	x
ARG	CB	x	x
	CG	x	x
ASN	CB	x	x
ASP	CB	x	x
CYS	CB	x	x
GLN	CB	x	x
	CG	x	x
GLU	CB	x	x
	CG	x	x
HIS	CB	x	x
ILE	CB	x	x
	CG1	x	x
	CG2	x	x
	CD1	x	x
LEU	CB	x	x
	CG	x	x
	CD1	x	x
	CD2	x	x
LYS	CB	x	x
	CG	x	x
	CD	x	x
MET	CB	x	x
	CG	x	x
	CD	x	x
PHE	CB	x	x
	CG		x
	CD1		x
	CD2		x
	CE1		x
	CE2		x
	CZ		x
PRO	CB	x	x
	CG	x	x
	CD	x	
THR	CG2		x
TRP	CB	x	x
	CG		x
	CD2		x
	CE3		x
	CZ2		x
	CZ3		x
	CH2		x
TYR	CB	x	x
	CG		x
	CD1		x
	CD2		x
	CE1		x
	CE2		x
VAL	CB	x	x
	CG1	x	x
	CG2	x	x

Tabela 2.1: Átomos considerados hidrofóbicos segundo [Sobolev et al., 1999] (Sob) e [Alexandre V. Fassio, 2017] (Ring). Os demais átomos foram suprimidos.

A conformação da proteína é, então, definida por três ângulos diédricos (um ângulo diédrico é o ângulo formado pela interseção de dois planos). Convencionou-se

chamar esses ângulos de torção de ϕ (phi), ψ (psi) e ω (ômega), ângulos esses que descrevem as rotações possíveis em torno de cada uma das três ligações repetidas na cadeia principal. Vide figura 2.1.

O ângulo ϕ envolve as ligações $C - N - C_\alpha - C$ (com a rotação ocorrendo em torno da ligação $N - C_\alpha$) e ψ envolve as ligações $N - C_\alpha - C - N$ (com a rotação ocorrendo em torno da ligação $C_\alpha - C$) [Richardson, 1981].

O terceiro ângulo diédrico, o ω, envolve as ligações $C_\alpha - C - N - C_\alpha$ (com a angulação ocorrendo em torno da ligação $C - N$) e não é frequentemente considerado [Pauling et al., 1951]. A ligação central, neste caso, é a ligação peptídica, em que a rotação é limitada. O ângulo ω na ligação peptídica é normalmente $\pm 180°$ (configuração comum *trans*) ou $0°$ (no caso raro de uma configuração *cis*), uma vez que o caráter de dupla ligação parcial mantém o peptídeo planar.

Em geral, a distância de separação entre C_α de aminoácidos adjacentes numa proteína é de cerca de 3.8 Å [Laskowski et al., 1993]). As demais distâncias são: $C_\alpha - C = 1.52$ Å, $C - N = 1.32$ Å e $N - C_\alpha = 1.46$ Å [Laskowski et al., 1993])

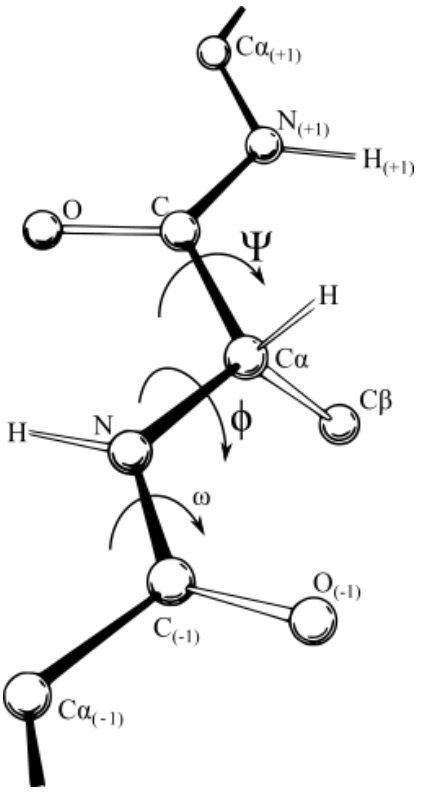


Figura 2.1: (a) Três ligações de carbonos alfa sequenciais numa cadeia polipeptídica. As ligações $N - C_\alpha$ e $C_\alpha - C$ podem rotacionar, descritas por ângulos diédricos designados ϕ e ψ, respectivamente. A ligação peptídica $C_\alpha - N$ não é livre para rotacionar. Figura retirada de [Dcrjsr & Redzikowski, 2011].

2.1.1.3 Ilhas apolares

As ilhas apolares foram caracterizadas como grupamentos de átomos apolares (2.1.1.1) que estivessem suficientemente próximos em termos de distâncias interatômicas. Foram considerados suficientemente próximos, átomos cuja distância interatômica fosse menor que 3.6Å, que é o dobro do raio de Van der Walls do enxofre (1.8Å[Bondi, 1964]), átomo de maior raio dentre os átomos que compõem os aminoácidos comumente encontrados nos seres vivos. Nessa distância-limite, átomos representados por esferas de raio de 1.8Å e com distância interatômica de 3.6Å tangenciariam-se.

2.1.1.4 Interseção volumétrica entre ilhas apolares

Outra caracterização que se faz necessária é a da interseção volumétrica entre ilhas apolares. Essa interseção é determinada pelo somatório das interseções volumétricas esféricas individuais entre todos os átomos de alguma ilha apolar, para todos os demais átomos de alguma outra ilha apolar. O volume de interseção entre duas esferas de raios idênticos r e distância entre centros d é dado pela fórmula [Weisstein, 2017]:

$$V = 1/12\pi(4r + d)(2r - d)^2 \tag{2.1}$$

Utilizou-se a fórmula apresentada com $r = 1.8$ e d igual à distância interatômica.

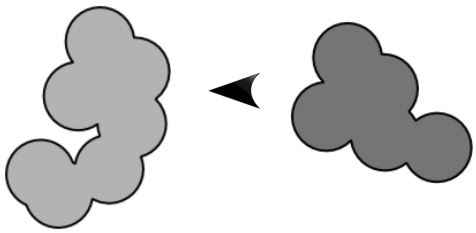


Figura 2.2: Duas ilhas antes da sobreposição. À esquerda, a ilha-alvo, e, à direita, a ilha-móvel.

Figura 2.3: Após a sobreposição, é possível calcular a interseção volumétrica. Essa interseção é dada pelo somatório de interseções esféricas entre todos os átomos da ilha-móvel para todos os átomos da ilha-fixa.

2.1.1.5 Arquipélagos apolares

Para demonstração da conservação posicional e volumétrica das ilhas apolares entre os monômeros alinhados de uma mesma família, faz-se necessário a definição de um critério de casamento entre ilhas apolares. Após a sobreposição entre o monômero-móvel e o monômero-fixo, cada ilha-móvel (ilha apolar do monômero móvel) deverá ser casada à ilha-fixa (ilha apolar do monômero fixo) de maior interseção volumétrica (2.1.1.4). Uma mesma ilha-fixa poderá conter múltiplas ilhas-móveis fundidas a ela. Uma ilha-móvel, porém, só poderá se fundir a uma única ilha-fixa (ou nenhuma), a saber, aquela de maior interseção volumétrica diferente de zero. Dessa forma, cada ilha-fixa é um pivô de um possível arquipélago apolar. Construímos, então, o algoritmo 1 para formalizar esse critério. As figuras 2.2 e 2.3 ilustram esse conceito. As figuras 2.4, 2.5 e 2.6 ilustram alguns arquipélagos reais.

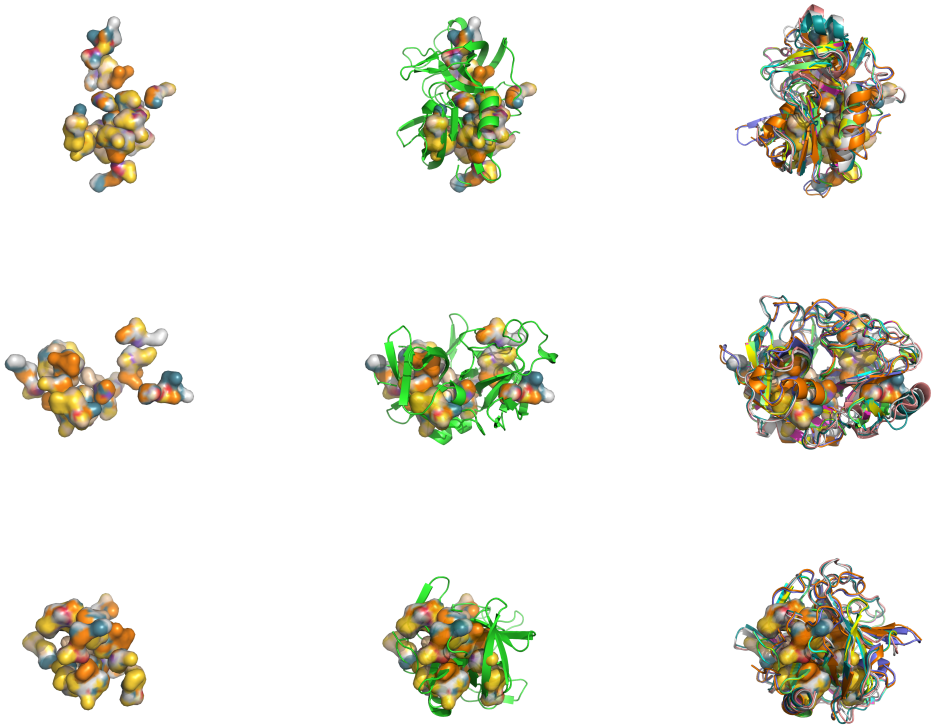
Algorithm 1 Hydrophobic archipelagos algorithm

```
procedure BuildArchipelagos(fixed, mobiles)
    archs ← {}                                   ▷ Archipelagos map.
    for all mobile in mobiles do
        vMap ← {}                                ▷ Volume intersection map.
        Align mobile to fixed
        for all mPatch in mobile.patches do
            for all fPatch in fixed.patches do
                v ← volume(mPatch, fPatch)
                vMap(mPatch, fPatch) ← v
        l ← Sorted list by v values of vMap
        for all (mPatch, fPatch) in l do
            if mPatch already visited then
                continue
            else if fPatch in arch then
                arch ← archs(fPatch)
                archs(fPatch) ← arch ∪ mPatch
            else
                archs(fPatch) ← set(mPatch)
    return archs.values                          ▷ List of patches set
```

2.1.2 Conceitos computacionais

2.1.2.1 Cutoff Scanning Matrix - CSM

Cutoff Scanning Matrix - *CSM* - é um modelo de classificação e predição de função que utiliza padrões de distâncias entre resíduos [Pires et al., 2011]. A *CSM* original



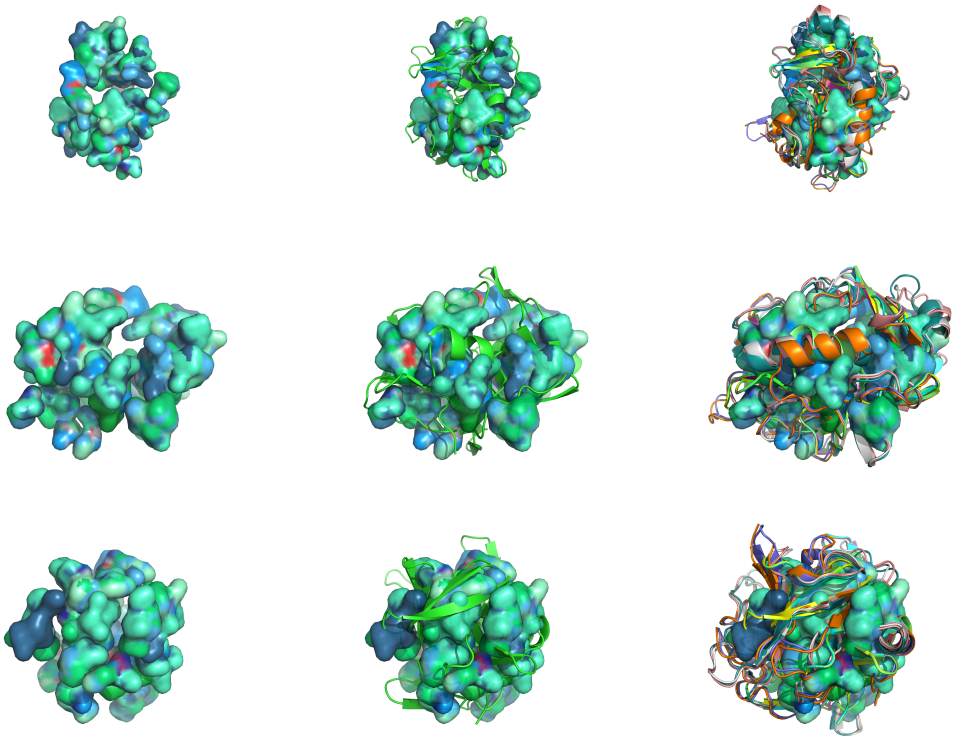
(a) (b) (c)

Figura 2.4: Três ângulos diferentes para o alinhamento de 10 monômeros da família *b.47.1.1*, segundo o critério *Sobolev*. Foram exibidos apenas arquipélagos apolares com 100% de conservação para melhor acuidade visual. A coluna (a) exibe somente os arquipélagos. A coluna (b), a posição desses em relação à estrutura fixa. A coluna (c), exibe os 10 monômeros alinhados.

gera vetores de características que representam padrões de distância entre resíduos de proteínas e utiliza esses vetores de características como evidência para a classificação. A decomposição em valores singulares - SVD - (2.1.2.2) é utilizada como um passo de pré-processamento para reduzir a dimensionalidade e o ruído.

Em seu artigo original, *CSM* foi capaz de atingir uma precisão de até 95% em uma experiência, usando todo o conjunto de domínios encontrados na última versão do *SCOP*. O método é eficaz em tarefas de classificação estrutural. Os padrões derivados por *CSMs* poderiam, efetivamente, ser usados para prever função de proteína e ajudar na anotação automatizada de funções. Esses fatos reforçam a ideia de que o padrão de distâncias entre resíduos é um componente importante das assinaturas estruturais das famílias proteicas.

Optou-se por essa técnica por ser o estado da arte em classificação de estruturas e totalmente independente de algoritmos de alinhamento estrutural. Até onde sabemos,



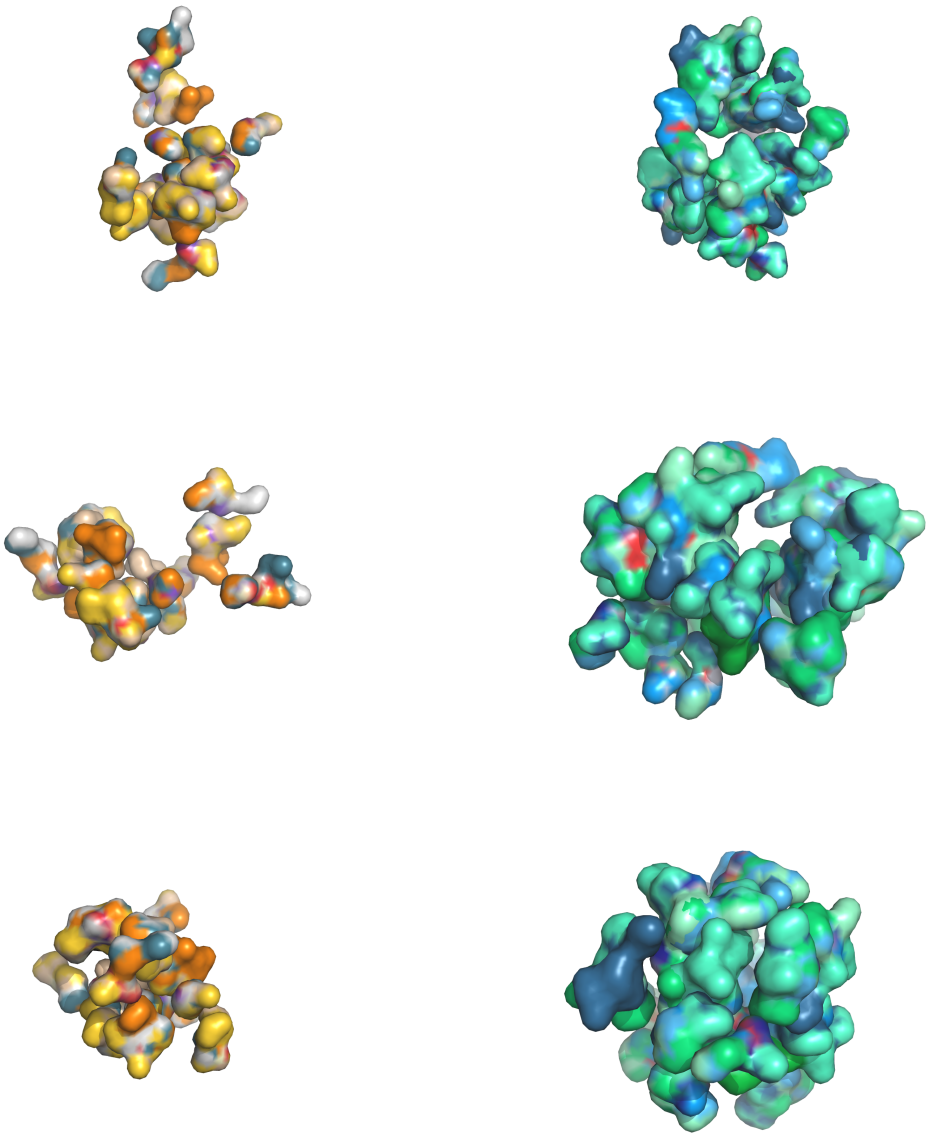
(a) (b) (c)

Figura 2.5: Três ângulos diferentes para o alinhamento de 10 monômeros da família *b.47.1.1*, segundo o critério *Ring*. Foram exibidos apenas arquipélagos apolares com 100% de conservação para melhor acuidade visual. A coluna (a) exibe somente os arquipélagos. A coluna (b), a posição desses em relação à estrutura fixa. A coluna (c), exibe os 10 monômeros alinhados.

é o único classificador capaz de operar com base de dados tão grande quanto o SCOP.

As *CSMs* foram construídas da seguinte forma: para cada um dos conjuntos atômicos, gera-se um vetor de características. Primeiro, calcula-se a distância euclidiana entre todos os pares de átomos do conjunto utilizado e define-se um intervalo de distâncias (pontos de corte) a ser considerado e um passo de distância. Examina-se essas distâncias, calculando a frequência de pares atômicos que sejam próximos de acordo com esse limiar de distância. O algoritmo 2 formaliza a função que calcula a *CSM*.

No corrente trabalho, assim como na *CSM* original, varia-se o limiar de distância de 0.0 Å a 30.0 Å, com um passo de 0.2 Å, representado por um vetor de 151 entradas para cada conjunto atômico. Juntos, esses vetores compõem a *CSM*. Em suma, cada linha da matriz representa uma proteína (ou um conjunto atômico derivado de uma proteína) e cada coluna representa a frequência de pares de átomos dentro de uma certa distância.



(a) (b)

Figura 2.6: Dois ângulos diferentes para o alinhamento de 10 monômeros da família *b.47.1.1*, segundo os critérios *Sobolev* e *Ring*. Os arquipélagos gerados pelo critério *Ring* são melhor definidos. Foram exibidos apenas arquipélagos apolares com 100% de conservação para melhor acuidade visual.

Algorithm 2 Cutoff Scanning Matrix calculation

```
procedure GENERATE-CSM(AtomsSets, CSM, Dist_MIN, Dist_MAX, Dist_STEP)
    for all AtomSet i in AtomsSets do
        j = 0
        Calculate the distances between all pairs of atoms
        for dist ← Dist_MIN to Dist_MAX step Dist_STEP do
            CSM[i][j] ← Get frequency of pairs of atoms within a distance dist
            j++
    return CSM
```

A motivação para o uso desse tipo de informação advém de que proteínas com diferentes enovelamentos e funções apresentam diferenças significativas em seus empacotamentos. Por outro lado, pode-se esperar que as proteínas com estruturas semelhantes também tenham empacotamentos semelhantes, informação que é capturada em uma *CSM*.

A variação de corte (varredura) agrega informações importantes relacionadas ao empacotamento da proteína e capta, implicitamente, a forma da proteína. A topologia da superfície, e mesmo cavidades do núcleo, são explicados bem por essa varredura.

Um exemplo de distribuição de contatos é mostrado na Figura 2.7. Três proteínas com formas muito diferentes foram selecionadas (uma globina, PDB: 1A6M, uma porina, PDB: 2ZFG e um colágeno, PDB: 1BKV) e a topologia do gráfico de contato obtido com diferentes pontos de corte (6.0 Å, 9.0 Å, 12.0 Å). As distribuições de densidade cumulativa e normalizada para os vetores de características CSM para estes representantes são também exibidas. Pode-se ver, a partir desses exemplos, que uma expressiva diferença de forma é contabilizada na CSM. Isso implica que a CSM está manipulando dois níveis essenciais de informação estrutural: contatos locais e não-locais relevantes. Também pode-se ver que as formas das proteínas interferem diretamente na rede de contato subjacente, que se reflete na dobra da proteína, como apontado por [Soundararajan et al., 2010].

Nesta tese, adaptou-se a técnica para encontrar padrões de distâncias atômicas e não só de resíduos.

2.1.2.2 Decomposição em valores singulares - SVD

A decomposição em valores singulares (*SVD*) é uma técnica de análise numérica que objetiva representar uma matriz A qualquer, composta por m linhas e n colunas, por um conjunto de matrizes derivadas [Pires et al., 2011]. Essa decomposição é uma forma diferente de representar os dados originais sem prejuízos semânticos. *SVD* é capaz de

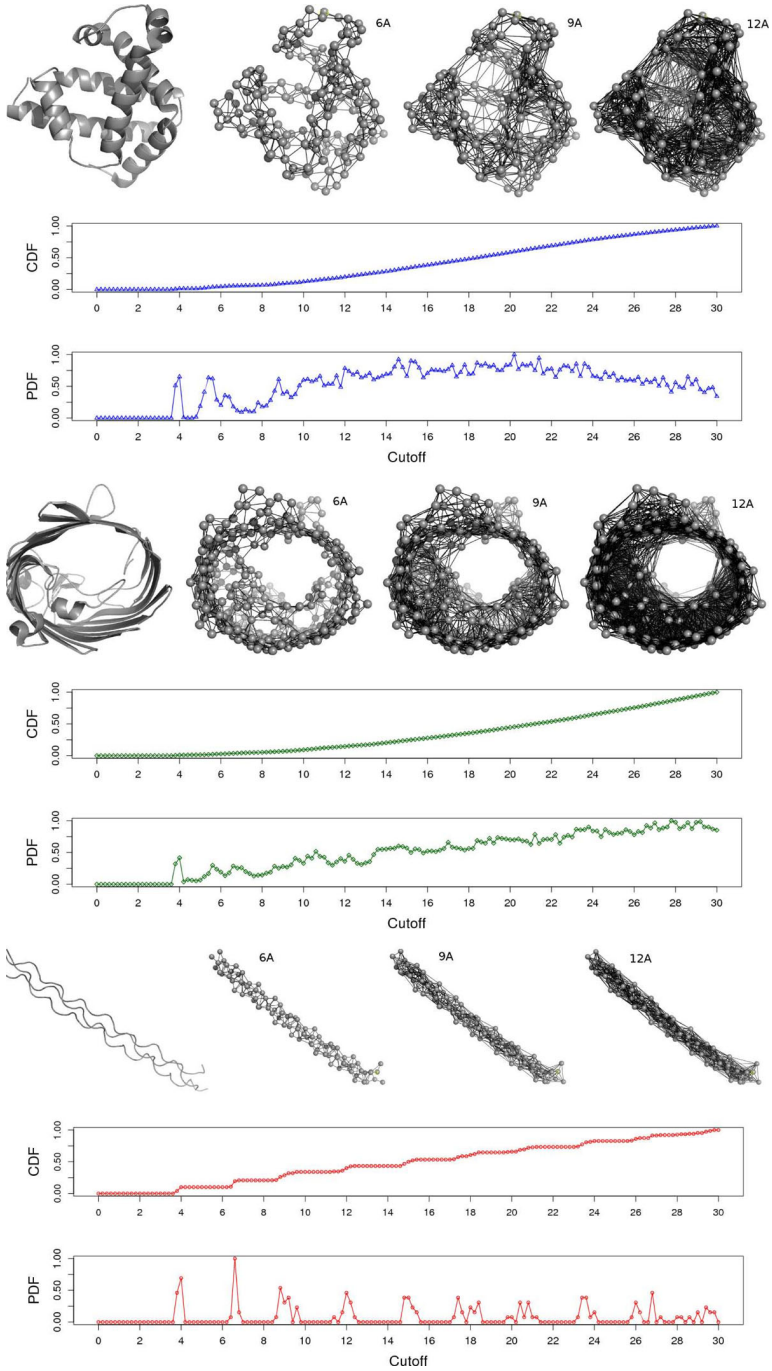


Figura 2.7: Grafo de contato da topologia por corte para proteínas com enovelamentos diferentes. São mostradas as topologias dos grafos de contato de três estruturas distintas (de cima para baixo: globina, porina e colágeno) com diferentes valores de corte: 6.0 Å, 9.0 Å, 12.0 Å. A distribuição cumulativa normalizada e distribuição de densidade do perfil de varredura de corte destas proteínas também são mostradas. Imagem retirada do artigo [Pires et al., 2011]

estabelecer relacionamentos não óbvios, porém, relevantes, entre elementos agrupados [Berry et al., 1995, Eldén, 2007].

Formalmente:

$$A = TSD^T \tag{2.2}$$

Onde T é uma matriz ortonormal de dimensões $m\ x\ m$, S é uma matriz diagonal de dimensões $m\ x\ n$ e D uma matriz ortonormal de dimensões $n\ x\ n$. Os valores diagonais de S são os valores singulares de A, ordenados do mais para o menos significativo.

Quando utilizados apenas alguns dos valores singulares, geramos uma matriz A_k, uma aproximação da matriz original A, onde $k \leq p$, p é o posto da matriz A e k o posto da matriz A_k:

$$A \approx A_k = T_k S_k D_k^T \tag{2.3}$$

A qualidade da aproximação depende da quantidade de valores singulares utilizados [del Castillo-Negrete et al., 2007]. Um conjunto de dados representado por uma quantidade menor de valores singulares que o conjunto original possui uma tendência de agrupar alguns itens que não seriam agrupados se utilizados os dados originais [Berry et al., 1995]. Neste trabalho, utilizou-se apenas duas das matrizes da fatoração de A_k, que, multiplicadas, geram a matriz V_k [Eldén, 2007]. Nesse contexto:

$$A_k = T_k S_k D_k^T = T_k(S_k D_k^T) = T_k V_k \tag{2.4}$$

A justificativa para se usar somente V_k é que os relacionamentos entre as colunas de A_k são preservados em V_k, porque T_k é uma base para as colunas de A_k.

No corrente trabalho, assim como no artigo original da CSM [Pires et al., 2011], a decomposição em valores singulares foi utilizada como um passo de pré-processamento para reduzir a dimensionalidade e o ruído dos classificadores. Utilizou-se, também, apenas 9 valores singulares [Pires et al., 2011].

2.1.2.3 k-nearest neighbors algorithm - k-NN

O algoritmo k-Nearest Neighbours (k-NN) é um método não-paramétrico utilizado para classificações e regressões [Altman, 1992].

Na classificação k-NN, a saída é uma associação de classe. Um objeto é classificado por um voto majoritário de seus vizinhos, sendo atribuído à classe mais comum entre seus k vizinhos mais próximos. Se k = 1, então o objeto é simplesmente atribuído à classe desse único vizinho mais próximo.

Apesar do algoritmo k-NN estar entre os mais simples de todos os algoritmos de aprendizado de máquina, ele é comumente utilizado graças à sua eficiência e simplici-

dade de implementação.

O algoritmo passa por duas fases: uma primeira, de treinamento e uma segunda, de classificação. A fase de treinamento consiste apenas em armazenar os vetores de recurso e rótulos de classe das amostras de treinamento.

Na fase de classificação, k é uma constante definida pelo usuário e um vetor pergunta (uma consulta ou ponto de teste) é classificado atribuindo o rótulo que é mais frequente entre as k amostras de treinamento mais próximas a esse ponto de consulta.

Uma métrica de distância comumente utilizada para variáveis contínuas é a distância euclidiana. Para variáveis discretas, como para a classificação de texto, pode ser usada outra métrica, como a métrica de sobreposição (ou a distância de Hamming [Hamming, 1950]). No corrente trabalho, utilizou-se a distância euclidiana. O algoritmo utilizado foi o *k-NN*, por ter sido reportado como o de melhor desempenho em [Pires et al., 2011].

2.1.2.4 Alinhamento estrutural

O alinhamento estrutural é uma ferramenta indispensável para comparação e classificação de proteínas [Eidhammer et al., 2000]. Porém, a despeito de sua extrema importância, não existe, ainda, nenhuma solução que encontre, simultaneamente, uma resposta rápida e precisa para esse problema [Poleksic, 2009]. Embora algumas funções de pontuação de semelhança estrutural possam ser aproximadas em tempo polinomial, não houve procedimento (de qualquer tempo de execução) capaz de encontrar a solução ótima utilizando qualquer métrica de alinhamento estrutural usualmente utilizada [Poleksic, 2009]. Em seu artigo de revisão sobre os progressos no campo da comparação de estruturas, Taylor e colaboradores escrevem: "Em comparação de estruturas, nós nem sequer temos um algoritmo que garante uma resposta ideal para os pares de estruturas."[Eidhammer et al., 2000]. Portanto, os algoritmos existentes são heurísticas.

O processo de alinhamento estrutural consiste em tentar estabelecer uma correlação entre duas ou mais estruturas de polímero, baseado em suas formas tridimensionais, de maneira a minimizar o valor da métrica de comparação [Eidhammer et al., 2000]. Várias métricas podem ser utilizadas [Godzik, 1996], porém, a mais comum é o *RMSD - Root Mean Square Deviation* [Hoehn & Niven, 1985]. Alinhadores que utilizam o *RMSD* como métrica objetivam minimizar as distâncias euclidianas entre os elementos correlacionados pelo alinhamento [Oldfield, 2007, Ortiz et al., 2002, Ye & Godzik, 2003].

Apesar de relacionados, os conceitos de alinhamento estrutural e de sobreposição não se confundem. A sobreposição, diferentemente do alinhamento, não faz nenhum tipo de tentativa de descobrir quais relações existem entre as estruturas comparadas. A sobreposição requer um alinhamento pré-calculado como entrada para, então, determinar as rotações e translações ótimas, de forma a minimizar a métrica entre os pontos relacionados pelo alinhamento [McLachlan, 1982].

Quando a métrica utilizada é a soma dos quadrados das distâncias (distância euclidiana) entre os pontos relacionados pelo alinhamento, a sobreposição pode ser modelada como uma instância do *problema da sobreposição de pontos* (2.1.2.6), que é um problema muito mais simples que o alinhamento e possui diversas soluções conhecidas na literatura, que são descritas e comparadas em [Sabata & Aggarwal, 1991, Istrail, 2003, Eggert et al., 1997].

2.1.2.5 Casamento de Padrões de Pontos - *Point Pattern Matching*

O problema do casamento de padrões de pontos (*Point Pattern Matching*) é um tópico importante para diversas áreas da Ciência da Computação, tais como Visão Computacional e Mineração de Dados [Zhang et al., 2003], podendo ser definido da seguinte forma: sejam $\{\mathcal{A}, \mathcal{B}\}$ dois conjuntos finitos de pontos pertencentes a um espaço vetorial real finito $\mathbb{R}^d$ contendo, respectivamente, a e b pontos. O problema consiste em encontrar uma transformação a ser aplicada no conjunto móvel (pergunta) $\mathcal{B}$, de tal forma a minimizar a diferença entre $\mathcal{B}$ e o conjunto estático (alvo) $\mathcal{A}$. Em outras palavras, o mapeamento desejado de $\mathbb{R}^d$ para $\mathbb{R}^d$ é aquele que melhor alinha o conjunto-pergunta transformado $\mathcal{B}'$ com o conjunto-alvo para alguma métrica pertencente a esse espaço [Jian & Vemuri, 2011]. No corrente trabalho, o espaço usado é o $\mathbb{R}^3$ (as 3 dimensões espaciais da posição atômica, (x, y, z)), a métrica utilizada foi a distância euclidiana e as transformações permitidas são a rotação e a translação.

O algoritmo exato mais rápido conhecido para o problema do casamento de padrões de pontos possui complexidade temporal da ordem $O(ba^d)$[de Rezende & Lee, 1995], sendo NP-Difícil para dimensões arbitrárias (d arbitrário)[Cabello et al., 2008] e pouco prático para a terceira dimensão ($d = 3$), obrigando a utilização de heurísticas para sua solução.

2.1.2.6 Sobreposição de Pontos - *Orthogonal Procruste Problem*

O problema da sobreposição de pontos pode ser modelado matematicamente como um problema de mínimos quadrados, que busca transformar uma dada matriz B em outra matriz A, por uma matriz de transformação ortogonal $\mathcal{T}$, de modo que a soma

dos quadrados da matriz residual $E = B\mathcal{T} - A$ é mínimo [Schönemann, 1966] (em nomenclatura já adaptada para esse trabalho). A solução computacional construída foi fortemente baseada na solução apresentada em [Eggert et al., 1997]: suponha que existam dois conjuntos de pontos correspondentes matricialmente representados por A e B, $i = 1..N$ (onde o número de linhas é igual ao número de pontos e o número de colunas é igual à dimensão do espaço vetorial) de modo que eles estão relacionados por:

$$A = RB + T + V \tag{2.5}$$

onde R é uma matriz de rotação padrão $d \times d$ (no caso deste trabalho $d = 3$), T é um vetor de translação dD e V um vetor de ruído. Resolver para a transformação ideal $\mathcal{T} = [\hat{R}, \hat{T}]$, que mapeia o conjunto B para A, normalmente requer uma minimização dos mínimos quadrados sob o critério de erro dada por:

$$\Sigma^2 = \sum_{i=1}^{N} \|A[i] - \hat{R}B[i] - \hat{T}\|^2 \tag{2.6}$$

Para encontrar a solução de mínimos quadrados da equação 2.6, os conjuntos de pontos A e B devem ter o mesmo centroide. Usando esta restrição, um novo conjunto de equações pode ser gerado. Definindo:

$$\bar{a} = \frac{1}{N}\sum_{i=1}^{N} A[i], \qquad a_{c_i} = A[i] - \bar{a} \tag{2.7}$$

$$\bar{b} = \frac{1}{N}\sum_{i=1}^{N} B[i], \qquad b_{c_i} = B[i] - \bar{b} \tag{2.8}$$

A equação 2.6 pode ser reescrita e reduzida para:

$$\Sigma^2 = \sum_{i=1}^{N} \|a_{c_i} - \hat{R}b_{c_i}\|^2 \tag{2.9}$$

$$= \sum_{i=1}^{N} (a_{c_i}a_{c_i}^T + b_{c_i}b_{c_i}^T - 2a_{c_i}^T\hat{R}b_{c_i}) \tag{2.10}$$

Esta equação é minimizada quando o último termo é maximizado, o que é equivalente a maximizar o traço da matriz $\hat{R}H$, onde H é uma matriz de correlação definida por:

$$H = \sum_{i=1}^{N} b_{c_i}a_{c_i}^T \tag{2.11}$$

Se a decomposição por valores singulares de H é dada por $H = U \wedge V^T$, então a matriz de rotação ótima, R, que maximiza o traço desejado, é a equação 2.12, conforme mostrado por [Eggert et al., 1997].

$$\hat{R} = U \begin{pmatrix} 1 & & \\ & 1 & \\ & & det(UV^T) \end{pmatrix} V^T \tag{2.12}$$

A translação ideal alinha o baricentro do conjunto A com o baricentro rotacionado do conjunto B, como mencionado anteriormente. Isto é:

$$\hat{T} = \bar{a} - \hat{R}\bar{b} \tag{2.13}$$

2.2 Desenho experimental

Uma vez apresentados os conceitos utilizados, pode-se detalhar os passos experimentais.

