UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
CENTRO DE BIOTECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOLOGIA CELULAR E MOLECULAR

LEONARDO ALVES SANTOS
LEONARDOAS95@GMAIL.COM

Inclusão de Contatos Evolutivamente Conservados em Métodos para Predição da Estrutura 3D de Polipeptídios

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre em Biologia Celular e Molecular

Orientador: Prof. Dr. Márcio Dorn
Co-orientador: Prof. Dr. Rodrigo Ligabue Braun

Porto Alegre
2019

CIP — CATALOGAÇÃO NA PUBLICAÇÃO

Santos, Leonardo Alves

Inclusão de Contatos Evolutivamente Conservados em Métodos para Predição da Estrutura 3D de Polipeptídios / Leonardo Alves Santos. – Porto Alegre: PPGBCM da UFRGS, 2019.

135 f.: il.

Dissertação (mestrado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Biologia Celular e Molecular, Porto Alegre, BR–RS, 2019. Orientador: Márcio Dorn; Co-orientador: Rodrigo Ligabue Braun.

1. Acoplamento de Resíduos. 2. Angulos Diedrais. 3. Bioinformática Estrutural. 4. Predição de Estrutura de Proteínas. I. Dorn, Márcio. II. Braun, Rodrigo Ligabue. III. Título.

AGRADECIMENTOS

Agradeço aos meus orientadores, Márcio Dorn e Rodrigo Ligabue Braun, por aceitarem e embarcarem nesta jornada de formação de um novo cientista. Durante nossa trajetória, IC e mestrado, vocês dois sempre me auxiliaram de maneira complementar, me indicando meios computacionais e achando sentido biológico naquilo que encontramos, o que quase nunca foi fácil. Muito obrigado por confiarem em mim durante esse tempo e, sem dúvidas, minhas conquistas são o reflexo da orientação de ambos.

Agradeço aos membros da comissão de acompanhamento, Hugo Verli e Guido Lenz, por sempre me fazerem enxergar além do óbvio, assim como por sempre estarem com as portas abertas para qualquer dúvida/problema. Ao Prof. Gang Fang, por ter me recebido em Shanghai por 3 meses e pela orientação em um área em que eu nunca havia trabalhado. Estendo aos demais professores que tive o prazer de encontrar ao frequentar as disciplinas cursadas ao longo do mestrado.

Aos colegas de laboratório (SBCB) que me auxiliaram, desde o início enquanto ainda estava na graduação, inclusive quando eu ainda não sabia como abrir um arquivo em Python. Aos meus amigos próximos, agradeço por segurarem a barra em momentos de reclamações, aguentar minhas oscilações de humor, me dar total apoio quando necessitei e também não me odiarem por furar vários roles durante a reta final do mestrado.

Alguns agradecimentos nominais, pois conheço meus amigos e vai dar ruim se não constar alguns nomes aqui: ao Bruno Grisci, por ter se tornado mais que um colega de trabalho, um Freund; ao Pablo, por ter sido um exemplo de pesquisador, revisor de artigo e amigo, apesar de não me convidar para o casamento dele; ao Lucas, por todas as discussões filosóficas, em mesa de bar ou no sofá da sala, jantinhas e suporte emocional; ao grupo BEDEUS, por todo apoio, saídas, ano novo, desespero, alegria, noites e tudo mais que passamos juntos, amo vocês; ao Pedro Narloch, por todas as ajudas, compartilhamentos de código, além de sempre me socorrer quando meus códigos não funcionavam de jeito nenhum.

Ao meu sensoito, Juliano, por todo apoio, puxão de orelha, me aturar por 3 meses dividindo um quarto, estar do meu lado (inclusive no outro lado do mundo), querer ir embora, me apresentar seriados maravilhosos que eu nunca reclamei e por ser essa pessoa maravilhosa que eu tive a sorte de encontrar. À minha parceira Elisa, por ter feito minha vida no vale menos miserável, me induzir a comer deliciosos docinhos após o RU, me defender quando necessário, me lembrar que as coisas valem a pena, por me fazer gastar

dinheiro com jogos de computador e também sempre estar do meu lado e fazer os passeios impulsivos mais desastrosos do mundo.

Rodrigo, eu nem sei por onde começar, muito menos terminar. Então eu te coloco nesse parágrafo pra dizer que sem tua ajuda eu definitivamente não estaria defendendo este mestrado, obrigado por ir além das obrigações de orientador e se tornar um amigo! Desculpa por te fazer lidar com meus problemas e ansiedades, mas ao mesmo tempo não me arrependo porque deu certo.

Aos meus pais, Caroena e Fernando, muito obrigado! Vocês são a base de tudo que eu sou, por isso só tenho a agradecer. Obrigado por serem sempre compreensíveis, sempre acreditarem em mim sem pensar duas vezes, por me amarem incondicionalmente, enfrentarem as barras comigo, se preocuparem, me suprir em tudo que eu necessitei, enfim por serem meu tudo. Eu amo muito vocês. À minha família estendida, os Kampmmann, obrigado por me receberem na família, por todos os finais de semana que eu praticamente morei na casa de vocês. Obrigado por todo apoio, alegria e momentos difíceis compartilhados. Eu também amo muito vocês.

Por fim agradeço ao Pedro Kampmmann. Durante os últimos três anos e alguns meses tu tem sido peça fundamental do meu ser. Obrigado por ser meu companheiro, melhor amigo e porto seguro. Tu sempre me apoiou em todas as decisões que tomei, sempre me deu suporte, aguentou comigo as barras mais pesadas que eu já tive, enfim, esteve do meu lado em todos os momentos. Queria te agradecer especificamente por ter me ensinado o significado de companheirismo, mesmo nesse tempo em que estamos fisicamente longe um do outro. Pedrinho, eu te amo. Obrigado por ser quem tu é e me deixar fazer parte da tua vida.

RESUMO

A predição de estrutura 3D de proteínas ainda permanece como um dos maiores desafios a ser superado pela Bioinformática Estrutural. O número de conformações 3D que uma dada sequência de aminoácidos pode assumir é praticamente infinita, classificando o problema como NP-Completo. Devido a sua alta complexidade, métodos exatos não são capazes de encontrar a solução ótima para tal problema em tempo de execução plausível. Portanto, meta-heurísticas são ótimas candidatas para resolução do problema de predição de estruturas, ainda que não sejam capazes de garantir que a melhor solução seja sempre alcançada. A utilização de informações biológicas durante o processo de otimização já foi constatada por autores na literatura como fatores que auxiliam na obtenção de soluções melhores. Por exemplo, o uso de preferências conformacionais de ângulos de torção sob estruturas secundárias específicas para geração de indivíduos que compõem a população inicial de meta-heurísticas populacionais. Assim como, a utilização de informação a cerca de acoplamento de pares de aminoácidos através de análises estatísticas inversas, as quais permitem inferir aproximação 3D de aminoácidos, os quais não precisam estar linearmente próximos na sequência de aminoácidos. Os resultados das últimas edições do *The Critical Assessment of Protein Structure Prediction* mostram que métodos que utilizam esta inferência de contato entre aminoácidos para predição de estrutura de proteínas tiveram um aumento de precisão de predição. O presente trabalho tem por objetivo propor uma metodologia baseada em conhecimento para geração de modelos estruturais com características estruturais próximas às estruturas experimentalmente determinadas, assim como utilizar estes modelos como integrantes da população inicial de algoritmos de otimização baseados em população. A geração de modelos utilizará dois limitantes, primeiro os ângulos de torção para cada aminoácido sob uma estrutura secundária específica, calculadas a partir do *Protein Data Bank*, e segundo, informações de contatos 3D entre aminoácidos preditos a partir de análises de alinhamentos múltiplos de sequência. Os resultados obtidos pela avaliação estrutural dos modelos gerados mostram que o método é capaz de gerar estruturas próximas às estruturas determinadas experimentalmente, enquanto que para o processo de otimização fica claro o aumento de precisão de predição ao utilizar candidatos iniciais de alta qualidade.

Palavras-chave: Acoplamento de Resíduos. Angulos Diedrais. Bioinformática Estrutural. Predição de Estrutura de Proteínas.

The Inclusion of Evolutionary Conserved Contacts in Polypeptidic Structural Prediction Methods

ABSTRACT

The prediction of the 3D structure of proteins is one of the milestones yet to be overcome in Structural Bioinformatics. The number of 3D viable structures a single amino acid sequence can assume is humongous, classifying the problem as an NP-Complete. Due to its high complexity, exact methods could not find the optimal solution in feasible process time to the PSP problem. Hence, metaheuristics are an interesting way to approach the problem and find good solutions to it, even though they do not guarantee the finding of the best solution in the search space. The usage of biological information throughout the optimization process has already been demonstrated in the literature as a valuable addition. For example, the usage of the conformational preference of amino acid torsion angles under specific secondary structures during the assembly of the initial candidate solutions for population-based metaheuristics. In addition to this, the usage of co-evoltuion between aminoacid pairs, inferred from inverse statistical analyses, which allow to infer the 3D proximity in protein structure of residues that are not necessarily close to each other in sequence length. The results from the last editions of The Critical Assessment of Protein Structure Prediction have shown an increase of accuracy for those methods using the contact prediction information in their methodology. The present work aims to propose a knowledge-based approach to generate structural models with structural features close to the experimentally determined structure, as well as the usage of these assembled models to compose the initial solutions of optimization algorithms. The generation of structural models takes in consideration two biological constraints, first the torsion angles conformations for each amino acid residue under a specific secondary structure, retrieved from the Protein Data Bank, and second, the 3D predicted contact information retrieved from multiple sequence alignments outputs. The results obtained from the structural evaluation shows that the proposed method is able to generate individuals much like the ones experimentally determined, as for the result of the optimization process shows it is clear that the generation of high-quality initial structural models play an important role on the results.

Keywords: Amino Acid Residues Coupling, Dihedral Angles, Protein Structure Prediction, Structural Bioinformatics.

LISTA DE ABREVIATURAS E SIGLAS

3D	Tridimensional
AE	Algoritmo Evolutivo
APL	*Angle Probability List* (Lista de Probabilidade de Ângulos)
CASP	*Critical Assessment of Protein Structure Prediction* (Avaliação Crítica da Predição da Estrutura Proteica)
CATH	*Class, Architecture, Topology and Homologous* (Classe, Arquitetura, Topologia e Homólogos)
DCA	*Direct Coupling Analysis* (Análise de Acoplamento Direto)
DE	*Differential Evolution* (Evolução Diferencial)
DEMO	*Differential Evolution Multi-Objective* (Evolução Diferencial Multi-Objetivo)
DM	Dinâmica Molecular
DNA	*Deoxyribonucleic Acid* (Ácido Desoxirribonucleico)
DSSP	*Define Secondary Structure of Proteins* (Definir Estrutura Secundária de Proteínas)
ES	Estrutura Secundária
FD-MD	*Fragment-Guided Molecular Dynamics* (Dinâmica Molecular Guiada por Fragmentos)
FM	*Free Modeling* (Modelagem Livre)
IM	Informação Mútua
MSA	*Multiple Sequence Alignment* (Alinhamneto Múltiplo de Sequências)
ME	Microscopia Eletrônica
MO	Multiobjetivo
NCBI	*National Center for Biotechnology Information* (Centro Nacional de Informação Biotecnológica)
PDB	*Protein Data Bank* (Banco de Dados de Proteínas)
PSP	*Protein Structure Prediction* (Predição de Estrutura de Proteínas)

RefSeq *Reference Sequence* (Sequência Referência)

REMC *Replica-Exchange Monte Carlo*

RG Raio de Giro

RMN Resonância Nuclear Magnética

RMSD *Root-Mean-Square Deviation* (Desvio da Raiz Quadrada Média)

RNA *Ribonucleic Acid* (Ácido Ribonucleico)

SASA *Solvent-Accessible Surface Area* (Área de Superfície Acessível ao Solvente)

SCOP *Structural Classification of Proteins* (Classificação Estrutural de Proteínas)

TBM *Template-Based Modeling* (Modelagem Baseada em Modelo)

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O aumento da quantidade de dados biológicos gerados ao longo das últimas décadas, causados principalmente pela redução de custos de sequenciamento e aumento de tamanho de bancos de dados, torna o uso de abordagens computacionais fundamentais para a interpretação e análise eficazes desta grande quantidade de informação (MUIR et al., 2016). Este cenário impulsiona o crescimento da Bioinformática, uma ciência composta por profissionais oriundos de diversas áreas do conhecimento, tais como biologia, ciência da computação, física, matemática, química, entre outras, visando a elucidação de problemas biológicos através do emprego de técnicas computacionais (VERLI, 2014; BAXEVANIS; OUELLETTE, 2004). A vasta abrangência da Bioinformática permite com que seja possível separá-la em dois grupos, baseando-se em seu principal objeto de estudo: sequências biológicas ou estrutura de macromoléculas.

A primeira vertente refere-se ao estudo de sequências de importância biológica, tais como Ácido Desoxirribonucleico (DNA, sigla em inglês), Ácido Ribonucleico (RNA, sigla em inglês) e/ou sequências de aminoácidos. A técnica mais comumente empregada para a análise de sequências biológicas consiste no alinhamento entre duas ou mais destas (CHENNA et al., 2003; CAMACHO et al., 2009). A comparação entre sequências pode ser realizada de maneira a tentar discriminar o maior número de caracteres (nucleotídeos ou aminoácidos) semelhantes entre ambas as sequências, caracterizando o alinhamento global (NEEDLEMAN; WUNSCH, 1970). Para fins análogos, ao invés de comparar sequências de maneira global, estas são fragmentadas e são comparadas as semelhanças localmente de acordo com cada fragmento comparado a um segundo fragmento (ALTSCHUL et al., 1990). A partir dos resultados provenientes de alinhamentos, uma série de questões podem ser redirecionadas, ou até mesmo resolvidas, baseando-se nas semelhanças e diferenças encontradas a partir da comparação realizada. Inferências a respeito da relação evolutiva entre sequências, tais como conservação, aquisição ou perda de características, sejam elas genotípicas ou fenotípicas, podem ser realizadas a partir de análises proveniente do alinhamento de sequências biológicas (LIGABUE-BRAUN et al., 2013).

A segunda grande linha a qual a Bioinformática engloba, reporta-se ao estudo do que diz respeito a estruturas tridimensionais (3D) de macromoléculas biológicas (CHOU, 2004). Diversas análises podem ser realizadas a partir do estudo de estruturas 3D destas moléculas, por exemplo, a elucidação da interação molécula-ligante através de estudos de

atracamento molecular (KITCHEN et al., 2004), ou avaliação do padrão comportamental de moléculas de acordo com o micro-ambiente no qual ela está inserida através de dinâmicas moleculares (DM) (KARPLUS; MCCAMMON, 2002). A unidade básica para análises estruturais é a estrutura 3D dos objetos de estudo. Contudo, a obtenção destas estruturas não trata-se de uma tarefa trivial. Atualmente, as abordagens mais comumente empregadas para obtenção de estruturas 3D de proteínas são aquelas dependentes de procedimentos experimentais de bancada, dentre as quais destacam-se: a Cristalografia obtidas por Raio-X (KALLEN et al., 1998; JOHANNSSON; NEUMANN; FICNER, 2018), a Ressonância Nuclear Magnética (RMN) (HOU et al., 2018; FATTORUSSO et al., 1999) e a Microscopia Eletrônica (ME) (YANG et al., 2016; GE et al., 2015).

Metodologias experimentais, apesar de serem as abordagens mais implementadas para obtenção de estruturas 3D de proteínas, apresentam certos limitantes durante sua execução. Dentre os fatores que dificultam os métodos apresentados, destacam-se o alto custo, grande quantidade de tempo necessário para execução de protocolo, dificuldade de expressão e purificação da proteína de interesse, assim como fatores externos que podem inviabilizar a obtenção da estrutura ao término da execução (VERLI, 2014; EDWARDS et al., 2000). Ao comparar o número de estruturas 3Ds depositadas no principal banco de dados de estrutura de biomoléculas, o *Protein Data Bank*[1] (PDB) (BERMAN et al., 2000), e o número de sequências genômicas depositadas do banco de dados *NCBI Reference Sequence*[2] (RefSeq) (PRUITT; TATUSOVA; MAGLOTT, 2006), nota-se a discrepância na quantidade de dados, na qual o número de estruturas 3D corresponde a apenas 1% do número de sequências. Sabe-se que proteínas estão presentes em todas unidades celulares e que estão presentes em praticamente todos os processos celulares, portanto à elucidação da estrutura 3D de proteínas é de extrema importância, uma vez que a função de uma dada proteína está intimamente associada à sua estrutura 3D (NELSON; LEHNINGER; COX, 2008).

Justificam-se então os esforços voltados ao desenvolvimento de abordagens computacionais a fim de que possam ser utilizados para a elucidação de estruturas 3D de proteínas. Dentro da Bioinformática Estrutural, o problema de Predição de Estrutura de Proteínas (PSP, sigla em inglês) caracteriza-se por um dos desafios mais importantes, contudo ainda permanece sem uma resolução definitiva (DILL; MACCALLUM, 2012). Segundo a teoria de complexidade computacional (COOK, 1983), o problema de PSP é classificado como um problema NP-difícil (UNGER; MOULT, 1993), por suas caracterís-

[1] <https://www.rcsb.org/>
[2] <https://www.ncbi.nlm.nih.gov/refseq/>

ticas de alta dimensionalidade e espaço de busca multimodal. A dificuldade de resolução do problema aumenta à medida em que resíduos de aminoácidos são adicionados a cadeia polipeptídica, uma vez que o número de possíveis conformações 3D que uma única sequência de aminoácidos pode assumir explode conforme o tamanho da sequência aumenta (BAXEVANIS; OUELLETTE, 2004).

Uma vasta variedade de metodologias e abordagens computacionais tem sido propostas ao longo das últimas décadas a fim de tentar solucionar o problema de PSP. Estes métodos podem ser classificados em quatro grupos distintos de acordo com o tipo de abordagem utilizado (FLOUDAS et al., 2006; DORN et al., 2014): métodos de primeiros princípios (*i*) que não utilizam informações proveniente de estruturas de proteínas depositadas previamente em banco de dados (OSGUTHORPE, 2000); (*ii*) que utilizam informações depositadas em bacos de dados (ROHL et al., 2004); (*iii*) métodos baseados em modelagem comparativa (*comparative modeling*) (MARTÍ-RENOM et al., 2000) e (*iv*) alinhavamento de estruturas (*fold recognition*) (BOWIE; LUTHY; EISENBERG, 1991).

As abordagens classificadas como pertencentes ao grupo (*i*), métodos *ab initio*, baseiam-se nos princípios termodinâmicos de que as conformações 3D correspondentes ao estado nativo de uma dada proteína também correspondem aos estados de energia livre mínima (ANFINSEN, 1973). Este grupo de metodologias visa a predição das estruturas 3D de proteínas apenas utilizando informações provenientes da sequência de aminoácidos da proteína de interesse. A partir deste dado, estes métodos tentam mimetizar computacionalmente os processos físico-químicos que guiam o processo de enovelamento de proteínas em uma célula (OSGUTHORPE, 2000). Esta categoria é considerada uma abordagem ideal pela capacidade de predizer conformações nunca antes observadas, uma vez que não depende de nenhuma informação prévia além da sequência de aminoácidos. Contudo, possui dois graves limitantes: o primeiro deles diz respeito ao conhecimento parcial de todas as leis físico-químicos envolvidas no processo de enovelamento; o segundo limitante refere-se a alta dimensionalidade do espaço de busca de mínimos energéticos, aumentando a complexidade do problema, impossibilitando sua resolução de maneira computacional. Este, é agravado pela estrutura nativa não-estática apresentada por proteínas em suas conformações nativas, portanto uma única proteína pode apresentar um grande número de estruturas consideradas nativas, e consequentemente, um grande número de mínimos locais energéticos (ANFINSEN, 1973).

Em divergência aos métodos pertencentes ao grupo (*i*), os restantes enquadram-

se aos grupos de abordagens que se beneficiam de informações estruturais estabelecidas previamente. Os grupos (*ii*), (*iii*) e (*iv*) utilizam informações a fim de contornar a alta complexidade do problema de PSP. Contudo, estes métodos dependem de informações estruturais previamente obtidas através de procedimentos experimentais, sejam elas estruturas completas ou fragmentos. Portanto, destaca-se que a eficácia destas abordagens está diretamente relacionada a quantidade de dados disponíveis em bancos de dados (DORN et al., 2014). Abordagens pertencentes ao grupo (*iii*) de modelagem comparativa baseiam-se na busca por proteínas, com alta similaridade de sequência, cuja estrutura 3D esteja depositada em bancos de dados para servir de modelo para a determinação da estrutura da proteína-alvo (BLUNDELL et al., 1987). Métodos englobados por *fold recognition* (*iv*) apropriam-se do fato de que proteínas são mais evolutivamente conservadas em níveis estruturais quando comparados a níveis de composição de sequência de aminoácidos (FLOUDAS et al., 2006). Juntamente partem do princípio de que o número de enovelamento proteico é finito (WANG, 1998). Portanto, esta classe visa tentar inserir aminoácidos de maneira sequencial (de acordo com a proteína-alvo) em um enovelamento proteico previamente conhecido.

Ao analisar as estruturas 3D de proteínas depositadas em bancos de dados estruturais é possível perceber que, apesar de uma grande variedade de conformações globais, estas estruturas são compostas por fragmentos que se repetem em diferentes proteínas. Aproveitando-se deste fato, os métodos pertencentes ao grupo *ii* visam a comparação entre fragmentos da sequência de aminoácidos contra fragmentos estruturais depositados em bancos de dados (DORN et al., 2014). Uma vez encontrados fragmentos de sequências e estruturais homólogos, algoritmos de otimização são utilizados a fim de unificar os fragmentos de forma que a estrutura final seja correspondente ao mínimo energético da molécula (SIMONS et al., 1997). Ao que diz respeito ao processo de otimização, esta abordagem assemelha-se ao grupo *i*, contudo, devido à utilização de fragmentos retirados de bancos de dados, não podem ser classificadas como *ab initio* (FLOUDAS et al., 2006). Estruturas complexas podem ser geradas a partir da junção dos fragmentos encontrados, porém este processo deve ser guiado por critérios de avaliação. Sabe-se que interações não covalentes são responsáveis por conferir estabilidade a estruturas 3D de proteínas (VOET; VOET, 2010). Portanto, critérios de avaliação estrutural necessitam ser implementados a fim de avaliar a probabilidade de que uma dada conformação local encontrada realmente corresponda à conformação nativa (DORN et al., 2014). Estes métodos têm obtido destaque por serem utilizados pelos preditores que alcançam os melhores resultados, com

maior acurácia, nas últimas edições do *Critical Assessment of Protein Structure Prediction* (CASP) [3] (MOULT et al., 2016; MOULT et al., 2018).

1.1 Motivação

Os esforços dedicados ao desenvolvimento de métricas capazes de elucidar o problema de PSP datam algumas décadas, porém este problema ainda representa um grande marco a ser resolvido pelo que compreende a bioinformática estrutural. Uma série de abordagens e metodologias têm sido apresentadas ao longo dos anos, contudo destacam-se aquelas capazes de obter êxito em combinar informações previamente estabelecidas e métodos *de novo*, a fim de mitigar possíveis impedimentos decorrentes da quantidade limitada de dados depositados em bancos de dados estruturais (FLOUDAS et al., 2006; DORN et al., 2014).

A resolução deste problema depende de uma série de etapas que necessitam ser superadas a fim de alcançar resultados próximos referentes aos estados nativos apresentados pela proteína de interesse. Em Dorn et al. (2014) são apontados três principais pontos os quais devem ser cuidadosamente elaborados para que o método proposto tenha sucesso na resolução daquilo que se propõe. Primeiramente, a representação computacional de modelos estruturais 3D de cada proteína. Uma vez proposto um estado estrutural, o segundo desafio a ser superado diz respeito a uma maneira fidedigna de avaliação do modelo proposto. Baseando-se na hipótese termodinâmica postulada por Anfinsen (1973), a qual diz que valores de energia livre mínimos correspondem à conformações nativas assumidas por proteínas. Contudo, sabe-se que uma única proteína apresenta diversos estados conformacionais, logo, diversos valores mínimos energéticos podem ser obtidos a partir de um único peptídeo. Esta liberdade de possíveis conformações assumidas por proteínas em seu estado nativo endereça o terceiro desafio, uma maneira eficiente de explorar o espaço de busca conformacional por um estado energético mínimo.

Desta maneira, este trabalho visa em primeira instância propor uma maneira de representação de modelos estruturais. Aproveitando-se de dados previamente depositados em bancos de dados, a geração destes modelos será guiada por dois limitantes de relevância biológica. Para tal fim, adotaremos a definição de limitante semelhante à proposta por Worth, Gong and Blundell (2009), o qual não apresenta cunho obrigatório, porém representa um fator de aceitação da condição a ser avaliada. Portanto, o limitante estrutural

[3] <http://www.predictioncenter.org/>

adotado será composto por uma condição composta de duas etapas para fins de aceitação de um modelo estrutural gerado. O primeiro limitante diz respeito ao padrão conformacional adotado por um determinado aminoácido quando inserido em uma dada estrutura secundária (LIGABUE-BRAUN et al., 2018); enquanto que o segundo, refere-se a proximidade espacial entre dois aminoácidos distintos em uma sequência petídica (WEIGT et al., 2009). Os modelos gerados serão avaliados através de função de energia livre de resolução atômica, assim como parâmetros capazes de indicar o estado de enovelamento (*protein folding*) de um dado modelo. Por fim, serão avaliados os desempenhos de diferentes algoritmos de busca na otimização destes modelos gerados a partir dos limitantes propostos. O problema de PSP é considerado como um problema NP-difícil (UNGER; MOULT, 1993), portanto sua resolução não é possível através de métodos determinísticos. Como alternativa, métodos estocásticos são capazes de encontrar soluções aproximadas em tempos de execuções aceitáveis (TALBI, 2009b). Tais métodos já vem sendo amplamente implementados na tentativa de propor uma solução para o problema de PSP, pois são capazes de contornar tamanha complexidade, além de permitir a inclusão de conhecimento prévio ao longo de sua execução a fim de diminuir o espaço de busca conformacional (DORN et al., 2014).

1.2 Objetivos

O objetivo geral do presente trabalho é a incorporação de duas informações com relevância biológica inferida e retirada de dados biológicos previamente depositados em bancos de dados, tendo em vista a geração de modelos mais próximos do estado nativo da proteína de interesse. As metas, cujo alcance são necessárias para que se cumpram os objetivos gerais e específicos da presente dissertação são:

1. Realização de um levantamento bibliográfico a fim de analisar a situação do estado da arte em relação ao problema de PSP;
2. Analisar a utilização das informações biológicas depositadas em bancos de dados para resolução do problema de PSP, a utilização individual de cada limitante, assim como a união de ambas métricas em uma mesma metodologia;
3. Propor um método de implementação para construção de modelos estruturais de maneira que sejam garantidos: primeiramente, os ângulos de torção retirados de bancos de dados; segundo, garantir que sejam respeitados os contatos preditos entre

aminoácidos distintos;

4. Implementação de algoritmos de busca para a otimização das estruturas geradas a partir do método composto por ambos limitantes propostos;
5. Adição de termos adicionais à função de avaliação dos modelos estruturais referentes aos limitantes biológicos utilizados, a fim de garantir que a informação obtida através dos contatos preditos não sejam perdidos durante o processo de otimização;
6. Utilização de um algoritmo multiobjetivo para avaliação estrutural, através de perfis energéticos e de enovelamento, por meio de uma função de avaliação energética e uma função de avaliação de aminoácidos em proximidade no espaço 3D.

Concluído o presente trabalho, espera-se que o método proposto de construção de modelos estruturais guiados por dois fatores limitantes de relevância biológica seja eficaz e acurado. Além disto, almeja-se evidenciar que a construção proposta seja capaz de direcionar e otimizar o processo de busca conformacional realizado por algoritmos de busca já utilizados para o problema de PSP.