A metodologia deste trabalho dividiu-se em **dois grupos de experimentos** principais. O **primeiro grupo de experimentos**, denominado **Ilhas Hidrofóbicas**, objetiva conferir, em exemplos de famílias SCOP distintas, a existência de grupamentos atômicos apolares nas proteinas dessas famílias. Esse passo consistiu em adquirir os monômeros de todas famílias *SCOP* e assinalar a polaridade de seus átomos (2.1.1.1). Em seguida, determinou-se suas ilhas apolares (2.1.1.3) e executou-se o alinhamento estrutural, par a par, de todas as estruturas em relação a uma estrutura alvo fixa, escolhida ao acaso dentre os monômeros da família. Determinou-se os arquipélagos apolares (2.1.1.5) para posterior cálculo das métricas de similaridade apolar (2.2.2.1). Finalmente, foram visualmente esses resultados como seções do programa *PyMol* [Schrödinger, LLC, 2015] para 5 famílias escolhidas ao acaso (material suplementar). Optou-se por um número reduzido de famílias uma vez que o objetivo é apenas o de exibição de resultados visuais.

O **segundo grupo de experimentos**, de mais larga escala e automatizado, objetiva demonstrar que um modelo que utiliza as informações de polaridade atômica é conservado e discriminativo entre famílias. Para tanto, efetuaram-se diversos experimentos de classificação que utilizaram assinaturas estruturais formadas por conjuntos atômicos variados. Tais classificadores seguiram a estratégia da *CSM*[Pires et al., 2011]. Optou-se por essa técnica por ser o estado da arte em classificação de estruturas em larga escala e totalmente independente de algoritmos de alinhamento estrutural (vide 2.1.2.1). Nesse artigo original [Pires et al., 2011], as matrizes *CSM* eram construídas a

partir da posição geométrica somente dos C_α. Esse resultado foi utilizado como grupo controle para comparação com o uso de outros grupamentos atômicos. Denominou-se tal grupo de experimentos de **classificadores**.

Todos os experimentos comungam dos passos de leitura e filtro dos arquivos PDB para extração dos posicionamentos espaciais dos átomos que constituem a estrutura. O PDB (http://www.rcsb.org/pdb/) é um repositório que mantém os dados estruturais de macromoléculas biológicas. Os dados depositados no PDB são obtidos por meio de várias técnicas, tais como difração de Raio-X, NMR, microscopia eletrônica por criogenia e modelagem teórica [Berman et al., 2000]. Existem hoje diversas bibliotecas que se prestam a trabalhar com arquivos PDB, provendo automatizações de tarefas comuns como leitura/escrita de arquivos desse tipo, bem como funções auxiliares, a exemplo de cálculo de distâncias atômicas. Uma que se destaca e foi amplamente utilizada neste trabalho é a Biopython (http://www.biopython.org/). Ela possui a capacidade de ler um arquivo PDB e devolver um objeto Python contendo as informações do arquivo, facilitando a programação [Hamelryck & Manderick, 2003].

O passo de assinalamento da polaridade atômica consiste em atribuir a polaridade de cada um desses átomos. Esse cálculo foi feito de maneira tabelada para os experimentos de classificação e de casos ilustrativos (2.1.1.1)

Os dados utilizados em cada um dos grupos experimentais são apresentados na seção seguinte, Dados (2.2.1). As técnicas utilizadas em cada um desses experimentos são detalhadas na seção Técnica (2.2.2).

2.2.1 Dados

2.2.1.1 Ilhas Hidrofóbicas

Nesse primeiro experimento, adquiriu-se os monômeros (mínimo de 10 e máximo de 100 monômeros por família) de todas famílias *SCOP* distintas e assinaladas as polaridades de seus átomos (2.1.1.1). Esse limite (máximo de 100 monômeros por família) foi traçado por questões de tempo de execução do algoritmo. A listagem completa dos arquivos PDB utilizados encontra-se nos materiais suplementares. Para a apresentação visual dos casos ilustrativos, foram escolhidas 5 famílias do *SCOP* ao acaso, e selecionados 20 de seus monômeros, apresentados como seções do programa *PyMol* (materiais suplementares). Na sessão 3.1 são apresentados e discutidos os resultados obtidos pelo experimento. As famílias apresentadas são: b.47.1.1 - *Prokaryotic proteases*, b.60.1.1 - *Retinol binding protein-like*, b.6.1.1 - *Plastocyanin/azurin-like*, a.1.1.2 - *Globins* e c.47.1.1 - *Thioltransferase*.

2.2.1.2 Classificadores

Para os experimentos de classificação, utilizou-se algumas das bases de dados apresentadas na *CSM* [Pires et al., 2011]. As bases de dados utilizadas foram a gold-standard dataset enzymes [Brown et al., 2006], o *SCOP* versão *1.75* e as bases 6SSE, 5SSE, 4SSE e 3SSE [Jain & Hirst, 2010]. Tais bases serviram como grupo controle para a demonstração da melhoria na capacidade de classificação utilizando outras assinaturas estruturais. Vide tabela 2.2.

A princípio, todas as cadeias diferentes presentes nessas bases de dados eram candidatas a compor o conjunto de testes. Em arquivos PDB contendo mais de um modelo, utilizamos apenas o primeiro modelo. Em arquivos contendo mais de uma cadeia, as cadeias foram separadas segundo a classificação SCOP e tratadas individualmente.

Algumas dessas estruturas incorreram em exceções, devido ao arquivo estar obsoleto ou a cadeia ser muito pequena (menos de 10 resíduos ou átomos). Encontra-se no material suplementar a listagem completa das cadeias utilizadas, bem como de todos os arquivos que geraram exceções e a razão delas.

Tabela 2.2: Estatísticas das cadeias das bases de dados utilizadas

Dataset	Total	Utilizadas	% Perda
Gold-Standard	899	895	0.44%
6SSE	2,315	2,303	0.52%
5SSE	2,930	2,853	2.63%
4SSE	1,756	1,720	2.05%
3SSE	880	866	1.59%
Full-SCOP 1.75	207,890	201,771	2.94%

2.2.2 Técnica

2.2.2.1 Ilhas Hidrofóbicas

Para os casos ilustrativos, determinou-se as ilhas apolares contidas nos núcleos monoméricos (2.1.1.3) e executou-se o alinhamento estrutural, par a par, de todas as estruturas em relação a uma estrutura alvo fixa, escolhida ao acaso dentre os monômeros da família. Esses alinhamentos estruturais par a par foram efetuados através do programa *TM-Align* [Zhang & Skolnick, 2005], um alinhador moderno, referência na literatura. Uma vez alinhados, foi possível encontrar a interseção volumétrica entre suas ilhas apolares (2.1.1.4) e efetuar a construção dos arquipélagos apolares (2.1.1.5).

Em seguida, determinou-se os arquipélagos com pelo menos 90% de conservação entre todos os monômeros alinhados para o cálculo das métricas de similaridade apolar.

Um arquipélago possui ao menos 90% de conservação quando é composto por ilhas presentes em 90% ou mais dos monômeros alinhados. Entendeu-se que esse é um valor de corte conservador e significativo. Essas métricas consistem no cálculo da média e variância da quantidade de átomos das ilhas que compõem algum arquipélago, bem como da distância média dos centroides das ilhas para o centroide do referido arquipélago. Os valores são apresentados na Tabela 3.1.

2.2.2.2 Classificadores

A fim de investigar se modelos de assinaturas estruturais que utilizam informações posicionais em nível atômico são conservados e discriminativos entre famílias, construiu-se classificadores.

O **grupo de experimentos de classificação** dividiu-se em três **subgrupos de experimentos**: o **primeiro subgrupo**, chamado de **polaridade atômica**, objetiva investigar se um modelo que utiliza informações de polaridade atômica é mais conservado e discriminativo entre famílias do que o modelo que utiliza somente os C_α.

O **segundo subgrupo**, chamado de **cadeia principal**, objetiva demonstrar que um modelo que utiliza informações posicionais dos átomos da cadeia principal é mais conservado e discriminativo entre famílias do que os modelos que utilizam informações de polaridade atômica.

O **terceiro subgrupo de experimentos** é o subgrupo mais importante da tese. Chamado de **poligonal geométrica**, objetiva demonstrar que não são, necessariamente, a posição dos átomos que compõem a cadeia principal o fator de melhoria discriminativa entre famílias, mas, sim, a geometria poligonal característica da cadeia principal. Um objetivo adicional desse terceiro subgrupo de experimentos é a demonstração de que a posição dos átomos que compõem a cadeia principal e os ângulos que esses átomos formam entre si (ângulos ϕ e ψ) só conseguem influenciar o grau discriminativo dos classificadores quando são adicionados pontos intermediários e fortalecida sua disposição poligonal.

Para cada um desses experimentos realizados, construiu-se uma *CSM* distinta. Nas matrizes *CSM*, cada linha representa o vetor de característica dos padrões de distância entre os átomos do conjunto selecionado de uma das cadeias utilizadas no respectivo experimento. Esses vetores (linhas da matriz) possuem 151 posições, onde a posição i do vetor contém a frequência de átomos com distâncias iguais a $0.2 \times i$ entre si [Pires et al., 2011]. De posse da matriz *CSM*, efetuou-se a redução de dimensionalidade da mesma e redução de ruído, através de sua decomposição por valores singulares (2.1.2.1). Esse conjunto vetorial pode, então, ser usado como entrada para o algoritmo

classificador. O algoritmo utilizado foi o *KNN*, por ter sido reportado como o de melhor desempenho em [Pires et al., 2011].

Todos os classificadores construídos foram testados utilizando-se a biblioteca *Weka* [Hall et al., 2009]. A estratégia de validação utilizada foi a *10-fold cross-validation* e, por isso, somente grupos de monômeros com dez ou mais representantes foram utilizados. Os desempenhos das classificações foram avaliados utilizando-se as métricas de *precisão* ($precision = TP/(TP+FP)$), revocação ($recall = TP/(TP+FN)$), score *F1* (média harmônica entre a precisão e a revocação: $2(\frac{Precision \times Recall}{Precision+Recall})$) e Área Sob a Curva ROC (AUC). Os resultados serão apresentados na sessão 3.

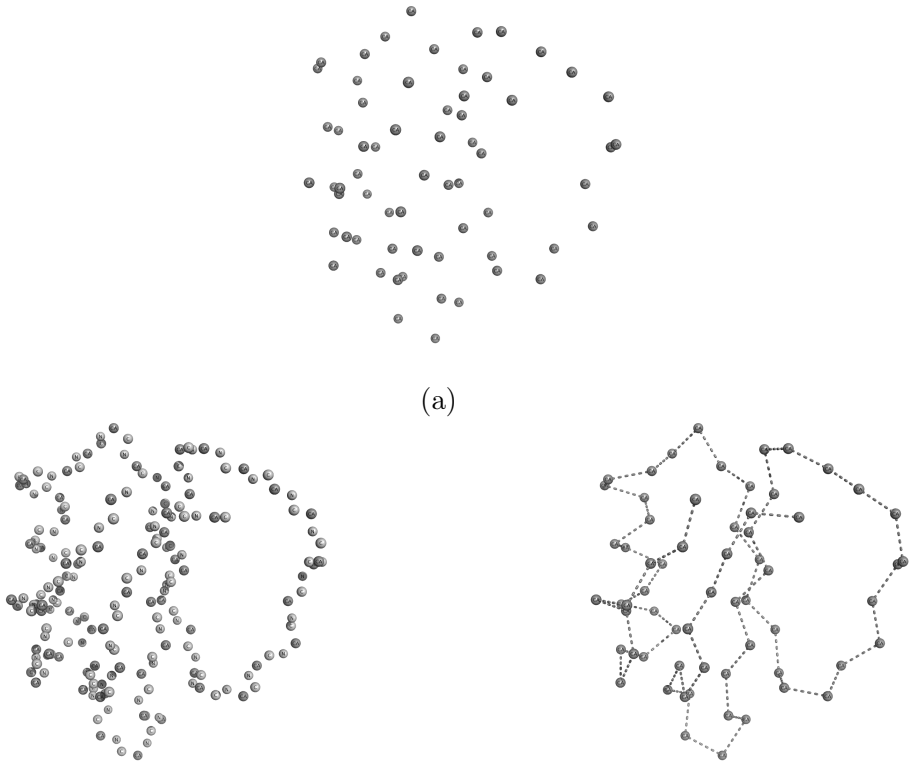
A principal contribuição do presente trabalho é a investigação do uso da geometria poligonal da cadeia principal (ou dos próprios C_α, caso sejam adicionados pontos intermediários entre esses) como uma possível assinatura estrutural. Entendemos, por esse trabalho, que a adição de pontos geométricos, reforçou essa característica geométrica da cadeia principal.

Dessa forma, é a relativa maior distância inter-pontos do modelo que utiliza somente os C_α (em geral, a distância de separação entre C_α de aminoácidos adjacentes numa proteína é de cerca de 3.8 Å [Laskowski et al., 1993]), em relação ao modelo que utiliza também os átomos C e N ($C_\alpha - C = 1.52$ Å, $C - N = 1.32$ Å e $N - C_\alpha = 1.46$ Å [Laskowski et al., 1993]), que traz inespecificidades para a sequência correta dos átomos do conjunto e, consequentemente, dificuldades para a tarefa de classificação. Essa mesma característica geométrica pode ser obtida (ou simulada) com a inclusão de pontos intermediários artificiais entre os C_α, sem prejuízos na precisão dessa classificação. A Figura 2.8 ilustra esse conceito.

Para evidenciarmos a importância da informação da geometria poligonal da cadeia principal como uma assinatura estrutural, efetuamos alguns experimentos em que foram inseridos pontos artificiais entre os pontos utilizados no modelo. Respeitamos a ordem com que os átomos aparecem na cadeia principal (seja o C_α somente ou a sequência completa $C_\alpha - C - N$), pois acreditamos que, de certa forma, essa informação de sequência é capturada de maneira indireta pela *CSM*. No corrente trabalho, efetuamos testes variando a distância entre os pontos intermediários em 0.2 Å, 0.4 Å, 0.6 Å e 0.8 Å. Optamos por esse intervalo (0.2 Å-0.8 Å), uma vez que o passo da CSM é de 0.2 Å e, na distância limite 0.8 Å, ainda é possível inserir um ponto intermediário entre os átomos do backbone, relativamente próximo ao ponto médio dessas distâncias. O algoritmo 3 formaliza a inserção dos pontos intermediários.

Como o **terceiro subgrupo de experimentos (poligonal geométrica)** é o mais importante, de modo a comparar as médias dos valores das métricas com maior segurança, executou-se cada um de seus classificadores 30 vezes. Para cada execução

de mesmo índice (exemplo: décima execução), utilizou-se a mesma semente aleatória para todos os classificadores. Para execuções de índices diferentes, utilizou-se sementes aleatórias diferentes. Uma vez garantido o uso da mesma semente para índices iguais e sementes diferentes para índices diferentes, torna-se possível a comparação de maneira mais assertiva dos valores dos resultados de cada classificador para o mesmo índice de execução. Em termos estatísticos, interpreta-se os valores dos resultados da classificação de mesmo índice de cada um dos classificadores como sendo o desempenho (valores das métricas) de um mesmo indivíduo (mesmo conjunto de cadeias a serem classificadas e mesma divisão populacional efetuada pelo *10-fold cross-validation*, uma vez que a semente aleatória é a mesma) medido após receber tratamentos diferentes (cada uma das classificações). Dessa forma, é possível efetuar um teste de hipótese para amostras pareadas onde, por definição, a hipótese estatística nula é de que as médias dos valores das métricas das 30 execuções de cada um dos classificadores são



(a) (b) (c)

Figura 2.8: Átomos da cadeia principal da estrutura *1TEC:I* do PDB. (a) Somente os Átomos C_α. Geometricamente, assemelha-se a um conjunto desordenado de pontos. (b) Átomos C_α, C e N. A característica geométrica poligonal da cadeia principal fica mais explícita. (c) Uma poligonal similar pode ser "obtida"adicionando pontos intermediários entre os átomos C_α.

Algorithm 3 Insertion of intermediate points

procedure INSERT-POINTS($\vec{p_1}, \vec{p_2}, step$)

$\vec{v} = \vec{p_2} - \vec{p_1}$

$\vec{u} = \dfrac{\vec{v}}{|\vec{v}|} \cdot step$

$points \leftarrow$ empty list

$last\_step = int\left(\dfrac{|\vec{v}|}{step}\right)$

for all i **in** 0 **to** $last\_step$ **do**
    $\vec{p} = \vec{p_1} + \vec{u} \cdot i$
    $points[i] = \vec{p}$

$points[last\_step] = \vec{p_2}$

return $points$

idênticas. Utiliza-se, então, um teste *t-Student para amostras pareadas*, com 30 graus de liberdade, para encontrar os respectivos *p-valores*. Aceita-se a hipótese nula de médias idênticas quando $p > 0.05$. Rejeita-se a hipótese nula caso contrário. Por serem muito extensas, as tabelas dos *p-valores* encontram-se nos materiais suplementares.

Capítulo 3

Resultados e discussões

3.1 Ilhas apolares

Para evidenciar a conservação posicional e volumétrica das ilhas hidrofóbicas presentes nos núcleos dos monômeros de uma mesma família, apresentamos a Tabela 3.1, que sumariza os dados dos casos ilustrativos. Lembrando que os alinhamentos foram feitos utilizando-se o algoritmo *TM-Align*.

Por esses dados, e visualmente, através das Figuras 2.4, 2.5 e 2.6, a disposição espacial globular das ilhas apolares nos monômeros das famílias estudadas. As famílias apresentaram, em geral, perfis de quantidade, volume e grau de conservação diferentes para seus respectivos arquipélagos. Porém, devido a essa mesma disposição globular, eventualmente, famílias diferentes podem apresentar a disposição de seus arquipélagos de maneira similar. Nota-se, também, que as ilhas formadas por átomos apolares, segundo o critério *Ring* ([Alexandre V. Fassio, 2017]), são mais bem definidas. Em geral, são mais conservadas, possuem maior quantidade de átomos constituintes e menor variância nessa quantidade de átomos (proporcionalmente). Conforme discutido na seção 3.2, o critério de *Ring* ([Alexandre V. Fassio, 2017]), em geral, obteve melhores resultados que o critério de *Sobolev* ([Sobolev et al., 1999]), inclusive nos experimentos de classificação.

Family	nr		Sobolev				Ring			
		Índice	Conservação(%)	Média nr átomos	Variância nr átomos	Distância	Conservação(%)	Média nr átomos	Variância nr átomos	Distância
a.1.1.2	100	0	100.0	11.68	3.2	1.29	100.0	40.55	4.43	0.56
		1	100.0	10.64	1.94	0.75	100.0	21.02	3.85	0.98
		2	99.0	8.18	2.37	1.38	100.0	10.63	5.15	1.65
		3	99.0	7.91	1.61	0.68	99.0	7.01	2.0	1.47
		4	99.0	4.85	1.74	0.88	98.0	12.5	4.34	1.26
b.47.1.1	47	0	100.0	9.32	2.97	1.62	100.0	17.02	6.95	1.53
		1	100.0	6.0	2.68	2.29	100.0	13.3	7.89	2.0
		2	97.87	5.41	3.83	1.57	100.0	11.68	3.43	1.61
		3	97.87	5.26	1.92	1.36	100.0	9.68	5.36	1.67
		4	97.87	3.93	2.17	0.8	100.0	9.23	4.14	1.3
b.6.1.1	100	0	99.0	7.27	3.35	1.91	100.0	22.77	12.67	1.9
		1	97.0	7.82	3.42	2.14	99.0	7.95	4.21	2.28
		2	96.0	4.46	3.41	1.97	99.0	7.63	3.37	2.15
		3	95.0	6.09	2.25	1.48	97.0	4.42	3.26	1.8
		4	94.0	4.09	2.31	1.19	95.0	6.29	3.08	1.48
b.60.1.1	100	0	98.0	5.03	2.15	1.61	100.0	21.16	19.56	2.75
		1	98.0	2.88	0.61	1.51	100.0	11.07	7.77	2.3
		2	97.0	6.02	3.63	1.65	97.0	4.41	2.55	1.82
		3	97.0	4.6	1.86	1.41	95.0	11.33	7.78	1.76
		4	97.0	4.03	1.91	1.28	95.0	10.73	4.37	2.43
c.47.1.1	78	0	100.0	20.24	12.57	2.79	100.0	36.28	21.67	2.34
		1	96.15	11.89	6.11	1.94	100.0	26.46	15.07	2.32
		2	93.59	8.52	6.33	2.79	91.03	18.14	10.61	2.4
		3	92.31	7.79	5.13	2.6	-	-	-	-
		4	91.03	4.08	1.71	1.73	-	-	-	-

Tabela 3.1: Arquipélogos hidrofóbicos. Comparação entre os arquipélagos gerados utilizando-se o critério de *Sobolev* e o critério *Ring*. Listamos apenas os cinco arquipélagos mais conservados por questão de acuidade visual. As demais famílias e valores encontram-se nos materiais suplementares. A coluna *Family* é o identificador da família *SCOP*. A coluna *Nr* é a quantidade de monômeros utilizados. A coluna *Índice*, o índice do arquipélogo. A coluna *Conservação(%)*, a porcentagem de conservação das ilhas que compõem o arquipélogo (100% significa que todos monômeros possuem ilhas pertencentes ao arquipélogo). A coluna *média nr átomos* é a média da quantidade de átomos que formam as ilhas daquele arquipélago. A coluna *variância nr átomos*, a variância dessa quantidade de átomos e a coluna *Distância*, a distância média dos centroides das ilhas para o centroide do arquipélago.

3.2 Classificadores

O **primeiro subgrupo de experimentos**, **polaridade atômica**, objetiva demonstrar que modelos que utilizam informações de assinaturas estruturais em nível atômico são mais conservados e discriminativos entre famílias do que o modelo que utiliza somente os C_α. Nesse grupo, quatro classificadores foram executados e comparados:

1. Utilizando-se apenas os carbonos alfa das cadeias, chamados neste trabalho de C_α, para servir de grupo controle e possibilitar a comparação com a *CSM* original.

2. Utilizando-se todos os átomos da estrutura, chamados de *All*, para possibilitar a comparação com o resultado da classificação utilizando os átomos polares e átomos apolares.

3. Utilizando-se os átomos apolares, segundo o critério de *Sobolev* (2.1.1.1), chamados de *SobNonPolar*, para averiguar a variação no grau de precisão da classificação quando utilizados os átomos apolares.

4. Utilizando-se apenas os átomos considerados como não-apolares (chamados de maneira simplificada no corrente trabalho como polares), segundo o critério de *Sobolev* (2.1.1.1), chamados de *SobPolar*, para averiguar a variação no grau de precisão da classificação quando utilizados os átomos polares.

Considerando todas as bases utilizadas, obtivemos uma melhoria de precisão quando comparada a classificação utilizando os átomos polares (*SobPolar*) com a classificação utilizando somente os C_α. Vide tabelas 3.2, 3.3 e 3.4. A classificação utilizando somente os C_α foi tomada como grupo controle, uma vez que foi reportada como a de melhor resultado em [Pires et al., 2011].

Tabela 3.2: Comparação de predição de função para a base gold-standard. Grupo controle C_α. Melhor assinatura *SobPolar*.

Superfamily	Ca				SobPolar				SobNonPolar				All			
	Prec	Recall	F1	ROC	Prec	Recall	F1	AUC	Prec	Recall	F1	AUC	Prec	Recall	F1	AUC
Amidohydrolase	0.997	0.989	0.983	0.975	**+0.3%**	**+0.8%**	**+1.2%**	**+1.5%**	+0.0%	-0.5%	-0.8%	-0.8%	-0.4%	-0.2%	-0.3%	-0.5%
Crotonase	0.977	0.971	0.968	0.964	**+2.4%**	**+3.0%**	**+3.3%**	**+3.7%**	-3.6%	-1.9%	-2.1%	-5.9%	-3.6%	-1.9%	-2.1%	-5.1%
Enolase	0.980	0.988	0.983	0.983	**+1.0%**	**+0.5%**	**+0.7%**	**+0.9%**	-0.3%	-0.1%	-0.3%	-0.1%	-0.6%	-0.6%	-1.0%	+0.0%
Haloacid Dehalogenase	0.962	0.968	0.965	0.955	**+4.0%**	**+3.3%**	**+3.6%**	**+4.7%**	+2.6%	+0.7%	+0.7%	+0.2%	+4.0%	+0.1%	+0.1%	-1.6%
Isoprenoid Synthase Type I	**1.000**	**1.000**	**1.000**	**1.000**	-2.9%	-2.9%	-3.1%	-7.0%	-8.8%	-4.6%	-4.7%	-8.3%	-8.8%	-8.8%	-9.2%	-7.9%
Vicinal Oxygen Chelate	**1.000**	**1.000**	**1.000**	**1.000**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	-4.5%	-5.0%	-5.6%	-11.7%	-2.2%	-1.1%	-1.3%	-2.0%
All	0.987	0.987	0.987	0.991	**+0.9%**	**+0.9%**	**+0.9%**	**+0.5%**	-1.0%	-1.0%	-1.1%	-0.9%	-0.9%	-0.9%	-0.9%	-0.4%

Tabela 3.3: Comparação de classificação estrutural para a base Full-SCOP. Grupo controle C_α. Melhor assinatura *SobPolar*.

SCOP Level	Ca				SobPolar				SobNonPolar				All			
	Prec	Recall	F1	ROC	Prec	Recall	F1	AUC	Prec	Recall	F1	AUC	Prec	Recall	F1	AUC
Class	0.940	0.940	0.940	0.961	**+0.7%**	**+0.7%**	**+0.7%**	**+0.5%**	-2.8%	-2.8%	-2.8%	-1.7%	-1.4%	-1.4%	-1.4%	-0.8%
Fold	0.885	0.886	0.884	0.943	**+2.9%**	**+2.9%**	**+3.1%**	**+1.4%**	-1.9%	-1.8%	-1.8%	-1.0%	-0.8%	-0.8%	-0.7%	-0.4%
Superfamily	0.876	0.877	0.876	0.938	**+3.4%**	**+3.4%**	**+3.4%**	**+1.6%**	-1.5%	-1.3%	-1.5%	-0.6%	-0.3%	-0.3%	-0.3%	-0.2%
Family	0.829	0.831	0.829	0.916	**+5.1%**	**+5.1%**	**+5.1%**	**+2.2%**	+0.4%	+0.6%	+0.4%	+0.2%	+0.6%	+0.6%	+0.6%	+0.2%

Diante do resultado obtido, em que a classificação utilizando-se os átomos polares (*SobPolar*) obteve melhor desempenho em praticamente todas as métricas comparadas, efetuamos o **segundo subgrupo de experimentos (cadeia principal)**. O objetivo foi a verificação de dois efeitos:

1. O grau de influência da cadeia principal, uma vez que ela é sabidamente polar.

2. O desempenho de classificação dos átomos cuja polaridade fosse determinada através de outro critério de polaridade, o critério *Ring* (2.1.1.1).

Tabela 3.4: Comparação de predição de função para as bases SSEs. Grupo controle C_α. Melhor assinatura *SobPolar*.

DataSet	SCOP Level	Ca				SobPolar				SobNonPolar				All			
		Prec	Recall	F1	ROC	Prec	Recall	F1	AUC	Prec	Recall	F1	AUC	Prec	Recall	F1	AUC
3SSE	Class	0.974	0.973	0.972	0.980	**+0.5%**	**+0.5%**	**+0.6%**	**+0.3%**	-1.5%	-1.6%	-1.6%	-1.2%	-0.2%	-0.3%	-0.2%	-0.3%
	Fold	0.906	0.903	0.899	0.959	**+4.2%**	**+4.4%**	**+4.9%**	**+1.4%**	+1.0%	+0.9%	+1.1%	-0.6%	+3.3%	+3.5%	+3.9%	+0.7%
	Superfamily	0.908	0.908	0.906	0.947	**+4.6%**	**+4.5%**	**+4.6%**	**+2.5%**	+0.0%	-0.6%	-0.8%	+1.1%	+3.4%	+3.3%	+3.4%	+1.8%
	Family	0.862	0.860	0.851	0.965	**+7.5%**	**+7.6%**	**+8.6%**	**+0.1%**	+3.1%	+3.1%	+3.8%	-2.8%	+6.4%	+6.3%	+7.3%	-1.3%
4SSE	Class	0.973	0.973	0.973	0.985	**+0.6%**	**+0.6%**	**+0.6%**	-0.1%	-2.0%	-1.8%	-2.0%	-2.2%	-0.2%	-0.2%	-0.2%	-0.7%
	Fold	0.923	0.921	0.920	0.959	**+2.7%**	**+3.0%**	**+3.0%**	**+1.5%**	-2.6%	-2.7%	-2.6%	-1.0%	+2.0%	+2.2%	+2.1%	+1.0%
	Superfamily	0.922	0.920	0.919	0.959	**+2.3%**	**+2.4%**	**+2.4%**	**+1.0%**	-4.7%	-4.8%	-4.9%	-2.5%	+1.0%	+1.0%	+1.0%	+0.3%
	Family	0.891	0.888	0.887	0.940	**+5.2%**	**+5.5%**	**+5.5%**	**+2.9%**	-0.8%	-0.7%	-0.8%	+0.0%	+4.7%	+5.1%	+4.8%	+2.8%
5SSE	Class	**0.964**	**0.964**	**0.964**	**0.975**	-0.5%	-0.6%	-0.6%	-0.4%	-2.2%	-2.2%	-2.2%	-1.4%	-0.9%	-0.9%	-0.9%	-0.7%
	Fold	0.927	0.926	0.925	0.963	**+1.0%**	**+1.0%**	**+1.0%**	**+0.5%**	-2.9%	-2.9%	-3.1%	-1.6%	+0.1%	+0.1%	+0.1%	-0.1%
	Superfamily	0.915	0.914	0.912	0.959	**+3.1%**	**+3.0%**	**+3.1%**	**+1.3%**	-1.3%	-1.3%	-1.4%	-1.0%	+1.3%	+1.2%	+1.3%	+0.3%
	Family	0.918	0.916	0.914	0.957	**+2.0%**	**+2.1%**	**+2.2%**	**+1.0%**	-2.1%	-1.9%	-2.1%	-0.8%	+1.0%	+1.0%	+1.1%	+0.3%
6SSE	Class	0.976	0.976	0.976	**0.991**	**+0.7%**	**+0.7%**	**+0.7%**	-0.4%	-0.7%	-0.7%	-0.7%	-1.6%	+0.2%	+0.2%	+0.2%	-0.7%
	Fold	0.943	0.941	0.941	0.969	**+2.3%**	**+2.4%**	**+2.4%**	**+1.3%**	-1.3%	-1.2%	-1.3%	-0.6%	+1.0%	+1.2%	+1.1%	+0.7%
	Superfamily	0.937	0.936	0.935	0.967	**+2.3%**	**+2.5%**	**+2.5%**	**+1.1%**	-0.4%	-0.4%	-0.4%	-0.4%	+1.4%	+1.3%	+1.3%	+0.8%
	Family	0.928	0.927	0.926	0.963	**+3.3%**	**+3.5%**	**+3.5%**	**+1.7%**	-0.5%	-0.4%	-0.4%	-0.1%	+2.3%	+2.3%	+2.3%	+1.1%

Dessa forma, utilizamos o resultado da classificação dos átomos polares como novo grupo controle (*SobPolar*) e construímos os seguintes classificadores:

1. Utilizando os átomos da cadeia principal (C_α, C, N), chamado neste trabalho de *Backbone*, para avaliar seu grau de conservação e capacidade discriminativa inter-famílias. Suprimimos o oxigênio (O) da carbonila, uma vez que sua posição geométrica envolve apenas um deslocamento em relação à posição geométrica do N.

2. Utilizando os átomos polares, segundo o critério *Ring* (2.1.1.1), chamados de *RingPolar*, para averiguar a variação no grau de precisão da classificação quando utilizados os átomos polares segundo tal critério.

3. Utilizando os átomos apolares, segundo o critério *Ring* (2.1.1.1), chamados de *RingNonPolar*, para contrastarmos o grau de precisão da classificação polar/apolar segundo esse mesmo critério (*Ring*).

4. Utilizando somente os átomos das cadeias laterais, chamado de *Side*, para possibilitar a comparação com o resultado da classificação utilizando os átomos da cadeia principal.

5. Utilizando todos os átomos da estrutura, chamado de *All*, para contrastarmos o resultado da classificação *Side* com o resultado da classificação *Backbone* e evidenciarmos a influência da cadeia principal na capacidade discriminativa desses dois outros classificadores. Esse grupo serviu, também, para contrastarmos o

resultado das classificações utilizando os átomos polares e átomos apolares, uma vez que o conjunto *All* é formado pela união dos conjuntos dos átomos polares com o conjunto dos átomos apolares.

Listamos na sessão 2.1.1.1 os átomos considerados apolares segundo cada um desses dois critérios. Reforçamos que consideramos de maneira simplificada, no corrente trabalho, como polares, todos os átomos que não fossem classificados como hidrofóbicos.

Após essa segunda rodada de experimentos, constatamos que a cadeia principal se mostrou como uma assinatura estrutural estritamente dominante em relação às demais assinaturas (vide tabelas 3.5, 3.6, 3.7, 3.8, 3.9 e 3.10). Questionou-se, então, que seria a geometria poligonal da cadeia principal o principal fator nessa melhoria discriminativa de classificação (vide 2.1.1.2 e 2.2.2.2), e não a posição geométrica dos átomos que a compõem.

Tabela 3.5: Comparação de predição de função para a base gold-standard. Grupo controle SobPolar.

Superfamily	SobPolar				Backbone				RingPolar				RingNonPolar			
	Prec	Recall	F1	ROC	Prec	Recall	F1	AUC	Prec	Recall	F1	AUC	Prec	Recall	F1	AUC
Amidohydrolase	**1.000**	**0.997**	**0.995**	**0.990**	**+0.0%**	-0.2%	-0.2%	-0.1%	**+0.0%**	-0.2%	-0.2%	-0.5%	-0.3%	-1.0%	-1.5%	-1.9%
Crotonase	**1.000**	**1.000**	**1.000**	**1.000**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	-0.6%	-0.6%	-0.6%	+0.0%	-0.6%	-0.6%	-0.6%
Enolase	**0.990**	0.993	0.990	0.992	**+0.0%**	**+0.2%**	**+0.3%**	**+0.2%**	**+0.0%**	**+0.2%**	**+0.3%**	**+0.2%**	-1.9%	-1.1%	-1.7%	-1.6%
Haloacid Dehalogenase	**1.000**	**1.000**	**1.000**	**1.000**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	-1.3%	-0.6%	-0.7%	-1.1%	**+0.0%**	-0.6%	-0.7%	-0.6%
Isoprenoid Synthase Type I	0.971	0.971	0.969	0.930	**+3.0%**	**+3.0%**	**+3.2%**	**+7.5%**	**+3.0%**	**+3.0%**	**+3.2%**	**+7.5%**	**+3.0%**	**+3.0%**	**+3.2%**	**+7.5%**
Vicinal Oxygen Chelate	**1.000**	**1.000**	**1.000**	**1.000**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	-2.2%	-1.7%	-1.9%	-3.7%
All	0.996	0.996	0.996	0.996	**+0.1%**	**+0.1%**	**+0.1%**	**+0.1%**	+0.0%	+0.0%	+0.0%	+0.0%	-0.9%	-0.9%	-0.9%	-0.6%

Tabela 3.6: Comparação de predição de função para a base Full-SCOP. Grupo controle SobPolar.

SCOP Level	SobPolar				Backbone				RingPolar				RingNonPolar			
	Prec	Recall	F1	ROC	Prec	Recall	F1	AUC	Prec	Recall	F1	AUC	Prec	Recall	F1	AUC
Class	0.947	0.947	0.947	0.966	**+3.0%**	**+3.0%**	**+3.0%**	**+1.9%**	+0.7%	+0.7%	+0.7%	+0.4%	-3.0%	-2.9%	-2.9%	-1.9%
Fold	0.911	0.912	0.911	0.956	**+4.5%**	**+4.4%**	**+4.5%**	**+2.1%**	+0.4%	+0.4%	+0.4%	+0.2%	-3.6%	-3.5%	-3.6%	-1.8%
Superfamily	0.906	0.907	0.906	0.953	**+4.6%**	**+4.5%**	**+4.6%**	**+2.2%**	+0.4%	+0.6%	+0.4%	+0.2%	-3.6%	-3.4%	-3.5%	-1.6%
Family	0.871	0.873	0.871	0.936	**+4.6%**	**+4.5%**	**+4.5%**	**+2.0%**	+0.1%	+0.1%	+0.1%	+0.0%	-2.9%	-2.7%	-3.0%	-1.3%

Até o presente momento, notam-se os seguintes resultados interessantes:

1. Os resultados da classificação utilizando-se os átomos polares (para ambos critérios) superaram os resultados do grupo controle para praticamente todas as métricas comparadas. O critério *Ring* mostrou-se ligeiramente melhor que o critério *Sobolev* para praticamente todas as métricas comparadas. Vide tabelas 3.5, 3.7 e 3.6.

Tabela 3.7: Comparação de predição de função para as bases SSEs. Grupo controle SobPolar.

DataSet	SCOP Level	SobPolar				Backbone				RingPolar				RingNonPolar			
		Prec	Recall	F1	ROC	Prec	Recall	F1	AUC	Prec	Recall	F1	AUC	Prec	Recall	F1	AUC
3SSE	Class	0.979	0.978	0.978	0.983	**+0.6%**	**+0.6%**	**+0.5%**	**+0.5%**	-0.6%	-0.5%	-0.5%	-0.3%	-2.2%	-2.1%	-2.2%	-1.5%
	Fold	0.944	0.943	0.943	**0.972**	**+0.6%**	**+0.6%**	**+0.5%**	-0.1%	+0.3%	+0.3%	+0.3%	-0.4%	+0.6%	+0.6%	+0.4%	-0.2%
	Superfamily	**0.950**	**0.949**	**0.948**	**0.971**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	-0.3%	-0.3%	-0.2%	-0.3%	**+0.0%**	**+0.0%**	-0.1%	-0.2%
	Family	0.927	0.925	0.924	**0.966**	**+0.8%**	**+0.9%**	**+0.8%**	-0.5%	+0.5%	+0.5%	+0.5%	-0.8%	+0.4%	+0.2%	+0.1%	-0.9%
4SSE	Class	0.979	0.979	0.979	0.984	**+0.7%**	**+0.6%**	**+0.7%**	**+0.5%**	+0.0%	+0.0%	+0.0%	+0.3%	-2.2%	-2.2%	-2.2%	-1.5%
	Fold	0.948	0.949	0.948	0.973	**+1.1%**	**+0.7%**	**+0.8%**	**+0.4%**	-0.4%	-0.7%	-0.6%	-0.3%	-4.3%	-4.5%	-4.5%	-2.3%
	Superfamily	0.943	0.942	0.941	0.969	**+0.8%**	**+0.6%**	**+0.7%**	**+0.4%**	-0.5%	-0.6%	-0.5%	-0.4%	-5.1%	-5.3%	-5.2%	-2.6%
	Family	0.937	0.937	0.936	0.967	**+1.7%**	**+1.5%**	**+1.6%**	**+0.9%**	-0.9%	-1.2%	-1.1%	-0.4%	-3.7%	-3.7%	-3.7%	-1.9%
5SSE	Class	0.959	0.958	0.958	0.971	**+2.1%**	**+2.2%**	**+2.2%**	**+1.6%**	+0.9%	+1.0%	+1.0%	+0.6%	-0.8%	-0.8%	-0.8%	-0.7%
	Fold	0.936	0.935	0.934	0.968	**+3.1%**	**+3.1%**	**+3.2%**	**+1.4%**	+0.6%	+0.5%	+0.6%	+0.0%	-2.1%	-2.0%	-2.0%	-1.2%
	Superfamily	0.943	0.941	0.940	0.971	**+1.8%**	**+1.9%**	**+2.0%**	**+1.0%**	-0.4%	-0.4%	-0.3%	-0.2%	-3.1%	-2.7%	-2.9%	-1.4%
	Family	0.936	0.935	0.934	0.967	**+2.5%**	**+2.6%**	**+2.6%**	**+1.2%**	-0.1%	-0.1%	-0.2%	-0.2%	-2.2%	-1.9%	-2.2%	-0.9%
6SSE	Class	0.983	0.983	0.983	0.987	**+0.8%**	**+0.8%**	**+0.8%**	**+0.7%**	+0.3%	+0.3%	+0.3%	+0.2%	-1.5%	-1.6%	-1.6%	-1.1%
	Fold	0.965	0.964	0.964	0.982	**+1.0%**	**+0.9%**	**+0.9%**	**+0.6%**	+0.0%	+0.0%	+0.0%	-0.1%	-2.1%	-2.2%	-2.2%	-1.2%
	Superfamily	0.959	0.959	0.958	0.978	**+1.4%**	**+1.1%**	**+1.3%**	**+0.6%**	+0.5%	+0.4%	+0.4%	+0.3%	-1.4%	-1.6%	-1.6%	-0.8%
	Family	0.959	0.959	0.958	0.979	**+1.1%**	**+1.0%**	**+1.0%**	**+0.4%**	+0.3%	+0.2%	+0.2%	+0.2%	-1.9%	-2.1%	-2.1%	-1.0%

Tabela 3.8: Comparação de predição de função para a base gold-standard. Grupo controle C_α. Comparação com *All* e *Side*.

Superfamily	C_α				Backbone				All				Side			
	Prec	Recall	F1	ROC	Prec	Recall	F1	AUC	Prec	Recall	F1	AUC	Prec	Recall	F1	AUC
Amidohydrolase	0.993	0.985	0.978	0.968	**+0.7%**	**+1.2%**	**+1.7%**	**+2.3%**	+0.0%	+0.5%	+0.7%	+1.2%	-0.6%	-1.6%	-2.6%	-2.9%
Crotonase	0.977	0.977	0.974	0.970	**+2.4%**	**+2.4%**	**+2.7%**	**+3.1%**	-3.6%	-2.5%	-2.7%	-6.2%	+0.0%	+1.1%	+1.3%	+0.9%
Enolase	0.984	0.989	0.983	0.980	**+0.9%**	**+0.8%**	**+1.2%**	**+1.6%**	+0.0%	+0.1%	+0.2%	+0.9%	-3.4%	-2.7%	-4.1%	-4.1%
Haloacid dehalogenase	0.949	0.961	0.957	0.943	**+5.4%**	**+4.1%**	**+4.5%**	**+6.0%**	**+5.4%**	+0.8%	+0.9%	-1.1%	**+5.4%**	+2.1%	+2.4%	+1.5%
Isoprenoid Synthase Type I	**1.000**	**1.000**	**1.000**	**1.000**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	-2.9%	-4.3%	-4.5%	-5.8%	-8.8%	-4.6%	-4.7%	-8.3%
Vicinal Oxygen Chelate	**1.000**	**1.000**	**1.000**	**1.000**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	-2.2%	-1.1%	-1.3%	-2.0%	-5.6%	-4.5%	-5.0%	-8.9%
All	0.986	0.985	0.985	0.989	**+1.2%**	**+1.3%**	**+1.3%**	**+0.8%**	-0.2%	-0.2%	-0.2%	+0.0%	-1.8%	-1.7%	-1.7%	-1.6%

Tabela 3.9: Classificação estrutural para a base Full-SCOP. Grupo controle C_α. Comparação com *All* e *Side*.

SCOP Level	C_α				Backbone				All				Side			
	Prec	Recall	F1	ROC	Prec	Recall	F1	AUC	Prec	Recall	F1	AUC	Prec	Recall	F1	AUC
Class	0.940	0.940	0.940	0.961	**+3.8%**	**+3.8%**	**+3.8%**	**+2.5%**	-1.4%	-1.4%	-1.4%	-0.8%	-5.3%	-5.3%	-5.3%	-3.3%
Fold	0.885	0.886	0.884	0.943	**+7.7%**	**+7.7%**	**+7.8%**	**+3.6%**	-0.8%	-0.8%	-0.7%	-0.4%	-8.9%	-8.8%	-8.8%	-4.2%
Superfamily	0.876	0.877	0.876	0.938	**+8.3%**	**+8.3%**	**+8.3%**	**+3.9%**	-0.3%	-0.3%	-0.3%	-0.2%	-8.9%	-8.8%	-8.9%	-4.2%
Family	0.829	0.831	0.829	0.916	**+10.3%**	**+10.1%**	**+10.1%**	**+4.5%**	+0.6%	+0.6%	+0.5%	+0.2%	-8.4%	-8.1%	-8.4%	-3.7%

Tabela 3.10: Comparação de predição de função para as bases SSEs. Grupo controle C_α. Comparação com *All* e *Side*.

DataSet	SCOP Level	C_α				Backbone				All				Side			
		Prec	Recall	F1	ROC	Prec	Recall	F1	AUC	Prec	Recall	F1	AUC	Prec	Recall	F1	AUC
3SSE	Class	0.973	0.973	0.973	0.980	**+0.9%**	**+0.8%**	**+0.8%**	**+0.7%**	-0.1%	-0.3%	-0.3%	-0.1%	-1.8%	-2.0%	-2.0%	-1.2%
	Fold	0.907	0.908	0.906	0.953	**+4.6%**	**+4.2%**	**+4.4%**	**+1.7%**	+2.4%	+2.1%	+2.2%	+0.8%	+2.6%	+2.4%	+2.6%	+0.8%
	Superfamily	0.910	0.911	0.908	0.951	**+4.3%**	**+3.8%**	**+4.2%**	**+1.9%**	+2.2%	+1.8%	+2.1%	+0.7%	+2.3%	+2.1%	+2.4%	+1.2%
	Family	0.890	0.889	0.885	0.955	**+4.4%**	**+4.3%**	**+4.6%**	**+0.3%**	+2.0%	+1.6%	+2.1%	-0.9%	+0.3%	+0.0%	+0.6%	-1.9%
4SSE	Class	0.975	0.974	0.974	0.986	**+1.4%**	**+1.4%**	**+1.4%**	**+0.6%**	-0.8%	-0.8%	-0.7%	-1.0%	-1.9%	-1.8%	-1.8%	-2.0%
	Fold	0.926	0.923	0.922	0.958	**+4.6%**	**+4.9%**	**+5.0%**	**+2.7%**	+1.2%	+1.3%	+1.3%	+0.8%	-0.4%	-0.5%	-0.5%	+0.0%
	Superfamily	0.926	0.922	0.920	0.959	**+4.0%**	**+4.3%**	**+4.5%**	**+2.1%**	+0.3%	+0.4%	+0.5%	+0.1%	-1.2%	-1.1%	-1.1%	-0.3%
	Family	0.904	0.899	0.897	0.947	**+6.7%**	**+7.2%**	**+7.4%**	**+3.7%**	+2.8%	+3.0%	+3.0%	+1.6%	+1.8%	+2.1%	+2.1%	+1.3%
5SSE	Class	0.954	0.954	0.953	0.982	**+2.8%**	**+2.8%**	**+2.9%**	**+0.7%**	+0.6%	+0.5%	+0.6%	-0.9%	-2.7%	-2.8%	-2.7%	-3.3%
	Fold	0.920	0.918	0.918	0.958	**+4.8%**	**+4.9%**	**+4.9%**	**+2.5%**	+1.6%	+1.7%	+1.7%	+0.8%	-5.9%	-5.9%	-6.1%	-3.1%
	Superfamily	0.912	0.910	0.909	0.955	**+5.2%**	**+5.4%**	**+5.4%**	**+2.6%**	+2.4%	+2.5%	+2.5%	+1.3%	-5.7%	-5.6%	-5.7%	-2.7%
	Family	0.910	0.907	0.904	0.952	**+5.4%**	**+5.7%**	**+6.0%**	**+2.7%**	+2.6%	+2.9%	+3.1%	+1.2%	-5.5%	-5.4%	-5.4%	-2.8%
6SSE	Class	0.974	0.974	0.974	0.991	**+1.5%**	**+1.5%**	**+1.5%**	**+0.2%**	+0.5%	+0.5%	+0.5%	-0.6%	-0.5%	-0.5%	-0.5%	-1.3%
	Fold	0.944	0.942	0.942	0.969	**+2.9%**	**+3.0%**	**+3.0%**	**+1.5%**	+1.2%	+1.3%	+1.3%	+0.8%	-0.2%	-0.1%	-0.1%	+0.2%
	Superfamily	0.940	0.938	0.937	0.968	**+3.2%**	**+3.3%**	**+3.3%**	**+1.5%**	+1.4%	+1.5%	+1.4%	+0.8%	+0.1%	+0.2%	+0.3%	+0.1%
	Family	0.928	0.927	0.925	0.962	**+4.3%**	**+4.3%**	**+4.5%**	**+2.2%**	+2.5%	+2.5%	+2.6%	+1.2%	+1.1%	+1.1%	+1.2%	+0.5%

2. O resultado da classificação utilizando-se os átomos apolares (para ambos critérios) ficou aquém de seus pares comparados. Acredita-se que a já comentada distribuição globular dificultaria uma diferenciação inter famílias. Vide tabelas 3.2, 3.3, 3.4, 3.5, 3.6 e 3.7.

3. Os resultados das classificações utilizando-se os átomos do backbone mostraram-se estritamente dominantes em relação aos resultados das classificações utilizando os carbonos alfa e utilizando átomos polares. Acredita-se que isso se deva à sua disposição geométrica mais poligonal. Vide tabelas 3.5, 3.6, 3.7, 3.8, 3.9, 3.10 e sessões 2.1.1.2 e 2.2.2.2.

4. A precisão das classificações utilizando-se todos os átomos da estrutura (classificador *All*) obteve melhor desempenho do que a classificação utilizando somente os átomos da cadeia lateral (classificador *Side*). Vide tabelas 3.8, 3.9 e 3.10. Acredita-se que isso ocorra porque o conjunto de pontos *All* contém os átomos da cadeia principal, ao passo que o conjunto atômico *Side* não contém os átomos do backbone. O resultado de menor desempenho da classificação *Side* em relação até mesmo os resultados das classificações utilizando somente os C_α (em geral) corrobora com nossa análise. Acredita-se que isso se deve à disposição geométrica menos poligonal do conjunto atômico *Side*.

Diante do resultado obtido por esse **segundo subgrupo de experimentos (cadeia principal)**, em que a classificação utilizando-se os átomos da cadeia principal (*Backbone*) obteve melhor desempenho em praticamente todas as métricas comparadas,

efetuamos um **terceiro subgrupo de experimentos (poligonal geométrica)** de classificação para verificar se:

1. É a característica geométrica poligonal da cadeia principal o efeito preponderante na melhoria da capacidade discriminativa dos classificadores (devido à mera inclusão de pontos intermediários entre os C_α) ou;

2. Os posicionamentos geométricos dos átomos C e N são indispensáveis para explicar a melhoria da capacidade discriminativa dos classificadores, uma vez que esses posicionamentos poderiam, inclusive, capturar indiretamente os ângulos ϕ e ψ.

O **terceiro subgrupo de experimentos (poligonal geométrica)** objetiva verificar se os posicionamentos dos átomos que compõem a cadeia principal são o principal fator de melhoria discriminativa entre famílias ou se bastaria a disposição geométrica poligonal, característica da cadeia principal. Dessa forma, construímos os seguintes classificadores:

1. Inserindo pontos intermediários fictícios entre os C_α, com distâncias de 0.2 Å, 0.4 Å, 0.6 Å e 0.8 Å, chamados neste trabalho, respectivamente, de $C_\alpha^{0.2}$, $C_\alpha^{0.4}$, $C_\alpha^{0.6}$ e $C_\alpha^{0.8}$. O objetivo é obter uma poligonal "artificial", similar à poligonal obtida pela inclusão dos átomos C e N no experimento *Backbone*. Dessa forma, obtêm-se o efeito geométrico da poligonal da cadeia principal sem a influência do posicionamento dos átomos C e N e, indiretamente, sem a influência dos ângulos ϕ e ψ (vide 2.1.1.2 e 2.2.2.2). A melhoria na precisão de classificação foi avaliada em relação ao grupo controle C_α.

2. Inserindo pontos intermediários fictícios entre os átomos da cadeia principal, com distâncias de 0.2 Å, 0.4 Å, 0.6 Å e 0.8 Å, chamados neste trabalho, respectivamente, de $Backbone^{0.2}$, $Backbone^{0.4}$, $Backbone^{0.6}$ e $Backbone^{0.8}$. O objetivo é avaliar a melhoria na precisão de classificação em relação ao grupo controle *Backbone*.

Acaso o desempenho dos classificadores da família C_α ($C_\alpha^{0.2}$, $C_\alpha^{0.4}$, $C_\alpha^{0.6}$ e $C_\alpha^{0.8}$) fosse similar aos classificadores da família *Backbone* ($Backbone^{0.2}$, $Backbone^{0.4}$, $Backbone^{0.6}$ e $Backbone^{0.8}$), isso sugeriria que é a geometria poligonal da cadeia principal, e não o posicionamento dos átomos C e N, o principal fator de melhoria da capacidade discriminativa dos classificadores.

Por esse **terceiro subgrupo de experimentos (poligonal geométrica)**, obtivemos os seguintes resultados:

1. Em média, a distância inter-pontos de melhor desempenho para a inclusão de pontos intermediários foi a de 0.8 Å (vide tabelas 3.11, 3.12, 3.14, 3.15, 3.17, 3.18). Na maioria dos casos, os resultados das classificações incluindo pontos intermediários com distâncias de 0.8 Å ($C_\alpha^{0.8}$ e $Backbone^{0.8}$) mostraram-se dominantes em relação aos resultados das classificações sem a inclusão de pontos intermediários (C_α e $Backbone$, respectivamente). Esse comportamento pode ser observado tanto para o $C_\alpha^{0.8}$ quanto para o $Backbone^{0.8}$, sendo mais evidente para o $C_\alpha^{0.8}$, uma vez que a geometria poligonal do grupo atômico C_α é menor que do grupo $Backbone$. Por esses experimentos, podemos observar que a inclusão de pontos intermediários com distância inter-pontos de 0.8 Å foi, em geral, a melhor assinatura estrutural utilizada para a tarefa de classificação. Essa dominância é sujeita a alguma flutuação estatística, dependendo da base utilizada e do experimento.

2. Não houve diferenças muito significativas entre o desempenho dos classificadores $C_\alpha^{0.8}$ e $Backbone$, tampouco os classificadores $C_\alpha^{0.8}$ e $Backbone^{0.8}$ (os melhores classificadores). Vide tabelas 3.13, 3.16 e 3.19. Dessa forma, evidencia-se que as posições dos átomos C e N (o que captura de maneira indireta os ângulos ϕ e ψ) são fatores menos determinantes no incremento da qualidade da classificação que o fortalecimento do caráter geométrico poligonal da cadeia principal (vide 2.1.1.2 e 2.2.2.2).

Diante do resultado obtido por esse **terceiro subgrupo de experimentos (poligonal geométrica)**, nota-se que as posições dos átomos C e N são fatores para a melhoria do grau discriminativo do modelo que utiliza os átomos do $Backbone$, mas a geometria poligonal da cadeia principal é um fator preponderante.

Como esse subgrupo de experimentos é o mais importante, de modo a comparar as médias dos valores das métricas com maior segurança, executamos cada um de seus classificadores 30 vezes. Utilizamos, então, um teste *t-Student para amostras pareadas*, com 30 graus de liberdade, para encontrar os respectivos *p-valores* (vide 2.2.2.2). Aceitamos a hipótese nula de médias idênticas quando $p > 0.05$. Rejeitamos a hipótese nula caso contrário. Por serem muito extensas, as tabelas dos *p-valores* encontram-se nos materiais suplementares. Conforme ficou figurado pelas tabelas de $p - valores$ (materiais suplementares, 5.1, 5.2, 5.3, 5.4, 5.5, 5.6, 5.7, 5.8, 5.9, 5.10, 5.11, 5.12, 5.13, 5.14, 5.15, 5.16, 5.17, 5.18 , 5.19, 5.20, 5.21, 5.22, 5.23, 5.24), só houve aceitação da hipótese nula quando os valores das médias das métricas comparadas estavam muito próximas. Para todas as hipóteses nulas aceitas, a diferença percentual entre os valores das médias obtidas foi menor ou igual a 0.3%. Em geral, os $p-valores$

encontrados foram extremamente baixos, e assim, as hipóteses nulas, de que as médias seriam iguais, puderam ser rejeitadas. Não houve, portanto, diferenças significativas entre o desempenho dos classificadores $C_\alpha^{0.8}$ e *Backbone*, nem entre os classificadores $C_\alpha^{0.8}$ e $Backbone^{0.8}$ (os melhores classificadores). Vide tabelas 3.13, 3.16 e 3.19).

A análise dos resultados da classificação sugere que a assinatura da geometria poligonal da cadeia principal das proteínas de uma mesma família é melhor diferencial de classificação que a assinatura que utiliza somente seus carbonos alfas. Ou seja, o caráter sequencial da cadeia polipeptídica e sua caracterítica poligonal são mais relevantes que apenas o empacotamento dos resíduos, como usado na *CSM* original.

Os átomos polares mostraram-se melhores como assinatura estrutural para comparações inter- famílias quando comparados com os átomos apolares. Apesar dos átomos apolares mostrarem possuir conservação por família, eles não necessariamente diferem significativamente entre famílias. Uma vez que os átomos apolares possuem a capacidade de se aglomerarem formando ilhas, provavelmente, famílias diferentes guardam alguma similaridade nas características dessas ilhas. Acreditamos que isso (em parte) se deve à sua distribuição mais globular, ao passo que a distribuição polar é mais poligonal (seguindo o backbone), o que explicaria o baixo desempenho do classificador apolar em relação ao classificador polar. Essa característica globular dificultaria uma tentativa de classificação seguindo apenas critérios estruturais, uma vez que famílias diferentes podem vir a guardar similaridades em suas assinaturas apolares.

Concluímos, assim, que a melhoria da precisão de classificação obtida pela CSM, ao se utilizar um modelo com os átomos da cadeia principal (C_α, C, N), em relação ao uso exclusivo dos carbonos alfa (C_α) dessa mesma cadeia, deve-se, principalmente, pela evidenciação do caráter geométrico poligonal da cadeia, caráter esse obtido pela inclusão dos átomos C e N. Dessa forma, é a relativa maior distância inter-pontos do modelo que utiliza somente os C_α, em relação ao modelo que utiliza também os átomos C e N, que traz inespecificidades para a sequência correta dos átomos do conjunto e, consequentemente, dificuldades para a tarefa de classificação. Vide seções 2.1.1.2 e 2.2.2.2 e figura 2.8.

Essa mesma geometria poligonal pode ser obtida (ou simulada) com a inclusão de pontos intermediários artificiais entre os C_α, sem prejuízos na precisão da classificação (em relação ao *Backbone*). Dessa forma, pudemos evidenciar que não são as posições dos átomos C e N (o que captura de maneira indireta os ângulos ϕ e ψ) os fatores determinantes no incremento da qualidade da classificação, mas, sim, o fortalecimento do caráter poligonal da cadeia principal. Demonstrou-se, também, que o grau de influência dos ângulos ϕ e ψ na capacidade discriminativa dos classificadores, além de menos significativa que a disposição poligonal da cadeia principal, é dependente da

adição de pontos intermediários artificiais para se fazer perceber (vide o resultado dos classificadores $Backbone^{0.8}$).

Pires [Pires et al., 2011] reporta que experiências foram conduzidas com outros centroides em vez do C_α, como o C_β ou o último átomo pesado (LHA) da cadeia lateral. O C_α obteve melhor desempenho em todos os experimentos, um fato que ele registrou como "exigindo uma investigação mais aprofundada". Pelo corrente trabalho, acreditamos que esses centroides obtiveram mais baixo desempenho, justamente, por estarem ainda mais afastados da disposição geométrica poligonal da cadeia principal.

Tabela 3.11: Comparação de predição de função para a base gold-standard. Pontos intermediários. Grupo controle C_α.

Superfamily	C_α				$C_\alpha^{0.2}$				$C_\alpha^{0.4}$				$C_\alpha^{0.6}$				$C_\alpha^{0.8}$			
	Prec	Recall	F1	ROC	Prec	Recall	F1	AUC	Prec	Recall	F1	AUC	Prec	Recall	F1	AUC	Prec	Recall	F1	AUC
Amidohydrolase	0.993	0.985	0.978	0.968	+0.4%	+0.8%	+1.2%	+1.7%	+0.4%	+0.8%	+1.2%	+1.7%	+0.7%	+1.0%	+1.5%	+2.2%	**+0.7%**	**+1.2%**	**+1.7%**	**+2.3%**
Crotonase	0.977	0.977	0.974	0.970	+0.0%	+1.1%	+1.3%	+1.1%	+1.1%	+1.7%	+2.1%	+2.1%	+0.0%	+1.1%	+1.3%	+1.1%	**+2.4%**	**+2.4%**	**+2.7%**	**+3.1%**
Enolase	0.984	0.989	0.983	0.980	**+0.6%**	+0.4%	+0.7%	+1.2%	**+0.6%**	+0.4%	+0.7%	+1.2%	**+0.6%**	**+0.6%**	**+1.0%**	**+1.4%**	**+0.6%**	+0.4%	+0.7%	+1.2%
Haloacid Dehalogenase	0.949	0.961	0.957	0.943	**+5.4%**	+2.7%	+3.0%	+4.7%	+5.4%	+3.4%	+3.8%	**+5.4%**	**+5.4%**	+2.7%	+3.0%	+4.7%	**+5.4%**	**+4.1%**	**+4.5%**	**+6.0%**
Isoprenoid Synthase Type I	**1.000**	**1.000**	**1.000**	**1.000**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	-2.9%	-2.9%	-3.1%	-7.0%	-2.9%	-2.9%	-3.1%	-7.0%
Vicinal Oxygen Chelate	**1.000**	**1.000**	**1.000**	**1.000**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**
All	0.986	0.985	0.985	0.989	+0.7%	+0.8%	+0.8%	+0.6%	+0.8%	+0.9%	+0.9%	+0.6%	+0.7%	+0.8%	+0.8%	+0.7%	**+1.0%**	**+1.1%**	**+1.1%**	**+0.7%**

Tabela 3.12: Comparação de predição de função para a base gold-standard. Pontos intermediários. Grupo controle Backbone.

Superfamily	*Backbone*				$Backbone^{0.2}$				$Backbone^{0.4}$				$Backbone^{0.6}$				$Backbone^{0.8}$			
	Prec	Recall	F1	ROC	Prec	Recall	F1	AUC	Prec	Recall	F1	AUC	Prec	Recall	F1	AUC	Prec	Recall	F1	AUC
Amidohydrolase	**1.000**	**0.997**	**0.995**	**0.990**	**+0.0%**	-0.2%	-0.2%	-0.5%	**+0.0%**	-0.2%	-0.2%	-0.5%	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	-0.4%	-0.5%	-0.9%
Crotonase	**1.000**	**1.000**	**1.000**	**1.000**	-1.2%	-0.6%	-0.6%	-1.0%	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**
Enolase	**0.993**	**0.997**	**0.995**	**0.996**	-0.3%	-0.2%	-0.2%	-0.2%	-0.3%	-0.2%	-0.2%	-0.2%	-0.3%	-0.2%	-0.2%	-0.2%	-0.6%	-0.4%	-0.5%	-0.5%
Haloacid Dehalogenase	**1.000**	**1.000**	**1.000**	**1.000**	**+0.0%**	-0.6%	-0.7%	-0.6%	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**
Isoprenoid Synthase TypeI	**1.000**	**1.000**	**1.000**	**1.000**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	-1.4%	-1.5%	-4.4%	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**
Vicinal Oxygen Chelate	**1.000**	**1.000**	**1.000**	**1.000**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**
All	**0.998**	**0.998**	**0.998**	**0.997**	-0.2%	-0.2%	-0.2%	-0.1%	-0.1%	-0.1%	-0.1%	-0.1%	-0.1%	-0.1%	-0.1%	+0.0%	-0.2%	-0.2%	-0.2%	-0.2%

Tabela 3.13: Comparação de predição de função para a base gold-standard. Melhores resultados. Grupo controle Backbone. O grupo C_α foi repetido para manter o padrão de exibição da tabela.

Superfamily	*Backbone*				$Backbone^{0.8}$				C_α				$C_\alpha^{0.8}$			
	Prec	Recall	F1	ROC	Prec	Recall	F1	AUC	Prec	Recall	F1	AUC	Prec	Recall	F1	AUC
Amidohydrolase	**1.000**	**0.997**	**0.995**	**0.990**	**+0.0%**	-0.4%	-0.5%	-0.9%	-0.7%	-1.2%	-1.7%	-2.2%	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**
Crotonase	**1.000**	**1.000**	**1.000**	**1.000**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	-2.3%	-2.3%	-2.6%	-3.0%	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**
Enolase	**0.993**	**0.997**	**0.995**	**0.996**	-0.6%	-0.4%	-0.5%	-0.5%	-0.9%	-0.8%	-1.2%	-1.6%	-0.3%	-0.4%	-0.5%	-0.4%
Haloacid Dehalogenase	**1.000**	**1.000**	**1.000**	**1.000**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	-5.1%	-3.9%	-4.3%	-5.7%	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**
Isoprenoid Synthase TypeI	**1.000**	**1.000**	**1.000**	**1.000**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	-2.9%	-2.9%	-3.1%	-7.0%
Vicinal Oxygen Chelate	**1.000**	**1.000**	**1.000**	**1.000**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**
All	**0.998**	**0.998**	**0.998**	**0.997**	-0.2%	-0.2%	-0.2%	-0.2%	-1.2%	-1.3%	-1.3%	-0.8%	-0.2%	-0.2%	-0.2%	-0.1%

Tabela 3.14: Classificação estrutural para a base Full-SCOP. Pontos intermediários. Grupo controle C_α.

Superfamily	C_α				$C_\alpha^{0.2}$				$C_\alpha^{0.4}$				$C_\alpha^{0.6}$				$C_\alpha^{0.8}$			
	Prec	Recall	F1	ROC	Prec	Recall	F1	AUC	Prec	Recall	F1	AUC	Prec	Recall	F1	AUC	Prec	Recall	F1	AUC
Class	0.940	0.940	0.940	0.961	+3.3%	+3.3%	+3.3%	+2.2%	+3.6%	+3.6%	+3.6%	+2.3%	+3.6%	+3.7%	+3.7%	+2.4%	**+3.8%**	**+3.8%**	**+3.8%**	**+2.5%**
Fold	0.885	0.886	0.884	0.943	+6.9%	+6.8%	+7.0%	+3.2%	+7.3%	+7.2%	+7.4%	+3.4%	+7.6%	+7.4%	+7.7%	+3.5%	**+7.9%**	**+7.8%**	**+8.0%**	**+3.6%**
Superfamily	0.876	0.877	0.876	0.938	+7.4%	+7.4%	+7.4%	+3.5%	+7.9%	+7.9%	+7.9%	+3.7%	+8.1%	+8.1%	+8.1%	+3.8%	**+8.6%**	**+8.4%**	**+8.6%**	**+4.1%**
Family	0.829	0.831	0.829	0.916	+8.8%	+8.7%	+8.8%	+3.8%	+9.7%	+9.4%	+9.5%	+4.1%	+9.9%	+9.7%	+9.9%	+4.4%	**+10.4%**	**+10.2%**	**+10.4%**	**+4.6%**

Tabela 3.15: Classificação estrutural para a base Full-SCOP. Pontos intermediários. Grupo controle *Backbone*.

Superfamily	*Backbone*				$Backbone^{0.2}$				$Backbone^{0.4}$				$Backbone^{0.6}$				$Backbone^{0.8}$			
	Prec	Recall	F1	ROC	Prec	Recall	F1	AUC	Prec	Recall	F1	AUC	Prec	Recall	F1	AUC	Prec	Recall	F1	AUC
Class	0.976	0.976	0.976	**0.985**	-0.2%	-0.2%	-0.2%	-0.1%	-0.2%	-0.2%	-0.2%	-0.1%	-0.1%	-0.1%	-0.1%	-0.1%	**+0.0%**	**+0.1%**	**+0.1%**	**+0.0%**
Fold	0.953	0.954	0.953	**0.977**	-0.3%	-0.3%	-0.3%	-0.2%	-0.2%	-0.3%	-0.2%	-0.2%	-0.1%	-0.2%	-0.1%	-0.1%	**+0.2%**	**+0.1%**	**+0.1%**	**+0.0%**
Superfamily	0.949	0.950	0.949	0.975	-0.3%	-0.3%	-0.3%	-0.2%	-0.2%	-0.3%	-0.2%	-0.1%	-0.2%	-0.2%	-0.2%	-0.1%	**+0.2%**	**+0.1%**	**+0.2%**	**+0.1%**
Family	0.914	0.915	0.913	0.957	-0.4%	-0.5%	-0.4%	-0.2%	-0.4%	-0.4%	-0.4%	-0.2%	-0.3%	-0.3%	-0.2%	-0.1%	**+0.2%**	**+0.1%**	**+0.2%**	**+0.1%**

Tabela 3.16: Classificação estrutural para a base Full-SCOP. Melhores resultados. Grupo controle Backbone. O grupo C_α foi repetido para manter o padrão de exibição da tabela.