1.3 Estrutura da dissertação

A presente dissertação está organizada de acordo com a estrutura:

- Capítulo 2: *Fundamentação Teórica*, neste capítulo estão descritos os conceitos gerais abordados durante a execução do presente trabalho. Estes conceitos referem-se a proteínas, sua composição química, organização e classificação estruturais e níveis de abstração, assim como a classificação estrutural de proteínas. No que diz respeito ao escopo computacional, este capítulo descreve o cenário atual do problema de predição de proteínas, sendo eles a modelagem computacional do problema biológico, definição teórica das informações utilizadas para elaborar a abordagem computacional proposta por este trabalho, funções de avaliação empregadas para avaliação dos modelos gerados e métodos de otimização utilizados. Por fim, o capítulo tem por objetivo situar o leitor em relação aos principais conceitos necessários para o entendimento da dissertação, assim como apresentar um panorama da abordagem proposta após a conclusão do trabalho;
- Capítulo 3: *Revisão teórica*, este capítulo tem por objetivo situar o presente trabalho dentro do cenário atual do problema de PSP e como este se relaciona com os demais trabalhos. Primeiramente, apresentando um panorama do estado da arte relacionado

a resolução do problema de PSP, seguido pela apresentação da utilização de meta-heurísticas aplicadas ao problema de PSP;

- Capítulo 4: *Materiais e Métodos*, este capítulo relata as metodologias empregadas para o método proposto. Explica o método elaborado para a construção de modelos estruturais utilizando ambos limitantes propostos, assim como os algoritmos utilizados nos processos de otimizações de cada modelo estrutural gerado. O objetivo deste capítulo é esclarecer a metodologia utilizada e empregada durante a elaboração e a execução do trabalho.
- Capítulo 5: *Resultados*, este capítulo apresentará os resultados obtidos pela metodologia proposta neste trabalho. Em uma primeira etapa serão apresentados os resultados das avaliações estruturais realizadas nos modelos estruturais gerados utilizando os limitantes propostos neste trabalho. Em seguida, os resultados dos processos de otimização realizados por ambas meta-heurísticas escolhidas utilizando os modelos gerados previamente na primeira etapa do trabalho. O capítulo tem por objetivo avaliar a eficácia do método proposto na geração de modelos estruturais a partir dos limitantes propostos, assim como a análise da utilização destes modelos como indivíduos em algoritmos de buscas empregados na otimização de estruturas de proteínas. O capítulo está estruturado em formato de artigo científico para futura publicação;
- Capítulo 6: *Discussão Geral*, o capítulo apresentará uma discussão da importância na inclusão de informações relevantes no processo de resolução do problema de PSP, assim como a eficácia dos limitantes propostos. Serão discutidos os resultados obtidos pela amostragem de indivíduos com parâmetros estruturais semelhantes aos descritos pelas estruturas 3D depositadas no PDB. Também será discutida a melhora de performance de algoritmos de busca quando adicionadas informações ao inicializar as populações iniciais, assim como a integração das informações de contato durante processo de busca no espaço conformacional.
- Capítulo 7: *Conclusões e Perspectivas*, neste capítulo constarão as considerações finais obtidas após o término do trabalho. Também neste capítulo serão indicadas diferentes possíveis aplicações dos modelos gerados utilizando o método proposto, assim como recomendações para continuação do trabalho.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Proteínas

Proteínas são macromoléculas de grande importância biológica, envolvidas virtualmente em todos os processos celulares. Sabe-se que suas funções estão diretamente ligadas ao seu estado nativo, ou seja, dependem das estruturas 3D assumidas pelas sequência de aminoácidos (ANFINSEN, 1973). Nota-se que também são descritos casos de proteínas funcionais que apesar de apresentarem padrões estruturais desordenados obrigatórios para o desempenho correto de suas funções (GUNASEKARAN et al., 2003), estas ainda adquirem conformações específicas ao encontrarem diferentes ligantes (DUNKER et al., 2001). Contudo, as interações moleculares presentes em conformações 3D de proteínas exercem extrema influência no desempenho de suas funções, portanto a descrição da estrutura 3D de proteínas é de extrema importância para o total entendimento do papel de uma dada proteína dentro do complexo esquema biológico ao qual esta está inserida (LASKOWSKI; WATSON; THORNTON, 2005).

2.1.1 Aminoácidos: blocos construtores

Aminoácidos são pequenas moléculas que atuam como blocos de construção durante o processo de síntese proteica. De maneira geral, são encontrados 20 aminoácidos canônicos (Tabela 2.1) ubíquos em diferentes tipos celulares de diferentes organismos. Independente do tipo de aminoácido, todos compartilham uma porção idêntica de cadeia, chamada de cadeia principal (*main chain*). Esta porção compartilhada é composta por um grupamento amina ($-NH2$), carboxila ($-COOH$), um átomo de hidrogênio (H) e uma cadeia R, ligados à um carbono central (C_α). O que difere cada um dos aminoácidos é a composição e o arranjo de suas cadeias R, as cadeias laterais (*side chain*) (Figura 2.1). As características físico-químicas de cada um são definidas de acordo com a composição atômica, carga elétrica e polaridade da cadeia lateral (NELSON; LEHNINGER; COX, 2008).

Tabela 2.1: Aminoácidos Canônicos

Aminoácido	**3 Letras**	**1 Letra**	**Aminoácido**	**3 Letras**	**1 Letra**
Cadeia lateral - Apolar e alifática			Cadeia lateral - Polar e sem carga		
Alanina	Ala	A	Treonina	Thr	T
Glicina	Gly	G	Cadeia lateral - Aromática		
Isoleucina	Ile	I	Fenilalanina	Phe	F
Leucina	Leu	L	Tirosina	Tyr	Y
Metionina	Met	M	Triptofano	Trp	W
Valina	Val	V	Cadeia lateral - Carregada positiva		
Prolina	Pro	P	Arginina	Arg	R
Cadeia lateral - Polar e sem carga			Histidina	His	H
Asparagina	Asn	N	Lisina	Lys	K
Cisteína	Cys	C	Cadeia lateral - Carregada negativa		
Glutamina	Gln	Q	Aspartato	Asp	D
Serina	Ser	S	Glutamato	Glu	E

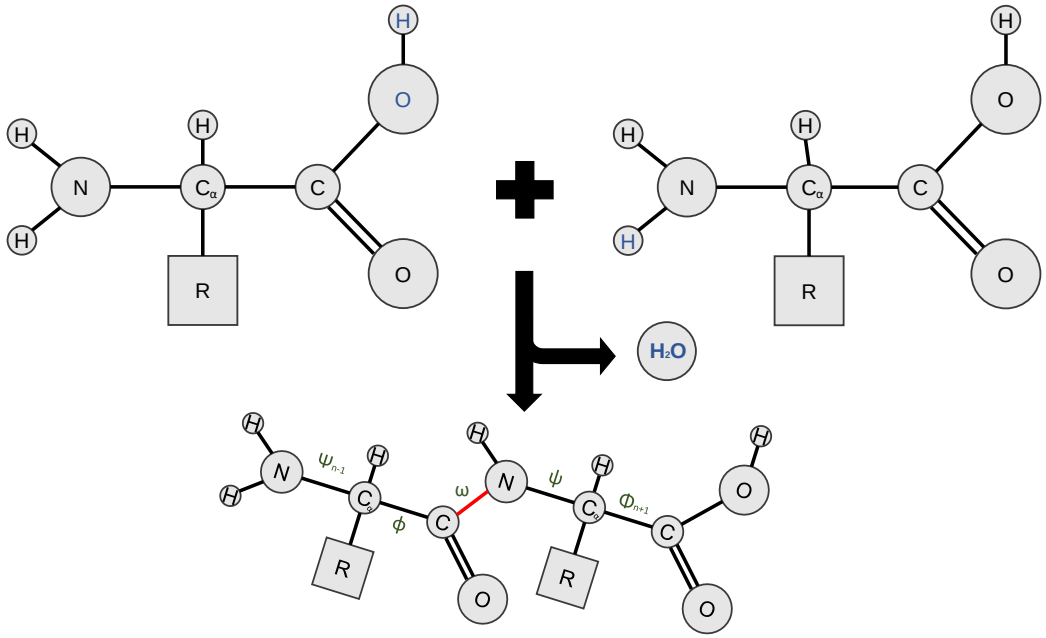
Lista dos 20 aminoácidos canônicos com seus respectivos códigos de três e uma letra, separados segundo as características físico-químicas de suas respectivas cadeias laterais.

2.1.2 Estrutura de proteínas: níveis hierárquicos

As possíveis funções biológicas exercidas por cada proteína podem ser inferidas e entendidas a partir de suas estruturas 3D. Desta forma, é possível descrever a conformação 3D de uma dada proteína a partir de quatro níveis organizacionais (Figura 2.2). A relação exercida por cada nível sobre o nível posterior é descrita como hierárquica, ou seja, as informações obtidas a partir do nível anterior são de alta importância para a caracterização do nível seguinte (VERLI, 2014).

O nível de complexidade basal de estrutura proteica, também conhecido como *estrutura primária*, refere-se à sequência linear de aminoácidos ligados através de ligações peptídicas (Figura 2.2). Cadeias peptídicas são formadas a partir da polimerização de aminoácidos através da liberação de uma molécula de água (H_2O). Esta liberação decorre da reação entre o grupamento carboxílico de um aminoácido com o grupamento amina de outro aminoácido ($C - N$) (Figura 2.1). Peptídeos são formados pela junção de dois ou mais aminoácidos ligados de modo linear (VOET; VOET, 2010). A medida em que o número de aminoácidos aumenta, peptídeos passam a ser classificados como polipeptídeos. Proteínas, por sua vez, são compostas por uma ou mais cadeias polipeptídicas. Cada proteína é composta por uma sequência única de aminoácidos, responsáveis por conferir características físico-químicas exclusivas e específicas para cada polipeptídeo,

Figura 2.1: Aminoácido e Ligação Peptídica



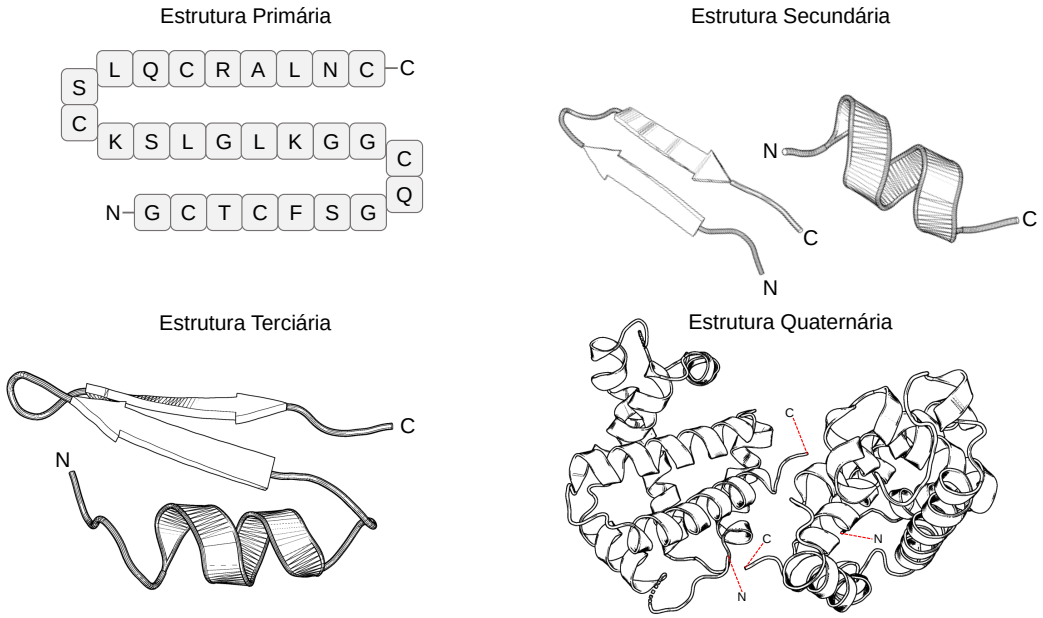
Representação simplificada de dois aminoácidos genéricos formando uma ligação peptídica (indicada pela cor vermelha). Indicados na imagem (em cor verde) estão presentes os ângulos de rotação PHI (ϕ), PSI (ψ) e OMEGA (ω) presentes no esqueleto peptídico.

conferindo então sua função fisiológica (LESK, 2010).

As estruturas pertencentes ao segundo nível de organização estrutural são definidas por conformações locais assumidas pelo esqueleto peptídico, chamadas de *estrutura secundária* (Figura 2.2). Este arranjo 3D adotado localmente por segmentos da cadeia de aminoácidos independe das conformações adotadas pelos átomos que compõem as cadeias laterais do aminoácidos e, salvo β-folhas (descritas posteriormente), também independem da interação entre fragmentos de sequência peptídica distintos (NELSON; LEHNINGER; COX, 2008).

As conformações adotadas pelo esqueleto peptídico de uma determinada sequência de aminoácidos pode ser descrita a partir de seus ângulos de torção, também chamados de ângulos diedrais. Contudo, destaca-se de antemão que devido ao caráter de ligação dupla parcial assumida pra ligação peptídica ($N - C$), o ângulo de torção ômega (ω) é definido como planar e assume valores rotacionais: conformação *trans*, ω=180° e *cis*, ω=0°, sendo a conformação *trans* mais comumente encontrada (PAULING; COREY; BRANSON, 1951). Desta maneira as diversas conformações adotadas por uma cadeia polipeptídica pode ser definida pelos ângulos de torção restantes, PHI (ϕ) entre a ligação $C_\alpha - N$ e PSI (ψ) entre $C_\alpha - C$ de cada aminoácido (Figura 2.1) (BRANDEN; TO-

Figura 2.2: Relação Estrutural Hierárquica de Proteínas



Representação simplificada da relação hierárquica entre níveis estruturais de proteínas. No quadrante superior esquerdo está representada a estrutura primária, referente à sequência linear de aminoácidos. O quadrante superior direito exemplifica as duas estruturas secundárias mais recorrentes, β-folha (esquerda) e α-hélice (direita). Um modelo de estrutura terciária (PDB ID: 1ACW) representado no quadrante inferior esquerdo, enquanto no inferior direito está representado o nível de estrutura quaternário (PDB ID: 4HHB).

OZE, 2012). Os ângulos ϕ e ψ podem assumir qualquer valor dentro do espaço rotacional $[-180°,180°]$, contudo algumas conformações são estericamente impossibilitadas. Este impedimento é decorrente da aproximação de átomos não-ligados à uma distância menor do que a distância de van der Waals correspondentes, sejam estes átomos pertencente ao esqueleto peptídico ou a cadeia lateral (VOET; VOET, 2010). As combinações de valores assumidos por ϕ e ψ podem ser vizualidas em um diagrama de Ramachandran (Figura 2.3) (RAMACHANDRAN, 1963; RAMACHANDRAN; SASISEKHARAN, 1968).