Superfamily	*Backbone*				$Backbone^{0.8}$				C_α				$C_\alpha^{0.8}$			
	Prec	Recall	F1	ROC	Prec	Recall	F1	AUC	Prec	Recall	F1	AUC	Prec	Recall	F1	AUC
Class	**0.976**	0.976	0.976	**0.985**	**+0.0%**	**+0.1%**	**+0.1%**	**+0.0%**	-3.7%	-3.7%	-3.7%	-2.4%	**+0.0%**	+0.0%	+0.0%	**+0.0%**
Fold	0.953	0.954	0.953	**0.977**	**+0.2%**	**+0.1%**	+0.1%	**+0.0%**	-7.1%	-7.1%	-7.2%	-3.5%	**+0.2%**	**+0.1%**	**+0.2%**	**+0.0%**
Superfamily	0.949	0.950	0.949	0.975	**+0.2%**	**+0.1%**	**+0.2%**	**+0.1%**	-7.7%	-7.7%	-7.7%	-3.8%	**+0.2%**	**+0.1%**	**+0.2%**	**+0.1%**
Family	0.914	0.915	0.913	0.957	**+0.2%**	**+0.1%**	**+0.2%**	**+0.1%**	-9.3%	-9.2%	-9.2%	-4.3%	+0.1%	+0.1%	+0.2%	+0.1%

Tabela 3.17: Comparação de predição de função para as bases SSEs. Pontos intermediários. Grupo controle C_α.

DataSet	SCOP Level	C_α				$C_\alpha^{0.2}$				$C_\alpha^{0.4}$				$C_\alpha^{0.6}$				$C_\alpha^{0.8}$			
		Prec	Recall	F1	ROC	Prec	Recall	F1	AUC	Prec	Recall	F1	AUC	Prec	Recall	F1	AUC	Prec	Recall	F1	AUC
3SSE	Class	0.973	0.973	0.973	0.980	+1.2%	+1.1%	+1.0%	+0.8%	+1.2%	+1.1%	+1.0%	+0.8%	**+1.4%**	**+1.4%**	**+1.3%**	**+1.0%**	+0.9%	+0.8%	+0.7%	+0.6%
	Fold	0.907	0.908	0.906	0.953	**+4.0%**	**+3.9%**	**+4.0%**	**+1.5%**	+3.0%	+2.8%	+2.6%	+1.7%	+2.6%	+2.4%	+2.3%	+1.8%	+2.4%	+2.4%	+2.3%	+1.7%
	Superfamily	0.910	0.911	0.908	0.951	**+3.6%**	**+3.5%**	**+3.9%**	+1.7%	+2.4%	+2.1%	+2.1%	+2.0%	+2.5%	+2.1%	+2.0%	**+3.0%**	+2.9%	+2.4%	+2.3%	+2.6%
	Family	0.890	0.889	0.885	0.955	**+5.2%**	**+5.2%**	**+5.5%**	+0.7%	+4.0%	+4.0%	+4.3%	+0.0%	+3.8%	+3.7%	+4.0%	+0.8%	+4.7%	+4.6%	+4.9%	**+1.0%**
4SSE	Class	0.975	0.974	0.974	0.986	+1.7%	+1.8%	+1.8%	+0.8%	+1.6%	+1.7%	+1.7%	+0.8%	+1.8%	+2.0%	+2.0%	+0.9%	+1.4%	+1.5%	+1.5%	+0.8%
	Fold	0.926	0.923	0.922	0.958	**+5.2%**	**+5.4%**	**+5.4%**	**+2.9%**	+5.0%	+5.2%	+5.3%	+2.8%	+5.0%	+5.2%	+5.3%	+2.8%	+4.5%	+4.8%	+4.9%	+2.6%
	Superfamily	0.926	0.922	0.920	0.959	**+4.4%**	**+4.7%**	**+4.9%**	**+2.4%**	+4.2%	+4.6%	+4.8%	+2.3%	**+4.4%**	**+4.7%**	**+4.9%**	**+2.4%**	+4.3%	+4.7%	+4.9%	+2.3%
	Family	0.904	0.899	0.897	0.947	+6.9%	+7.3%	+7.5%	+3.8%	**+7.2%**	**+7.6%**	**+7.8%**	**+4.0%**	+7.0%	+7.5%	+7.7%	**+4.0%**	+6.9%	+7.3%	+7.6%	+3.9%
5SSE	Class	0.954	0.954	0.953	0.982	+2.6%	+2.6%	+2.7%	+0.4%	+2.9%	+2.9%	+3.0%	+0.5%	**+3.1%**	**+3.1%**	**+3.3%**	**+0.8%**	+2.9%	+2.9%	+3.0%	+0.6%
	Fold	0.920	0.918	0.918	0.958	+3.9%	+4.1%	+4.0%	+1.9%	+4.5%	+4.6%	+4.5%	+2.2%	+4.8%	+4.9%	+4.9%	**+2.4%**	**+4.9%**	**+5.0%**	**+5.0%**	**+2.4%**
	Superfamily	0.912	0.910	0.909	0.955	+4.8%	+5.1%	+5.1%	+2.5%	+4.9%	+5.2%	+5.2%	+2.5%	**+5.4%**	**+5.5%**	**+5.5%**	**+2.7%**	**+5.4%**	**+5.5%**	**+5.5%**	**+2.7%**
	Family	0.910	0.907	0.904	0.952	+4.8%	+5.2%	+5.4%	+2.4%	+4.9%	+5.2%	+5.4%	+2.6%	**+5.4%**	+5.6%	**+5.9%**	+2.7%	**+5.4%**	**+5.7%**	**+5.9%**	**+2.9%**
6SSE	Class	0.974	0.974	0.974	0.991	+1.8%	+1.8%	+1.8%	+0.3%	+1.8%	+1.8%	+1.8%	+0.5%	+1.7%	+1.7%	+1.7%	+0.4%	**+2.1%**	**+2.1%**	**+2.1%**	**+0.6%**
	Fold	0.944	0.942	0.942	0.969	+3.0%	+3.1%	+3.1%	+1.8%	+3.5%	+3.6%	+3.6%	+1.9%	**+3.7%**	**+3.9%**	**+3.9%**	+2.0%	+3.6%	+3.7%	+3.8%	**+2.1%**
	Superfamily	0.940	0.938	0.937	0.968	+3.1%	+3.3%	+3.3%	+1.7%	+3.6%	+3.7%	+3.8%	+1.8%	+3.8%	+3.9%	+4.1%	+2.0%	**+4.0%**	**+4.2%**	**+4.3%**	**+2.1%**
	Family	0.928	0.927	0.925	0.962	+4.0%	+4.0%	+4.1%	+2.1%	+4.5%	+4.5%	+4.6%	+2.2%	+4.8%	+4.9%	+5.1%	+2.4%	**+5.1%**	**+5.1%**	**+5.3%**	**+2.6%**

Tabela 3.18: Comparação de predição de função para as bases SSEs. Pontos intermediários. Grupo controle *Backbone*.

DataSet	SCOP Level	*Backbone*				*Backbone*$^{0.2}$				*Backbone*$^{0.4}$				*Backbone*$^{0.6}$				*Backbone*$^{0.8}$			
		Prec	Recall	F1	ROC	Prec	Recall	F1	AUC	Prec	Recall	F1	AUC	Prec	Recall	F1	AUC	Prec	Recall	F1	AUC
3SSE	Class	0.982	0.981	0.981	0.987	+0.0%	+0.0%	-0.1%	-0.1%	+0.0%	+0.0%	-0.1%	-0.1%	**+0.3%**	**+0.3%**	**+0.2%**	**+0.2%**	+0.0%	+0.0%	+0.0%	+0.0%
	Fold	**0.949**	0.946	0.946	0.969	-1.4%	-1.2%	-1.4%	-0.7%	-2.2%	-2.0%	-2.3%	-0.1%	-1.1%	-0.8%	-1.0%	-0.6%	**+0.0%**	**+0.3%**	**+0.2%**	**+0.2%**
	Superfamily	**0.949**	0.946	0.946	0.969	-1.8%	-1.7%	-2.0%	-0.3%	-2.4%	-2.2%	-2.6%	+0.0%	-1.1%	-0.8%	-1.0%	-0.5%	**+0.0%**	**+0.3%**	**+0.2%**	**+0.2%**
	Family	0.929	0.927	0.926	0.958	+0.1%	+0.3%	+0.2%	+0.0%	+0.2%	+0.3%	+0.2%	+0.0%	+0.2%	+0.3%	+0.3%	+0.0%	**+0.5%**	**+0.6%**	**+0.5%**	**+0.2%**
4SSE	Class	0.989	0.988	0.988	0.992	**+0.2%**	**+0.3%**	**+0.3%**	**+0.2%**	**+0.2%**	**+0.3%**	**+0.3%**	**+0.2%**	+0.1%	+0.2%	+0.2%	+0.1%	+0.0%	+0.1%	+0.1%	+0.1%
	Fold	0.969	0.968	0.968	0.984	+0.4%	+0.4%	+0.4%	+0.2%	+0.4%	+0.3%	+0.3%	+0.1%	+0.3%	+0.2%	+0.2%	+0.0%	**+1.0%**	**+0.9%**	**+0.9%**	**+0.5%**
	Superfamily	0.963	0.962	0.961	0.979	+0.4%	+0.3%	+0.4%	+0.2%	+0.4%	+0.3%	+0.4%	+0.2%	+0.4%	+0.3%	+0.4%	+0.3%	**+0.9%**	**+0.8%**	**+0.9%**	**+0.5%**
	Family	0.965	0.964	0.963	0.982	+0.4%	+0.3%	+0.4%	+0.3%	+0.4%	+0.4%	+0.5%	+0.4%	+0.2%	+0.1%	+0.2%	+0.2%	**+0.8%**	**+0.7%**	**+0.8%**	**+0.4%**
5SSE	Class	0.981	0.981	0.981	0.989	+0.0%	+0.0%	+0.0%	-0.3%	+0.0%	+0.0%	+0.0%	-0.2%	+0.0%	+0.0%	+0.0%	+0.1%	**+0.4%**	**+0.4%**	**+0.4%**	**+0.1%**
	Fold	**0.964**	**0.963**	**0.963**	**0.982**	-0.5%	-0.5%	-0.5%	-0.4%	-0.2%	-0.2%	-0.2%	-0.3%	-0.4%	-0.3%	-0.4%	-0.3%	-0.1%	-0.1%	-0.1%	-0.3%
	Superfamily	**0.959**	**0.959**	**0.958**	**0.980**	-0.2%	-0.3%	-0.3%	-0.1%	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	-0.1%	-0.1%	**+0.0%**	+0.2%	+0.2%	+0.2%	+0.1%
	Family	**0.959**	**0.959**	**0.958**	**0.978**	-0.4%	-0.5%	-0.6%	-0.2%	-0.2%	-0.3%	-0.3%	**+0.0%**	-0.2%	-0.3%	-0.3%	-0.2%	**+0.0%**	**+0.0%**	-0.1%	**+0.0%**
6SSE	Class	0.989	0.989	0.989	0.993	+0.2%	+0.2%	+0.2%	+0.3%	+0.2%	+0.2%	+0.2%	+0.3%	**+0.3%**	**+0.3%**	**+0.3%**	**+0.4%**	+0.2%	+0.2%	+0.2%	+0.2%
	Fold	0.971	0.970	0.970	0.984	+0.6%	+0.6%	+0.6%	+0.3%	+0.5%	+0.6%	+0.6%	+0.3%	**+0.9%**	**+0.9%**	**+0.9%**	**+0.5%**	+0.7%	+0.8%	+0.8%	+0.4%
	Superfamily	0.970	0.969	0.968	0.983	+0.4%	+0.4%	+0.5%	+0.3%	**+0.5%**	+0.6%	+0.6%	+0.4%	**+0.5%**	+0.5%	+0.6%	+0.3%	**+0.5%**	**+0.6%**	**+0.6%**	**+0.4%**
	Family	0.968	0.967	0.967	0.983	+0.2%	+0.2%	+0.2%	+0.0%	+0.2%	+0.3%	+0.2%	+0.1%	**+0.5%**	**+0.5%**	**+0.4%**	**+0.2%**	**+0.4%**	**+0.4%**	+0.4%	**+0.2%**

Tabela 3.19: Comparação de predição de função para as bases SSEs. Melhores resultados. Grupo controle Backbone. O grupo C_α foi repetido para manter o padrão de exibição da tabela.

DataSet	SCOP Level	*Backbone*				*Backbone*$^{0.8}$				C_α				$C_\alpha^{0.8}$			
		Prec	Recall	F1	ROC	Prec	Recall	F1	AUC	Prec	Recall	F1	AUC	Prec	Recall	F1	AUC
3SSE	Class	**0.982**	**0.981**	**0.981**	**0.987**	**+0.0%**	**+0.0%**	**+0.0%**	**+0.0%**	-0.9%	-0.8%	-0.8%	-0.7%	**+0.0%**	**+0.0%**	-0.1%	-0.1%
	Fold	**0.949**	0.946	0.946	0.969	**+0.0%**	**+0.3%**	**+0.2%**	**+0.2%**	-4.4%	-4.0%	-4.2%	-1.7%	-2.1%	-1.7%	-2.0%	+0.0%
	Superfamily	**0.949**	0.946	0.946	0.969	**+0.0%**	**+0.3%**	**+0.2%**	**+0.2%**	-4.1%	-3.7%	-4.0%	-1.9%	-1.4%	-1.4%	-1.8%	+0.7%
	Family	0.929	0.927	0.926	0.958	**+0.5%**	**+0.6%**	**+0.5%**	+0.2%	-4.2%	-4.1%	-4.4%	-0.3%	+0.3%	+0.3%	+0.2%	**+0.7%**
4SSE	Class	**0.989**	0.988	0.988	0.992	**+0.0%**	**+0.1%**	**+0.1%**	+0.1%	-1.4%	-1.4%	-1.4%	-0.6%	**+0.0%**	**+0.1%**	**+0.1%**	**+0.2%**
	Fold	0.969	0.968	0.968	0.984	**+1.0%**	**+0.9%**	**+0.9%**	**+0.5%**	-4.4%	-4.6%	-4.8%	-2.6%	-0.1%	-0.1%	-0.1%	-0.1%
	Superfamily	0.963	0.962	0.961	0.979	**+0.9%**	**+0.8%**	**+0.9%**	**+0.5%**	-3.8%	-4.2%	-4.3%	-2.0%	+0.3%	+0.3%	+0.4%	+0.2%
	Family	0.965	0.964	0.963	0.982	**+0.8%**	**+0.7%**	**+0.8%**	**+0.4%**	-6.3%	-6.7%	-6.9%	-3.6%	+0.1%	+0.1%	+0.2%	+0.2%
5SSE	Class	0.981	0.981	0.981	0.989	**+0.4%**	**+0.4%**	**+0.4%**	**+0.1%**	-2.8%	-2.8%	-2.9%	-0.7%	+0.1%	+0.1%	+0.1%	-0.1%
	Fold	0.964	0.963	0.963	**0.982**	-0.1%	-0.1%	-0.1%	-0.3%	-4.6%	-4.7%	-4.7%	-2.4%	**+0.1%**	**+0.1%**	**+0.1%**	-0.1%
	Superfamily	0.959	0.959	0.958	0.980	**+0.2%**	**+0.2%**	**+0.2%**	**+0.1%**	-4.9%	-5.1%	-5.1%	-2.6%	**+0.2%**	+0.1%	+0.1%	+0.1%
	Family	**0.959**	**0.959**	**0.958**	0.978	**+0.0%**	**+0.0%**	-0.1%	+0.0%	-5.1%	-5.4%	-5.6%	-2.7%	**+0.0%**	**+0.0%**	-0.1%	**+0.2%**
6SSE	Class	0.989	0.989	0.989	0.993	+0.2%	+0.2%	+0.2%	+0.2%	-1.5%	-1.5%	-1.5%	-0.2%	**+0.5%**	**+0.5%**	**+0.5%**	**+0.4%**
	Fold	0.971	0.970	0.970	0.984	**+0.7%**	**+0.8%**	**+0.8%**	+0.4%	-2.8%	-2.9%	-2.9%	-1.5%	**+0.7%**	+0.7%	**+0.8%**	**+0.5%**
	Superfamily	0.970	0.969	0.968	0.983	+0.5%	+0.6%	+0.6%	+0.4%	-3.1%	-3.2%	-3.2%	-1.5%	**+0.8%**	**+0.8%**	**+0.9%**	**+0.5%**
	Family	0.968	0.967	0.967	0.983	+0.4%	+0.4%	+0.4%	+0.2%	-4.1%	-4.1%	-4.3%	-2.1%	**+0.7%**	**+0.7%**	**+0.7%**	**+0.4%**

Capítulo 4

Conclusões e trabalhos futuros

A motivação por trás de nossa abordagem é a busca por padrões conservados em átomos com polaridades similares nas proteínas globulares de uma mesma família. Acreditamos que tais proteínas, após os respectivos processos de enovelamentos, adquiram núcleos semelhantes.

Por meio de estudos de caso e de um algoritmo de detecção de ilhas e arquipélagos apolares, executado para toda a base do SCOP, evidenciamos a existência de grupamentos atômicos com polaridade similar nessas estruturas. Verificamos, através de casos ilustrativos, conservação posicional de átomos apolares, mesmo que, algumas vezes, esses átomos fossem de elementos diferentes.

Acreditávamos que poderíamos utilizar as ilhas hidrofóbicas (representadas pelos átomos polares que as compõem) como uma assinatura estrutural para diferenciação inter-famílias. Percebemos, porém, que, apesar das estruturas de mesma família guardarem muita semelhança na forma, volume e posicionamento dessas ilhas apolares, não havia discriminação suficiente entre estruturas de famílias diferentes. Uma vez que os átomos apolares possuem a capacidade de se aglomerarem formando ilhas, provavelmente, famílias diferentes guardam alguma similaridade nas características dessas ilhas. Essa característica globular dificultaria uma tentativa de classificação seguindo apenas critérios estruturais, pois famílias diferentes podem vir a guardar similaridades em suas assinaturas apolares.

Para nossa surpresa, os átomos hidrofílicos mostraram-se como uma assinatura estrutural muito mais discriminativa entre famílias do que os átomos hidrofóbicos. Analisando diversos experimentos de classificação, constatamos que a "assinatura polar" das famílias é melhor diferencial de classificação que seus carbonos alfa e seus átomos apolares. Para a base 5SSE, obtivemos melhoria de até 7.8% na precisão de classificação em relação ao grupo controle. Em grandes conjuntos de dados, como na base Full-

SCOP, a estratégia de utilizar átomos polares na classificação também obteve bons resultados. Acreditamos que a diferença de precisão de classificação entre a assinatura apolar em relação à precisão de classificação da assinatura polar (em parte) deve-se à sua distribuição espacial mais globular, ao passo que a distribuição polar é mais poligonal (seguindo o backbone).

Essas características de distribuições espaciais dos átomos e suas respectivas polaridades podem, eventualmente, levar-nos a novos *insights* sobre formas de classificação dessas estruturas. No nosso entendimento, consideramos importante agregar informações fisico-químicas nos processos de classificação e alinhamento estrutural. Merece, portanto, futura investigação de novas formas de alinhamento e classificação estrutural, que levem em consideração as características de distribuição das polaridades atômicas.

Como sabíamos de antemão que a cadeia principal é polar, isso nos levou a questionar a influência dessa cadeia na melhoria de precisão da classificação da assinatura polar. Na tentativa de elucidar esse questionamento, conseguimos, inclusive, discriminar os átomos do backbone como a melhor assinatura estrutural estudada. Utilizando os átomos da cadeia principal como uma assinatura estrutural, fomos capazes de melhorar em até 10.3% a precisão da classificação de famílias para a base *SCOP* em relação à precisão obtida pela técnica *CSM* original, que utiliza somente o posicionamento dos carbonos alfa.

Uma vez que a classificação utilizando-se os átomos da cadeia principal obteve melhor desempenho em praticamente todas as métricas comparadas, nos interessamos em verificar se os átomos C e N (átomos esses que compõem a cadeia principal) são indispensáveis para explicar a melhoria da capacidade discriminativa dos classificadores, eis que o uso de tais átomos poderia, inclusive, capturar indiretamente os ângulos ϕ e ψ. Nossa hipótese era que a adição de pontos geométricos promovida pelo uso dos demais átomos do backbone reforçou a característica geométrica poligonal da cadeia principal, o que auxiliou a diferenciação inter-famílias pela *CSM*.

Verificamos, então, se os posicionamentos dos átomos que compõem a cadeia principal são o principal fator de melhoria discriminativa entre famílias ou se bastaria a disposição geométrica poligonal, característica da cadeia principal. Inserindo pontos intermediários fictícios entre os C_α, obtivemos uma geometria poligonal "artificial", similar à geometria poligonal obtida pela inclusão dos átomos C e N. Dessa forma, conseguimos o efeito da poligonal da cadeia principal sem a influência do posicionamento dos átomos C e N e, indiretamente, sem a influência dos ângulos ϕ e ψ.

Concluímos, assim, que a melhoria da precisão de classificação, ao se utilizar um modelo com os átomos da cadeia principal, em relação ao uso exclusivo dos carbonos alfa dessa mesma cadeia, deve-se, principalmente, pela evidenciação do caráter geo-

métrico poligonal da cadeia. Dessa forma, o corrente estudo sugere que não são as posições dos átomos C e N (o que captura de maneira indireta os ângulos ϕ e ψ) os fatores determinantes no incremento da precisão da classificação, mas, sim, o fortalecimento do caráter geométrico poligonal da cadeia principal. Concluímos, também, que a assinatura da geometria poligonal da cadeia principal das proteínas de uma mesma família é melhor diferencial de classificação que a assinatura que utiliza somente seus carbonos alfas. Ou seja, o caráter sequencial da cadeia polipeptídica e sua geometria poligonal são mais relevantes que apenas o empacotamento dos resíduos, como usado na *CSM* original.

Pires [Pires et al., 2011] reporta que experiências foram conduzidas com outros centroides em vez do C_α, como o C_β ou o último átomo pesado (LHA) da cadeia lateral. O C_α obteve melhor desempenho em todos os experimentos, um fato que ele registrou como "exigindo uma investigação mais aprofundada". Pelo corrente trabalho, acreditamos que esses centroides obtiveram mais baixo desempenho, justamente, por estarem ainda mais afastados da disposição geométrica poligonal da cadeia principal.

Portanto, as contribuições do presente trabalho são a indicação da existência de ilhas hidrofóbicas nas estruturas globulares de mesma família e a investigação do uso da disposição poligonal da cadeia principal (ou dos próprios C_α, caso sejam adicionados pontos intermediários fictícios entre esses) como uma assinatura estrutural viável. Comparou-se a capacidade discriminativa das assinaturas dos átomos polares e da poligonal da cadeia principal para as tarefas de classificação e de predição funcional com a assinatura original promovida pelo uso exclusivo dos C_α, constatando-se melhorias.

Eventualmente, a técnica apresentada de inclusão de pontos intermediários entre os pontos dos conjuntos utilizados na classificação, de maneira a reforçar suas características geométricas, poderia ser utilizada para melhoria da *CSM* em outros contextos.

4.1 Direções de trabalhos futuros

Após análise dos resultados, percebeu-se que algumas melhorias poderiam aprimorar e expandir a metodologia proposta.

4.1.1 Estudo das condições necessárias e suficientes sobre o uso da poligonal da cadeia principal como uma assinatura estrutural viável

Os indícios de que a poligonal da cadeia principal, baseada no posicionamento dos C_α, é suficiente para uma classificação viável é um bom começo. Porém, não podemos indicar, ainda, que os posicionamentos dos C_α são também necessários. Investigações futuras serão importantes para determinação, também, das condições necessárias.

Nossa intuição arrisca dizer que a classificação continuaria viável para qualquer átomo da cadeia principal, especialmente os carbonos não-alfas (o C da carbonila) e nitrogênios (o N da amina) da ligação peptídica. Uma hipótese incial seria a de que é importante escolher um átomo do backbone. Por isso, acreditamos que a presença do C_α seria suficiente, mas não necessária, para gerar a assinatura estrutural da poligonal geométrica. Talvez, sejam suficientes, também, o C da carbonila ou o N da amina, átomos que ajudam a montar o "fio" do backbone. O oxigênio da carbonila envolve apenas um deslocamento em relação ao N. O "fio" construído com O seria apenas "deslocado" em relação ao "fio" com N. Se N produzir uma classificação viável, espera-se que o O também deva produzi-la.

Logo, uma hipótese maior poderia ser: para uma classificação viável, é necessário e suficiente que seja um átomo do backbone (do "fio" mais "central").

Experimentos poderiam ser conduzidos:

- Com somente os carbonos betas C_β: nesse caso, haveria apenas um "deslocamento" do "fio" C_β em relação ao "fio" C_α (o que, possivelmente, produziria um resultado similar). Glicinas causariam uma "descontinuidade" no "fio" C_β, o que poderia interferir na precisão da classificação, quando comparada com a classificação do backbone.

- Com o próximo átomo depois do C_β: nesse caso, continuaria um "deslocamento" do backbone, mas alaninas abririam mais uma "descontinuidade" no "fio", comprometendo mais um pouco a classificação.

- Com o centro geométrico C_g da cadeia lateral: o "deslocamento" do "fio" não seria tão ruidoso, mas ainda poderia haver a inserção de algum ruído, comprometendo a classificação.

- Com um e somente um átomo aleatório da cadeia lateral: os "deslocamentos" em relação à cadeia principal adquiririam ruídos e estes, possivelmente, impactariam na precisão da classificação. Esse resultado poderia ser usado para demonstrar a

importância do "fio" como assinatura estrutural, uma vez que os átomos utilizados estariam afastados deste.

Nossa hipótese atual é a de que qualquer outro conjunto formado por algum dos átomos do backbone (C, N ou O) obteria resultado indistinto em relação ao conjunto dos átomos C_α. É uma hipótese a ser testada no futuro.

4.1.2 Uso de informações de cargas parciais

Utilizamos neste trabalho informações de polaridade atômica de maneira estática através dos critérios de Sobolev e [Alexandre V. Fassio, 2017]. Agora que temos em mãos os valores dos resultados para esses critérios, podemos repetir os experimentos utilizando valores calculados de cargas parciais atômicas para comparações. Acreditamos que um cálculo refinado de cargas parciais possa trazer-nos informações de contexto espacial que valores de polaridade estáticos não são capazes de depreender.

Capítulo 5

Suplementar

Lista de arquivos adicionais:

- monomers.txt **Arquivo adicional 1:** Lista completa dos arquivos PDB contendo monômeros. Os monômeros foram usados nos casos ilustrativos.
- gold\gold_standard_dataset.csv **Arquivo adicional 2:** Enzyme gold-standard dataset. Lista dos identificadores PDB que compõe o gold-standard dataset além da atribuição de suas famílias e superfamílias.
- gold\gold_experiments.xls **Arquivo adicional 3:** Resultados dos 30 experimentos de classificação para a base Enzyme gold-standard dataset.
- gold\gold_pdbs_candidates.txt **Arquivo adicional 4:** Lista de todas as cadeias da base Enzyme gold-standard dataset candidatas a serem usadas na classificação.
- gold\gold_pdbs_efectives.txt **Arquivo adicional 5:** Lista de todas as cadeias efetivamente utilizadas nas classificações da base Enzyme gold-standard dataset.
- gold\gold_pdbs_errors.txt **Arquivo adicional 6:** Todas as cadeias utilizadas nas classificações da base Enzyme gold-standard dataset que geraram exceções.
- sses\Xsse\Xsse.txt **Arquivo adicional 7:** Xsse dataset (onde Xsse=3sse,4sse,5sse,6sse). Domínios dos XSSEs da versão 1.69 do SCOP. Esse arquivo lista os identificadores dos XSSEs que contém domínios no SCOP versão 1.69.
- sses\Xsse_experiments.xls **Arquivo adicional 8:** (onde Xsse=3sse,4sse,5sse,6sse) Resultados dos 30 experimentos de classificação para as bases XSSEs.

- sses\Xsse\Xsse_pdbs_candidates.txt **Arquivo adicional 9:** Lista de todas as cadeias candidatas a serem utilizadas nas classificações da base XSSE dataset (onde Xsse=3sse,4sse,5sse,6sse).

- sses\Xsse\Xsse_pdbs_efectives.txt **Arquivo adicional 10:** Lista de todas as cadeias efetivamente utilizadas nas classificações da base XSSE dataset (onde Xsse=3sse,4sse,5sse,6sse).

- sses\Xsse\Xsse_pdbs_errors.txt **Arquivo adicional 11:** Todas as cadeias utilizadas nas classificações da base XSSE que geraram exceções.

- sses\Xsse\Xsse_short_csm_classes.txt **Arquivo adicional 12:** Famílias SCOP com menos de 10 representantes nas classificações da base XSSE (onde Xsse=3sse,4sse,5sse,6sse).

- scop\scop_experiments.xls **Arquivo adicional 13:** Resultados dos 30 experimentos de classificação para a base Full-SCOP 1.75.

- scop\scop_pdbs_candidates.txt **Arquivo adicional 14:** Lista de todas as cadeias candidatas a serem utilizadas nas classificações da base Full-SCOP 1.75.

- scop\scop_pdbs_efectives.txt **Arquivo adicional 15:** Lista de todas as cadeias efetivamente utilizadas nas classificações da base XSSE Full-SCOP 1.75.

- scop\scop_pdbs_errors.txt **Arquivo adicional 16:** Todas as cadeias utilizadas nas classificações da base Full-SCOP 1.75 que geraram exceções.

- scop\scop_short_csm_classes.txt **Arquivo adicional 17:** Famílias SCOP com menos de 10 representantes nas classificações da base Full-SCOP 1.75.

- pymol\sobolev\family\session.pse **Arquivo adicional 18:** Sessões pymol ilustrando arquipélogos apolares segundo o critério *Sobolev* (onde familia=a.1.1.2, b.6.1.1, b.47.1.1, b.60.1.1, c.47.1.1).

- pymol\sobring\family\session.pse **Arquivo adicional 19:** Sessões pymol ilustrando arquipélogos apolares segundo o critério *Ring* (onde familia=a.1.1.2, b.6.1.1, b.47.1.1, b.60.1.1, c.47.1.1).

Tabela 5.1: *P-valores* para a classificação estrutural para a bases gold-standard. Grupo controle C_α. Hipóteses nulas rejeitadas estão realçadas em negrito.

Dataset	Superfamily		C_α		*Backbone*		*All*		*Side*	
			Prec	Recall	Prec	Recall	Prec	Recall	Prec	Recall
Gold	Amidohydrolase	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.253×10^{-17}	8.227×10^{-18}	1.305×10^{-02}	3.117×10^{-13}	1.663×10^{-13}	3.558×10^{-23}
		Backbone	2.253×10^{-17}	8.227×10^{-18}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	3.075×10^{-26}	1.418×10^{-13}	5.049×10^{-23}	2.651×10^{-28}
		All	1.305×10^{-02}	3.117×10^{-13}	3.075×10^{-26}	1.418×10^{-13}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.073×10^{-16}	1.814×10^{-26}
		Side	1.663×10^{-13}	3.558×10^{-23}	5.049×10^{-23}	2.651×10^{-28}	1.073×10^{-16}	1.814×10^{-26}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Crotonase	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.088×10^{-11}	1.880×10^{-14}	$\mathbf{6.752 \times 10^{-01}}$	8.368×10^{-07}	4.639×10^{-08}	3.183×10^{-10}
		Backbone	1.088×10^{-11}	1.880×10^{-14}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.590×10^{-09}	5.913×10^{-21}	5.104×10^{-03}	3.958×10^{-03}
		All	$\mathbf{6.752 \times 10^{-01}}$	8.368×10^{-07}	2.590×10^{-09}	5.913×10^{-21}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	7.434×10^{-09}	1.752×10^{-19}
		Side	4.639×10^{-08}	3.183×10^{-10}	5.104×10^{-03}	3.958×10^{-03}	7.434×10^{-09}	1.752×10^{-19}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Enolase	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	5.810×10^{-08}	1.439×10^{-14}	1.150×10^{-03}	1.488×10^{-02}	8.498×10^{-23}	5.498×10^{-25}
		Backbone	5.810×10^{-08}	1.439×10^{-14}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	3.555×10^{-09}	3.310×10^{-12}	5.124×10^{-24}	8.218×10^{-29}
		All	1.150×10^{-03}	1.488×10^{-02}	3.555×10^{-09}	3.310×10^{-12}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	4.060×10^{-21}	5.063×10^{-25}
		Side	8.498×10^{-23}	5.498×10^{-25}	5.124×10^{-24}	8.218×10^{-29}	4.060×10^{-21}	5.063×10^{-25}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Haloacid Dehalogenase	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	3.935×10^{-18}	4.657×10^{-19}	3.935×10^{-18}	2.481×10^{-13}	4.678×10^{-16}	3.170×10^{-13}
		Backbone	3.935×10^{-18}	4.657×10^{-19}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	8.828×10^{-09}	$\mathbf{6.855 \times 10^{-02}}$	4.040×10^{-09}
		All	3.935×10^{-18}	2.481×10^{-13}	$\mathbf{1.000 \times 10^{+00}}$	8.828×10^{-09}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{6.855 \times 10^{-02}}$	$\mathbf{8.209 \times 10^{-01}}$
		Side	4.678×10^{-16}	3.170×10^{-13}	$\mathbf{6.855 \times 10^{-02}}$	4.040×10^{-09}	$\mathbf{6.855 \times 10^{-02}}$	$\mathbf{8.209 \times 10^{-01}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Isoprenoid Synthase Type I	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	5.330×10^{-18}	1.271×10^{-21}	4.831×10^{-30}	4.755×10^{-26}
		Backbone	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	5.330×10^{-18}	1.271×10^{-21}	4.831×10^{-30}	4.755×10^{-26}
		All	5.330×10^{-18}	1.271×10^{-21}	5.330×10^{-18}	1.271×10^{-21}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.602×10^{-20}	$\mathbf{4.237 \times 10^{-01}}$
		Side	4.831×10^{-30}	4.755×10^{-26}	4.831×10^{-30}	4.755×10^{-26}	2.602×10^{-20}	$\mathbf{4.237 \times 10^{-01}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Vicinal Oxygen Chelate	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	6.723×10^{-04}	6.723×10^{-04}	6.205×10^{-18}	1.034×10^{-16}	5.111×10^{-21}	1.200×10^{-23}
		Backbone	6.723×10^{-04}	6.723×10^{-04}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$0.000 \times 10^{+00}$	$0.000 \times 10^{+00}$	2.098×10^{-22}	4.466×10^{-24}
		All	6.205×10^{-18}	1.034×10^{-16}	$0.000 \times 10^{+00}$	$0.000 \times 10^{+00}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	4.076×10^{-17}	2.949×10^{-21}
		Side	5.111×10^{-21}	1.200×10^{-23}	2.098×10^{-22}	4.466×10^{-24}	4.076×10^{-17}	2.949×10^{-21}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	All	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	4.412×10^{-20}	4.759×10^{-20}	$\mathbf{1.111 \times 10^{-01}}$	$\mathbf{1.715 \times 10^{-01}}$	2.210×10^{-23}	3.108×10^{-23}
		Backbone	4.412×10^{-20}	4.759×10^{-20}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.412×10^{-21}	1.129×10^{-20}	2.178×10^{-31}	5.148×10^{-31}
		All	$\mathbf{1.111 \times 10^{-01}}$	$\mathbf{1.715 \times 10^{-01}}$	1.412×10^{-21}	1.129×10^{-20}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	7.437×10^{-26}	7.657×10^{-25}
		Side	2.210×10^{-23}	3.108×10^{-23}	2.178×10^{-31}	5.148×10^{-31}	7.437×10^{-26}	7.657×10^{-25}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$

Tabela 5.2: *P-valores* para a classificação estrutural para a bases Full-SCOP. Grupo controle C_α. Hipóteses nulas rejeitadas estão realçadas em negrito.