Análogo aos ângulos pertencentes à cadeia principal previamente descritos, os átomos das cadeias laterais dos aminoácidos também possuem ângulos de torção, os ângulos CHI (χ). O número de ângulos χ da cadeia lateral varia de 0 a 4 ângulos dependendo do tipo de aminoácido e, semelhante aos ângulos ϕ e ψ, possuem liberdade rotacional dentro do espaço $[-180°,180°]$ (KESSEL; BEN-TAL, 2010). Portanto, a estrutura 3D de uma proteína pode ser descrita a partir do conjunto de ângulos de rotação, sendo ϕ, ψ e ω definindo a conformação adotada pela cadeia principal, enquanto que o grupo de ângulos χ descrevendo a conformação da cadeia lateral de cada aminoácido pertencente

Figura 2.3: Diagrama de Ramachandram

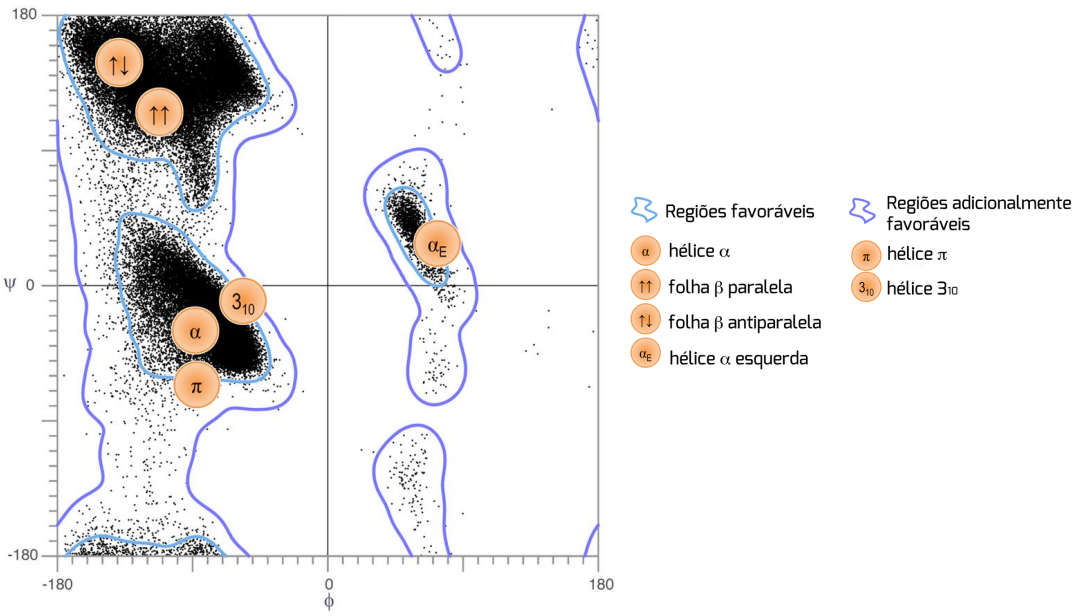


Diagrama de Ramachandram para aminoácidos em geral calculados a partir de estruturas de proteínas experimentalmente determinadas. Adaptado de Verli (2014).

ao peptídio (HOVMÖLLER; ZHOU; OHLSON, 2002).

Dentre as diferentes conformações assumidas por fragmentos, destacam-se algumas por ocorrerem amplamente entre diversas proteínas distintas. Estas estruturais secundárias mais comumente encontradas são:

Hélices Descrita pela primeira vez há mais de meio século (PAULING; COREY; BRANSON, 1951), α-hélices possuem um dos arranjos conformacionais mais rígidos encontrados entre as diferentes estruturas secundárias (Figura 2.2). Esta conformação é o resultado da ligação de hidrogênio entre a ligação $C = O$ da porção carboxílica do aminoácido de posição *n* e a ligação $N - H$ da porção amina do aminoácido de posição *n+4* da cadeia peptídica (LESK, 2010). As combinações de ângulos ϕ/ψ geralmente assumidas por aminoácidos sob α-hélices estão dentro do intervalo ϕ = $(-70.0°,-60.0°)$ e ψ = $(-45.0°,-39.0°)$, caracterizando estas conformações como pertencentes ao grupo de maior restrição (LIGABUE-BRAUN et al., 2018). Sua estrutura é estabilizada pelo alto número de ligações de hidrogênio encontradas em uma única hélice (NELSON; LEHNINGER; COX, 2008). Apesar de α-hélices serem mais frequentemente encontradas, ainda existem diferentes tipos de hélices ao que diz respeito ao número de aminoácidos e átomos em cada volta da hélice. Esta classificação possui a notação n_m, onde n refere-se ao número de aminoácidos

por volta, enquanto que $_m$ representa o número de átomos por volta. Seguindo esta nomenclatura, α-hélices são representadas por 3.6_{16}, enquanto que diferentes hélices possuem diferentes descrições, por exemplo, 3_{10} e 4.4_{16} hélices (VOET; VOET, 2010).

Folhas Este padrão de estruturas secundárias foi descrito pela primeira vez no mesmo ano em que as α-hélices (PAULING; COREY, 1951). β-Folhas possuem segmentos de sequência praticamente estendidas e, diferente das estruturas de hélice, os valores possíveis adotados pelos ângulos de rotação ϕ e ψ são mais abrangentes, ocupando praticamente todo o quadrante superior esquerdo do mapa de Ramachandran (RICHARDSON, 1981). As ligações de hidrogênio ocorrem entre aminoácidos provenientes de segmentos distintos da cadeia peptídica, inclusive provenientes de outras moléculas (Figura 2.2) (KESSEL; BEN-TAL, 2010). As folhas podem possuir dois sentidos distintos: paralelo, quando ambas as fitas possuem o mesmo sentido amina-carboxila, ou anti-paralelo, quando uma fita apresenta sentido amina-caboxila e a outra fita está no sentido relativo carboxila-amina (BRANDEN; TOOZE, 2012). As regiões de maior densidade no mapa de Ramachandran para folhas paralelas são mais abrangentes quando comparadas as regiões de folhas anti-paralelas (RICHARDSON, 1981). De forma abrangente, o intervalo de ângulos referentes às β-folhas é ϕ = ($-139.0°$,$-119.0°$) e ψ = ($-135.0°$,$-113.0°$), sendo ϕ = $-119.0°$, ψ = $-113.0°$ para paralelas e ϕ = $-139.0°$, ψ = $-135.0°$ para anti-paralelas (Figura 2.3) (LIGABUE-BRAUN et al., 2018).

Padrões não-repetitivos Também conhecidos como regiões de *Coil*, estes fragmentos de estrutura secundária possuem formas irregulares e mais complexas de serem descritas. Estas regiões geralmente apresentam padrões de densidade eletrônica muito menores do que as estruturas que apresentam uma maior regularidade. Podendo estar completamente ausente em alguns casos (RICHARDSON, 1981). Nota-se que *Coil* e *Random Coil* possuem semelhanças, porém diferem em detalhamento, o primeiro refere-se a regiões naturalmente desordenadas, enquanto que o segundo é a definição dada a conformações altamente flutuantes comumente atribuídas à proteínas em processo de desnaturação (VOET; VOET, 2010). Devido a maior liberdade conformacional, as combinações de ângulos ϕ e ψ permitidas para estas estruturas secundárias possuem o maior grau de liberdade, resultando em uma ocupação mais abrangente do mapa de Ramachandran (LIGABUE-BRAUN et al., 2018)

Enquanto o segundo nível de organização estrutural de proteínas refere-se ao ar-

ranjo 3D local da sequência peptídica, o terceiro nível hierárquico, *estrutura terciária*, diz respeito à conformação de todos os átomos do peptídeo. Em termos gerais, refere-se ao arranjos de todas as estruturas secundárias presentes na macromolécula (Figura 2.2) (LESK, 2010). Neste nível de conformação, as estruturas passam a também ser chamadas de estruturas nativas. O arranjo 3D assumido pelas moléculas neste estágio é decorrente de uma série de interações realizadas pelos átomos constituintes, sejam elas intramoleculares, ligações de hidrogênio, hidrofobicidade, entre outras (RICHARDSON, 1981). Semelhante ao conceito de composição 3D de fragmentos conformacionais, a *estrutura quaternária* de uma proteína é definida pela formação de arranjos 3D de complexos estruturais entre mais de uma cadeia polipeptídica ou subunidades, sejam elas idênticas ou distintas (Figura 2.2) (NELSON; LEHNINGER; COX, 2008).

2.2 Classificação estrutural de proteínas

A tendência evolutiva sobre a maior conservação estrutural em comparação à sequência de aminoácidos é um fato consolidado e amplamente descrito na literatura (LESK; CHOTHIA, 1980; ROST, 1999; FOX; BRENNER; CHANDONIA, 2015). Considerando esta alta taxa de conservação, uma série de modelos de classificação estruturais tem sido propostos ao longo dos últimos anos. O modelo de classificação CATH [1] (DAWSON et al., 2016) (sigla em inglês, *Class, Architecture, Topology and Homologous*) categoriza proteínas em um esquema hierárquico composto por quatro níveis de classificação de acordo com características de cada proteína. O primeiro, diz respeito à Classe a qual a proteína pertence embasando-se na composição da estrutura secundária apresentada pela proteína em questão, sendo elas: principalmente α, principalmente β, α/β (possuindo ambas α hélices e β folhas) e poucas estruturas secundárias. Em segunda estância, temos a descrição geral do arranjo de estrutura secundária, denominada Arquitetura, seguida pela Topologia a qual descreve a conectividade entre as estruturas secundárias. Em quarto e último nível, temos a classificação de Homologia de acordo com a estruturas homólogas a proteína de interesse.

De forma similar, o banco de dados SCOP [2] (MURZIN et al., 1995) (sigla em inglês, *Structural Classification of Proteins*) categoriza as estruturas 3D de proteínas depositadas no PDB. Sua classificação também começa de maneira mais abrangente a partir

[1] <http://www.cathdb.info/>
[2] <http://scop.mrc-lmb.cam.ac.uk/scop/>

do grupo de classes (*Class*), de acordo com as descrições: classe de hélices (All-α), classe de folhas (All-β), α/β (onde há a presença de ambas estruturas secundárias de maneira homogênea na estrutura da proteína), α+β (quando ambas estruturas secundárias ocorrem de forma heterogênea) e multi-domínio (quando há presença de mais de um domínio pertencente à classes distintas). Em seguida, estas proteínas são classificadas de acordo com seu Enovelamento (*Fold*), ou seja, grupos que compartilham arranjos de estrutura secundária semelhantes. Superfamília (*Superfamily*) e Família (*Family*), sendo este classificado de acordo com relações evolutivas próximas conforme a similaridade de sequência e estrutura de proteínas, e aquele de acordo com relações evolutivas distantes classificadas de acordo com critérios estruturais e funcionais. Em última estância são classificadas de acordo com a Proteína e Espécie de origem.

Desta maneira, considerando o padrão de classificação de proteínas adotados por dois grandes classificadores, CATH e SCOP, o atual trabalho propõe uma classificação mais simplificada de proteínas. A divisão escolhida dá-se, principalmente, de acordo com os dois limitantes biológicos investigados ao longo da dissertação. A classificação será baseada na estrutura secundária predominante descritas para cada proteína utilizada em nosso grupo de teste. As classes utilizadas ao longo dessa dissertação seguem os padrões propostos pelo SCOP e Chou and Zhang (1995), sendo elas:

- Classe de hélices: Abrangendo proteínas constituídas por mais de 60% de seus aminoácidos sob a estrutura secundária de hélices (independente do tipo de hélice);
- Classe de folhas: Integrando proteínas que possuam mais de 60% dos aminoácidos em formação de folhas β (independente do sentido de suas folhas);
- Classe híbrida: Englobando proteínas contendo um misto das duas classes propostas anteriormente de acordo com sua composição de estrutura secundária, incluindo regiões de *Coil*.

2.3 Representação computacional de proteínas

A estrutura 3D de uma dada proteína é um resultado de uma série de forças físico-químicas, tais como temperatura, salinidade, entre outros, que regem o processo espontâneo de enovelamento de uma sequência de aminoácidos (ANFINSEN, 1973; ANFINSEN et al., 1961). Esta série de fatores são responsáveis por guiar o processo de definição da conformação espacial destas moléculas e acabam dificultando a representação computa-

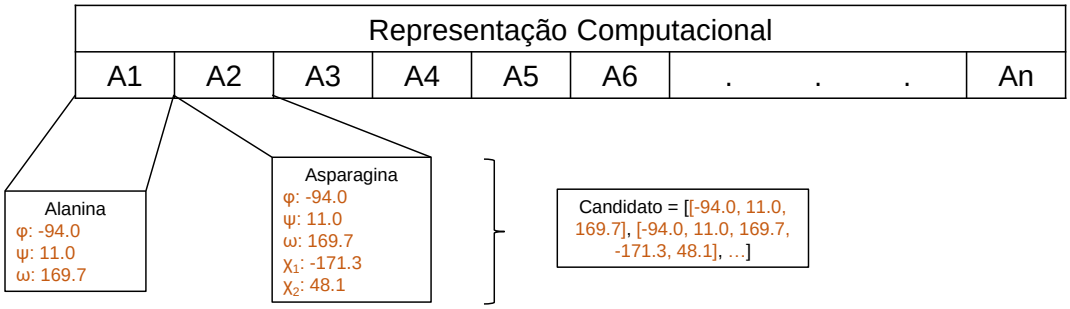
cional de proteínas, não sendo suficiente apenas a representação de posições atômicas, mas também a descrição de interações moleculares e forças estabilizadoras presentes na estrutura nativa, por exemplo, através do emprego de uma função de energia potencial. Desta forma, o nível de detalhamento com o qual o modelo é representado é de extrema importância, pois quanto maior o nível descritivo, maior será a fidelidade ao modelo nativo, contudo, também representa um aumento na complexidade computacional do problema. Por exemplo, modelos de resolução atômica, onde todos os átomos da molécula são descritos (*all atom*). Estratégias distintas são utilizadas para diminuir a complexidade computacional, porém há perda de informação e redução da capacidade descritiva do modelo quando comparado à estrutura nativa (MIRNY; SHAKHNOVICH, 2001).

Em decorrência da influência direta em relação ao aumento ou diminuição da complexidade computacional do problema de PSP, a escolha de representação geométrica 3D de proteínas é um dos passos mais importantes no processo de desenvolvimento de métodos computacionais para resolução do problema de PSP. Devido ao alto custo computacional atrelado a representações com um alto grau de detalhamento, representações com um nível de abstração maior geralmente são escolhidas com maior frequência em decorrência da diminuição de complexidade (CHIVIAN et al., 2003; CORRÊA; DORN, 2017). A representação cartesiana (x,y,z) dos átomos de cada de aminoácido é dada de maneira que cada cadeia peptídica seja representada por um conjunto A de átomos a em um espaço 3D, onde $\{a|a \in \mathbb{R}^3\}$. Como alternativa, cadeias polipeptídicas podem ser representadas igualmente por seus ângulos de diedro, uma vez que os comprimentos de ligações químicas entre os átomos dos aminoácidos se mantém praticamente constante ao longo de toda proteína (NEUMAIER, 1997). Desta maneira é possível diminuir o nível de complexidade do problema, pois o aminoácido com o menor número de átomos (Glicina) possui 10 átomos, enquanto que os aminoácidos com o maior número de ângulos diedrais (Lisina e Arginina) possuem 7 ângulos de torção. Contudo, nota-se que modificações mínimas a nível de ângulos torcionais podem representar drásticas mudanças conformacionais a nível global da molécula, enquanto que mudanças pontuais em posições atômicas representam pouca mudanças globais na estrutura 3D da proteína.

Desta maneira, para fins de realização do presente trabalho, o modelo de representação computacional de estrutura 3D de polipeptídios durante os processos de otimização dos indivíduos gerados a partir da métrica proposta será utilizando os ângulos diedrais de cada aminoácido presente na sequência de cada proteína (Figura 2.4). Espera-se uma redução da complexidade computacional, em decorrência da não-utilização da resolução

full atom, porém ainda mantendo um grau de precisão considerável dos modelos gerados e otimizados quando comparados as estruturas depositadas no PBD. Portanto, a estrutura 3D do peptídio P possuindo n aminoácidos, pode ser definida segundo a Eq. 2.1, onde apenas os valores dos ângulos de diedro são atribuídos para cada aminoácido compondo a proteína.

Figura 2.4: Representação Computacional de Proteínas



Representação visual da representação computacional adotada neste trabalho. Cada modelo estrutural será representado como um vetor, onde cada posição deste vetor corresponde à um aminoácido da sequência peptídica. Por sua vez, cada aminoácido é representado por um vetor contendo os ângulos de torção respectivos.

$$P = (aa_1, aa_2, aa_3, ..., aa_{n-1}, aa_n) \tag{2.1}$$

onde,

$$aa_i = (\phi_i, \psi_i, \omega_i, \chi_{0...4}) \tag{2.2}$$

A cadeia principal de P é representada pelo modelo com n aminoácidos, onde cada um possui 3 $\times n$ graus de liberdade ou variáveis a serem otimizadas. Igualmente, adicionando os ângulos da cadeia lateral, é possível calcular a dimensionalidade das variáveis, ou seja, a cardinalidade do conjunto de ângulos (Eq. 2.3). Nota-se que o ângulo ϕ, na posição N-terminal, e o ângulo ψ, na porção C-terminal, são inexistentes, portanto devem ser excluídos do cálculo.

$$|P| = n \times 3 \times (\sum_{1}^{n} |\chi_i|) - 2 \tag{2.3}$$

O problema de PSP passa a poder ser descrito como um problema de otimização matemática (LEUNG; WANG, 2001). Sendo $f(x)$ a função utilizada para avaliar as soluções encontradas, de maneira que $f(x)$ seja minimizada em relação ao intervalo

de números pertencentes ao conjunto de número reais delimitado por $l \leq x \leq u$, onde $x = P$ (Eq. 2.1), que por sua vez também é constituído por um vetor de variáveis de valores de ângulos torcionais de um determinado aminoácido (Eq. 2.2).

2.4 Funções de avaliação

Funções de energia são amplamente empregadas para a resolução do problema de PSP a fim de avaliar o estado de enovelamento do modelo gerado, de tal maneira que sejam considerados os modelos mais semelhantes à estrutura considerada nativa da proteína (FARAGGI; KLOCZKOWSKI, 2014). Durante os processos de otimização, procura-se a minimização do valor energético calculado para um determinado modelo, uma vez que, segundo a teoria termodinâmica, estruturas nativas correspondem a estruturas com maior grau de estabilidade, logo também correspondem a estados energéticos mínimos (ANFINSEN, 1973; LAZARIDIS; KARPLUS, 2000). Contudo, a totalidade de processos e forças presentes no processo de enovelamento de proteínas ainda permanece parcialmente conhecida e descrita, portanto as funções de energia disponíveis não são capazes de mimetizar de maneira absoluta o estado conformacional de proteínas em seu estado nativo (KIM et al., 2009). Sabe-se que valores semelhantes de mínimos energéticos podem representar estruturas 3D diferentes para uma mesma proteína de interesse (ANFINSEN, 1973).

Frequentemente, funções de energia desenvolvidas para serem utilizadas na avaliação de modelos estruturais são consideradas como *knowledge-based energy functions*, ou seja, possuem funções potenciais derivadas a partir de estrutura 3D previamente descritas e depositadas no PDB (HAO; SCHERAGAT, 1999; CHIVIAN et al., 2003). De maneira geral, funções de energia são o resultado do somatório, ponderado ou não, de parâmetros que mimetizam as interações físico-químicas necessárias para a representação da proteína em sua conformação 3D (JR, 2004). Para fins de avaliação de modelos gerados a partir do método proposto, neste trabalho emprega-se a função de energia de resolução atômica (*all atom*) implementada pelo Rosetta[3] (ROHL et al., 2004).

[3]<https://www.rosettacommons.org/>

2.4.1 Função de Energia do Rosetta

Os modelos gerados utilizando o método proposto foram avaliados segundo seus valores energéticos calculados a partir da função de energia de resolução atômica implementada no PyRosetta[4] (CHAUDHURY; LYSKOV; GRAY, 2010). Este, refere-se à uma implementação na linguagem de programação Python[5] do programa de modelagem molecular do Rosetta (ROHL et al., 2004). Desta forma sendo possível considerar equivalentes as implementações de função de energia do PyRosetta às implementações do Rosetta. A função de energia do Rosetta foi desenvolvida inicialmente apenas com potenciais estatísticos de aminoácidos individuais e frequência de interação de pares de aminoácidos a partir de análises de estruturas previamente depositadas no PDB (SIMONS et al., 1997). Posteriormente, termos adicionais foram sendo incorporados à função de energia, tais como termos de empacotamento, de ligação de hidrogênio, estrutura secundária, interações de van der Waals (SIMONS et al., 1999). Esses termos apenas permitiam modelagens de baixa resolução, considerando apenas coordenadas da cadeia principal, enquanto que as cadeias laterais eram tratadas de maneira implícita. No início dos anos 2000 uma função de energia com maior resolução atômica foi desenvolvida por Kuhlman and Baker (2000), através da adição de novos termos, tais como Lennard-Jones (NERIA; FISCHER; KARPLUS, 1996), modelos de solvatação (LAZARIDIS; KARPLUS, 1999) e preferência de rotâmeros dependentes da cadeia principal (JR; COHEN, 1997).

A resolução dos átomos de uma proteína pode ser representada de duas maneiras distintas no Rosetta: modelo centróide e modelo *all-atom* (LEUNG; WANG, 2001). Na primeira representação, o nível de detalhamento da estrutura 3D da proteína de interesse é reduzido, pois os átomos pertencentes à cadeia lateral dos aminoácido que compõem a proteína são representados apenas pela localização correspondente ao centro de massa da cadeia lateral. Em contraste, o modelo *all-atom* disponibiliza uma maior resolução, descrição e detalhamento das estruturas 3D dos modelos avaliados, pois considera individualmente todos os átomos presentes na cadeia lateral de cada aminoácido, incluindo os átomos de hidrogênio (ROHL et al., 2004). A função de avaliação *all-atom* utilizada para avaliar as estruturas geradas durante a execução do presente trabalho é a função REF2015, a qual incorpora um total de 19 termos de energia para o cálculo de energia livre de cada modelo 3D (Tabela 2.2) (ALFORD et al., 2017). Nesta função, as for-

[4] <http://www.pyrosetta.org/>
[5] <https://www.python.org/>

ças de atração e repulsão atuando sobre os pares de átomos de acordo com a distância entre átomos, decorrentes das interações de van der Waals, são calculadas utilizando o potencial de Lennard-Jones 6-12 (JONES; CHAPMAN, 1924; JONES, 1924). Interações eletroestáticas não-ligadas de átomos, totalmente ou parcialmente carregados, são calculados com base na lei de Coulomb retiradas e adaptadas do CHARMM (PARK et al., 2016; BROOKS et al., 1983). Com base no modelo de exclusão Gaussiano implícito de Lazaridis-Karplus, as aproximações de solvatação são estimadas utilizando um modelo de água *bulk* (LAZARIDIS; KARPLUS, 1999). As ligações de hidrogênios são calculadas a partir da junção dos termos eletroestáticos e a avaliação de estruturas cristalográficas de alta resolução (O'MEARA et al., 2015). Os termos utilizados para a conformação de ângulos de torção ϕ/ψ da cadeia principal são calculados embasados em análises de mapas de Ramachandran, assim como as conformações das cadeias laterais são calculadas utilizando probabilidades de acordo com uma biblioteca de rotâmeros dependentes da cadeia principal (SHAPOVALOV; JR, 2011). A energia potencial calculada pelo PyRosetta é o resultado da somatória linear de todos os termos ponderados de acordo com seus respectivos pesos (ALFORD et al., 2017).

2.4.2 Função de avaliação final

Juntamente com a função REF2015 (ALFORD et al., 2017) disponibilizada pelo Rosetta, descrita na seção anterior, dois termos adicionais foram incorporados ao valor de energia calculado. O primeiro termo refere-se ao cálculo de área total de superfície acessível ao solvente (SASA, sigla em inglês *Solvent-Accessible Surface Area*) (LEE; RICHARDS, 1971; CONNOLLY, 1983). Este foi calculado utilizando a implementação também oferecida pelo PyRosetta, com o raio atômico de 1.5Å. A interpretação do valor de SASA pode ser feita, de uma maneira simplificada, de tal maneira: quanto menor o valor de SASA calculado, menor o número de átomos expostos ao solvente, logo pode-se inferir que a estrutura 3D da proteína encontra-se em um grau de enovelamento maior, ou seja, o peptídio encontra-se em uma conformação mais compacta (ROSE et al., 1985). Através da adição deste termo à função de energia avaliadora dos modelos estruturais, espera-se que as estruturas que possuam um nível maior de compactação sejam favorecidas durante o processo de otimização, culminando no encaminhamento do processo à estruturas que possuam um maior grau de empacotamento.

O segundo termo adicional (Eq. 2.4), refere-se à preservação de estrutura secun-

Tabela 2.2: Termos da Função de Energia *full atom* do Rosetta

Descrição	Referência
Atração de átomos de aminoácidos distintos	Jones and Chapman (1924), Jones (1924)
Repulsão de átomos de aminoácidos distintos	Jones and Chapman (1924), Jones (1924)
Repulsão de átomos do mesmo aminoácido	Jones and Chapman (1924), Jones (1924)
Solvatação de aminoácidos diferentes	Lazaridis and Karplus (1999)
Solvatação de átomos polares	Park et al. (2016), Yanover and Bradley (2011)
Solvatação mesmo aminoácido	Lazaridis and Karplus (1999)
Interação entre átomos carregados não-ligados	Park et al. (2016)
Ligação de hidrogênio de curta distância	Kortemme, Morozov and Baker (2003), O'Meara et al. (2015)
Ligação de hidrogênio de longa distância	Kortemme, Morozov and Baker (2003), O'Meara et al. (2015)
Ligação de hidrogênio cadeia lateral-principal	Kortemme, Morozov and Baker (2003), O'Meara et al. (2015)
Ligação de hidrogênio entre cadeias laterias	Kortemme, Morozov and Baker (2003), O'Meara et al. (2015)
Energia de pontes dissulfeto	O'Meara et al. (2015)
Probabilidade de ângulos ϕ/ψ de cada aminoácido	Park et al. (2016), Leaver-Fay et al. (2013)
Probabilidade de um aminoácido dado ϕ/ψ	Leaver-Fay et al. (2013)
Probabilidade de rotâmero dados ϕ/ψ	Shapovalov and Jr (2011)
Penalidades por desvio de ângulo ω	Berkholz et al. (2012)
Penalidade abertura de um anel e ligação de Prolina	Leaver-Fay et al. (2013)
Penalidade por ângulo χ_3 não planar de Tirosinas	O'Meara et al. (2015)
Energia de referência por tipo de aminoácidos	Kuhlman and Baker (2000), Leaver-Fay et al. (2013)

Lista dos 19 termos de energia utilizados para o cálculo de energia livre do PyRosetta. Adaptado de Alford et al. (2017).

dária (ES) dos modelos. Desta maneira, uma constante negativa é adicionada ao valor de energia ($-const$), a fim de reforçar positivamente cada aminoácido pertencente à proteína P, todas as vezes em que a ES (es_n^b) descrita pra o aminoácido n (aa_n) de um modelo seja equivalente à ES (es_n^a) do mesmo resíduo fornecida como entrada ao algoritmo. Da mesma forma, se ambas ES forem divergentes, uma constante positiva ($+const$) é adicionada ao valor de energia, de maneira a penalizar ES diferentes entre os mesmos resíduos. A cada avaliação a ES de cada aminoácido de um modelo é descrita utilizando a implementação do algoritmo Definindo a Estrutura Secundária de Proteínas (DSSP, sigla do inglês *Define Secondary Structure of Proteins*) (KABSCH; SANDER, 1983), também implementado pelo PyRosetta.

$$Energy_{ES} = \sum_{aa \in P}^{n+1} \Delta_{aa,es^a,es^b}(aa_n, es_n^a, es_n^b) \tag{2.4}$$

onde,

$$\Delta_{aa,es^a,es^b}(aa_n, es_n^a, es_n^b) = \begin{cases} -const, & es_n^a = es_n^b \\ +const, & es_n^a \neq es_n^b \end{cases} \tag{2.5}$$

Os modelos estruturais foram avaliados a partir da função de energia final (Eq. 2.6), a qual resulta na soma de ambos os termos adicionais descritos anteriormente à função de energia *all-atom* implementada pelo PyRosetta. Este modelo de avaliação, composto pela adição de termos à função de energia, foi utilizada previamente por Correa et al. (2016) e Borguesan et al. (2018). Ao mencionarmos os valores energéticos dos modelos estruturais durante o decorrer do trabalho, estes referem-se ao valor calculado pela equação de energia composta.