Dataset	SCOP Level		C_α		*Backbone*		*All*		*Side*	
			Prec	Recall	Prec	Recall	Prec	Recall	Prec	Recall
Full-SCOP	Class	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.986×10^{-61}	9.563×10^{-64}	2.737×10^{-51}	2.737×10^{-51}	1.099×10^{-59}	7.366×10^{-62}
		Backbone	2.986×10^{-61}	9.563×10^{-64}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.031×10^{-65}	4.792×10^{-68}	1.739×10^{-66}	1.309×10^{-68}
		All	2.737×10^{-51}	2.737×10^{-51}	1.031×10^{-65}	4.792×10^{-68}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.466×10^{-54}	2.788×10^{-57}
		Side	1.099×10^{-59}	7.366×10^{-62}	1.739×10^{-66}	1.309×10^{-68}	1.466×10^{-54}	2.788×10^{-57}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Fold	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	3.096×10^{-64}	4.114×10^{-64}	8.309×10^{-30}	1.172×10^{-28}	3.852×10^{-62}	1.884×10^{-61}
		Backbone	3.096×10^{-64}	4.114×10^{-64}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	6.007×10^{-71}	2.485×10^{-61}	4.244×10^{-73}	2.377×10^{-69}
		All	8.309×10^{-30}	1.172×10^{-28}	6.007×10^{-71}	2.485×10^{-61}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	6.179×10^{-63}	2.197×10^{-60}
		Side	3.852×10^{-62}	1.884×10^{-61}	4.244×10^{-73}	2.377×10^{-69}	6.179×10^{-63}	2.197×10^{-60}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Superfamily	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	9.706×10^{-64}	5.419×10^{-62}	8.952×10^{-21}	1.416×10^{-24}	6.234×10^{-62}	2.523×10^{-64}
		Backbone	9.706×10^{-64}	5.419×10^{-62}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	4.861×10^{-65}	6.420×10^{-61}	3.687×10^{-72}	1.440×10^{-69}
		All	8.952×10^{-21}	1.416×10^{-24}	4.861×10^{-65}	6.420×10^{-61}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.163×10^{-58}	5.345×10^{-59}
		Side	6.234×10^{-62}	2.523×10^{-64}	3.687×10^{-72}	1.440×10^{-69}	1.163×10^{-58}	5.345×10^{-59}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Family	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	4.602×10^{-65}	5.048×10^{-65}	5.366×10^{-31}	1.147×10^{-29}	5.944×10^{-64}	7.264×10^{-63}
		Backbone	4.602×10^{-65}	5.048×10^{-65}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	3.980×10^{-62}	1.779×10^{-63}	4.113×10^{-70}	3.572×10^{-73}
		All	5.366×10^{-31}	1.147×10^{-29}	3.980×10^{-62}	1.779×10^{-63}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.079×10^{-63}	4.743×10^{-62}
		Side	5.944×10^{-64}	7.264×10^{-63}	4.113×10^{-70}	3.572×10^{-73}	1.079×10^{-63}	4.743×10^{-62}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$

Tabela 5.3: *P-valores* para a classificação estrutural para a bases 3SSE. Grupo controle C_α. Hipóteses nulas rejeitadas estão realçadas em negrito.

Dataset	SCOP Level		C_α		*Backbone*		*All*		*Side*	
			Prec	Recall	Prec	Recall	Prec	Recall	Prec	Recall
3SSE	Class	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	9.566×10^{-15}	3.189×10^{-13}	1.544×10^{-02}	4.561×10^{-04}	3.032×10^{-15}	4.001×10^{-15}
		Backbone	9.566×10^{-15}	3.189×10^{-13}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	3.016×10^{-14}	2.551×10^{-15}	9.861×10^{-22}	4.357×10^{-21}
		All	1.544×10^{-02}	4.561×10^{-04}	3.016×10^{-14}	2.551×10^{-15}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.103×10^{-18}	3.129×10^{-18}
		Side	3.032×10^{-15}	4.001×10^{-15}	9.861×10^{-22}	4.357×10^{-21}	1.103×10^{-18}	3.129×10^{-18}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Fold	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	6.101×10^{-26}	1.542×10^{-26}	1.558×10^{-21}	8.020×10^{-20}	1.919×10^{-14}	9.474×10^{-13}
		Backbone	6.101×10^{-26}	1.542×10^{-26}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	7.227×10^{-16}	1.173×10^{-16}	2.268×10^{-18}	6.147×10^{-18}
		All	1.558×10^{-21}	8.020×10^{-20}	7.227×10^{-16}	1.173×10^{-16}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	8.298×10^{-07}	3.561×10^{-05}
		Side	1.919×10^{-14}	9.474×10^{-13}	2.268×10^{-18}	6.147×10^{-18}	8.298×10^{-07}	3.561×10^{-05}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Superfamily	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	8.738×10^{-26}	2.833×10^{-26}	3.378×10^{-17}	5.644×10^{-15}	3.047×10^{-11}	3.374×10^{-09}
		Backbone	8.738×10^{-26}	2.833×10^{-26}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.383×10^{-19}	2.147×10^{-18}	1.063×10^{-16}	2.963×10^{-16}
		All	3.378×10^{-17}	5.644×10^{-15}	2.383×10^{-19}	2.147×10^{-18}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	7.223×10^{-06}	1.082×10^{-04}
		Side	3.047×10^{-11}	3.374×10^{-09}	1.063×10^{-16}	2.963×10^{-16}	7.223×10^{-06}	1.082×10^{-04}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Family	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.167×10^{-25}	1.015×10^{-25}	9.709×10^{-18}	2.610×10^{-14}	1.885×10^{-07}	$\mathbf{2.223 \times 10^{-01}}$
		Backbone	2.167×10^{-25}	1.015×10^{-25}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	8.426×10^{-26}	1.020×10^{-26}	6.164×10^{-25}	1.616×10^{-25}
		All	9.709×10^{-18}	2.610×10^{-14}	8.426×10^{-26}	1.020×10^{-26}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.780×10^{-15}	8.895×10^{-15}
		Side	1.885×10^{-07}	$\mathbf{2.223 \times 10^{-01}}$	6.164×10^{-25}	1.616×10^{-25}	1.780×10^{-15}	8.895×10^{-15}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$

Tabela 5.4: *P-valores* para a classificação estrutural para a bases 4SSE. Grupo controle C_α. Hipóteses nulas rejeitadas estão realçadas em negrito.

Dataset	SCOP Level		C_α		*Backbone*		*All*		*Side*	
			Prec	Recall	Prec	Recall	Prec	Recall	Prec	Recall
4SSE	Class	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	4.565×10^{-25}	1.263×10^{-24}	1.701×10^{-17}	2.758×10^{-17}	1.170×10^{-21}	5.966×10^{-21}
		Backbone	4.565×10^{-25}	1.263×10^{-24}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	5.969×10^{-34}	9.591×10^{-34}	6.435×10^{-30}	9.369×10^{-30}
		All	1.701×10^{-17}	2.758×10^{-17}	5.969×10^{-34}	9.591×10^{-34}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	5.482×10^{-14}	7.817×10^{-13}
		Side	1.170×10^{-21}	5.966×10^{-21}	6.435×10^{-30}	9.369×10^{-30}	5.482×10^{-14}	7.817×10^{-13}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Fold	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	6.806×10^{-34}	1.830×10^{-34}	6.075×10^{-17}	4.470×10^{-19}	3.883×10^{-02}	$\mathbf{4.954 \times 10^{-01}}$
		Backbone	6.806×10^{-34}	1.830×10^{-34}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	5.289×10^{-32}	3.956×10^{-33}	9.021×10^{-35}	2.159×10^{-34}
		All	6.075×10^{-17}	4.470×10^{-19}	5.289×10^{-32}	3.956×10^{-33}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.083×10^{-20}	5.502×10^{-19}
		Side	3.883×10^{-02}	$\mathbf{4.954 \times 10^{-01}}$	9.021×10^{-35}	2.159×10^{-34}	1.083×10^{-20}	5.502×10^{-19}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Superfamily	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.564×10^{-33}	2.111×10^{-34}	1.821×10^{-17}	4.451×10^{-18}	$\mathbf{9.041 \times 10^{-01}}$	$\mathbf{7.369 \times 10^{-02}}$
		Backbone	2.564×10^{-33}	2.111×10^{-34}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	6.212×10^{-34}	7.999×10^{-34}	3.508×10^{-36}	2.431×10^{-36}
		All	1.821×10^{-17}	4.451×10^{-18}	6.212×10^{-34}	7.999×10^{-34}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	7.907×10^{-23}	2.047×10^{-21}
		Side	$\mathbf{9.041 \times 10^{-01}}$	$\mathbf{7.369 \times 10^{-02}}$	3.508×10^{-36}	2.431×10^{-36}	7.907×10^{-23}	2.047×10^{-21}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Family	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.246×10^{-37}	1.922×10^{-38}	1.375×10^{-27}	2.480×10^{-28}	4.863×10^{-17}	3.902×10^{-18}
		Backbone	2.246×10^{-37}	1.922×10^{-38}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.978×10^{-32}	6.222×10^{-34}	2.680×10^{-33}	7.816×10^{-34}
		All	1.375×10^{-27}	2.480×10^{-28}	2.978×10^{-32}	6.222×10^{-34}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.342×10^{-20}	3.612×10^{-20}
		Side	4.863×10^{-17}	3.902×10^{-18}	2.680×10^{-33}	7.816×10^{-34}	1.342×10^{-20}	3.612×10^{-20}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$

Tabela 5.5: *P-valores* para a classificação estrutural para a bases 5SSE. Grupo controle C_α. Hipóteses nulas rejeitadas estão realçadas em negrito.

Dataset	SCOP Level		C_α		*Backbone*		*All*		*Side*	
			Prec	Recall	Prec	Recall	Prec	Recall	Prec	Recall
5SSE	Class	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	3.012×10^{-32}	2.024×10^{-32}	2.899×10^{-05}	3.947×10^{-04}	1.224×10^{-33}	9.915×10^{-35}
		Backbone	3.012×10^{-32}	2.024×10^{-32}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.633×10^{-30}	2.009×10^{-30}	7.732×10^{-43}	5.097×10^{-44}
		All	2.899×10^{-05}	3.947×10^{-04}	2.633×10^{-30}	2.009×10^{-30}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.791×10^{-33}	5.550×10^{-35}
		Side	1.224×10^{-33}	9.915×10^{-35}	7.732×10^{-43}	5.097×10^{-44}	1.791×10^{-33}	5.550×10^{-35}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Fold	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.065×10^{-39}	7.705×10^{-39}	1.803×10^{-23}	2.137×10^{-23}	4.049×10^{-35}	3.460×10^{-34}
		Backbone	1.065×10^{-39}	7.705×10^{-39}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	3.038×10^{-36}	5.334×10^{-36}	2.384×10^{-44}	1.024×10^{-43}
		All	1.803×10^{-23}	2.137×10^{-23}	3.038×10^{-36}	5.334×10^{-36}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	7.473×10^{-39}	1.267×10^{-38}
		Side	4.049×10^{-35}	3.460×10^{-34}	2.384×10^{-44}	1.024×10^{-43}	7.473×10^{-39}	1.267×10^{-38}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Superfamily	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	8.138×10^{-41}	9.461×10^{-43}	8.559×10^{-28}	3.531×10^{-28}	4.977×10^{-37}	5.075×10^{-39}
		Backbone	8.138×10^{-41}	9.461×10^{-43}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.871×10^{-34}	3.425×10^{-35}	5.167×10^{-45}	1.187×10^{-47}
		All	8.559×10^{-28}	3.531×10^{-28}	1.871×10^{-34}	3.425×10^{-35}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.766×10^{-42}	7.533×10^{-44}
		Side	4.977×10^{-37}	5.075×10^{-39}	5.167×10^{-45}	1.187×10^{-47}	1.766×10^{-42}	7.533×10^{-44}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Family	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	4.437×10^{-38}	4.566×10^{-39}	2.017×10^{-27}	3.500×10^{-27}	4.797×10^{-35}	6.445×10^{-36}
		Backbone	4.437×10^{-38}	4.566×10^{-39}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	9.607×10^{-36}	3.702×10^{-35}	1.009×10^{-45}	1.120×10^{-47}
		All	2.017×10^{-27}	3.500×10^{-27}	9.607×10^{-36}	3.702×10^{-35}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.313×10^{-43}	3.903×10^{-44}
		Side	4.797×10^{-35}	6.445×10^{-36}	1.009×10^{-45}	1.120×10^{-47}	2.313×10^{-43}	3.903×10^{-44}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$

Tabela 5.6: *P-valores* para a classificação estrutural para a bases 6SSE. Grupo controle C_α. Hipóteses nulas rejeitadas estão realçadas em negrito.

Dataset	SCOP Level		C_α		*Backbone*		*All*		*Side*	
			Prec	Recall	Prec	Recall	Prec	Recall	Prec	Recall
6SSE	Class	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	3.591×10^{-31}	3.534×10^{-30}	8.836×10^{-11}	8.321×10^{-11}	5.845×10^{-23}	6.008×10^{-22}
		Backbone	3.591×10^{-31}	3.534×10^{-30}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	5.098×10^{-28}	4.762×10^{-28}	2.181×10^{-34}	4.626×10^{-33}
		All	8.836×10^{-11}	8.321×10^{-11}	5.098×10^{-28}	4.762×10^{-28}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	4.638×10^{-28}	7.577×10^{-27}
		Side	5.845×10^{-23}	6.008×10^{-22}	2.181×10^{-34}	4.626×10^{-33}	4.638×10^{-28}	7.577×10^{-27}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Fold	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.101×10^{-33}	7.191×10^{-33}	4.608×10^{-24}	9.257×10^{-24}	4.923×10^{-02}	$\mathbf{2.592 \times 10^{-01}}$
		Backbone	2.101×10^{-33}	7.191×10^{-33}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.095×10^{-31}	1.953×10^{-30}	1.242×10^{-35}	8.240×10^{-36}
		All	4.608×10^{-24}	9.257×10^{-24}	1.095×10^{-31}	1.953×10^{-30}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	7.083×10^{-28}	1.810×10^{-28}
		Side	4.923×10^{-02}	$\mathbf{2.592 \times 10^{-01}}$	1.242×10^{-35}	8.240×10^{-36}	7.083×10^{-28}	1.810×10^{-28}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Superfamily	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.319×10^{-36}	5.413×10^{-37}	1.890×10^{-25}	3.675×10^{-26}	$\mathbf{2.999 \times 10^{-01}}$	9.299×10^{-04}
		Backbone	1.319×10^{-36}	5.413×10^{-37}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.002×10^{-34}	2.559×10^{-32}	8.789×10^{-39}	5.403×10^{-39}
		All	1.890×10^{-25}	3.675×10^{-26}	1.002×10^{-34}	2.559×10^{-32}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	5.045×10^{-27}	2.852×10^{-28}
		Side	$\mathbf{2.999 \times 10^{-01}}$	9.299×10^{-04}	8.789×10^{-39}	5.403×10^{-39}	5.045×10^{-27}	2.852×10^{-28}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Family	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	3.063×10^{-35}	1.305×10^{-35}	5.540×10^{-26}	5.100×10^{-26}	1.686×10^{-12}	3.943×10^{-13}
		Backbone	3.063×10^{-35}	1.305×10^{-35}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.020×10^{-30}	2.624×10^{-30}	5.602×10^{-41}	8.318×10^{-40}
		All	5.540×10^{-26}	5.100×10^{-26}	1.020×10^{-30}	2.624×10^{-30}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	6.061×10^{-27}	1.942×10^{-28}
		Side	1.686×10^{-12}	3.943×10^{-13}	5.602×10^{-41}	8.318×10^{-40}	6.061×10^{-27}	1.942×10^{-28}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$

Tabela 5.7: *P-valores* para a predição de função para a base gold-standard. Pontos intermediários. Grupo controle C_α. Hipóteses nulas rejeitadas estão realçadas em negrito.

Dataset	SCOP Level		C_α		$C_\alpha^{0.2}$		$C_\alpha^{0.4}$		$C_\alpha^{0.6}$		$C_\alpha^{0.8}$	
			Prec	Recall	Prec	Recall	Prec	Recall	Prec	Recall	Prec	Recall
Gold	Amidohydrolase	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	3.008×10^{-13}	3.585×10^{-14}	7.235×10^{-13}	4.945×10^{-16}	2.253×10^{-17}	3.204×10^{-22}	2.253×10^{-17}	1.533×10^{-21}
		$C_\alpha^{0.2}$	3.008×10^{-13}	3.585×10^{-14}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.165×10^{-02}	2.902×10^{-03}	1.342×10^{-18}	7.429×10^{-07}	1.342×10^{-18}	2.050×10^{-10}
		$C_\alpha^{0.4}$	7.235×10^{-13}	4.945×10^{-16}	1.165×10^{-02}	2.902×10^{-03}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	8.053×10^{-10}	8.056×10^{-06}	8.053×10^{-10}	2.949×10^{-09}
		$C_\alpha^{0.6}$	2.253×10^{-17}	3.204×10^{-22}	1.342×10^{-18}	7.429×10^{-07}	8.053×10^{-10}	8.056×10^{-06}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.337×10^{-07}
		$C_\alpha^{0.8}$	2.253×10^{-17}	1.533×10^{-21}	1.342×10^{-18}	2.050×10^{-10}	8.053×10^{-10}	2.949×10^{-09}	$\mathbf{1.000 \times 10^{+00}}$	1.337×10^{-07}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Crotonase	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.343×10^{-02}	1.172×10^{-06}	1.510×10^{-03}	9.139×10^{-10}	3.596×10^{-02}	1.335×10^{-06}	1.088×10^{-11}	1.880×10^{-14}
		$C_\alpha^{0.2}$	2.343×10^{-02}	1.172×10^{-06}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	4.340×10^{-02}	1.646×10^{-02}	$\mathbf{1.608 \times 10^{-01}}$	$\mathbf{1.608 \times 10^{-01}}$	4.367×10^{-14}	3.568×10^{-11}
		$C_\alpha^{0.4}$	1.510×10^{-03}	9.139×10^{-10}	4.340×10^{-02}	1.646×10^{-02}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.174×10^{-02}	4.658×10^{-03}	1.342×10^{-18}	1.342×10^{-18}
		$C_\alpha^{0.6}$	3.596×10^{-02}	1.335×10^{-06}	$\mathbf{1.608 \times 10^{-01}}$	$\mathbf{1.608 \times 10^{-01}}$	1.174×10^{-02}	4.658×10^{-03}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	3.773×10^{-18}	3.272×10^{-13}
		$C_\alpha^{0.8}$	1.088×10^{-11}	1.880×10^{-14}	4.367×10^{-14}	3.568×10^{-11}	1.342×10^{-18}	1.342×10^{-18}	3.773×10^{-18}	3.272×10^{-13}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Enolase	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	3.748×10^{-07}	1.556×10^{-11}	3.066×10^{-12}	5.857×10^{-15}	3.833×10^{-20}	7.983×10^{-21}	3.833×10^{-20}	8.587×10^{-19}
		$C_\alpha^{0.2}$	3.748×10^{-07}	1.556×10^{-11}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	4.386×10^{-02}	6.348×10^{-03}	1.557×10^{-02}	4.283×10^{-06}	1.557×10^{-02}	1.603×10^{-02}
		$C_\alpha^{0.4}$	3.066×10^{-12}	5.857×10^{-15}	4.386×10^{-02}	6.348×10^{-03}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{8.307 \times 10^{-02}}$	4.385×10^{-05}	$\mathbf{8.307 \times 10^{-02}}$	$\mathbf{3.405 \times 10^{-01}}$
		$C_\alpha^{0.6}$	3.833×10^{-20}	7.983×10^{-21}	1.557×10^{-02}	4.283×10^{-06}	$\mathbf{8.307 \times 10^{-02}}$	4.385×10^{-05}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.982×10^{-10}
		$C_\alpha^{0.8}$	3.833×10^{-20}	8.587×10^{-19}	1.557×10^{-02}	1.603×10^{-02}	$\mathbf{8.307 \times 10^{-02}}$	$\mathbf{3.405 \times 10^{-01}}$	$\mathbf{1.000 \times 10^{+00}}$	1.982×10^{-10}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Haloacid Dehalogenase	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.163×10^{-17}	2.905×10^{-14}	3.935×10^{-18}	1.815×10^{-16}	1.163×10^{-17}	1.886×10^{-14}	3.935×10^{-18}	4.657×10^{-19}
		$C_\alpha^{0.2}$	1.163×10^{-17}	2.905×10^{-14}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.608 \times 10^{-01}}$	1.669×10^{-02}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.608 \times 10^{-01}}$	$\mathbf{1.608 \times 10^{-01}}$	3.101×10^{-10}
		$C_\alpha^{0.4}$	3.935×10^{-18}	1.815×10^{-16}	$\mathbf{1.608 \times 10^{-01}}$	1.669×10^{-02}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.608 \times 10^{-01}}$	5.193×10^{-03}	$\mathbf{1.000 \times 10^{+00}}$	1.342×10^{-18}
		$C_\alpha^{0.6}$	1.163×10^{-17}	1.886×10^{-14}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.608 \times 10^{-01}}$	$\mathbf{1.608 \times 10^{-01}}$	5.193×10^{-03}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.608 \times 10^{-01}}$	7.149×10^{-12}
		$C_\alpha^{0.8}$	3.935×10^{-18}	4.657×10^{-19}	$\mathbf{1.608 \times 10^{-01}}$	3.101×10^{-10}	$\mathbf{1.000 \times 10^{+00}}$	1.342×10^{-18}	$\mathbf{1.608 \times 10^{-01}}$	7.149×10^{-12}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Isoprenoid Synthase Type I	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$0.000 \times 10^{+00}$	$0.000 \times 10^{+00}$	$0.000 \times 10^{+00}$	2.462×10^{-22}
		$C_\alpha^{0.2}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$0.000 \times 10^{+00}$	$0.000 \times 10^{+00}$	$0.000 \times 10^{+00}$	2.462×10^{-22}
		$C_\alpha^{0.4}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$0.000 \times 10^{+00}$	$0.000 \times 10^{+00}$	$0.000 \times 10^{+00}$	2.462×10^{-22}
		$C_\alpha^{0.6}$	$0.000 \times 10^{+00}$	$0.000 \times 10^{+00}$	$0.000 \times 10^{+00}$	$0.000 \times 10^{+00}$	$0.000 \times 10^{+00}$	$0.000 \times 10^{+00}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.261×10^{-02}
		$C_\alpha^{0.8}$	$0.000 \times 10^{+00}$	2.462×10^{-22}	$0.000 \times 10^{+00}$	2.462×10^{-22}	$0.000 \times 10^{+00}$	2.462×10^{-22}	$\mathbf{1.000 \times 10^{+00}}$	2.261×10^{-02}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Vicinal Oxygen Chelate	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{9.526 \times 10^{-01}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{7.800 \times 10^{-01}}$	6.723×10^{-04}	6.723×10^{-04}	6.723×10^{-04}	6.723×10^{-04}
		$C_\alpha^{0.2}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{9.526 \times 10^{-01}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{3.256 \times 10^{-01}}$	2.261×10^{-02}	1.328×10^{-02}	2.261×10^{-02}	1.328×10^{-02}
		$C_\alpha^{0.4}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{7.800 \times 10^{-01}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{3.256 \times 10^{-01}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.261×10^{-02}	2.261×10^{-02}	2.261×10^{-02}	2.261×10^{-02}
		$C_\alpha^{0.6}$	6.723×10^{-04}	6.723×10^{-04}	2.261×10^{-02}	1.328×10^{-02}	2.261×10^{-02}	2.261×10^{-02}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
		$C_\alpha^{0.8}$	6.723×10^{-04}	6.723×10^{-04}	2.261×10^{-02}	1.328×10^{-02}	2.261×10^{-02}	2.261×10^{-02}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	All	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	3.768×10^{-15}	3.332×10^{-15}	4.958×10^{-17}	4.494×10^{-17}	4.281×10^{-19}	3.147×10^{-19}	8.531×10^{-22}	8.789×10^{-22}
		$C_\alpha^{0.2}$	3.768×10^{-15}	3.332×10^{-15}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.799×10^{-04}	1.799×10^{-04}	5.193×10^{-03}	5.193×10^{-03}	1.231×10^{-07}	1.231×10^{-07}
		$C_\alpha^{0.4}$	4.958×10^{-17}	4.494×10^{-17}	1.799×10^{-04}	1.799×10^{-04}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{7.590 \times 10^{-01}}$	$\mathbf{7.590 \times 10^{-01}}$	1.080×10^{-05}	1.080×10^{-05}
		$C_\alpha^{0.6}$	4.281×10^{-19}	3.147×10^{-19}	5.193×10^{-03}	5.193×10^{-03}	$\mathbf{7.590 \times 10^{-01}}$	$\mathbf{7.590 \times 10^{-01}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.919×10^{-09}	1.919×10^{-09}
		$C_\alpha^{0.8}$	8.531×10^{-22}	8.789×10^{-22}	1.231×10^{-07}	1.231×10^{-07}	1.080×10^{-05}	1.080×10^{-05}	1.919×10^{-09}	1.919×10^{-09}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$

Tabela 5.8: *P-valores* para a predição de função para a base gold-standard. Pontos intermediários. Grupo controle *Backbone*. Hipóteses nulas rejeitadas estão realçadas em negrito.

Dataset	Superfamily		*Backbone*		*Backbone*$^{0.2}$		*Backbone*$^{0.4}$		*Backbone*$^{0.6}$		*Backbone*$^{0.8}$	
			Prec	Recall	Prec	Recall	Prec	Recall	Prec	Recall	Prec	Recall
Gold	Amidohydrolase	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	8.093×10^{-03}	$\mathbf{1.000 \times 10^{+00}}$	2.312×10^{-03}	$\mathbf{1.000 \times 10^{+00}}$	1.909×10^{-02}	$\mathbf{1.000 \times 10^{+00}}$	8.368×10^{-05}
		Backbone$^{0.2}$	$\mathbf{1.000 \times 10^{+00}}$	8.093×10^{-03}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{8.307 \times 10^{-02}}$	$\mathbf{1.000 \times 10^{+00}}$	2.039×10^{-12}	$\mathbf{1.000 \times 10^{+00}}$	7.769×10^{-04}
		Backbone$^{0.4}$	$\mathbf{1.000 \times 10^{+00}}$	2.312×10^{-03}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{8.307 \times 10^{-02}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.737×10^{-11}	$\mathbf{1.000 \times 10^{+00}}$	1.224×10^{-02}
		Backbone$^{0.6}$	$\mathbf{1.000 \times 10^{+00}}$	1.909×10^{-02}	$\mathbf{1.000 \times 10^{+00}}$	2.039×10^{-12}	$\mathbf{1.000 \times 10^{+00}}$	1.737×10^{-11}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.058×10^{-17}
		Backbone$^{0.8}$	$\mathbf{1.000 \times 10^{+00}}$	8.368×10^{-05}	$\mathbf{1.000 \times 10^{+00}}$	7.769×10^{-04}	$\mathbf{1.000 \times 10^{+00}}$	1.224×10^{-02}	$\mathbf{1.000 \times 10^{+00}}$	1.058×10^{-17}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Crotonase	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$0.000 \times 10^{+00}$	$0.000 \times 10^{+00}$	4.340×10^{-02}	4.340×10^{-02}	$\mathbf{8.307 \times 10^{-02}}$	$\mathbf{8.307 \times 10^{-02}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
		Backbone$^{0.2}$	$0.000 \times 10^{+00}$	$0.000 \times 10^{+00}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	3.218×10^{-14}	3.218×10^{-14}	4.879×10^{-16}	4.879×10^{-16}	$0.000 \times 10^{+00}$	$0.000 \times 10^{+00}$
		Backbone$^{0.4}$	4.340×10^{-02}	4.340×10^{-02}	3.218×10^{-14}	3.218×10^{-14}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{3.256 \times 10^{-01}}$	$\mathbf{3.256 \times 10^{-01}}$	4.340×10^{-02}	4.340×10^{-02}
		Backbone$^{0.6}$	$\mathbf{8.307 \times 10^{-02}}$	$\mathbf{8.307 \times 10^{-02}}$	4.879×10^{-16}	4.879×10^{-16}	$\mathbf{3.256 \times 10^{-01}}$	$\mathbf{3.256 \times 10^{-01}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{8.307 \times 10^{-02}}$	$\mathbf{8.307 \times 10^{-02}}$
		Backbone$^{0.8}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$0.000 \times 10^{+00}$	$0.000 \times 10^{+00}$	4.340×10^{-02}	4.340×10^{-02}	$\mathbf{8.307 \times 10^{-02}}$	$\mathbf{8.307 \times 10^{-02}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Enolase	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{2.319 \times 10^{-01}}$	$\mathbf{5.270 \times 10^{-02}}$	$\mathbf{1.316 \times 10^{-01}}$	1.841×10^{-02}	$\mathbf{8.715 \times 10^{-01}}$	$\mathbf{3.925 \times 10^{-01}}$	7.927×10^{-06}	9.822×10^{-05}
		Backbone$^{0.2}$	$\mathbf{2.319 \times 10^{-01}}$	$\mathbf{5.270 \times 10^{-02}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{8.307 \times 10^{-02}}$	$\mathbf{8.307 \times 10^{-02}}$	9.981×10^{-03}	9.792×10^{-03}	4.592×10^{-13}	6.521×10^{-09}
		Backbone$^{0.4}$	$\mathbf{1.316 \times 10^{-01}}$	1.841×10^{-02}	$\mathbf{8.307 \times 10^{-02}}$	$\mathbf{8.307 \times 10^{-02}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.585×10^{-02}	1.076×10^{-02}	5.921×10^{-09}	3.454×10^{-05}
		Backbone$^{0.6}$	$\mathbf{8.715 \times 10^{-01}}$	$\mathbf{3.925 \times 10^{-01}}$	9.981×10^{-03}	9.792×10^{-03}	1.585×10^{-02}	1.076×10^{-02}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	6.585×10^{-18}	2.130×10^{-11}
		Backbone$^{0.8}$	7.927×10^{-06}	9.822×10^{-05}	4.592×10^{-13}	6.521×10^{-09}	5.921×10^{-09}	3.454×10^{-05}	6.585×10^{-18}	2.130×10^{-11}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Haloacid Dehalogenase	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$0.000 \times 10^{+00}$	$\mathbf{1.000 \times 10^{+00}}$	4.340×10^{-02}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{8.307 \times 10^{-02}}$	$\mathbf{1.608 \times 10^{-01}}$	$\mathbf{1.608 \times 10^{-01}}$
		Backbone$^{0.2}$	$\mathbf{1.000 \times 10^{+00}}$	$0.000 \times 10^{+00}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	3.218×10^{-14}	$\mathbf{1.000 \times 10^{+00}}$	4.879×10^{-16}	$\mathbf{1.608 \times 10^{-01}}$	1.342×10^{-18}
		Backbone$^{0.4}$	$\mathbf{1.000 \times 10^{+00}}$	4.340×10^{-02}	$\mathbf{1.000 \times 10^{+00}}$	3.218×10^{-14}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{3.256 \times 10^{-01}}$	$\mathbf{1.608 \times 10^{-01}}$	$\mathbf{3.256 \times 10^{-01}}$
		Backbone$^{0.6}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{8.307 \times 10^{-02}}$	$\mathbf{1.000 \times 10^{+00}}$	4.879×10^{-16}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{3.256 \times 10^{-01}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.608 \times 10^{-01}}$	$\mathbf{5.725 \times 10^{-01}}$
		Backbone$^{0.8}$	$\mathbf{1.608 \times 10^{-01}}$	$\mathbf{1.608 \times 10^{-01}}$	$\mathbf{1.608 \times 10^{-01}}$	1.342×10^{-18}	$\mathbf{1.608 \times 10^{-01}}$	$\mathbf{3.256 \times 10^{-01}}$	$\mathbf{1.608 \times 10^{-01}}$	$\mathbf{5.725 \times 10^{-01}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Isoprenoid Synthase Type I	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	8.330×10^{-13}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
		Backbone$^{0.2}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	8.330×10^{-13}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
		Backbone$^{0.4}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	8.330×10^{-13}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
		Backbone$^{0.6}$	$\mathbf{1.000 \times 10^{+00}}$	8.330×10^{-13}	$\mathbf{1.000 \times 10^{+00}}$	8.330×10^{-13}	$\mathbf{1.000 \times 10^{+00}}$	8.330×10^{-13}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	8.330×10^{-13}
		Backbone$^{0.8}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	8.330×10^{-13}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Vicinal Oxygen Chelate	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.261×10^{-02}	2.261×10^{-02}	2.261×10^{-02}	2.261×10^{-02}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
		Backbone$^{0.2}$	2.261×10^{-02}	2.261×10^{-02}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.261×10^{-02}	2.261×10^{-02}	2.261×10^{-02}	2.261×10^{-02}
		Backbone$^{0.4}$	2.261×10^{-02}	2.261×10^{-02}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.261×10^{-02}	2.261×10^{-02}	2.261×10^{-02}	2.261×10^{-02}
		Backbone$^{0.6}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.261×10^{-02}	2.261×10^{-02}	2.261×10^{-02}	2.261×10^{-02}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
		Backbone$^{0.8}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.261×10^{-02}	2.261×10^{-02}	2.261×10^{-02}	2.261×10^{-02}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	All	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	7.236×10^{-06}	7.236×10^{-06}	5.193×10^{-03}	4.948×10^{-03}	$\mathbf{5.194 \times 10^{-01}}$	$\mathbf{4.746 \times 10^{-01}}$	5.799×10^{-05}	4.943×10^{-05}
		Backbone$^{0.2}$	7.236×10^{-06}	7.236×10^{-06}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	6.057×10^{-10}	1.138×10^{-10}	3.487×10^{-09}	5.547×10^{-09}	$\mathbf{4.421 \times 10^{-01}}$	$\mathbf{5.861 \times 10^{-01}}$
		Backbone$^{0.4}$	5.193×10^{-03}	4.948×10^{-03}	6.057×10^{-10}	1.138×10^{-10}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	4.484×10^{-04}	4.484×10^{-04}	1.333×10^{-03}	1.782×10^{-03}
		Backbone$^{0.6}$	$\mathbf{5.194 \times 10^{-01}}$	$\mathbf{4.746 \times 10^{-01}}$	3.487×10^{-09}	5.547×10^{-09}	4.484×10^{-04}	4.484×10^{-04}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	3.981×10^{-14}	4.592×10^{-13}
		Backbone$^{0.8}$	5.799×10^{-05}	4.943×10^{-05}	$\mathbf{4.421 \times 10^{-01}}$	$\mathbf{5.861 \times 10^{-01}}$	1.333×10^{-03}	1.782×10^{-03}	3.981×10^{-14}	4.592×10^{-13}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$

Tabela 5.9: *P-valores* para a classificação estrutural para a base Full-Scop. Pontos intermediários. Grupo controle C_α. Hipóteses nulas rejeitadas estão realçadas em negrito.