$$FinalEnergy = Energy_{PyRosetta} + Energy_{SASA} + Energy_{ES} \tag{2.6}$$

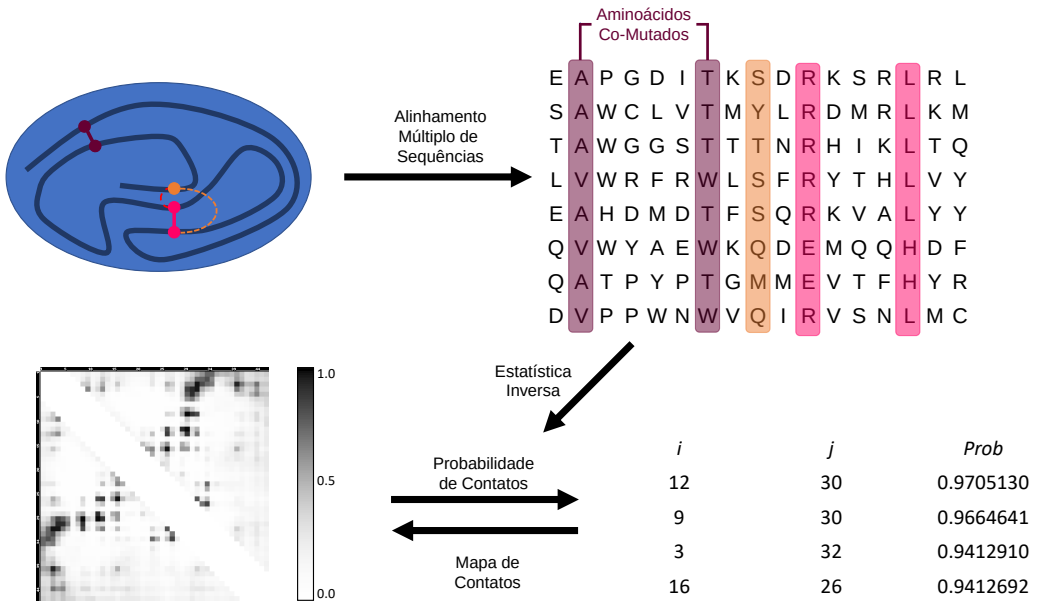
2.5 Contato 3D entre aminoácidos

O processo evolutivo é dirigido por mutações aleatórias nas sequências genômicas de forma que proteínas com alta taxa de conservação de estrutura e função possuem uma alta variabilidade na sequência de aminoácidos (ILLERGÅRD; ARDELL; ELOFSSON, 2009). Contudo, ao analisar o conteúdo de dados depositados em bancos de dados genômicos, é possível identificar um certo padrão não-aleatório nas mutações conservadas durante o processo evolutivo. Assim, é possível que haja algum tipo de coordenação entre diferentes mutações pontuais de maneira que ocorra a conservação de estrutura, logo, a preservação de função que permita a sobrevivência de um determinado organismo (SIKOSEK; CHAN, 2014). O avanço nas tecnologias de sequenciamento em larga escala com baixos custos disponibiliza uma enorme quantidade de dados genômicos para análises evolutivas (MUKHERJEE et al., 2018).

Aminoácidos localizados em regiões *específicas*, tais como sítio ativos ou outras posições críticas para a conservação de estrutura, tendem a possuir um grau maior de conservação, uma vez que uma mutação nestas posições poderia levar à perda completa de estrutura e função. De maneira similar, aminoácidos pertencentes a posições não críticas, possuem uma maior variabilidade, contudo, ainda apresentam efeitos desestabilizantes. Por exemplo, por perda de interação com aminoácidos vizinhos no espaço

3D (Figura 2.5). Entretanto, mutações em conjunto entre dois aminoácidos podem ser mutualmente compensatórias, mantendo ou até mesmo melhorando a estabilidade da proteína (ZERIHUN; SCHUG, 2017). Estas mutações em conjunto provavelmente indicam um contato direto ou indireto entre os aminoácidos em questão. Classificam-se como contatos indiretos, os aminoácidos aa_1 e aa_2 próximos um do outro devido a um terceiro aminoácido aa_3. Por exemplo, como mostra a Figura 2.5, onde o aa_1, representado em cor alaranjada, encontra-se em proximidade 3D do aminoácido aa_2 (linha tracejada laranja), pela proximidade de aa_1 com aa_3 (linha tracejada vermelha). Em contraste, contatos diretos resultam da aproximação espacial entre os aminoácido aa_1 e aa_2 sem nenhuma influência de aminoácidos terceiros (JUAN; PAZOS; VALENCIA, 2013). Exemplificados na Figura 2.5 pela ligação direta (linhas contínuas magenta e roxa) entre os aminoácidos aa_1 e aa_2.

Figura 2.5: Esquema de Aminoácidos Preditos em Contato



Esquema geral da obtenção de informação evolutiva de acoplamento de aminoácidos. Primeiramente, a sequência de uma proteína-alvo (canto superior esquerdo) é comparada com um conjunto de proteínas a fim de identificar a família de tal proteína. Análises de estatística inversa são aplicadas ao resultado do alinhamento múltiplo de sequências (canto superior direito), para que então seja possível calcular a probabilidade de dois aminoácidos estarem em contato 3D (canto inferior direito). Uma vez calculadas as probabilidades de contato, é possível gerar os chamados mapas de contato (canto inferior esquerdo).

Um dos grandes desafios quando se trata da inferência de correlação entre aminoácidos é a diferenciação correta de contatos diretos e indiretos, desta forma em 2009 o método *Direct coupling analysis* (DCA) foi proposto e experimentalmente validado como

uma maneira eficaz para predição de contatos (WEIGT et al., 2009; CASINO; RUBIO; MARINA, 2009). Para tais fins, os autores propõem a utilização de um método diferente da Informação Mútua (IM), pois esta considera apenas um par de aminoácidos por vez, dificultando a diferenciação de contatos diretos ou indiretos.

A partir do resultado de um alinhamento múltiplo de sequências (MSA), é possível que o modelo estatístico assuma que as sequências em um MSA representam uma amostra de uma distribuição de Boltzmann (P), onde para cada sequência $\sigma = (aa_1, aa_2, ...aa_L)$ de tamanho L há uma função de energia $H(\sigma)$ (COCCO et al., 2018). A probabilidade de cada sequência dentro da distribuição é dada pela equação

$$P(\sigma) = \frac{1}{Z} exp\{-\beta H(\sigma)\} \tag{2.7}$$

onde Z refere-se à uma constante de normalização e β ao inverso da temperatura (β = 1 sem perda de generalização). Para cada posição aa_n de σ, são possíveis 21 variáveis, correspondentes aos 20 aminoácidos e *gap* (decorrentes de inserções ou deleções), resultando em 21^L possíveis conformações de σ. Este número gigantesco de possibilidades torna inviável o cálculo total de combinações, dado o número de sequências disponíveis em um dado MSA e a falta de de uma relação explícita da influência destes contatos na função de energia (ZERIHUN; SCHUG, 2017).

Weigt et al. (2009) propôs uma forma de expressar a energia $H(\sigma)$ de uma forma mais simples, porém considerando o acoplamento direto entre dois aminoácidos de uma dada sequência. Nos cálculos utilizados pelo DCA, a energia é expressa como

$$H(\sigma) = -\sum_{i=1}^{L-1} \sum_{j=i+1}^{L} - \sum_{i=1}^{L} h_i(aa_i) \tag{2.8}$$

onde são considerados ambos os acoplamentos de i, j e campos locais $h_i(aa_1)$. Apesar de ainda ser um cálculo custoso, dado a quantidade de dados disponíveis, o número de parâmetros é reduzido de 21^L para $\frac{1}{2}L(L-1)q^2$.

2.6 Resumo do capítulo

Neste capítulo foram abordados os principais conceitos teóricos necessários para o entendimento da metodologia de geração de modelos estruturais proposta por este trabalho, assim como os conceitos principais para o entendimento do problema de PSP.

Os conceitos apresentados referem-se a: (*i*) blocos construtores de proteínas, ou seja, composição química dos aminoácidos; (*ii*) níveis hierárquicos de organização estrutural de proteínas; (*iii*) classes estruturais de proteínas, dependendo da sua topologia, assim como a classificação utilizada para as proteínas pertencentes ao grupo de teste neste trabalho; (*iv*) representação computacional de modelos estruturais de proteínas, portanto definiu-se a representação adotada para realização do atual trabalho, utilizando os ângulos de torção; (*v*) panorama de funções de energia utilizadas para avaliar estruturas de macromoléculas biológicas, assim como a descrição dos termos utilizados pela função de avaliação utilizada; (*vi*) introdução ao conceito de acoplamento entre pares de aminoácidos inferido a partir de análises de alinhamentos múltiplos de sequências.

O próximo capítulo será responsável por contextualizar e inserir o trabalho proposto no cenário de resolução do problema de PSP. O capítulo faz uma revisão bibliográfica do estado da arte de PSP e também da utilização de meta-heurísticas aplicadas à resolução do problema de PSP.

3 REVISÃO BIBLIOGRÁFICA

Neste capítulo serão apresentadas metodologias encontradas na literatura, a fim de montar um panorama do atual estado-da-arte das metodologias propostas para resolução do problema de PSP, assim como serão descritas as duas metodologias vencedoras do principal consórcio de avaliação deste métodos. Juntamente, será feito um levantamento da utilização de contatos evolutivos entre aminoácidos, tanto no que diz respeitos aos métodos, quanto a sua utilização dentro do escopo do problema de PSP. Por último, será abordado a utilização de algoritmos de busca, as meta-heurísticas, também aplicadas ao problema de PSP.

3.1 Problema de PSP

A cada dois anos o Centro de Predição de Estruturas de Proteínas [1] (*Protein Structure Prediction Center*) é responsável pela organização e realização do *Critical Assessment of Structure Prediction*. Desde sua primeira edição em 1994, este evento tem por objetivo monitorar e avaliar o estado da arte no que diz respeito ao problema de PSP (MOULT et al., 1995). A competição procede da seguinte maneira: em um primeiro momento, cientistas experimentais são convidados a disponibilizar a sequência de aminoácidos de proteínas, as quais ainda não possuíram sua estrutura 3D depositadas em bancos de dados, mas já foram resolvidas experimentalmente. Em seguida, estas sequências são disponibilizadas à comunidade científica de modelagem para que possam testar seus métodos propostos para a elucidação do problema de PSP (MOULT et al., 2018). Os níveis de dificuldade de cada proteína-alvo a ser elucidada são medidos através da semelhança de sequência e estrutura com o molde mais próximo disponível em bancos de dados (KRYSHTAFOVYCH; FIDELIS; MOULT, 2014).

Diferentes categorias são utilizadas para avaliar os métodos de modelagem participantes de cada edição do CASP, tais como formação de estrutura terciária, predição de contato entre aminoácidos, e também formação do complexo da estrutura quaternária (MOULT et al., 2018). Em termos de predição da estrutura terciária, cada proteína-alvo é classificada de acordo com as métricas apresentadas em Abriata et al. (2018a). Esta classificação é realizada a partir de certo valor de semelhança de sequência e/ou estrutura entre a proteína-alvo e proteínas depositadas em bancos de dados. As proteínas

[1] <http://predictioncenter.org/index.cgi>

alvo podem ser calculadas em: *template-based modeling* (TBM), os quais são considerados quando há maior semelhança, *free modeling* (FM), quando as proteínas-alvo não possuem semelhança com proteínas já descritas; em adição à estas duas categorias, uma terceira categoria *template-based modeling/free modeling* (TBM/FM) é considerada, para proteínas que habitam a zona de transição entre ambas categorias descritas anteriormente.

Durante a realização do CASP12 (MOULT et al., 2018), os métodos de predição da categoria FM obtiveram um progresso significativo quando comparados com os resultados obtidos durante a edição anterior (CASP11) (MOULT et al., 2016). O aumento de eficácia de predição obtido por estes métodos é atribuído à incorporação de acoplamento de aminoácido durante o processo de predição. Isto ocorre possivelmente pelo aumento de cobertura nos alinhamentos múltiplos de sequência, o qual ajuda na predição de contatos e também na identificação de proteínas homólogas distantes (ABRIATA et al., 2018b). Os resultados alcançados na categoria de predição de contato entre aminoácidos no espaço 3D foram os de maior destaque durante o CASP12, pois esta categoria foi responsável pelos maiores avanços em comparação as edições anteriores, as quais alcançaram quase o dobro de precisão quando comparados com o melhor colocado da edição CASP11 (SCHAARSCHMIDT et al., 2018). Contudo, apesar do grande progresso apresentado para modelos FM, os métodos de predição TBM foram os que alcançaram a maior precisão. Esta alta acurácia é alcançada em decorrência de uma maior eficiência de métodos que combinam diferentes moldes, métodos de refinamento de estrutura mais eficazes, melhora em metodologias de reconstrução *ab initio* para regiões em que não são encontrados moldes compatíveis, assim como na melhora dos métodos de avaliação de precisão de estimação de modelos (KRYSHTAFOVYCH et al., 2018).

O método utilizado nesta dissertação para otimização estrutural de proteínas enquadra-se na categoria de FM, portanto serão descritas a seguir as metodologias empregadas pelos dois grupos vencedores da categoria FM do CASP12, sendo eles os métodos Rosetta e QUARK. Estes métodos serão utilizados como referência para fins de comparação [2] com os resultados obtidos após a otimização dos modelos gerados a partir do método de geração de estruturas utilizando limitantes de relevância biológica proposto.

[2]Destaca-se que os artigos contendo os resultados da 13^a edição do CASP (CASP13) ainda não foram disponibilizados até o momento de execução do atual trabalho, portanto tomamos como parâmetro os resultados da edição CASP12.

3.1.1 Rosetta

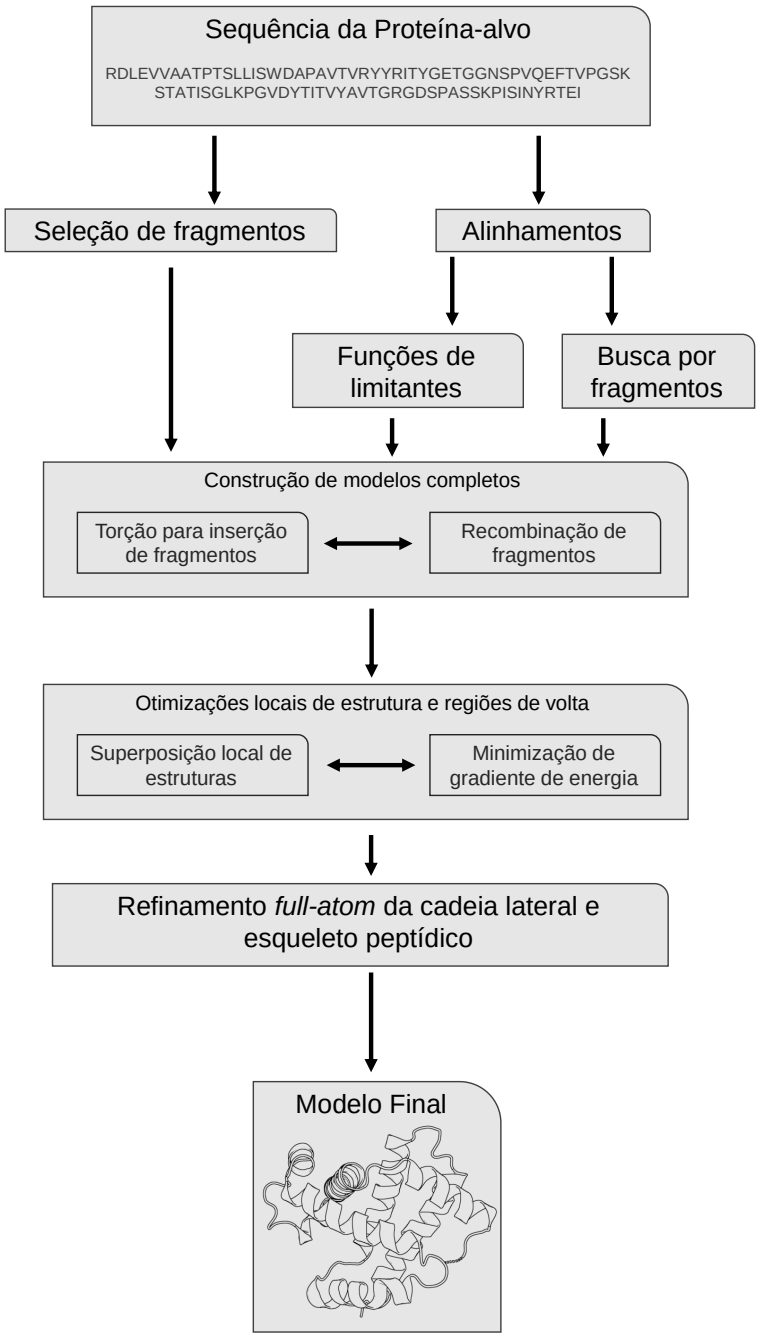
Os protocolos de predição de estrutura de proteínas disponíveis pelo consórcio Rosetta (ROHL et al., 2004), tanto para predições *ab initio* (RAMAN et al., 2009), quanto para predições de modelagem comparativa (SONG et al., 2013), estão disponíveis no servidor online Robetta [3] (KIM; CHIVIAN; BAKER, 2004). O protocolo de predição baseia-se no uso combinado de proteínas homólogas com estrutura previamente conhecida, juntamente com uma biblioteca de fragmentos de estruturas retirados do PDB (SIMONS et al., 1997).

Primeiramente, para cada proteína é construída uma biblioteca de 3 e 9 aminoácidos, gerados a partir de estruturas previamente depositadas no PDB. Esta janela é sobreposta para cada aminoácido e seus vizinhos. Então, são selecionados os 200 melhores fragmentos de acordo com uma função avaliadora de fragmentos (GRONT et al., 2011). Os modelos iniciais gerados por este método são construídos a partir da combinação de fragmentos provenientes das bibliotecas geradas. O método de otimização é composto por três passos principais, descritos em Song et al. (2013): (i) geração de uma grande quantidade de modelos a partir da junção de fragmentos e processos de otimização, partindo de uma precisão mais baixa, utilizando a função de avaliação centróide (*centroid*), progredindo gradativamente ao longo da execução até atingir a precisão atômica (*full atom*); (ii) Os modelos gerados pelo estágio inicial são otimizados a partir da substituição de seus ângulos de torção da cadeia principal pelos ângulos presentes na biblioteca de fragmentos, a fim de melhorar as regiões de união entre fragmentos distintos; por fim (iii) as estruturas tem suas cadeias laterais adicionadas e ajustadas ao modelo gerado, de maneira que a cada iteração as cadeias laterais tendam a se ajustar reduzindo a força de repulsão entre átomos. Nota-se que a cada passo apenas os modelos com menores valores energéticos são encaminhados para o passo seguinte. A otimização é efetuada a partir da implementação de uma simulação de Monte Carlo, onde milhares de estruturas são avaliadas dentro do espaço conformacional e otimizadas através da troca de parâmetros entre fragmentos estruturais e simulações individuais (Figura 3.1).

O método, apesar de estar entre os vencedores do CASP12, possui desafios a serem superados, tais como conformações locais não contempladas pela utilização de pequenos fragmentos de sequência similar, ou topologias complexas contendo muitas interações não-locais (OVCHINNIKOV et al., 2018). Assim como, a predição da estrutura secun-

[3] <http://new.robetta.org/>

Figura 3.1: Fluxograma de metodologia utilizado pelo Rosetta



Esquema geral da metodologia implementada pelo preditor de estrutura de proteínas Rosetta. Adaptado de Song et al. (2013).

dária feita por preditores, tais como PSIPRED (MCGUFFIN; BRYSON; JONES, 2000), possuem dificuldades de predizer a estrutura secundária de pequenos fragmentos que necessitam interações entre fragmentos distintos da cadeia peptídica. Contudo, os resultados obtidos por Ovchinnikov et al. (2018) e Ovchinnikov et al. (2017) demonstram a capaci-

dade de auxilio de dados de co-evolução, em conjunto com algoritmos evolutivos, para contornar tais dificuldades.

3.1.2 QUARK

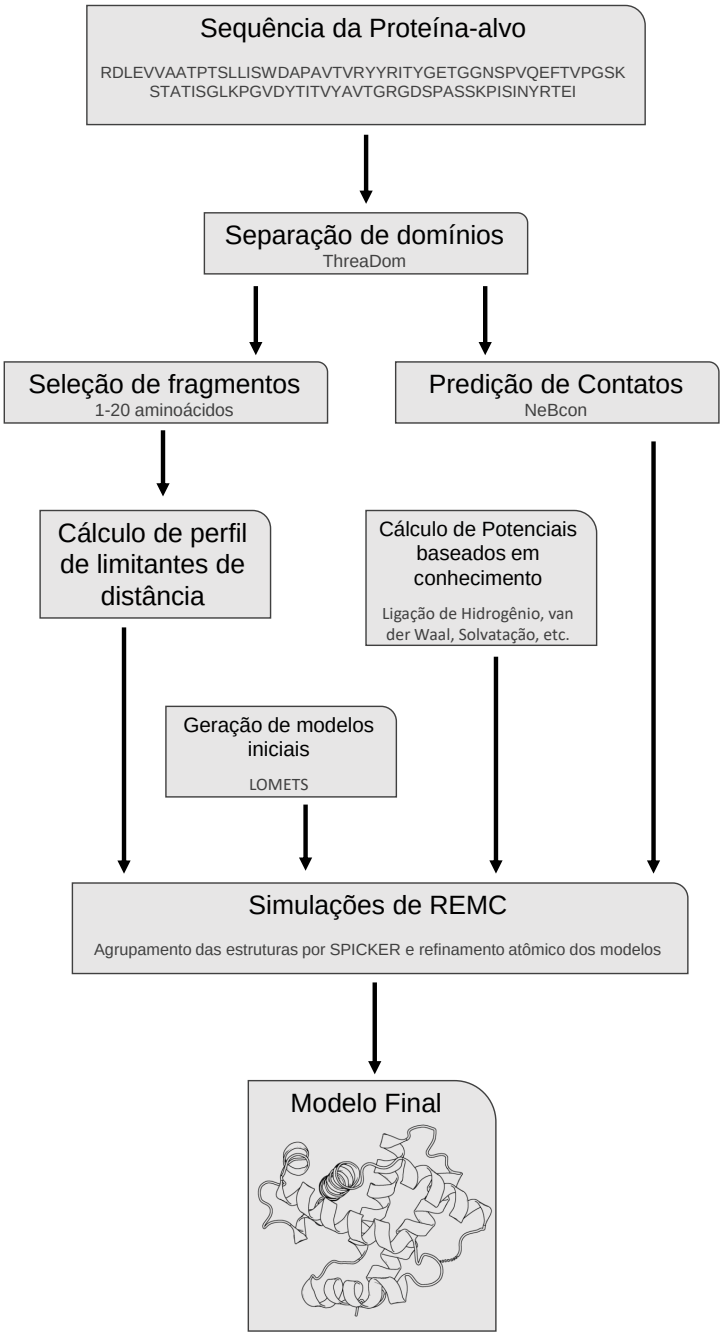
O servidor QUARK [4] (XU; ZHANG, 2012) utiliza um algoritmo de primeiros princípios para a predição de enovelamento de estrutura de proteínas. Este método tem por objetivo a construção de modelos 3D de proteínas a partir de sua sequência de aminoácidos, utilizando pequenos fragmentos de aminoácidos (entre 1 e 20 aminoácidos por fragmento) através de simulações de *Replica-Exchange Monte Carlo* utilizando um campo de forças baseado em conhecimento de resolução atômica.

Primeiramente, a sequência de aminoácidos da proteína de interesse é dividida de acordo com os domínios individuais preditos pela ferramenta ThreaDom (XUE et al., 2013). Em seguida, esses domínios são comparados aos fragmentos gerados a partir de estruturas depositadas no PDB a fim de gerar fragmentos estruturais contínuos, entre 1 e 20 aminoácidos, para cada posição específica na sequência de cada domínio. Os modelos gerados são avaliados utilizando uma função de energia a fim de comparar diferentes parâmetros experimentais, tais como, área acessível ao solvente, estrutura secundária e comparação perfil-perfil entre o modelos gerados e a proteína de referência (XU; ZHANG, 2013). Após esta avaliação, os 200 fragmentos melhor classificados são selecionados para montar um perfil de distância, considerando até 9 Å entre os aminoácidos. Concomitantemente, são preditos os contatos para a sequência da proteínas utilizando o preditor NeBcon (HE et al., 2017). Informações sobre a distância entre aminoácidos pertencentes aos modelos são utilizados para a formação de modelos estruturais a partir de fragmentos, juntamente com restrições de perfis e temos físicos potenciais, através de simulações de *replica-exchange Monte Carlo* (REMC). Então, os candidatos gerados são agrupados de acordo com suas semelhanças estruturais pelo SPICKER (ZHANG; SKOLNICK, 2004), para que o modelo central de cada grupo seja selecionado. Estes centros de cada grupamento são tratados como estruturas de estados mínimos de energia. Os modelos centrais dos cinco maiores grupos são selecionados para uma etapa final de refinamento utilizando ModRefiner (XU; ZHANG, 2011) ou dinâmica molecular guiada por fragemntos (FG-MD, sigla do inglês *Fragment-Guided Molecular Dynamics*) (ZHANG; LIANG; ZHANG, 2011). Em caso de proteínas com mais de um domínio, estes domínios são uni-

[4] <https://zhanglab.ccmb.med.umich.edu/QUARK/>

ficados através de simulações de corpo rígidos utilizando o algoritmo *Metropolis Monte Carlo* (ZHANG et al., 2018) (Figura 3.2).

Figura 3.2: Fluxograma de metodologia utilizado pelo QUARK



Esquema geral da metodologia implementada pelo preditor de estrutura de proteínas QUARK. Adaptado de Ovchinnikov et al. (2018).

Zhang et al. (2018) apontam que o método é extremamente dependente da precisão de predição de contato, o que apesar de limitante, pode também justificar os resultados

obtidos devido ao aumento da acurácia de predição de contatos (SCHAARSCHMIDT et al., 2018). Também são apontadas as dificuldades do método para predizer estrutura de proteínas multi-domínios, começando pela falta de modelos estruturais que possuem um alinhamento satisfatório. Assim como a dificuldade de montagem da estrutura multi-domínio final e também estruturas com uma taxa de β-Folhas, as quais podem possuir contatos não detectáveis pelo preditor NeBcon (HE et al., 2017).

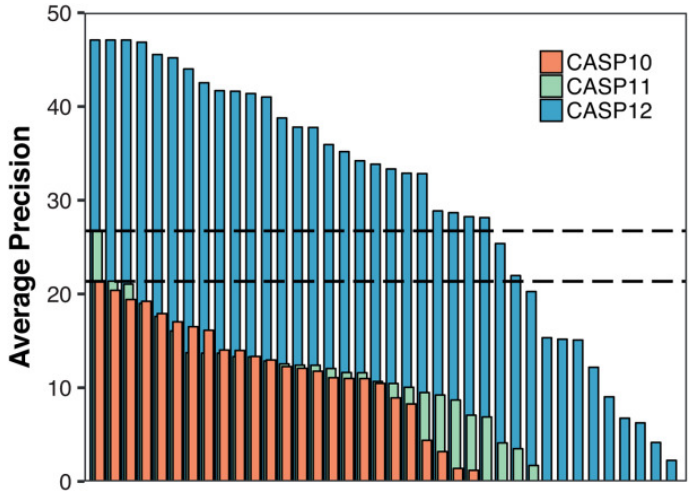
3.2 Metodos para determinação de contatos entre aminoácidos

Proposto por Weigt et al. (2009), o método DCA foi o marco inicial no aumento da acurácia de predição de contatos 3D entre pares de aminoácidos, sejam eles pertencentes à mesma cadeia peptídica ou cadeias distintas. Desde sua elaboração, uma série de novas abordagens foram propostas a fim de aumentar a acurácia de predição de maneira otimizada, considerando-se a complexidade de cálculos para a inferência do parâmetro de normalização Z (Equação 2.7). Desta forma, devido ao alto custo computacional $O(q^L)$, algoritmos de inferência inversa baseiam-se em aproximações (ZERIHUN; SCHUG, 2017). O algoritmo implementado para o DCA é chamado de *message-passing*, o qual procura pela melhor combinação de parâmetros através de processos iterativos a partir de combinações iniciais até que um critério de parada, ou convergência desejada, seja alcançado (MÉZARD; MORA, 2009). Contudo, esta abordagem é limitada ao que diz respeito ao tamanho da sequência e também ao número de sequências presentes no resultado do MSA. Portanto, diversos métodos foram propostos durante a última década, incluindo aprendizado de máquina de Boltzmann (ACKLEY; HINTON; SEJNOWSKI, 1985), aproximação Gaussiana (BALDASSI et al., 2014), ou redes Bayesianas (BURGER; NIMWEGEN, 2010).