Dataset	SCOP Level		C_α		$C_\alpha^{0.2}$		$C_\alpha^{0.4}$		$C_\alpha^{0.6}$		$C_\alpha^{0.8}$	
			Prec	Recall	Prec	Recall	Prec	Recall	Prec	Recall	Prec	Recall
Full-Scop	Class	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.153×10^{-59}	1.974×10^{-56}	5.017×10^{-63}	5.017×10^{-63}	6.656×10^{-55}	6.763×10^{-61}	2.830×10^{-61}	9.853×10^{-56}
		$C_\alpha^{0.2}$	2.153×10^{-59}	1.974×10^{-56}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	6.564×10^{-37}	3.436×10^{-26}	1.386×10^{-24}	3.551×10^{-29}	6.719×10^{-39}	2.175×10^{-30}
		$C_\alpha^{0.4}$	5.017×10^{-63}	5.017×10^{-63}	6.564×10^{-37}	3.436×10^{-26}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.286×10^{-05}	5.697×10^{-23}	3.547×10^{-32}	1.696×10^{-22}
		$C_\alpha^{0.6}$	6.656×10^{-55}	6.763×10^{-61}	1.386×10^{-24}	3.551×10^{-29}	2.286×10^{-05}	5.697×10^{-23}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.227×10^{-16}	5.181×10^{-17}
		$C_\alpha^{0.8}$	2.830×10^{-61}	9.853×10^{-56}	6.719×10^{-39}	2.175×10^{-30}	3.547×10^{-32}	1.696×10^{-22}	1.227×10^{-16}	5.181×10^{-17}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Fold	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.032×10^{-60}	1.680×10^{-60}	1.849×10^{-59}	2.355×10^{-61}	5.457×10^{-61}	2.172×10^{-63}	2.068×10^{-61}	6.030×10^{-64}
		$C_\alpha^{0.2}$	1.032×10^{-60}	1.680×10^{-60}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.556×10^{-26}	1.954×10^{-30}	6.410×10^{-33}	1.878×10^{-37}	1.073×10^{-36}	1.212×10^{-42}
		$C_\alpha^{0.4}$	1.849×10^{-59}	2.355×10^{-61}	1.556×10^{-26}	1.954×10^{-30}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	3.436×10^{-26}	9.258×10^{-32}	2.662×10^{-31}	2.195×10^{-43}
		$C_\alpha^{0.6}$	5.457×10^{-61}	2.172×10^{-63}	6.410×10^{-33}	1.878×10^{-37}	3.436×10^{-26}	9.258×10^{-32}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	7.453×10^{-25}	$0.000 \times 10^{+00}$
		$C_\alpha^{0.8}$	2.068×10^{-61}	6.030×10^{-64}	1.073×10^{-36}	1.212×10^{-42}	2.662×10^{-31}	2.195×10^{-43}	7.453×10^{-25}	$0.000 \times 10^{+00}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Superfamily	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.138×10^{-63}	2.330×10^{-64}	2.031×10^{-64}	1.167×10^{-64}	2.197×10^{-60}	2.161×10^{-70}	2.633×10^{-63}	1.402×10^{-63}
		$C_\alpha^{0.2}$	1.138×10^{-63}	2.330×10^{-64}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$0.000 \times 10^{+00}$	3.380×10^{-28}	4.189×10^{-36}	6.339×10^{-37}	1.332×10^{-45}	6.033×10^{-39}
		$C_\alpha^{0.4}$	2.031×10^{-64}	1.167×10^{-64}	$0.000 \times 10^{+00}$	3.380×10^{-28}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	3.150×10^{-23}	3.150×10^{-23}	4.296×10^{-39}	2.633×10^{-30}
		$C_\alpha^{0.6}$	2.197×10^{-60}	2.161×10^{-70}	4.189×10^{-36}	6.339×10^{-37}	3.150×10^{-23}	3.150×10^{-23}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	5.869×10^{-28}	7.637×10^{-28}
		$C_\alpha^{0.8}$	2.633×10^{-63}	1.402×10^{-63}	1.332×10^{-45}	6.033×10^{-39}	4.296×10^{-39}	2.633×10^{-30}	5.869×10^{-28}	7.637×10^{-28}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Family	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	6.301×10^{-67}	3.380×10^{-65}	6.528×10^{-64}	1.347×10^{-64}	1.262×10^{-71}	7.068×10^{-62}	3.169×10^{-72}	7.638×10^{-63}
		$C_\alpha^{0.2}$	6.301×10^{-67}	3.380×10^{-65}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	3.511×10^{-31}	5.195×10^{-38}	4.741×10^{-42}	3.270×10^{-37}	1.049×10^{-46}	7.445×10^{-40}
		$C_\alpha^{0.4}$	6.528×10^{-64}	1.347×10^{-64}	3.511×10^{-31}	5.195×10^{-38}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	6.216×10^{-22}	4.535×10^{-24}	7.298×10^{-34}	6.709×10^{-34}
		$C_\alpha^{0.6}$	1.262×10^{-71}	7.068×10^{-62}	4.741×10^{-42}	3.270×10^{-37}	6.216×10^{-22}	4.535×10^{-24}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	4.252×10^{-36}	8.156×10^{-28}
		$C_\alpha^{0.8}$	3.169×10^{-72}	7.638×10^{-63}	1.049×10^{-46}	7.445×10^{-40}	7.298×10^{-34}	6.709×10^{-34}	4.252×10^{-36}	8.156×10^{-28}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$

Tabela 5.10: *P-valores* para a predição de função para a base Full-Scop. Pontos intermediários. Grupo controle *Backbone*. Hipóteses nulas rejeitadas estão realçadas em negrito.

Dataset	SCOP Level		*Backbone*		*Backbone*$^{0.2}$		*Backbone*$^{0.4}$		*Backbone*$^{0.6}$		*Backbone*$^{0.8}$	
			Prec	Recall	Prec	Recall	Prec	Recall	Prec	Recall	Prec	Recall
Full-Scop	Class	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	9.258×10^{-32}	7.160×10^{-22}	3.294×10^{-23}	1.227×10^{-16}	5.697×10^{-23}	5.697×10^{-23}	6.889×10^{-08}	$0.000 \times 10^{+00}$
		Backbone$^{0.2}$	9.258×10^{-32}	7.160×10^{-22}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{8.307 \times 10^{-02}}$	2.939×10^{-03}	$0.000 \times 10^{+00}$	3.218×10^{-14}	1.428×10^{-23}	3.270×10^{-27}
		Backbone$^{0.4}$	3.294×10^{-23}	1.227×10^{-16}	$\mathbf{8.307 \times 10^{-02}}$	2.939×10^{-03}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	4.879×10^{-16}	3.154×10^{-07}	4.817×10^{-21}	1.696×10^{-22}
		Backbone$^{0.6}$	5.697×10^{-23}	5.697×10^{-23}	$0.000 \times 10^{+00}$	3.218×10^{-14}	4.879×10^{-16}	3.154×10^{-07}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	6.205×10^{-18}	9.258×10^{-32}
		Backbone$^{0.8}$	6.889×10^{-08}	$0.000 \times 10^{+00}$	1.428×10^{-23}	3.270×10^{-27}	4.817×10^{-21}	1.696×10^{-22}	6.205×10^{-18}	9.258×10^{-32}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Fold	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.172×10^{-24}	8.038×10^{-22}	2.807×10^{-28}	1.206×10^{-20}	$0.000 \times 10^{+00}$	1.105×10^{-06}	$0.000 \times 10^{+00}$	2.374×10^{-17}
		Backbone$^{0.2}$	1.172×10^{-24}	8.038×10^{-22}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.138×10^{-08}	1.426×10^{-03}	3.789×10^{-19}	1.302×10^{-15}	1.810×10^{-31}	9.671×10^{-28}
		Backbone$^{0.4}$	2.807×10^{-28}	1.206×10^{-20}	2.138×10^{-08}	1.426×10^{-03}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	3.482×10^{-20}	4.811×10^{-11}	9.737×10^{-37}	1.268×10^{-25}
		Backbone$^{0.6}$	$0.000 \times 10^{+00}$	1.105×10^{-06}	3.789×10^{-19}	1.302×10^{-15}	3.482×10^{-20}	4.811×10^{-11}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$0.000 \times 10^{+00}$	9.643×10^{-22}
		Backbone$^{0.8}$	$0.000 \times 10^{+00}$	2.374×10^{-17}	1.810×10^{-31}	9.671×10^{-28}	9.737×10^{-37}	1.268×10^{-25}	$0.000 \times 10^{+00}$	9.643×10^{-22}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Superfamily	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	6.564×10^{-37}	8.795×10^{-21}	1.428×10^{-23}	7.528×10^{-22}	6.884×10^{-14}	5.284×10^{-14}	2.807×10^{-28}	3.577×10^{-19}
		Backbone$^{0.2}$	6.564×10^{-37}	8.795×10^{-21}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.426×10^{-03}	6.723×10^{-04}	1.693×10^{-18}	1.731×10^{-15}	1.492×10^{-43}	5.114×10^{-28}
		Backbone$^{0.4}$	1.428×10^{-23}	7.528×10^{-22}	1.426×10^{-03}	6.723×10^{-04}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	3.349×10^{-12}	3.482×10^{-20}	1.810×10^{-31}	7.185×10^{-34}
		Backbone$^{0.6}$	6.884×10^{-14}	5.284×10^{-14}	1.693×10^{-18}	1.731×10^{-15}	3.349×10^{-12}	3.482×10^{-20}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.578×10^{-26}	3.551×10^{-29}
		Backbone$^{0.8}$	2.807×10^{-28}	3.577×10^{-19}	1.492×10^{-43}	5.114×10^{-28}	1.810×10^{-31}	7.185×10^{-34}	2.578×10^{-26}	3.551×10^{-29}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Family	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	4.644×10^{-31}	2.388×10^{-28}	1.705×10^{-27}	1.887×10^{-27}	8.017×10^{-21}	2.099×10^{-20}	1.359×10^{-16}	2.046×10^{-16}
		Backbone$^{0.2}$	4.644×10^{-31}	2.388×10^{-28}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	5.689×10^{-05}	1.349×10^{-04}	2.835×10^{-15}	5.439×10^{-16}	6.804×10^{-30}	2.698×10^{-34}
		Backbone$^{0.4}$	1.705×10^{-27}	1.887×10^{-27}	5.689×10^{-05}	1.349×10^{-04}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.993×10^{-17}	3.496×10^{-10}	2.465×10^{-30}	3.079×10^{-29}
		Backbone$^{0.6}$	8.017×10^{-21}	2.099×10^{-20}	2.835×10^{-15}	5.439×10^{-16}	2.993×10^{-17}	3.496×10^{-10}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.160×10^{-29}	7.939×10^{-30}
		Backbone$^{0.8}$	1.359×10^{-16}	2.046×10^{-16}	6.804×10^{-30}	2.698×10^{-34}	2.465×10^{-30}	3.079×10^{-29}	2.160×10^{-29}	7.939×10^{-30}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$

Tabela 5.11: *P-valores* para a classificação estrutural para a base 3SSE. Pontos intermediários. Grupo controle C_α. Hipóteses nulas rejeitadas estão realçadas em negrito.

Dataset	SCOP Level		C_α		$C_\alpha^{0.2}$		$C_\alpha^{0.4}$		$C_\alpha^{0.6}$		$C_\alpha^{0.8}$	
			Prec	Recall	Prec	Recall	Prec	Recall	Prec	Recall	Prec	Recall
3SSE	Class	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	6.727×10^{-15}	1.375×10^{-14}	4.448×10^{-15}	3.473×10^{-14}	3.430×10^{-17}	3.265×10^{-16}	7.665×10^{-14}	7.248×10^{-13}
		$C_\alpha^{0.2}$	6.727×10^{-15}	1.375×10^{-14}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	3.035×10^{-03}	$\mathbf{5.501 \times 10^{-02}}$	8.474×10^{-09}	3.016×10^{-08}	7.654×10^{-04}	1.793×10^{-07}
		$C_\alpha^{0.4}$	4.448×10^{-15}	3.473×10^{-14}	3.035×10^{-03}	$\mathbf{5.501 \times 10^{-02}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.540×10^{-08}	8.470×10^{-09}	2.601×10^{-08}	1.930×10^{-09}
		$C_\alpha^{0.6}$	3.430×10^{-17}	3.265×10^{-16}	8.474×10^{-09}	3.016×10^{-08}	2.540×10^{-08}	8.470×10^{-09}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	3.337×10^{-15}	2.094×10^{-15}
		$C_\alpha^{0.8}$	7.665×10^{-14}	7.248×10^{-13}	7.654×10^{-04}	1.793×10^{-07}	2.601×10^{-08}	1.930×10^{-09}	3.337×10^{-15}	2.094×10^{-15}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Fold	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	5.231×10^{-24}	5.484×10^{-25}	1.894×10^{-22}	1.682×10^{-23}	1.448×10^{-20}	2.810×10^{-22}	3.226×10^{-24}	3.049×10^{-25}
		$C_\alpha^{0.2}$	5.231×10^{-24}	5.484×10^{-25}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.131×10^{-09}	3.755×10^{-12}	2.600×10^{-08}	3.979×10^{-09}	1.232×10^{-02}	2.436×10^{-04}
		$C_\alpha^{0.4}$	1.894×10^{-22}	1.682×10^{-23}	1.131×10^{-09}	3.755×10^{-12}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{2.657 \times 10^{-01}}$	$\mathbf{4.844 \times 10^{-01}}$	7.262×10^{-10}	2.321×10^{-10}
		$C_\alpha^{0.6}$	1.448×10^{-20}	2.810×10^{-22}	2.600×10^{-08}	3.979×10^{-09}	$\mathbf{2.657 \times 10^{-01}}$	$\mathbf{4.844 \times 10^{-01}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	9.528×10^{-09}	7.891×10^{-08}
		$C_\alpha^{0.8}$	3.226×10^{-24}	3.049×10^{-25}	1.232×10^{-02}	2.436×10^{-04}	7.262×10^{-10}	2.321×10^{-10}	9.528×10^{-09}	7.891×10^{-08}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Superfamily	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	5.577×10^{-22}	2.316×10^{-23}	6.976×10^{-20}	6.345×10^{-21}	6.681×10^{-20}	2.252×10^{-21}	2.099×10^{-22}	1.295×10^{-23}
		$C_\alpha^{0.2}$	5.577×10^{-22}	2.316×10^{-23}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.780×10^{-07}	1.118×10^{-09}	1.210×10^{-06}	5.564×10^{-07}	$\mathbf{9.285 \times 10^{-01}}$	$\mathbf{2.983 \times 10^{-01}}$
		$C_\alpha^{0.4}$	6.976×10^{-20}	6.345×10^{-21}	1.780×10^{-07}	1.118×10^{-09}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{5.784 \times 10^{-01}}$	$\mathbf{1.453 \times 10^{-01}}$	4.779×10^{-12}	1.190×10^{-11}
		$C_\alpha^{0.6}$	6.681×10^{-20}	2.252×10^{-21}	1.210×10^{-06}	5.564×10^{-07}	$\mathbf{5.784 \times 10^{-01}}$	$\mathbf{1.453 \times 10^{-01}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.932×10^{-10}	1.508×10^{-08}
		$C_\alpha^{0.8}$	2.099×10^{-22}	1.295×10^{-23}	$\mathbf{9.285 \times 10^{-01}}$	$\mathbf{2.983 \times 10^{-01}}$	4.779×10^{-12}	1.190×10^{-11}	2.932×10^{-10}	1.508×10^{-08}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Family	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.237×10^{-25}	5.521×10^{-26}	1.107×10^{-26}	1.838×10^{-26}	1.449×10^{-26}	5.159×10^{-27}	1.002×10^{-29}	1.807×10^{-29}
		$C_\alpha^{0.2}$	1.237×10^{-25}	5.521×10^{-26}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.130×10^{-09}	1.066×10^{-10}	7.728×10^{-10}	4.263×10^{-10}	7.033×10^{-04}	4.384×10^{-05}
		$C_\alpha^{0.4}$	1.107×10^{-26}	1.838×10^{-26}	1.130×10^{-09}	1.066×10^{-10}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	3.895×10^{-02}	$\mathbf{6.969 \times 10^{-01}}$	5.245×10^{-08}	1.610×10^{-06}
		$C_\alpha^{0.6}$	1.449×10^{-26}	5.159×10^{-27}	7.728×10^{-10}	4.263×10^{-10}	3.895×10^{-02}	$\mathbf{6.969 \times 10^{-01}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.901×10^{-10}	5.855×10^{-08}
		$C_\alpha^{0.8}$	1.002×10^{-29}	1.807×10^{-29}	7.033×10^{-04}	4.384×10^{-05}	5.245×10^{-08}	1.610×10^{-06}	1.901×10^{-10}	5.855×10^{-08}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$

Tabela 5.12: *P-valores* para a classificação estrutural para a base 4SSE. Pontos intermediários. Grupo controle C_α. Hipóteses nulas rejeitadas estão realçadas em negrito.

Dataset	SCOP Level		C_α		$C_\alpha^{0.2}$		$C_\alpha^{0.4}$		$C_\alpha^{0.6}$		$C_\alpha^{0.8}$	
			Prec	Recall	Prec	Recall	Prec	Recall	Prec	Recall	Prec	Recall
4SSE	Class	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.897×10^{-25}	3.389×10^{-25}	7.572×10^{-26}	7.901×10^{-26}	9.166×10^{-27}	1.803×10^{-26}	1.097×10^{-24}	4.698×10^{-24}
		$C_\alpha^{0.2}$	2.897×10^{-25}	3.389×10^{-25}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.782×10^{-03}	1.782×10^{-03}	6.564×10^{-07}	6.564×10^{-07}	1.550×10^{-07}	1.799×10^{-07}
		$C_\alpha^{0.4}$	7.572×10^{-26}	7.901×10^{-26}	1.782×10^{-03}	1.782×10^{-03}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.504×10^{-06}	1.504×10^{-06}	6.618×10^{-13}	1.339×10^{-13}
		$C_\alpha^{0.6}$	9.166×10^{-27}	1.803×10^{-26}	6.564×10^{-07}	6.564×10^{-07}	1.504×10^{-06}	1.504×10^{-06}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.573×10^{-16}	3.084×10^{-18}
		$C_\alpha^{0.8}$	1.097×10^{-24}	4.698×10^{-24}	1.550×10^{-07}	1.799×10^{-07}	6.618×10^{-13}	1.339×10^{-13}	2.573×10^{-16}	3.084×10^{-18}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Fold	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.408×10^{-35}	3.960×10^{-36}	1.677×10^{-34}	9.656×10^{-35}	5.264×10^{-36}	1.741×10^{-36}	3.330×10^{-34}	5.697×10^{-35}
		$C_\alpha^{0.2}$	1.408×10^{-35}	3.960×10^{-36}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{5.086 \times 10^{-02}}$	1.396×10^{-02}	$\mathbf{7.826 \times 10^{-01}}$	$\mathbf{3.539 \times 10^{-01}}$	2.047×10^{-13}	1.395×10^{-12}
		$C_\alpha^{0.4}$	1.677×10^{-34}	9.656×10^{-35}	$\mathbf{5.086 \times 10^{-02}}$	1.396×10^{-02}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	4.948×10^{-03}	2.114×10^{-05}	7.174×10^{-13}	8.321×10^{-11}
		$C_\alpha^{0.6}$	5.264×10^{-36}	1.741×10^{-36}	$\mathbf{7.826 \times 10^{-01}}$	$\mathbf{3.539 \times 10^{-01}}$	4.948×10^{-03}	2.114×10^{-05}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.104×10^{-15}	1.763×10^{-16}
		$C_\alpha^{0.8}$	3.330×10^{-34}	5.697×10^{-35}	2.047×10^{-13}	1.395×10^{-12}	7.174×10^{-13}	8.321×10^{-11}	1.104×10^{-15}	1.763×10^{-16}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Superfamily	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	9.489×10^{-34}	1.577×10^{-34}	4.601×10^{-33}	6.721×10^{-34}	1.721×10^{-34}	1.040×10^{-34}	3.732×10^{-34}	6.725×10^{-35}
		$C_\alpha^{0.2}$	9.489×10^{-34}	1.577×10^{-34}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.547 \times 10^{-01}}$	$\mathbf{8.741 \times 10^{-02}}$	$\mathbf{1.431 \times 10^{-01}}$	$\mathbf{2.018 \times 10^{-01}}$	9.508×10^{-08}	9.508×10^{-08}
		$C_\alpha^{0.4}$	4.601×10^{-33}	6.721×10^{-34}	$\mathbf{1.547 \times 10^{-01}}$	$\mathbf{8.741 \times 10^{-02}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.240×10^{-04}	6.810×10^{-06}	9.151×10^{-06}	1.867×10^{-05}
		$C_\alpha^{0.6}$	1.721×10^{-34}	1.040×10^{-34}	$\mathbf{1.431 \times 10^{-01}}$	$\mathbf{2.018 \times 10^{-01}}$	1.240×10^{-04}	6.810×10^{-06}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.866×10^{-11}	3.514×10^{-11}
		$C_\alpha^{0.8}$	3.732×10^{-34}	6.725×10^{-35}	9.508×10^{-08}	9.508×10^{-08}	9.151×10^{-06}	1.867×10^{-05}	1.866×10^{-11}	3.514×10^{-11}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Family	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.753×10^{-37}	1.325×10^{-38}	1.100×10^{-37}	2.982×10^{-38}	7.951×10^{-38}	2.906×10^{-38}	1.160×10^{-37}	2.722×10^{-38}
		$C_\alpha^{0.2}$	2.753×10^{-37}	1.325×10^{-38}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	4.414×10^{-06}	1.848×10^{-05}	1.410×10^{-09}	1.327×10^{-10}	$\mathbf{1.303 \times 10^{-01}}$	$\mathbf{9.953 \times 10^{-02}}$
		$C_\alpha^{0.4}$	1.100×10^{-37}	2.982×10^{-38}	4.414×10^{-06}	1.848×10^{-05}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	3.316×10^{-02}	1.741×10^{-03}	2.945×10^{-08}	3.594×10^{-07}
		$C_\alpha^{0.6}$	7.951×10^{-38}	2.906×10^{-38}	1.410×10^{-09}	1.327×10^{-10}	3.316×10^{-02}	1.741×10^{-03}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.039×10^{-12}	1.100×10^{-12}
		$C_\alpha^{0.8}$	1.160×10^{-37}	2.722×10^{-38}	$\mathbf{1.303 \times 10^{-01}}$	$\mathbf{9.953 \times 10^{-02}}$	2.945×10^{-08}	3.594×10^{-07}	2.039×10^{-12}	1.100×10^{-12}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$

Tabela 5.13: *P-valores* para a classificação estrutural para a base 5SSE. Pontos intermediários. Grupo controle C_α. Hipóteses nulas rejeitadas estão realçadas em negrito.

Dataset	SCOP Level		C_α		$C_\alpha^{0.2}$		$C_\alpha^{0.4}$		$C_\alpha^{0.6}$		$C_\alpha^{0.8}$	
			Prec	Recall	Prec	Recall	Prec	Recall	Prec	Recall	Prec	Recall
5SSE	Class	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	4.764×10^{-31}	3.715×10^{-30}	1.812×10^{-33}	1.812×10^{-33}	3.078×10^{-35}	3.078×10^{-35}	5.911×10^{-35}	5.911×10^{-35}
		$C_\alpha^{0.2}$	4.764×10^{-31}	3.715×10^{-30}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.892×10^{-16}	5.439×10^{-16}	1.535×10^{-23}	1.017×10^{-22}	1.007×10^{-21}	1.259×10^{-21}
		$C_\alpha^{0.4}$	1.812×10^{-33}	1.812×10^{-33}	2.892×10^{-16}	5.439×10^{-16}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.975×10^{-19}	2.975×10^{-19}	6.591×10^{-13}	6.591×10^{-13}
		$C_\alpha^{0.6}$	3.078×10^{-35}	3.078×10^{-35}	1.535×10^{-23}	1.017×10^{-22}	2.975×10^{-19}	2.975×10^{-19}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.920×10^{-07}	2.920×10^{-07}
		$C_\alpha^{0.8}$	5.911×10^{-35}	5.911×10^{-35}	1.007×10^{-21}	1.259×10^{-21}	6.591×10^{-13}	6.591×10^{-13}	2.920×10^{-07}	2.920×10^{-07}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Fold	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	3.455×10^{-37}	6.086×10^{-37}	9.973×10^{-39}	5.392×10^{-38}	3.629×10^{-41}	1.322×10^{-40}	1.333×10^{-40}	7.435×10^{-40}
		$C_\alpha^{0.2}$	3.455×10^{-37}	6.086×10^{-37}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.771×10^{-13}	3.607×10^{-12}	4.718×10^{-20}	4.032×10^{-19}	1.524×10^{-21}	1.112×10^{-20}
		$C_\alpha^{0.4}$	9.973×10^{-39}	5.392×10^{-38}	1.771×10^{-13}	3.607×10^{-12}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.484×10^{-16}	1.984×10^{-17}	2.375×10^{-18}	3.214×10^{-17}
		$C_\alpha^{0.6}$	3.629×10^{-41}	1.322×10^{-40}	4.718×10^{-20}	4.032×10^{-19}	1.484×10^{-16}	1.984×10^{-17}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{3.050 \times 10^{-01}}$	$\mathbf{3.050 \times 10^{-01}}$
		$C_\alpha^{0.8}$	1.333×10^{-40}	7.435×10^{-40}	1.524×10^{-21}	1.112×10^{-20}	2.375×10^{-18}	3.214×10^{-17}	$\mathbf{3.050 \times 10^{-01}}$	$\mathbf{3.050 \times 10^{-01}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Superfamily	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	3.722×10^{-39}	9.441×10^{-40}	7.007×10^{-40}	1.286×10^{-40}	1.764×10^{-41}	1.061×10^{-42}	2.220×10^{-40}	8.120×10^{-43}
		$C_\alpha^{0.2}$	3.722×10^{-39}	9.441×10^{-40}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.892×10^{-10}	1.109×10^{-05}	4.279×10^{-18}	1.245×10^{-14}	4.159×10^{-17}	8.283×10^{-16}
		$C_\alpha^{0.4}$	7.007×10^{-40}	1.286×10^{-40}	1.892×10^{-10}	1.109×10^{-05}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.996×10^{-17}	9.323×10^{-17}	4.102×10^{-14}	3.535×10^{-15}
		$C_\alpha^{0.6}$	1.764×10^{-41}	1.061×10^{-42}	4.279×10^{-18}	1.245×10^{-14}	2.996×10^{-17}	9.323×10^{-17}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{2.061 \times 10^{-01}}$	$\mathbf{3.256 \times 10^{-01}}$
		$C_\alpha^{0.8}$	2.220×10^{-40}	8.120×10^{-43}	4.159×10^{-17}	8.283×10^{-16}	4.102×10^{-14}	3.535×10^{-15}	$\mathbf{2.061 \times 10^{-01}}$	$\mathbf{3.256 \times 10^{-01}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Family	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.751×10^{-36}	3.946×10^{-36}	2.099×10^{-36}	1.867×10^{-36}	1.548×10^{-37}	2.906×10^{-37}	1.332×10^{-37}	1.267×10^{-37}
		$C_\alpha^{0.2}$	1.751×10^{-36}	3.946×10^{-36}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	8.647×10^{-05}	4.975×10^{-02}	1.239×10^{-19}	1.028×10^{-19}	3.152×10^{-20}	2.550×10^{-18}
		$C_\alpha^{0.4}$	2.099×10^{-36}	1.867×10^{-36}	8.647×10^{-05}	4.975×10^{-02}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.906×10^{-16}	4.879×10^{-16}	3.916×10^{-16}	4.399×10^{-17}
		$C_\alpha^{0.6}$	1.548×10^{-37}	2.906×10^{-37}	1.239×10^{-19}	1.028×10^{-19}	1.906×10^{-16}	4.879×10^{-16}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{3.256 \times 10^{-01}}$	2.011×10^{-04}
		$C_\alpha^{0.8}$	1.332×10^{-37}	1.267×10^{-37}	3.152×10^{-20}	2.550×10^{-18}	3.916×10^{-16}	4.399×10^{-17}	$\mathbf{3.256 \times 10^{-01}}$	2.011×10^{-04}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$

Tabela 5.14: *P-valores* para a classificação estrutural para a base 6SSE. Pontos intermediários. Grupo controle C_α. Hipóteses nulas rejeitadas estão realçadas em negrito.

Dataset	SCOP Level		C_α		$C_\alpha^{0.2}$		$C_\alpha^{0.4}$		$C_\alpha^{0.6}$		$C_\alpha^{0.8}$	
			Prec	Recall	Prec	Recall	Prec	Recall	Prec	Recall	Prec	Recall
6SSE	Class	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	4.671×10^{-30}	1.034×10^{-29}	2.225×10^{-31}	1.552×10^{-30}	1.461×10^{-32}	7.906×10^{-32}	1.240×10^{-32}	7.349×10^{-32}
		$C_\alpha^{0.2}$	4.671×10^{-30}	1.034×10^{-29}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.524×10^{-03}	1.524×10^{-03}	$\mathbf{5.918 \times 10^{-02}}$	$\mathbf{5.918 \times 10^{-02}}$	2.191×10^{-12}	2.191×10^{-12}
		$C_\alpha^{0.4}$	2.225×10^{-31}	1.552×10^{-30}	1.524×10^{-03}	1.524×10^{-03}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{2.313 \times 10^{-01}}$	$\mathbf{2.313 \times 10^{-01}}$	7.459×10^{-10}	7.459×10^{-10}
		$C_\alpha^{0.6}$	1.461×10^{-32}	7.906×10^{-32}	$\mathbf{5.918 \times 10^{-02}}$	$\mathbf{5.918 \times 10^{-02}}$	$\mathbf{2.313 \times 10^{-01}}$	$\mathbf{2.313 \times 10^{-01}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	7.719×10^{-14}	7.719×10^{-14}
		$C_\alpha^{0.8}$	1.240×10^{-32}	7.349×10^{-32}	2.191×10^{-12}	2.191×10^{-12}	7.459×10^{-10}	7.459×10^{-10}	7.719×10^{-14}	7.719×10^{-14}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Fold	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	5.676×10^{-37}	3.197×10^{-37}	2.131×10^{-41}	3.215×10^{-40}	3.096×10^{-40}	4.419×10^{-39}	5.362×10^{-42}	5.144×10^{-40}
		$C_\alpha^{0.2}$	5.676×10^{-37}	3.197×10^{-37}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	7.285×10^{-18}	6.561×10^{-18}	1.239×10^{-21}	2.677×10^{-21}	1.371×10^{-21}	4.967×10^{-21}
		$C_\alpha^{0.4}$	2.131×10^{-41}	3.215×10^{-40}	7.285×10^{-18}	6.561×10^{-18}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.150×10^{-14}	5.105×10^{-16}	2.786×10^{-13}	3.275×10^{-14}
		$C_\alpha^{0.6}$	3.096×10^{-40}	4.419×10^{-39}	1.239×10^{-21}	2.677×10^{-21}	1.150×10^{-14}	5.105×10^{-16}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{8.454 \times 10^{-01}}$	$\mathbf{5.725 \times 10^{-01}}$
		$C_\alpha^{0.8}$	5.362×10^{-42}	5.144×10^{-40}	1.371×10^{-21}	4.967×10^{-21}	2.786×10^{-13}	3.275×10^{-14}	$\mathbf{8.454 \times 10^{-01}}$	$\mathbf{5.725 \times 10^{-01}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Superfamily	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	3.426×10^{-36}	3.801×10^{-36}	5.839×10^{-38}	1.720×10^{-38}	1.486×10^{-37}	3.905×10^{-39}	2.105×10^{-39}	2.537×10^{-39}
		$C_\alpha^{0.2}$	3.426×10^{-36}	3.801×10^{-36}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	7.124×10^{-21}	1.280×10^{-21}	1.411×10^{-25}	5.800×10^{-25}	2.276×10^{-27}	8.898×10^{-27}
		$C_\alpha^{0.4}$	5.839×10^{-38}	1.720×10^{-38}	7.124×10^{-21}	1.280×10^{-21}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	3.789×10^{-19}	8.750×10^{-19}	1.891×10^{-20}	5.044×10^{-19}
		$C_\alpha^{0.6}$	1.486×10^{-37}	3.905×10^{-39}	1.411×10^{-25}	5.800×10^{-25}	3.789×10^{-19}	8.750×10^{-19}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	8.716×10^{-08}	5.273×10^{-04}
		$C_\alpha^{0.8}$	2.105×10^{-39}	2.537×10^{-39}	2.276×10^{-27}	8.898×10^{-27}	1.891×10^{-20}	5.044×10^{-19}	8.716×10^{-08}	5.273×10^{-04}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Family	C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	3.977×10^{-36}	1.338×10^{-36}	1.597×10^{-38}	4.440×10^{-39}	1.403×10^{-39}	2.296×10^{-39}	9.278×10^{-39}	7.885×10^{-39}
		$C_\alpha^{0.2}$	3.977×10^{-36}	1.338×10^{-36}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.329×10^{-21}	4.091×10^{-22}	2.269×10^{-27}	2.475×10^{-28}	1.819×10^{-29}	1.254×10^{-31}
		$C_\alpha^{0.4}$	1.597×10^{-38}	4.440×10^{-39}	2.329×10^{-21}	4.091×10^{-22}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	9.524×10^{-22}	2.227×10^{-20}	1.853×10^{-23}	1.080×10^{-23}
		$C_\alpha^{0.6}$	1.403×10^{-39}	2.296×10^{-39}	2.269×10^{-27}	2.475×10^{-28}	9.524×10^{-22}	2.227×10^{-20}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	3.044×10^{-11}	6.945×10^{-10}
		$C_\alpha^{0.8}$	9.278×10^{-39}	7.885×10^{-39}	1.819×10^{-29}	1.254×10^{-31}	1.853×10^{-23}	1.080×10^{-23}	3.044×10^{-11}	6.945×10^{-10}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$

Tabela 5.15: *P-valores* para a classificação estrutural para a base 3SSE. Pontos intermediários. Grupo controle *Backbone*. Hipóteses nulas rejeitadas estão realçadas em negrito.

Dataset	SCOP Level		*Backbone*		*Backbone*$^{0.2}$		*Backbone*$^{0.4}$		*Backbone*$^{0.6}$		*Backbone*$^{0.8}$	
			Prec	Recall	Prec	Recall	Prec	Recall	Prec	Recall	Prec	Recall
3SSE	Class	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{2.054 \times 10^{-01}}$	$\mathbf{1.388 \times 10^{-01}}$	$\mathbf{3.256 \times 10^{-01}}$	$\mathbf{2.084 \times 10^{-01}}$	3.898×10^{-07}	9.331×10^{-07}	9.781×10^{-03}	4.531×10^{-03}
		Backbone$^{0.2}$	$\mathbf{2.054 \times 10^{-01}}$	$\mathbf{1.388 \times 10^{-01}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{7.687 \times 10^{-01}}$	$\mathbf{1.000 \times 10^{+00}}$	7.595×10^{-12}	2.886×10^{-10}	8.598×10^{-03}	4.070×10^{-03}
		Backbone$^{0.4}$	$\mathbf{3.256 \times 10^{-01}}$	$\mathbf{2.084 \times 10^{-01}}$	$\mathbf{7.687 \times 10^{-01}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	6.319×10^{-11}	5.545×10^{-10}	3.225×10^{-02}	1.801×10^{-02}
		Backbone$^{0.6}$	3.898×10^{-07}	9.331×10^{-07}	7.595×10^{-12}	2.886×10^{-10}	6.319×10^{-11}	5.545×10^{-10}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.117×10^{-04}	4.343×10^{-04}
		Backbone$^{0.8}$	9.781×10^{-03}	4.531×10^{-03}	8.598×10^{-03}	4.070×10^{-03}	3.225×10^{-02}	1.801×10^{-02}	2.117×10^{-04}	4.343×10^{-04}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Fold	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.832×10^{-12}	1.954×10^{-11}	2.023×10^{-16}	4.576×10^{-16}	1.555×10^{-11}	6.079×10^{-09}	3.983×10^{-07}	5.428×10^{-12}
		Backbone$^{0.2}$	2.832×10^{-12}	1.954×10^{-11}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	3.695×10^{-09}	3.487×10^{-09}	1.781×10^{-06}	1.540×10^{-07}	1.547×10^{-14}	3.128×10^{-16}
		Backbone$^{0.4}$	2.023×10^{-16}	4.576×10^{-16}	3.695×10^{-09}	3.487×10^{-09}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	3.003×10^{-19}	3.931×10^{-19}	5.553×10^{-18}	1.080×10^{-19}
		Backbone$^{0.6}$	1.555×10^{-11}	6.079×10^{-09}	1.781×10^{-06}	1.540×10^{-07}	3.003×10^{-19}	3.931×10^{-19}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.644×10^{-14}	2.453×10^{-15}
		Backbone$^{0.8}$	3.983×10^{-07}	5.428×10^{-12}	1.547×10^{-14}	3.128×10^{-16}	5.553×10^{-18}	1.080×10^{-19}	1.644×10^{-14}	2.453×10^{-15}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Superfamily	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	4.362×10^{-13}	1.420×10^{-12}	8.166×10^{-18}	1.441×10^{-17}	2.155×10^{-11}	1.173×10^{-09}	1.645×10^{-08}	9.698×10^{-13}
		Backbone$^{0.2}$	4.362×10^{-13}	1.420×10^{-12}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	5.544×10^{-10}	2.502×10^{-11}	1.425×10^{-05}	2.008×10^{-06}	1.081×10^{-17}	1.473×10^{-19}
		Backbone$^{0.4}$	8.166×10^{-18}	1.441×10^{-17}	5.544×10^{-10}	2.502×10^{-11}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.761×10^{-18}	3.682×10^{-19}	4.972×10^{-22}	6.400×10^{-24}
		Backbone$^{0.6}$	2.155×10^{-11}	1.173×10^{-09}	1.425×10^{-05}	2.008×10^{-06}	1.761×10^{-18}	3.682×10^{-19}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	3.532×10^{-16}	2.281×10^{-17}
		Backbone$^{0.8}$	1.645×10^{-08}	9.698×10^{-13}	1.081×10^{-17}	1.473×10^{-19}	4.972×10^{-22}	6.400×10^{-24}	3.532×10^{-16}	2.281×10^{-17}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Family	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	5.564×10^{-07}	6.842×10^{-06}	1.555×10^{-06}	6.355×10^{-06}	5.635×10^{-05}	7.835×10^{-05}	1.509×10^{-03}	$\mathbf{1.008 \times 10^{-01}}$
		Backbone$^{0.2}$	5.564×10^{-07}	6.842×10^{-06}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{4.021 \times 10^{-01}}$	$\mathbf{7.706 \times 10^{-02}}$	$\mathbf{7.265 \times 10^{-02}}$	$\mathbf{6.569 \times 10^{-01}}$	2.232×10^{-04}	2.946×10^{-05}
		Backbone$^{0.4}$	1.555×10^{-06}	6.355×10^{-06}	$\mathbf{4.021 \times 10^{-01}}$	$\mathbf{7.706 \times 10^{-02}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	4.164×10^{-02}	$\mathbf{9.827 \times 10^{-02}}$	5.341×10^{-04}	5.502×10^{-05}
		Backbone$^{0.6}$	5.635×10^{-05}	7.835×10^{-05}	$\mathbf{7.265 \times 10^{-02}}$	$\mathbf{6.569 \times 10^{-01}}$	4.164×10^{-02}	$\mathbf{9.827 \times 10^{-02}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.702×10^{-02}	1.068×10^{-03}
		Backbone$^{0.8}$	1.509×10^{-03}	$\mathbf{1.008 \times 10^{-01}}$	2.232×10^{-04}	2.946×10^{-05}	5.341×10^{-04}	5.502×10^{-05}	1.702×10^{-02}	1.068×10^{-03}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$

Tabela 5.16: *P-valores* para a classificação estrutural para a base 4SSE. Pontos intermediários. Grupo controle *Backbone*. Hipóteses nulas rejeitadas estão realçadas em negrito.

Dataset	SCOP Level		*Backbone*		*Backbone*$^{0.2}$		*Backbone*$^{0.4}$		*Backbone*$^{0.6}$		*Backbone*$^{0.8}$	
			Prec	Recall	Prec	Recall	Prec	Recall	Prec	Recall	Prec	Recall
4SSE	Class	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	6.866×10^{-09}	2.298×10^{-09}	6.866×10^{-09}	2.298×10^{-09}	4.608×10^{-05}	3.032×10^{-05}	$\mathbf{5.312 \times 10^{-02}}$	2.962×10^{-02}
		Backbone$^{0.2}$	6.866×10^{-09}	2.298×10^{-09}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.316×10^{-03}	2.316×10^{-03}	4.580×10^{-07}	4.580×10^{-07}
		Backbone$^{0.4}$	6.866×10^{-09}	2.298×10^{-09}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.316×10^{-03}	2.316×10^{-03}	4.580×10^{-07}	4.580×10^{-07}
		Backbone$^{0.6}$	4.608×10^{-05}	3.032×10^{-05}	2.316×10^{-03}	2.316×10^{-03}	2.316×10^{-03}	2.316×10^{-03}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	4.343×10^{-04}	4.343×10^{-04}
		Backbone$^{0.8}$	$\mathbf{5.312 \times 10^{-02}}$	2.962×10^{-02}	4.580×10^{-07}	4.580×10^{-07}	4.580×10^{-07}	4.580×10^{-07}	4.343×10^{-04}	4.343×10^{-04}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Fold	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	6.797×10^{-12}	2.286×10^{-11}	2.415×10^{-11}	3.203×10^{-09}	1.757×10^{-07}	2.037×10^{-06}	9.060×10^{-25}	6.943×10^{-24}
		Backbone$^{0.2}$	6.797×10^{-12}	2.286×10^{-11}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	7.033×10^{-04}	8.298×10^{-07}	1.244×10^{-02}	6.971×10^{-03}	5.660×10^{-20}	8.732×10^{-19}
		Backbone$^{0.4}$	2.415×10^{-11}	3.203×10^{-09}	7.033×10^{-04}	8.298×10^{-07}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{2.715 \times 10^{-01}}$	$\mathbf{4.282 \times 10^{-01}}$	1.702×10^{-21}	6.715×10^{-20}
		Backbone$^{0.6}$	1.757×10^{-07}	2.037×10^{-06}	1.244×10^{-02}	6.971×10^{-03}	$\mathbf{2.715 \times 10^{-01}}$	$\mathbf{4.282 \times 10^{-01}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	6.107×10^{-20}	4.606×10^{-20}
		Backbone$^{0.8}$	9.060×10^{-25}	6.943×10^{-24}	5.660×10^{-20}	8.732×10^{-19}	1.702×10^{-21}	6.715×10^{-20}	6.107×10^{-20}	4.606×10^{-20}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Superfamily	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	3.528×10^{-11}	2.951×10^{-11}	1.974×10^{-12}	1.184×10^{-11}	2.166×10^{-09}	1.582×10^{-08}	1.570×10^{-24}	3.243×10^{-26}
		Backbone$^{0.2}$	3.528×10^{-11}	2.951×10^{-11}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{2.582 \times 10^{-01}}$	4.070×10^{-03}	$\mathbf{7.011 \times 10^{-02}}$	$\mathbf{9.195 \times 10^{-02}}$	3.748×10^{-16}	6.206×10^{-16}
		Backbone$^{0.4}$	1.974×10^{-12}	1.184×10^{-11}	$\mathbf{2.582 \times 10^{-01}}$	4.070×10^{-03}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{2.094 \times 10^{-01}}$	$\mathbf{8.103 \times 10^{-01}}$	1.279×10^{-17}	1.492×10^{-19}
		Backbone$^{0.6}$	2.166×10^{-09}	1.582×10^{-08}	$\mathbf{7.011 \times 10^{-02}}$	$\mathbf{9.195 \times 10^{-02}}$	$\mathbf{2.094 \times 10^{-01}}$	$\mathbf{8.103 \times 10^{-01}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.055×10^{-17}	2.808×10^{-18}
		Backbone$^{0.8}$	1.570×10^{-24}	3.243×10^{-26}	3.748×10^{-16}	6.206×10^{-16}	1.279×10^{-17}	1.492×10^{-19}	1.055×10^{-17}	2.808×10^{-18}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Family	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.647×10^{-15}	2.529×10^{-15}	4.087×10^{-16}	5.740×10^{-16}	9.722×10^{-09}	1.228×10^{-07}	4.693×10^{-18}	4.186×10^{-17}
		Backbone$^{0.2}$	2.647×10^{-15}	2.529×10^{-15}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	8.068×10^{-03}	1.524×10^{-03}	1.170×10^{-05}	4.601×10^{-07}	1.544×10^{-07}	3.109×10^{-06}
		Backbone$^{0.4}$	4.087×10^{-16}	5.740×10^{-16}	8.068×10^{-03}	1.524×10^{-03}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	3.853×10^{-04}	9.838×10^{-06}	1.288×10^{-09}	2.073×10^{-08}
		Backbone$^{0.6}$	9.722×10^{-09}	1.228×10^{-07}	1.170×10^{-05}	4.601×10^{-07}	3.853×10^{-04}	9.838×10^{-06}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.951×10^{-11}	1.752×10^{-11}
		Backbone$^{0.8}$	4.693×10^{-18}	4.186×10^{-17}	1.544×10^{-07}	3.109×10^{-06}	1.288×10^{-09}	2.073×10^{-08}	2.951×10^{-11}	1.752×10^{-11}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$

Tabela 5.17: *P-valores* para a classificação estrutural para a base 5SSE. Pontos intermediários. Grupo controle *Backbone*. Hipóteses nulas rejeitadas estão realçadas em negrito.

Dataset	SCOP Level		*Backbone*		*Backbone*$^{0.2}$		*Backbone*$^{0.4}$		*Backbone*$^{0.6}$		*Backbone*$^{0.8}$	
			Prec	Recall	Prec	Recall	Prec	Recall	Prec	Recall	Prec	Recall
5SSE	Class	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	4.083×10^{-05}	4.083×10^{-05}	$\mathbf{1.176 \times 10^{-01}}$	$\mathbf{1.176 \times 10^{-01}}$	1.068×10^{-03}	1.068×10^{-03}	1.583×10^{-17}	4.561×10^{-18}
		Backbone$^{0.2}$	4.083×10^{-05}	4.083×10^{-05}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	3.710×10^{-11}	3.710×10^{-11}	4.232×10^{-14}	4.232×10^{-14}	5.252×10^{-19}	3.592×10^{-19}
		Backbone$^{0.4}$	$\mathbf{1.176 \times 10^{-01}}$	$\mathbf{1.176 \times 10^{-01}}$	3.710×10^{-11}	3.710×10^{-11}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.115×10^{-08}	1.115×10^{-08}	1.728×10^{-18}	1.198×10^{-18}
		Backbone$^{0.6}$	1.068×10^{-03}	1.068×10^{-03}	4.232×10^{-14}	4.232×10^{-14}	1.115×10^{-08}	1.115×10^{-08}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	3.438×10^{-12}	1.995×10^{-12}
		Backbone$^{0.8}$	1.583×10^{-17}	4.561×10^{-18}	5.252×10^{-19}	3.592×10^{-19}	1.728×10^{-18}	1.198×10^{-18}	3.438×10^{-12}	1.995×10^{-12}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Fold	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	6.524×10^{-21}	5.594×10^{-21}	7.900×10^{-13}	4.502×10^{-14}	3.056×10^{-17}	3.452×10^{-16}	2.103×10^{-06}	1.258×10^{-08}
		Backbone$^{0.2}$	6.524×10^{-21}	5.594×10^{-21}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	8.042×10^{-18}	8.017×10^{-21}	2.208×10^{-07}	8.203×10^{-06}	2.454×10^{-17}	2.827×10^{-16}
		Backbone$^{0.4}$	7.900×10^{-13}	4.502×10^{-14}	8.042×10^{-18}	8.017×10^{-21}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.815×10^{-10}	8.053×10^{-10}	7.677×10^{-06}	1.147×10^{-05}
		Backbone$^{0.6}$	3.056×10^{-17}	3.452×10^{-16}	2.208×10^{-07}	8.203×10^{-06}	1.815×10^{-10}	8.053×10^{-10}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	5.025×10^{-13}	4.468×10^{-12}
		Backbone$^{0.8}$	2.103×10^{-06}	1.258×10^{-08}	2.454×10^{-17}	2.827×10^{-16}	7.677×10^{-06}	1.147×10^{-05}	5.025×10^{-13}	4.468×10^{-12}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Superfamily	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	7.188×10^{-14}	1.415×10^{-15}	$\mathbf{5.518 \times 10^{-02}}$	3.165×10^{-03}	2.441×10^{-06}	6.845×10^{-09}	$\mathbf{5.047 \times 10^{-02}}$	$\mathbf{6.306 \times 10^{-01}}$
		Backbone$^{0.2}$	7.188×10^{-14}	1.415×10^{-15}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.201×10^{-18}	1.851×10^{-21}	5.944×10^{-10}	5.066×10^{-11}	6.897×10^{-14}	3.289×10^{-16}
		Backbone$^{0.4}$	$\mathbf{5.518 \times 10^{-02}}$	3.165×10^{-03}	1.201×10^{-18}	1.851×10^{-21}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.781×10^{-06}	6.057×10^{-10}	2.685×10^{-03}	2.241×10^{-03}
		Backbone$^{0.6}$	2.441×10^{-06}	6.845×10^{-09}	5.944×10^{-10}	5.066×10^{-11}	1.781×10^{-06}	6.057×10^{-10}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	6.004×10^{-08}	5.489×10^{-11}
		Backbone$^{0.8}$	$\mathbf{5.047 \times 10^{-02}}$	$\mathbf{6.306 \times 10^{-01}}$	6.897×10^{-14}	3.289×10^{-16}	2.685×10^{-03}	2.241×10^{-03}	6.004×10^{-08}	5.489×10^{-11}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Family	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	8.999×10^{-18}	7.692×10^{-22}	6.426×10^{-08}	7.938×10^{-11}	1.685×10^{-10}	2.686×10^{-15}	$\mathbf{5.312 \times 10^{-02}}$	8.598×10^{-03}
		Backbone$^{0.2}$	8.999×10^{-18}	7.692×10^{-22}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.894×10^{-19}	2.178×10^{-22}	1.150×10^{-14}	4.764×10^{-19}	2.002×10^{-15}	9.326×10^{-19}
		Backbone$^{0.4}$	6.426×10^{-08}	7.938×10^{-11}	1.894×10^{-19}	2.178×10^{-22}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.524×10^{-03}	1.792×10^{-09}	1.478×10^{-05}	1.305×10^{-06}
		Backbone$^{0.6}$	1.685×10^{-10}	2.686×10^{-15}	1.150×10^{-14}	4.764×10^{-19}	1.524×10^{-03}	1.792×10^{-09}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.239×10^{-08}	6.973×10^{-11}
		Backbone$^{0.8}$	$\mathbf{5.312 \times 10^{-02}}$	8.598×10^{-03}	2.002×10^{-15}	9.326×10^{-19}	1.478×10^{-05}	1.305×10^{-06}	1.239×10^{-08}	6.973×10^{-11}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$

Tabela 5.18: *P-valores* para a classificação estrutural para a base 6SSE. Pontos intermediários. Grupo controle *Backbone*. Hipóteses nulas rejeitadas estão realçadas em negrito.

Dataset	SCOP Level		*Backbone*		*Backbone*$^{0.2}$		*Backbone*$^{0.4}$		*Backbone*$^{0.6}$		*Backbone*$^{0.8}$	
			Prec	Recall	Prec	Recall	Prec	Recall	Prec	Recall	Prec	Recall
6SSE	Class	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.250×10^{-16}	1.219×10^{-16}	5.588×10^{-17}	4.442×10^{-17}	6.408×10^{-17}	4.404×10^{-17}	1.390×10^{-18}	4.114×10^{-17}
		Backbone$^{0.2}$	1.250×10^{-16}	1.219×10^{-16}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.092×10^{-04}	1.092×10^{-04}	$\mathbf{2.550 \times 10^{-01}}$	$\mathbf{2.550 \times 10^{-01}}$	2.005×10^{-03}	$\mathbf{8.130 \times 10^{-01}}$
		Backbone$^{0.4}$	5.588×10^{-17}	4.442×10^{-17}	1.092×10^{-04}	1.092×10^{-04}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	3.044×10^{-02}	3.044×10^{-02}	$\mathbf{7.122 \times 10^{-01}}$	1.679×10^{-02}
		Backbone$^{0.6}$	6.408×10^{-17}	4.404×10^{-17}	$\mathbf{2.550 \times 10^{-01}}$	$\mathbf{2.550 \times 10^{-01}}$	3.044×10^{-02}	3.044×10^{-02}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.058×10^{-02}	$\mathbf{5.219 \times 10^{-01}}$
		Backbone$^{0.8}$	1.390×10^{-18}	4.114×10^{-17}	2.005×10^{-03}	$\mathbf{8.130 \times 10^{-01}}$	$\mathbf{7.122 \times 10^{-01}}$	1.679×10^{-02}	2.058×10^{-02}	$\mathbf{5.219 \times 10^{-01}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Fold	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	6.561×10^{-18}	3.046×10^{-15}	5.172×10^{-20}	9.160×10^{-20}	9.037×10^{-21}	1.151×10^{-18}	5.487×10^{-16}	5.785×10^{-17}
		Backbone$^{0.2}$	6.561×10^{-18}	3.046×10^{-15}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	7.677×10^{-06}	1.215×10^{-07}	2.099×10^{-09}	1.174×10^{-11}	$\mathbf{6.817 \times 10^{-01}}$	1.852×10^{-03}
		Backbone$^{0.4}$	5.172×10^{-20}	9.160×10^{-20}	7.677×10^{-06}	1.215×10^{-07}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	8.811×10^{-05}	6.493×10^{-06}	5.208×10^{-03}	$\mathbf{2.649 \times 10^{-01}}$
		Backbone$^{0.6}$	9.037×10^{-21}	1.151×10^{-18}	2.099×10^{-09}	1.174×10^{-11}	8.811×10^{-05}	6.493×10^{-06}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.154×10^{-07}	3.454×10^{-05}
		Backbone$^{0.8}$	5.487×10^{-16}	5.785×10^{-17}	$\mathbf{6.817 \times 10^{-01}}$	1.852×10^{-03}	5.208×10^{-03}	$\mathbf{2.649 \times 10^{-01}}$	1.154×10^{-07}	3.454×10^{-05}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Superfamily	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	9.040×10^{-12}	7.052×10^{-10}	5.643×10^{-17}	1.240×10^{-16}	4.159×10^{-17}	7.558×10^{-17}	3.042×10^{-19}	3.592×10^{-19}
		Backbone$^{0.2}$	9.040×10^{-12}	7.052×10^{-10}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.216×10^{-09}	4.592×10^{-13}	7.252×10^{-14}	1.569×10^{-16}	1.018×10^{-10}	1.446×10^{-12}
		Backbone$^{0.4}$	5.643×10^{-17}	1.240×10^{-16}	2.216×10^{-09}	4.592×10^{-13}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	7.760×10^{-06}	7.427×10^{-06}	1.798×10^{-05}	7.689×10^{-06}
		Backbone$^{0.6}$	4.159×10^{-17}	7.558×10^{-17}	7.252×10^{-14}	1.569×10^{-16}	7.760×10^{-06}	7.427×10^{-06}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{3.514 \times 10^{-01}}$	$\mathbf{1.331 \times 10^{-01}}$
		Backbone$^{0.8}$	3.042×10^{-19}	3.592×10^{-19}	1.018×10^{-10}	1.446×10^{-12}	1.798×10^{-05}	7.689×10^{-06}	$\mathbf{3.514 \times 10^{-01}}$	$\mathbf{1.331 \times 10^{-01}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Family	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	7.429×10^{-03}	6.071×10^{-04}	3.851×10^{-08}	5.084×10^{-09}	4.366×10^{-12}	1.827×10^{-13}	3.218×10^{-14}	2.558×10^{-13}
		Backbone$^{0.2}$	7.429×10^{-03}	6.071×10^{-04}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	8.395×10^{-12}	4.411×10^{-11}	5.094×10^{-11}	7.279×10^{-11}	4.122×10^{-15}	1.582×10^{-14}
		Backbone$^{0.4}$	3.851×10^{-08}	5.084×10^{-09}	8.395×10^{-12}	4.411×10^{-11}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.077×10^{-05}	5.163×10^{-06}	1.044×10^{-10}	3.796×10^{-10}
		Backbone$^{0.6}$	4.366×10^{-12}	1.827×10^{-13}	5.094×10^{-11}	7.279×10^{-11}	1.077×10^{-05}	5.163×10^{-06}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	8.324×10^{-04}	9.781×10^{-03}
		Backbone$^{0.8}$	3.218×10^{-14}	2.558×10^{-13}	4.122×10^{-15}	1.582×10^{-14}	1.044×10^{-10}	3.796×10^{-10}	8.324×10^{-04}	9.781×10^{-03}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$

Tabela 5.19: *P-valores* para a classificação estrutural para a bases gold-standard. Melhores resultados. Grupo controle *Backbone*. Hipóteses nulas rejeitadas estão realçadas em negrito.

Dataset	SCOP Level		*Backbone*		*Backbone*$^{0.8}$		C_α		$C_\alpha^{0.8}$	
			Prec	Recall	Prec	Recall	Prec	Recall	Prec	Recall
Gold	Amidohydrolase	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	8.368×10^{-05}	2.253×10^{-17}	8.227×10^{-18}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.200 \times 10^{-01}}$
		Backbone$^{0.8}$	$\mathbf{1.000 \times 10^{+00}}$	8.368×10^{-05}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.253×10^{-17}	2.832×10^{-21}	$\mathbf{1.000 \times 10^{+00}}$	1.114×10^{-10}
		C_α	2.253×10^{-17}	8.227×10^{-18}	2.253×10^{-17}	2.832×10^{-21}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.253×10^{-17}	1.533×10^{-21}
		$C_\alpha^{0.8}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.200 \times 10^{-01}}$	$\mathbf{1.000 \times 10^{+00}}$	1.114×10^{-10}	2.253×10^{-17}	1.533×10^{-21}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Crotonase	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.088×10^{-11}	1.880×10^{-14}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
		Backbone$^{0.8}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.088×10^{-11}	1.880×10^{-14}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
		C_α	1.088×10^{-11}	1.880×10^{-14}	1.088×10^{-11}	1.880×10^{-14}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.088×10^{-11}	1.880×10^{-14}
		$C_\alpha^{0.8}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.088×10^{-11}	1.880×10^{-14}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Enolase	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	7.927×10^{-06}	9.822×10^{-05}	5.810×10^{-08}	1.439×10^{-14}	$\mathbf{4.710 \times 10^{-01}}$	1.781×10^{-03}
		Backbone$^{0.8}$	7.927×10^{-06}	9.822×10^{-05}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	3.015×10^{-05}	1.872×10^{-16}	6.951×10^{-19}	3.885×10^{-04}
		C_α	5.810×10^{-08}	1.439×10^{-14}	3.015×10^{-05}	1.872×10^{-16}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	3.833×10^{-20}	8.587×10^{-19}
		$C_\alpha^{0.8}$	$\mathbf{4.710 \times 10^{-01}}$	1.781×10^{-03}	6.951×10^{-19}	3.885×10^{-04}	3.833×10^{-20}	8.587×10^{-19}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Haloacid Dehalogenase	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.608 \times 10^{-01}}$	$\mathbf{1.608 \times 10^{-01}}$	3.935×10^{-18}	4.657×10^{-19}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
		Backbone$^{0.8}$	$\mathbf{1.608 \times 10^{-01}}$	$\mathbf{1.608 \times 10^{-01}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	9.355×10^{-19}	3.288×10^{-19}	$\mathbf{1.608 \times 10^{-01}}$	$\mathbf{1.608 \times 10^{-01}}$
		C_α	3.935×10^{-18}	4.657×10^{-19}	9.355×10^{-19}	3.288×10^{-19}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	3.935×10^{-18}	4.657×10^{-19}
		$C_\alpha^{0.8}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.608 \times 10^{-01}}$	$\mathbf{1.608 \times 10^{-01}}$	3.935×10^{-18}	4.657×10^{-19}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Isoprenoid Synthase Type I	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$0.000 \times 10^{+00}$	2.462×10^{-22}
		Backbone$^{0.8}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$0.000 \times 10^{+00}$	2.462×10^{-22}
		C_α	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$0.000 \times 10^{+00}$	2.462×10^{-22}
		$C_\alpha^{0.8}$	$0.000 \times 10^{+00}$	2.462×10^{-22}	$0.000 \times 10^{+00}$	2.462×10^{-22}	$0.000 \times 10^{+00}$	2.462×10^{-22}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Vicinal Oxygen Chelate	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	6.723×10^{-04}	6.723×10^{-04}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
		Backbone$^{0.8}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	6.723×10^{-04}	6.723×10^{-04}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
		C_α	6.723×10^{-04}	6.723×10^{-04}	6.723×10^{-04}	6.723×10^{-04}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	6.723×10^{-04}	6.723×10^{-04}
		$C_\alpha^{0.8}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	6.723×10^{-04}	6.723×10^{-04}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	All	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	5.799×10^{-05}	4.943×10^{-05}	4.412×10^{-20}	4.759×10^{-20}	2.168×10^{-04}	2.168×10^{-04}
		Backbone$^{0.8}$	5.799×10^{-05}	4.943×10^{-05}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	4.328×10^{-22}	4.328×10^{-22}	4.598×10^{-03}	5.358×10^{-03}
		C_α	4.412×10^{-20}	4.759×10^{-20}	4.328×10^{-22}	4.328×10^{-22}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	8.531×10^{-22}	8.789×10^{-22}
		$C_\alpha^{0.8}$	2.168×10^{-04}	2.168×10^{-04}	4.598×10^{-03}	5.358×10^{-03}	8.531×10^{-22}	8.789×10^{-22}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$

Tabela 5.20: *P-valores* para a classificação estrutural para a bases Full-Scop. Melhores resultados. Grupo controle *Backbone*. Hipóteses nulas rejeitadas estão realçadas em negrito.

Dataset	SCOP Level		*Backbone*		*Backbone*$^{0.8}$		C_α		$C_\alpha^{0.8}$	
			Prec	Recall	Prec	Recall	Prec	Recall	Prec	Recall
Full-Scop	Class	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	6.889×10^{-08}	$0.000 \times 10^{+00}$	2.986×10^{-61}	9.563×10^{-64}	$\mathbf{1.608 \times 10^{-01}}$	1.033×10^{-06}
		Backbone$^{0.8}$	6.889×10^{-08}	$0.000 \times 10^{+00}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	9.603×10^{-55}	4.320×10^{-64}	8.716×10^{-08}	5.689×10^{-05}
		C_α	2.986×10^{-61}	9.563×10^{-64}	9.603×10^{-55}	4.320×10^{-64}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.830×10^{-61}	9.853×10^{-56}
		$C_\alpha^{0.8}$	$\mathbf{1.608 \times 10^{-01}}$	1.033×10^{-06}	8.716×10^{-08}	5.689×10^{-05}	2.830×10^{-61}	9.853×10^{-56}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Fold	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$0.000 \times 10^{+00}$	2.374×10^{-17}	3.096×10^{-64}	4.114×10^{-64}	1.693×10^{-18}	3.789×10^{-19}
		Backbone$^{0.8}$	$0.000 \times 10^{+00}$	2.374×10^{-17}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.341×10^{-64}	1.114×10^{-59}	1.426×10^{-03}	1.426×10^{-03}
		C_α	3.096×10^{-64}	4.114×10^{-64}	1.341×10^{-64}	1.114×10^{-59}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.068×10^{-61}	6.030×10^{-64}
		$C_\alpha^{0.8}$	1.693×10^{-18}	3.789×10^{-19}	1.426×10^{-03}	1.426×10^{-03}	2.068×10^{-61}	6.030×10^{-64}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Superfamily	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.807×10^{-28}	3.577×10^{-19}	9.706×10^{-64}	5.419×10^{-62}	5.697×10^{-23}	1.225×10^{-16}
		Backbone$^{0.8}$	2.807×10^{-28}	3.577×10^{-19}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.329×10^{-63}	2.544×10^{-64}	4.340×10^{-02}	8.298×10^{-07}
		C_α	9.706×10^{-64}	5.419×10^{-62}	1.329×10^{-63}	2.544×10^{-64}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.633×10^{-63}	1.402×10^{-63}
		$C_\alpha^{0.8}$	5.697×10^{-23}	1.225×10^{-16}	4.340×10^{-02}	8.298×10^{-07}	2.633×10^{-63}	1.402×10^{-63}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Family	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.359×10^{-16}	2.046×10^{-16}	4.602×10^{-65}	5.048×10^{-65}	2.261×10^{-15}	1.144×10^{-11}
		Backbone$^{0.8}$	1.359×10^{-16}	2.046×10^{-16}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.739×10^{-66}	2.575×10^{-68}	8.699×10^{-06}	7.033×10^{-04}
		C_α	4.602×10^{-65}	5.048×10^{-65}	1.739×10^{-66}	2.575×10^{-68}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	3.169×10^{-72}	7.638×10^{-63}
		$C_\alpha^{0.8}$	2.261×10^{-15}	1.144×10^{-11}	8.699×10^{-06}	7.033×10^{-04}	3.169×10^{-72}	7.638×10^{-63}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$

Tabela 5.21: *P-valores* para a classificação estrutural para a bases 3SSE. Melhores resultados. Grupo controle *Backbone*. Hipóteses nulas rejeitadas estão realçadas em negrito.

Dataset	SCOP Level		*Backbone*		*Backbone*$^{0.8}$		C_α		$C_\alpha^{0.8}$	
			Prec	Recall	Prec	Recall	Prec	Recall	Prec	Recall
3SSE	Class	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	9.781×10^{-03}	4.531×10^{-03}	9.566×10^{-15}	3.189×10^{-13}	4.457×10^{-02}	1.540×10^{-02}
		Backbone$^{0.8}$	9.781×10^{-03}	4.531×10^{-03}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.245×10^{-16}	1.354×10^{-14}	3.875×10^{-04}	7.179×10^{-05}
		C_α	9.566×10^{-15}	3.189×10^{-13}	2.245×10^{-16}	1.354×10^{-14}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	7.665×10^{-14}	7.248×10^{-13}
		$C_\alpha^{0.8}$	4.457×10^{-02}	1.540×10^{-02}	3.875×10^{-04}	7.179×10^{-05}	7.665×10^{-14}	7.248×10^{-13}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Fold	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	3.983×10^{-07}	5.428×10^{-12}	6.101×10^{-26}	1.542×10^{-26}	9.489×10^{-09}	4.764×10^{-07}
		Backbone$^{0.8}$	3.983×10^{-07}	5.428×10^{-12}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	4.149×10^{-27}	7.712×10^{-28}	4.121×10^{-12}	2.144×10^{-14}
		C_α	6.101×10^{-26}	1.542×10^{-26}	4.149×10^{-27}	7.712×10^{-28}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	3.226×10^{-24}	3.049×10^{-25}
		$C_\alpha^{0.8}$	9.489×10^{-09}	4.764×10^{-07}	4.121×10^{-12}	2.144×10^{-14}	3.226×10^{-24}	3.049×10^{-25}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Superfamily	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.645×10^{-08}	9.698×10^{-13}	8.738×10^{-26}	2.833×10^{-26}	9.665×10^{-07}	1.874×10^{-05}
		Backbone$^{0.8}$	1.645×10^{-08}	9.698×10^{-13}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.131×10^{-28}	4.549×10^{-29}	1.439×10^{-12}	1.336×10^{-14}
		C_α	8.738×10^{-26}	2.833×10^{-26}	2.131×10^{-28}	4.549×10^{-29}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.099×10^{-22}	1.295×10^{-23}
		$C_\alpha^{0.8}$	9.665×10^{-07}	1.874×10^{-05}	1.439×10^{-12}	1.336×10^{-14}	2.099×10^{-22}	1.295×10^{-23}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Family	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.509×10^{-03}	$\mathbf{1.008 \times 10^{-01}}$	2.167×10^{-25}	1.015×10^{-25}	$\mathbf{6.001 \times 10^{-01}}$	$\mathbf{6.058 \times 10^{-01}}$
		Backbone$^{0.8}$	1.509×10^{-03}	$\mathbf{1.008 \times 10^{-01}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	5.483×10^{-27}	2.148×10^{-27}	2.328×10^{-02}	$\mathbf{4.103 \times 10^{-01}}$
		C_α	2.167×10^{-25}	1.015×10^{-25}	5.483×10^{-27}	2.148×10^{-27}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.002×10^{-29}	1.807×10^{-29}
		$C_\alpha^{0.8}$	$\mathbf{6.001 \times 10^{-01}}$	$\mathbf{6.058 \times 10^{-01}}$	2.328×10^{-02}	$\mathbf{4.103 \times 10^{-01}}$	1.002×10^{-29}	1.807×10^{-29}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$

Tabela 5.22: *P-valores* para a classificação estrutural para a bases 4SSE. Melhores resultados. Grupo controle *Backbone*. Hipóteses nulas rejeitadas estão realçadas em negrito.

Dataset	SCOP Level		*Backbone*		*Backbone*$^{0.8}$		C_α		$C_\alpha^{0.8}$	
			Prec	Recall	Prec	Recall	Prec	Recall	Prec	Recall
4SSE	Class	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{5.312 \times 10^{-02}}$	2.962×10^{-02}	4.565×10^{-25}	1.263×10^{-24}	1.815×10^{-03}	1.832×10^{-03}
		Backbone$^{0.8}$	$\mathbf{5.312 \times 10^{-02}}$	2.962×10^{-02}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	8.623×10^{-25}	9.285×10^{-25}	1.208×10^{-05}	3.918×10^{-06}
		C_α	4.565×10^{-25}	1.263×10^{-24}	8.623×10^{-25}	9.285×10^{-25}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.097×10^{-24}	4.698×10^{-24}
		$C_\alpha^{0.8}$	1.815×10^{-03}	1.832×10^{-03}	1.208×10^{-05}	3.918×10^{-06}	1.097×10^{-24}	4.698×10^{-24}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Fold	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	9.060×10^{-25}	6.943×10^{-24}	6.806×10^{-34}	1.830×10^{-34}	1.588×10^{-03}	2.571×10^{-03}
		Backbone$^{0.8}$	9.060×10^{-25}	6.943×10^{-24}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.166×10^{-37}	2.857×10^{-38}	3.583×10^{-28}	4.911×10^{-26}
		C_α	6.806×10^{-34}	1.830×10^{-34}	1.166×10^{-37}	2.857×10^{-38}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	3.330×10^{-34}	5.697×10^{-35}
		$C_\alpha^{0.8}$	1.588×10^{-03}	2.571×10^{-03}	3.583×10^{-28}	4.911×10^{-26}	3.330×10^{-34}	5.697×10^{-35}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Superfamily	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.570×10^{-24}	3.243×10^{-26}	2.564×10^{-33}	2.111×10^{-34}	$\mathbf{4.746 \times 10^{-01}}$	$\mathbf{1.540 \times 10^{-01}}$
		Backbone$^{0.8}$	1.570×10^{-24}	3.243×10^{-26}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.261×10^{-36}	3.289×10^{-37}	2.878×10^{-20}	9.338×10^{-21}
		C_α	2.564×10^{-33}	2.111×10^{-34}	1.261×10^{-36}	3.289×10^{-37}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	3.732×10^{-34}	6.725×10^{-35}
		$C_\alpha^{0.8}$	$\mathbf{4.746 \times 10^{-01}}$	$\mathbf{1.540 \times 10^{-01}}$	2.878×10^{-20}	9.338×10^{-21}	3.732×10^{-34}	6.725×10^{-35}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Family	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	4.693×10^{-18}	4.186×10^{-17}	2.246×10^{-37}	1.922×10^{-38}	1.886×10^{-07}	7.994×10^{-08}
		Backbone$^{0.8}$	4.693×10^{-18}	4.186×10^{-17}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.985×10^{-38}	4.547×10^{-39}	1.231×10^{-12}	3.784×10^{-11}
		C_α	2.246×10^{-37}	1.922×10^{-38}	1.985×10^{-38}	4.547×10^{-39}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.160×10^{-37}	2.722×10^{-38}
		$C_\alpha^{0.8}$	1.886×10^{-07}	7.994×10^{-08}	1.231×10^{-12}	3.784×10^{-11}	1.160×10^{-37}	2.722×10^{-38}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$

Tabela 5.23: *P-valores* para a classificação estrutural para a bases 5SSE. Melhores resultados. Grupo controle *Backbone*. Hipóteses nulas rejeitadas estão realçadas em negrito.

Dataset	SCOP Level		*Backbone*		*Backbone*$^{0.8}$		C_α		$C_\alpha^{0.8}$	
			Prec	Recall	Prec	Recall	Prec	Recall	Prec	Recall
5SSE	Class	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.583×10^{-17}	4.561×10^{-18}	3.012×10^{-32}	2.024×10^{-32}	5.335×10^{-08}	5.335×10^{-08}
		Backbone$^{0.8}$	1.583×10^{-17}	4.561×10^{-18}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.115×10^{-35}	1.281×10^{-35}	5.331×10^{-11}	9.468×10^{-11}
		C_α	3.012×10^{-32}	2.024×10^{-32}	1.115×10^{-35}	1.281×10^{-35}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	5.911×10^{-35}	5.911×10^{-35}
		$C_\alpha^{0.8}$	5.335×10^{-08}	5.335×10^{-08}	5.331×10^{-11}	9.468×10^{-11}	5.911×10^{-35}	5.911×10^{-35}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Fold	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.103×10^{-06}	1.258×10^{-08}	1.065×10^{-39}	7.705×10^{-39}	5.683×10^{-08}	5.218×10^{-10}
		Backbone$^{0.8}$	2.103×10^{-06}	1.258×10^{-08}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	4.249×10^{-39}	1.596×10^{-39}	$\mathbf{5.015 \times 10^{-01}}$	$\mathbf{9.659 \times 10^{-02}}$
		C_α	1.065×10^{-39}	7.705×10^{-39}	4.249×10^{-39}	1.596×10^{-39}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.333×10^{-40}	7.435×10^{-40}
		$C_\alpha^{0.8}$	5.683×10^{-08}	5.218×10^{-10}	$\mathbf{5.015 \times 10^{-01}}$	$\mathbf{9.659 \times 10^{-02}}$	1.333×10^{-40}	7.435×10^{-40}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Superfamily	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{5.047 \times 10^{-02}}$	$\mathbf{6.306 \times 10^{-01}}$	8.138×10^{-41}	9.461×10^{-43}	3.229×10^{-02}	$\mathbf{8.252 \times 10^{-01}}$
		Backbone$^{0.8}$	$\mathbf{5.047 \times 10^{-02}}$	$\mathbf{6.306 \times 10^{-01}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.281×10^{-39}	3.628×10^{-41}	$\mathbf{7.031 \times 10^{-01}}$	$\mathbf{7.826 \times 10^{-01}}$
		C_α	8.138×10^{-41}	9.461×10^{-43}	2.281×10^{-39}	3.628×10^{-41}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.220×10^{-40}	8.120×10^{-43}
		$C_\alpha^{0.8}$	3.229×10^{-02}	$\mathbf{8.252 \times 10^{-01}}$	$\mathbf{7.031 \times 10^{-01}}$	$\mathbf{7.826 \times 10^{-01}}$	2.220×10^{-40}	8.120×10^{-43}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Family	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{5.312 \times 10^{-02}}$	8.598×10^{-03}	4.437×10^{-38}	4.566×10^{-39}	$\mathbf{3.854 \times 10^{-01}}$	$\mathbf{4.551 \times 10^{-01}}$
		Backbone$^{0.8}$	$\mathbf{5.312 \times 10^{-02}}$	8.598×10^{-03}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	8.359×10^{-38}	8.887×10^{-38}	1.263×10^{-02}	3.701×10^{-02}
		C_α	4.437×10^{-38}	4.566×10^{-39}	8.359×10^{-38}	8.887×10^{-38}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.332×10^{-37}	1.267×10^{-37}
		$C_\alpha^{0.8}$	$\mathbf{3.854 \times 10^{-01}}$	$\mathbf{4.551 \times 10^{-01}}$	1.263×10^{-02}	3.701×10^{-02}	1.332×10^{-37}	1.267×10^{-37}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$

Tabela 5.24: *P-valores* para a classificação estrutural para a bases 6SSE. Melhores resultados. Grupo controle *Backbone*. Hipóteses nulas rejeitadas estão realçadas em negrito.

Dataset	SCOP Level		*Backbone*		*Backbone*$^{0.8}$		C_α		$C_\alpha^{0.8}$	
			Prec	Recall	Prec	Recall	Prec	Recall	Prec	Recall
6SSE	Class	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.390×10^{-18}	4.114×10^{-17}	3.591×10^{-31}	3.534×10^{-30}	2.740×10^{-21}	1.978×10^{-21}
		Backbone$^{0.8}$	1.390×10^{-18}	4.114×10^{-17}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	7.830×10^{-32}	1.413×10^{-30}	2.509×10^{-06}	1.919×10^{-09}
		C_α	3.591×10^{-31}	3.534×10^{-30}	7.830×10^{-32}	1.413×10^{-30}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.240×10^{-32}	7.349×10^{-32}
		$C_\alpha^{0.8}$	2.740×10^{-21}	1.978×10^{-21}	2.509×10^{-06}	1.919×10^{-09}	1.240×10^{-32}	7.349×10^{-32}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Fold	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	5.487×10^{-16}	5.785×10^{-17}	2.101×10^{-33}	7.191×10^{-33}	1.184×10^{-21}	1.230×10^{-19}
		Backbone$^{0.8}$	5.487×10^{-16}	5.785×10^{-17}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	4.762×10^{-37}	5.999×10^{-37}	2.758×10^{-10}	3.404×10^{-07}
		C_α	2.101×10^{-33}	7.191×10^{-33}	4.762×10^{-37}	5.999×10^{-37}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	5.362×10^{-42}	5.144×10^{-40}
		$C_\alpha^{0.8}$	1.184×10^{-21}	1.230×10^{-19}	2.758×10^{-10}	3.404×10^{-07}	5.362×10^{-42}	5.144×10^{-40}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Superfamily	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	3.042×10^{-19}	3.592×10^{-19}	1.319×10^{-36}	5.413×10^{-37}	4.535×10^{-24}	7.692×10^{-22}
		Backbone$^{0.8}$	3.042×10^{-19}	3.592×10^{-19}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	1.243×10^{-38}	7.796×10^{-39}	1.905×10^{-09}	1.917×10^{-04}
		C_α	1.319×10^{-36}	5.413×10^{-37}	1.243×10^{-38}	7.796×10^{-39}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.105×10^{-39}	2.537×10^{-39}
		$C_\alpha^{0.8}$	4.535×10^{-24}	7.692×10^{-22}	1.905×10^{-09}	1.917×10^{-04}	2.105×10^{-39}	2.537×10^{-39}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$
	Family	*Backbone*	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	3.218×10^{-14}	2.558×10^{-13}	3.063×10^{-35}	1.305×10^{-35}	4.100×10^{-21}	1.774×10^{-20}
		Backbone$^{0.8}$	3.218×10^{-14}	2.558×10^{-13}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	2.293×10^{-37}	6.397×10^{-37}	1.582×10^{-14}	4.793×10^{-11}
		C_α	3.063×10^{-35}	1.305×10^{-35}	2.293×10^{-37}	6.397×10^{-37}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$	9.278×10^{-39}	7.885×10^{-39}
		$C_\alpha^{0.8}$	4.100×10^{-21}	1.774×10^{-20}	1.582×10^{-14}	4.793×10^{-11}	9.278×10^{-39}	7.885×10^{-39}	$\mathbf{1.000 \times 10^{+00}}$	$\mathbf{1.000 \times 10^{+00}}$

Referências Bibliográficas

[Alexandre V. Fassio, 2017] Alexandre V. Fassio, Lucianna H. S. Santos, S. A. S. R. S. F. R. C. d. M.-M. (No prelo 2017). napoli: analysis of conserved protein-ligand interactions in large-scale. *Submitted in Bioinformatics*.

[Almonacid et al., 2010] Almonacid, D. E.; Yera, E. R.; Mitchell, J. B. & Babbitt, P. C. (2010). Quantitative comparison of catalytic mechanisms and overall reactions in convergently evolved enzymes: implications for classification of enzyme function. *PLoS Comput Biol*, 6(3):e1000700.

[Altman, 1992] Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175--185.

[Andreeva et al., 2004] Andreeva, A.; Howorth, D.; Brenner, S. E.; Hubbard, T. J.; Chothia, C. & Murzin, A. G. (2004). Scop database in 2004: refinements integrate structure and sequence family data. *Nucleic acids research*, 32(suppl 1):D226--D229.

[Anfinsen, 1972] Anfinsen, C. B. (1972). Studies on the principles that govern the folding of protein chains.

[Baker & Agard, 1994] Baker, D. & Agard, D. A. (1994). Kinetics versus thermodynamics in protein folding. *Biochemistry*, 33(24):7505--7509.

[Baldwin, 1986] Baldwin, R. L. (1986). Temperature dependence of the hydrophobic interaction in protein folding. *Proceedings of the National Academy of Sciences*, 83(21):8069--8072.

[Baldwin, 2007] Baldwin, R. L. (2007). Energetics of protein folding. *Journal of molecular biology*, 371(2):283--301.

[Baldwin, 2014] Baldwin, R. L. (2014). Dynamic hydration shell restores kauzmann's 1959 explanation of how the hydrophobic factor drives protein folding. *Proceedings of the National Academy of Sciences*, 111(36):13052--13056.

[Baldwin & Rose, 2016] Baldwin, R. L. & Rose, G. D. (2016). How the hydrophobic factor drives protein folding. *Proceedings of the National Academy of Sciences*, 113(44):12462--12466.

[Bellissent-Funel et al., 2016] Bellissent-Funel, M.-C.; Hassanali, A.; Havenith, M.; Henchman, R.; Pohl, P.; Sterpone, F.; van der Spoel, D.; Xu, Y. & Garcia, A. E. (2016). Water determines the structure and dynamics of proteins. *Chemical reviews*.

[Berman et al., 2000] Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.; Weissig, H.; Shindyalov, I. N. & Bourne, P. E. (2000). The protein data bank. *Nucleic acids research*, 28(1):235--242.

[Berry et al., 1995] Berry, M. W.; Dumais, S. T. & O'Brien, G. W. (1995). Using linear algebra for intelligent information retrieval. *SIAM review*, 37(4):573--595.

[Binkowski & Joachimiak, 2008] Binkowski, T. A. & Joachimiak, A. (2008). Protein functional surfaces: global shape matching and local spatial alignments of ligand binding sites. *BMC structural biology*, 8(1):45.

[Bondi, 1964] Bondi, A. (1964). van der waals volumes and radii. *The Journal of physical chemistry*, 68(3):441--451.

[Bork et al., 1993] Bork, P.; Sander, C. & Valencia, A. (1993). Convergent evolution of similar enzymatic function on different protein folds: the hexokinase, ribokinase, and galactokinase families of sugar kinases. *Protein Science*, 2(1):31--40.

[Bottini et al., 2013] Bottini, S.; Bernini, A.; De Chiara, M.; Garlaschelli, D.; Spiga, O.; Dioguardi, M.; Vannuccini, E.; Tramontano, A. & Niccolai, N. (2013). Prococoa: A quantitative approach for analyzing protein core composition. *Computational biology and chemistry*, 43:29--34.

[Brown et al., 2006] Brown, S. D.; Gerlt, J. A.; Seffernick, J. L. & Babbitt, P. C. (2006). A gold standard set of mechanistically diverse enzyme superfamilies. *Genome biology*, 7(1):1.

[Bu et al., 1999] Bu, W.-S.; Feng, Z.-P.; Zhang, Z. & Zhang, C.-T. (1999). Prediction of protein (domain) structural classes based on amino-acid index. *European Journal of Biochemistry*, 266(3):1043--1049.

[Cabello et al., 2008] Cabello, S.; Giannopoulos, P. & Knauer, C. (2008). On the parameterized complexity of d-dimensional point set pattern matching. *Information Processing Letters*, 105(2):73--77.

[Chandler, 2005] Chandler, D. (2005). Interfaces and the driving force of hydrophobic assembly. *Nature*, 437(7059):640--647.

[Chen et al., 2008a] Chen, C.; Chen, L.-X.; Zou, X.-Y. & Cai, P.-X. (2008a). Predicting protein structural class based on multi-features fusion. *Journal of theoretical biology*, 253(2):388--392.

[Chen et al., 2008b] Chen, K.; Kurgan, L. A. & Ruan, J. (2008b). Prediction of protein structural class using novel evolutionary collocation-based sequence representation. *Journal of computational chemistry*, 29(10):1596--1604.

[Cheung et al., 2002] Cheung, M. S.; García, A. E. & Onuchic, J. N. (2002). Protein folding mediated by solvation: water expulsion and formation of the hydrophobic core occur after the structural collapse. *Proceedings of the National Academy of Sciences*, 99(2):685--690.

[Choi & Kim, 2006] Choi, I.-G. & Kim, S.-H. (2006). Evolution of protein structural classes and protein sequence families. *Proceedings of the National Academy of Sciences*, 103(38):14056--14061.

[Chothia, 1992] Chothia, C. (1992). Proteins. one thousand families for the molecular biologist. *Nature*, 357(6379):543.

[Compiani & Capriotti, 2013] Compiani, M. & Capriotti, E. (2013). Computational and theoretical methods for protein folding. *Biochemistry*, 52(48):8601--8624.

[Consortium et al., 2014] Consortium, U. et al. (2014). Uniprot: a hub for protein information. *Nucleic acids research*, p. gku989.

[Daggett & Fersht, 2003] Daggett, V. & Fersht, A. R. (2003). Is there a unifying mechanism for protein folding? *Trends in biochemical sciences*, 28(1):18--25.

[Dai et al., 2013] Dai, Q.; Li, Y.; Liu, X.; Yao, Y.; Cao, Y. & He, P. (2013). Comparison study on statistical features of predicted secondary structures for protein structural class prediction: From content to position. *BMC bioinformatics*, 14(1):152.

[Dcrjsr & Redzikowski, 2011] Dcrjsr & Redzikowski, A. (2011). Protein backbone dihedral angles phi, psi, and omega. `https://upload.wikimedia.org/wikipedia/commons/thumb/9/97/Protein_backbone_PhiPsiOmega_drawing.svg/252px-Protein_backbone_PhiPsiOmega_drawing.svg.png`. Accessed: 2017-03-02.

[de Rezende & Lee, 1995] de Rezende, P. J. & Lee, D. (1995). Point set pattern matching ind-dimensions. *Algorithmica*, 13(4):387--404.

[Dehzangi et al., 2013] Dehzangi, A.; Paliwal, K.; Sharma, A.; Dehzangi, O. & Sattar, A. (2013). A combination of feature extraction methods with an ensemble of different classifiers for protein structural class prediction problem. *IEEE/ACM transactions on computational biology and bioinformatics*, 10(3):564--575.

[del Castillo-Negrete et al., 2007] del Castillo-Negrete, D.; Hirshman, S. P.; Spong, D. A. & D'Azevedo, E. F. (2007). Compression of magnetohydrodynamic simulation data using singular value decomposition. *Journal of Computational Physics*, 222(1):265--286.

[Deschavanne & Tuffery, 2008] Deschavanne, P. & Tuffery, P. (2008). Exploring an alignment free approach for protein classification and structural class prediction. *Biochimie*, 90(4):615--625.

[Dill, 1990] Dill, K. A. (1990). Dominant forces in protein folding. *Biochemistry*, 29(31):7133--7155.

[Dill, 1999] Dill, K. A. (1999). Polymer principles and protein folding. *Protein Science*, 8(06):1166--1180.

[Dill & MacCallum, 2012] Dill, K. A. & MacCallum, J. L. (2012). The protein-folding problem, 50 years on. *Science*, 338(6110):1042--1046.

[Dill et al., 2008] Dill, K. A.; Ozkan, S. B.; Shell, M. S. & Weikl, T. R. (2008). The protein folding problem. *Annual review of biophysics*, 37:289.

[Dill et al., 2007] Dill, K. A.; Ozkan, S. B.; Weikl, T. R.; Chodera, J. D. & Voelz, V. A. (2007). The protein folding problem: when will it be solved? *Current opinion in structural biology*, 17(3):342--346.

[Ding & Dokholyan, 2006] Ding, F. & Dokholyan, N. V. (2006). Emergence of protein fold families through rational design. *PLoS Comput Biol*, 2(7):e85.

[Ding et al., 2012] Ding, S.; Zhang, S.; Li, Y. & Wang, T. (2012). A novel protein structural classes prediction method based on predicted secondary structure. *Biochimie*, 94(5):1166--1171.

[Ding et al., 2007] Ding, Y.-S.; Zhang, T.-L. & Chou, K.-C. (2007). Prediction of protein structure classes with pseudo amino acid composition and fuzzy support vector machine network. *Protein and peptide letters*, 14(8):811--815.

[Eggert et al., 1997] Eggert, D. W.; Lorusso, A. & Fisher, R. B. (1997). Estimating 3-d rigid body transformations: a comparison of four major algorithms. *Machine Vision and Applications*, 9(5-6):272--290.

[Eidhammer et al., 2000] Eidhammer, I.; Jonassen, I. & Taylor, W. R. (2000). Structure comparison and structure patterns. *Journal of Computational Biology*, 7(5):685--716.

[Eldén, 2007] Eldén, L. (2007). *Matrix methods in data mining and pattern recognition*, volume 4. SIAM.

[Galperin et al., 1998] Galperin, M. Y.; Walker, D. R. & Koonin, E. V. (1998). Analogous enzymes: independent inventions in enzyme evolution. *Genome Research*, 8(8):779--790.

[Gherardini et al., 2007] Gherardini, P. F.; Wass, M. N.; Helmer-Citterich, M. & Sternberg, M. J. (2007). Convergent evolution of enzyme active sites is not a rare phenomenon. *Journal of molecular biology*, 372(3):817--845.

[Godzik, 1996] Godzik, A. (1996). The structural alignment between two proteins: Is there a unique answer? *Protein Science*, 5(7):1325--1338.

[Gonçalves-Almeida et al., 2012] Gonçalves-Almeida, V.; Pires, D. E.; de Melo-Minardi, R. C.; da Silveira, C. H.; Meira, W. & Santoro, M. M. (2012). Hydropace: understanding and predicting cross-inhibition in serine proteases through hydrophobic patch centroids. *Bioinformatics*, 28(3):342--349.

[Hall et al., 2009] Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P. & Witten, I. H. (2009). The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10--18.

[Hamelryck & Manderick, 2003] Hamelryck, T. & Manderick, B. (2003). Pdb file parser and structure class implemented in python. *Bioinformatics*, 19(17):2308--2310.

[Hamming, 1950] Hamming, R. W. (1950). Error detecting and error correcting codes. *Bell Labs Technical Journal*, 29(2):147--160.

[Heinz & Suter, 2004] Heinz, H. & Suter, U. W. (2004). Atomic charges for classical simulations of polar systems. *The Journal of Physical Chemistry B*, 108(47):18341--18352.

[Hoehn & Niven, 1985] Hoehn, L. & Niven, I. (1985). Averages on the move. *Mathematics Magazine*, pp. 151--156.

[Illergård et al., 2009] Illergård, K.; Ardell, D. H. & Elofsson, A. (2009). Structure is three to ten times more conserved than sequence—a study of structural response in protein cores. *Proteins: Structure, Function, and Bioinformatics*, 77(3):499--508.

[Istrail, 2003] Istrail, C. G. S. (2003). Mathematical methods for protein structure analysis and design.

[Jain & Hirst, 2010] Jain, P. & Hirst, J. D. (2010). Automatic structure classification of small proteins using random forest. *BMC bioinformatics*, 11(1):1.

[Jian & Vemuri, 2011] Jian, B. & Vemuri, B. C. (2011). Robust point set registration using gaussian mixture models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(8):1633--1645.

[Kauzmann, 1959] Kauzmann, W. (1959). Some factors in the interpretation of protein denaturation. *Advances in protein chemistry*, 14:1--63.

[Kedarisetti et al., 2006] Kedarisetti, K. D.; Kurgan, L. & Dick, S. (2006). Classifier ensembles for protein structural class prediction with varying homology. *Biochemical and biophysical research communications*, 348(3):981--988.

[Kennedy & Norman, 2005] Kennedy, D. & Norman, C. (2005). What don't we know? *Science*, 309(5731):78--102.

[Kurgan et al., 2008] Kurgan, L.; Cios, K. & Chen, K. (2008). Scpred: accurate prediction of protein structural class for sequences of twilight-zone similarity with predicting sequences. *BMC bioinformatics*, 9(1):226.

[Larson et al., 2002] Larson, S. M.; Ruczinski, I.; Davidson, A. R.; Baker, D. & Plaxco, K. W. (2002). Residues participating in the protein folding nucleus do not exhibit preferential evolutionary conservation. *Journal of molecular biology*, 316(2):225--233.

[Laskowski et al., 1993] Laskowski, R. A.; Moss, D. S. & Thornton, J. M. (1993). Main-chain bond lengths and bond angles in protein structures. *Journal of molecular biology*, 231(4):1049--1067.

[Lee et al., 2007] Lee, D.; Redfern, O. & Orengo, C. (2007). Predicting protein function from sequence and structure. *Nature Reviews Molecular Cell Biology*, 8(12):995--1005.

[Levy et al., 2006] Levy, E. D.; Pereira-Leal, J. B.; Chothia, C. & Teichmann, S. A. (2006). 3d complex: a structural classification of protein complexes. *PLoS Comput Biol*, 2(11):e155.

[Liu & Jia, 2010] Liu, T. & Jia, C. (2010). A high-accuracy protein structural class prediction algorithm using predicted secondary structural information. *Journal of theoretical biology*, 267(3):272--275.

[Liu et al., 2010] Liu, T.; Zheng, X. & Wang, J. (2010). Prediction of protein structural class using a complexity-based distance measure. *Amino acids*, 38(3):721--728.

[McLachlan, 1982] McLachlan, A. D. (1982). Rapid comparison of protein structures. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, 38(6):871--873.

[Nick Pace et al., 2014] Nick Pace, C.; Scholtz, J. M. & Grimsley, G. R. (2014). Forces stabilizing proteins. *FEBS letters*, 588(14):2177--2184.

[Nisius et al., 2012] Nisius, B.; Sha, F. & Gohlke, H. (2012). Structure-based computational analysis of protein binding sites for function and druggability prediction. *Journal of biotechnology*, 159(3):123--134.

[Oldfield, 2007] Oldfield, T. (2007). Caalign: a program for pairwise and multiple protein-structure alignment. *Acta Crystallographica Section D: Biological Crystallography*, 63(4):514--525.

[Ortiz et al., 2002] Ortiz, A. R.; 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.

[Pace et al., 2014] Pace, C. N.; Fu, H.; Fryar, K.; Landua, J.; Trevino, S. R.; Schell, D.; Thurlkill, R. L.; Imura, S.; Scholtz, J. M.; Gajiwala, K. et al. (2014). Contribution of hydrogen bonds to protein stability. *Protein Science*, 23(5):652--661.

[Pace et al., 1996] Pace, C. N.; Shirley, B. A.; McNutt, M. & Gajiwala, K. (1996). Forces contributing to the conformational stability of proteins. *The FASEB journal*, 10(1):75--83.

[Parasuram et al., 2010] Parasuram, R.; Lee, J. S.; Yin, P.; Somarowthu, S. & Ondrechen, M. J. (2010). Functional classification of protein 3d structures from predicted local interaction sites. *Journal of bioinformatics and computational biology*, 8(supp01):1--15.

[Pauling et al., 1951] Pauling, L.; Corey, R. B. & Branson, H. R. (1951). The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. *Proceedings of the National Academy of Sciences*, 37(4):205--211.

[Perunov & England, 2014] Perunov, N. & England, J. L. (2014). Quantitative theory of hydrophobic effect as a driving force of protein structure. *Protein Science*, 23(4):387--399.

[Piana et al., 2014] Piana, S.; Klepeis, J. L. & Shaw, D. E. (2014). Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations. *Current opinion in structural biology*, 24:98--105.

[Piana et al., 2013] Piana, S.; Lindorff-Larsen, K. & Shaw, D. E. (2013). Atomic-level description of ubiquitin folding. *Proceedings of the National Academy of Sciences*, 110(15):5915--5920.

[Pires et al., 2011] Pires, D. E.; de Melo-Minardi, R. C.; dos Santos, M. A.; da Silveira, C. H.; Santoro, M. M. & Meira, W. (2011). Cutoff scanning matrix (csm): structural classification and function prediction by protein inter-residue distance patterns. *BMC genomics*, 12(4):1.

[Poleksic, 2009] Poleksic, A. (2009). Algorithms for optimal protein structure alignment. *Bioinformatics*, 25(21):2751--2756.

[Privalov & Gill, 1988] Privalov, P. L. & Gill, S. J. (1988). Stability of protein structure and hydrophobic interaction. *Advances in protein chemistry*, 39:191--234.

[Rawlings et al., 2008] Rawlings, N. D.; Morton, F. R.; Kok, C. Y.; Kong, J. & Barrett, A. J. (2008). Merops: the peptidase database. *Nucleic acids research*, 36(suppl 1):D320--D325.

[Rentzsch & Orengo, 2009] Rentzsch, R. & Orengo, C. A. (2009). Protein function prediction–the power of multiplicity. *Trends in biotechnology*, 27(4):210--219.

[Richardson, 1981] Richardson, J. S. (1981). The anatomy and taxonomy of protein structure. *Advances in protein chemistry*, 34:167--339.

[Røgen & Fain, 2003] Røgen, P. & Fain, B. (2003). Automatic classification of protein structure by using gauss integrals. *Proceedings of the National Academy of Sciences*, 100(1):119--124.

[Rose et al., 1985] Rose, G. D.; Geselowitz, A. R.; Lesser, G. J.; Lee, R. H. & Zehfus, M. H. (1985). Hydrophobicity of amino acid residues in globular proteins. *Science*, 229(4716):834--838.

[Rost, 2002] Rost, B. (2002). Enzyme function less conserved than anticipated. *Journal of molecular biology*, 318(2):595--608.

[Russell, 1998] Russell, R. B. (1998). Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. *Journal of molecular biology*, 279(5):1211--1227.

[Sabata & Aggarwal, 1991] Sabata, B. & Aggarwal, J. (1991). Estimation of motion from a pair of range images: A review. *CVGIP: Image Understanding*, 54(3):309--324.

[Sael et al., 2012] Sael, L.; Chitale, M. & Kihara, D. (2012). Structure-and sequence-based function prediction for non-homologous proteins. *Journal of structural and functional genomics*, 13(2):111--123.

[Sahu & Panda, 2010] Sahu, S. S. & Panda, G. (2010). A novel feature representation method based on chou's pseudo amino acid composition for protein structural class prediction. *Computational biology and chemistry*, 34(5):320--327.

[Schafer et al., 2014] Schafer, N. P.; Kim, B. L.; Zheng, W. & Wolynes, P. G. (2014). Learning to fold proteins using energy landscape theory. *Israel journal of chemistry*, 54(8-9):1311--1337.

[Schönemann, 1966] Schönemann, P. H. (1966). A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1--10.

[Schrödinger, LLC, 2015] Schrödinger, LLC (2015). The PyMOL molecular graphics system, version 1.8.

[Sharma et al., 2013] Sharma, S.; Kumar, S. K.; Buldyrev, S. V.; Debenedetti, P. G.; Rossky, P. J. & Stanley, H. E. (2013). A coarse-grained protein model in a water-like solvent. *Scientific reports*, 3:1841.

[Sherwood, 2005] Sherwood, L. (2005). *Fundamentals of Physiology: A Human Perspective*, p. A7. Thomson Brooks/Cole, 3 edição.

[Sleator & Walsh, 2010] Sleator, R. D. & Walsh, P. (2010). An overview of in silico protein function prediction. *Archives of microbiology*, 192(3):151--155.

[Sobolev et al., 1999] Sobolev, V.; Sorokine, A.; Prilusky, J.; Abola, E. E. & Edelman, M. (1999). Automated analysis of interatomic contacts in proteins. *Bioinformatics*, 15(4):327--332.

[Soundararajan et al., 2010] Soundararajan, V.; Raman, R.; Raguram, S.; Sasisekharan, V. & Sasisekharan, R. (2010). Atomic interaction networks in the core of protein domains and their native folds. *PLoS One*, 5(2):e9391.

[Tseng & Liang, 2004] Tseng, Y. Y. & Liang, J. (2004). Are residues in a protein folding nucleus evolutionarily conserved? *Journal of molecular biology*, 335(4):869--880.

[Volkamer et al., 2013] Volkamer, A.; Kuhn, D.; Rippmann, F. & Rarey, M. (2013). Predicting enzymatic function from global binding site descriptors. *Proteins: Structure, Function, and Bioinformatics*, 81(3):479--489.

[Weisstein, 2017] Weisstein, E. W. (2017). Sphere-Sphere Intersection. from mathworld–a wolfram web resource. `http://mathworld.wolfram.com/Sphere-SphereIntersection.html`. Accessed: 2016-05-05.

[Xin & Radivojac, 2011] Xin, F. & Radivojac, P. (2011). Computational methods for identification of functional residues in protein structures. *Current Protein and Peptide Science*, 12(6):456--469.

[Ye & Godzik, 2003] Ye, Y. & Godzik, A. (2003). Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics*, 19(suppl 2):ii246--ii255.

[Zhang et al., 2003] Zhang, L.; Xu, W. & Chang, C. (2003). Genetic algorithm for affine point pattern matching. *Pattern Recognition Letters*, 24(1):9--19.

[Zhang & Skolnick, 2005] Zhang, Y. & Skolnick, J. (2005). Tm-align: a protein structure alignment algorithm based on the tm-score. *Nucleic acids research*, 33(7):2302--2309.

[Zheng et al., 2010] Zheng, X.; Li, C. & Wang, J. (2010). An information-theoretic approach to the prediction of protein structural class. *Journal of computational chemistry*, 31(6):1201--1206.

[Zhou, 1998] Zhou, G.-P. (1998). An intriguing controversy over protein structural class prediction. *Journal of protein chemistry*, 17(8):729--738.

[Zhou et al., 2004] Zhou, R.; Huang, X.; Margulis, C. J. & Berne, B. J. (2004). Hydrophobic collapse in multidomain protein folding. *Science*, 305(5690):1605--1609.