A predição de contato entre aminoácidos é uma categoria do CASP desde sua segunda edição em 1996 (LESK, 1997). Em decorrência das notáveis melhoras proporcionadas pelo acréscimo de informações evolutivas de contato entre pares de aminoácidos nos métodos de predição *de novo* (BOHR et al., 1993; SKOLNICK; KOLINSKI; ORTIZ, 1997), desde sua décima edição o CASP possui uma categoria de avaliação para métodos que incorporam esta informação para predição de estrutura de proteínas (TAYLOR et al., 2014). Em comparação, os métodos de predição de contato propostos durante a realização do CASP12 praticamente dobraram o grau de precisão em relação ao melhor colocado da edição CASP11 (SCHAARSCHMIDT et al., 2018) (Figura 3.3). A adição de informações

evolutivas e a maior acurácia na diferenciação de contatos diretos ou indiretos é atribuída como um dos fatores que impulsionaram este aumento de precisão.

Figura 3.3: Precisão de Predição de Contato no CASP12



Precisão média de predição de contatos separados por mais de 23 aminoácidos em uma mesma cadeia peptídica. As linhas pontilhadas indicam os valores médios alcançados pelos preditores mais bem colocados nas edições CASP10 e CASP11, respectivamente. Enquanto que o eixo x estão dispostos em ordem de ranqueamento os métodos de predição de contatos avaliados em cada edição. Retirado de Schaarschmidt et al. (2018).

Os preditores de estrutura de proteínas classificados em primeiro lugar no CASP12 também possuem informações evolutivas acerca de aminoácidos em contato 3D em suas metodologias. No método adotado pelo Rosetta os modelos são gerados levando em consideração a aproximação dos contatos preditos para os aminoácidos separados por mais de cinco aminoácidos na sequência linear do peptídio (OVCHINNIKOV et al., 2017). Integrados ao método utilizado por QUARK, são medidas as distâncias entre todos os aminoácidos dos modelos gerados e comparados aos contatos preditos para a proteína-alvo, sendo esta informação utilizada como critério de avaliação durante o processo de otimização (ZHANG et al., 2018).

3.3 Meta-heurísticas

Diferentes tipos de problemas das mais diversas áreas possuem diferentes complexidades de resolução, portanto sua resolução poderá ser alcançadas através de métodos determinísticos ou aproximados. Apesar de resultados mais precisos, os métodos determinísticos são incapazes de atingir soluções ótimas para problemas NP-difíceis (COOK,

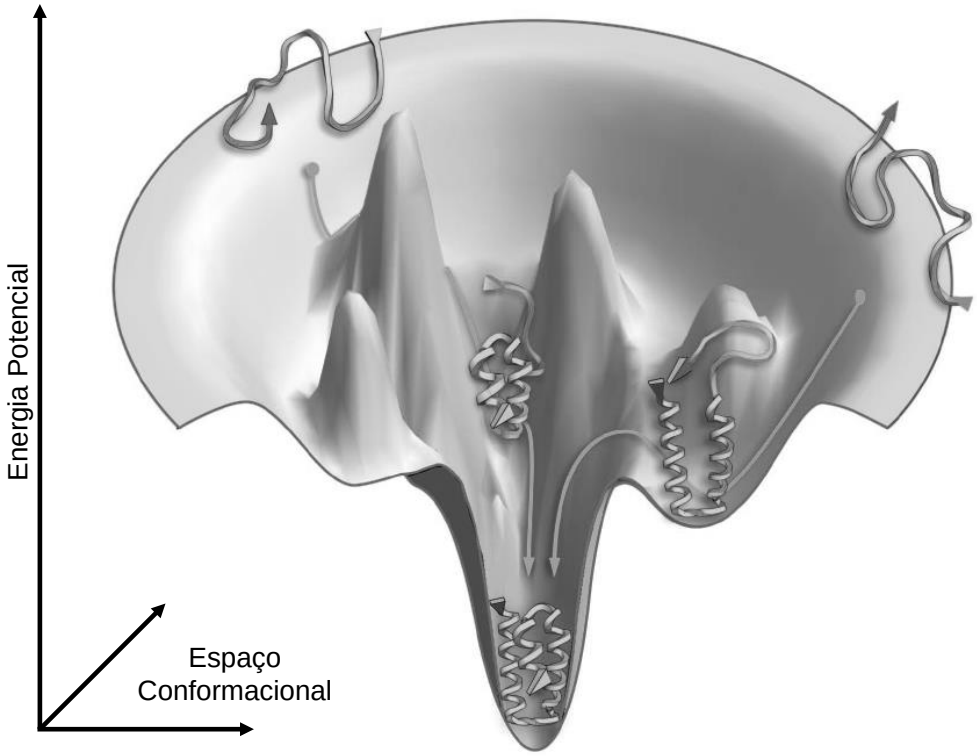
1983), pois apresentam tempo de execução não-polinomial. Em um processo de troca, métodos aproximados são capazes de chegar em soluções próximas da solução ideal em tempo de execução aceitável, porém não garantem a obtenção do resultado ótimo (TALBI, 2009a).

Meta-heurísticas são algoritmos desenvolvidos para resolução de problemas de otimização complexos. Estes métodos são formalmente definidos como uma heurística subordinada à um processo iterativo, o qual unifica diversos conceitos para uma busca eficiente dentro de um espaço de busca (OSMAN; LAPORTE, 1996). Segundo Boussaïd, Lepagnot and Siarry (2013), praticamente todos os métodos de meta-heurísticas compartilham as seguintes características: (*i*) são inspiradas por conceitos encontrados na natureza, tais como processos biológicos de mutação e *crossover*; (*ii*) utilizam processos estocásticos durante sua execução, por exemplo, variáveis com certos níveis de aleatoriedade; (*iii*) possuem uma série de parâmetros que devem ser incorporados de acordo com a finalidade com a qual tal método por implementado; (*iv*) não dependem de matriz hessiana ou gradiente da função de avaliação. Nota-se que meta-heurísticas são ideais para resolução de uma diversidade de problemas de otimização complexos, portanto permitem a incorporação de informações problema-específicas conforme a necessidade de sua aplicação. Uma meta-heurística é bem sucedida para a resolução de um problema, se esta proporciona um balanço ideal entre *exploration* e *exploitation*. A primeira diz respeito a diversidade do método, ou seja, é responsável pela identificação de regiões promissoras dentro do espaço de busca, enquanto que a segunda refere-se ao refinamento ou intensificação destas regiões promissoras (BIRATTARI et al., 2001).

Dentre as diferentes abordagens referentes à meta-heurísticas, podemos classificá-las em dois grupos: solução única (também conhecidos como métodos de trajetória) e baseadas em populações. Métodos de solução única, geralmente começam a execução com apenas uma solução e a partir dela são feitas as operações de mudanças a fim de otimizar a solução inicial, deste forma retornam apenas uma solução final para o problema (BOUSSAÏD; LEPAGNOT; SIARRY, 2013). Por outro lado, métodos baseados em população focam seus esforços na otimização de um problema mais complexo de forma a percorrer melhor o espaço de busca, retornando não apenas uma solução final, porém um conjunto de soluções ótimas. Desta maneira os algoritmos possuem maior liberdade em explorar diferentes pontos do espaço de busca, de maneira que pontos que possuam um menor desempenho não influenciem no desempenho global do algoritmo (DAS et al., 2011). Os inúmeros problemas, pelos quais tenta-se encontrar soluções ideias através da utiliza-

ção de meta-heurísticas, possuem funções objetivo complexas para tal fim (GLIBOVETS; GULAYEVA, 2013). Portanto, processos de otimização multimodais são de extrema complexidade, uma vez que sejam necessários a identificação de diferentes soluções-ótimas e não apenas uma única solução (DAS et al., 2011). Ao que se refere ao problema de PSP, sabe-se que o processo de enovelamento de uma proteína é caracterizado por um processo composto por diferentes estados conformacionais (Figura 3.4) (DILL; MACCALLUM, 2012). Assim como a característica não estática da estrutura nativa de uma dada proteína (ANFINSEN, 1973). Desta maneira, caracterizando as funções de energia utilizadas para avaliação estrutural de proteínas como uma função multimodal.

Figura 3.4: Espaço Conformacional Energético



Panorama do espaço de busca conformacional multimodal de estrutura de proteínas em função da energia potencial. Destaca-se que diferentes conformações 3D correspondem ao mesmo valor energético, assim como mínimos locais energéticos. Adaptado de Dill and MacCallum (2012).

Para tentar melhorar o processo de busca por soluções ideais, o emprego de métodos baseados em populações, tais como Algoritmos evolutivos (AE) (BACK, 1996), são de extrema importância. Em comparação, caso empregados para a resolução de um problema multimodal, os algoritmos baseados em uma única solução necessitam ser exe-

cutados uma série de vezes, afim de explorar o espaço de busca de maneira semelhante aos métodos populacionais. Contudo, as dificuldades destes métodos em manter a diversidade durante a execução é um importante fator a ser considerado, uma vez que estes podem convergir a população, através de uma deriva genética, à um único ótimo global (BELDA et al., 2007). Em decorrência do que concerne ao problema de PSP, é de extrema importância a geração de diferentes soluções-ideias, dada a multimodalidade do problema. Este conjunto de soluções pode ser submetido a etapas de avaliação posteriores para um possível refinamento das estruturas geradas. Isto é reforçado ao analisar os métodos vencedores do CASP12, Rosetta e QUARK, os quais possuem como resultado um conjunto de soluções-ótimas encontrados.

3.3.1 Meta-heurísticas e o problema de PSP

O problema de PSP representa um problema de alta dimensionalidade, em decorrência do grande número de variáveis embutidas ao problema (UNGER; MOULT, 1993), assim como é classificado como um problema NP-Completo de acordo com a teoria de complexidade computacional (COOK, 1983; CRESCENZI et al., 1998). Desta forma, o emprego de meta-heurísticas para tentar resolver este problema vem sendo amplamente empregados ao longo dos últimos anos (DORN et al., 2014). Geralmente, estes métodos alteram a conformação 3D de modelos estruturais, através das mais variadas operações de acordo com a meta-heurística empregada. Estas operações são responsáveis por alterar a orientação de átomos, através de operações matemáticas, a fim de minimizar a função de avaliação implementada com o objetivo de encontrar o indivíduo mais apto, ou seja, mínimos energéticos correspondentes aos mínimos globais próximos à solução ideal (DESJARLAIS; CLARKE, 1998).

Em Elofsson, Grand and Eisenberg (1995) utilizou-se algoritmo genético em conjunto com simulações de Monte Carlo para predição de enovelamento de proteínas através de alterações locais no ângulos de diedro da cadeia principal. Dorn et al. (2013), com o objetivo de diminuir o espaço conformacional de busca, utiliza informações previamente determinadas experimentalmente para aminoácidos semelhantes aos que compõem a proteína-alvo em um algoritmo genético baseado em conhecimento. Borguesan et al. (2015) utiliza a informação depositada no banco de dados do PDB no que diz respeito aos ângulos de rotação de aminoácidos com sua conformação 3D determinada experimentalmente. Neste trabalho, as Listas de Probabilidade de Ângulos (APL, singla em inglês

Angle Probability List), é utilizada como informação para geração de modelos estruturais, os quais foram submetidos à processos de otimização utilizando um algoritmo genético e também, para fins de comparação, um algoritmo de otimização por enxame de partículas (SHI et al., 2001). De forma similar, Correa et al. (2016) propõe um modelo ampliado da APL para geração de modelos 3D, para que sejam otimizados através da implementação de um algoritmo memético. Narloch and Dorn (2019a) demonstram a eficácia da utilização de APLs para a geração de indíviduos iniciais, otimizados posteriormente por um algoritmo de evolução diferencial (DE, sigla do inglês *Differential Evolution*). Em Narloch and Dorn (2019b), os autores utilizam uma implementação auto-adaptativa de um DE baseado em conhecimento, também utilizando informações provenientes das APLs, onde os parâmetros utilizados pelos algoritmo são modificados ao longo da execução, assim como os processos utilizados pelos mecanismos de mutação ao longo da otimização.

Comumente, problemas com um alto grau de complexidade, tais como o problema de PSP, possuem funções objetivo compostas por diversos termos. Estes, apesar de agrupados de maneira a calcular um valor de *fitness* para cada solução encontrada, muitas vezes possuem caráter conflitantes entre si, de maneira que a otimização simultânea entre todos os termos seja impossibilitada de ocorrer de maneira adequada (KONAK; COIT; SMITH, 2006). Portanto, a utilização de métodos multiobjetivos para otimização de soluções em caso de problemas de alta complexidade permite uma melhor aproximação de soluções reais. Desta forma, também é possível que sejam adicionadas um maior número de informações relevantes ao escopo do problema abordado, guiando o processo de otimização de maneira mais objetiva (ZHOU et al., 2011). Devido ao seu alto grau de complexidade, o problema de PSP é um ótimo candidato ao uso de abordagens multiobjetivas. Dois termos comumente utilizados em funções de energia aplicadas à avaliação de estrutura 3D de proteínas são referentes aos termos ligados e não-ligados, ambos conflitantes entre si ao que diz respeito de interações moleculares. Ambos os termos foram tratados como objetivos distintos durante o processo de otimização utilizando um algoritmo evolutivo proposto por Cutello, Narzisi and Nicosia (2005). Seguindo esta linha, Brasil, Delbem and Silva (2013) utilizam quatro termos para uma otimização multiobjetiva em uma abordagem *ab initio* para a resolução do problema de PSP, sendo os termos: van der Waals, interações eletrostáticas, termo referente à solvatação e ligações de hidrogênio.

Destaca-se que o problema de PSP, além de alta complexidade, apresenta um caráter multimodal, ao que diz respeito de valores energéticos semelhantes para soluções estruturais distintas (Figura 3.4). Apesar dos avanços já obtidos para resolução do pro-

blema de PSP, ainda há a necessidade de contornar uma série de fatores que aumentam consideravelmente a complexidade do problema, tais como a ineficiência do desenvolvimento de funções de energia capazes de descrever de maneira fidedigna o estado de um modelo 3D e baixa quantidade de informação estrutural depositada em bancos de dados (KIM et al., 2009). Desta forma, a geração de um conjuntos diverso de soluções é de extrema importância para resolução de problemas multimodais, de maneira que seja seja possível a obtenção de soluções finais de alta qualidade (DAS et al., 2011).

Neste trabalho, será proposta uma metodologia para geração de modelos estruturais 3D de proteínas utilizando dois limitantes de alta relevância ao escopo do problema de PSP. O primeiro, diz respeito as preferências conformacionais de aminoácidos sobre uma determinada estrutura secundária, a partir de dados depositados no PDB (BORGUESAN et al., 2015). Enquanto que o segundo, refere-se à inferência de acoplamento entre dois aminoácidos a partir da análise evolutiva da proteína-alvo (WEIGT et al., 2009). Estes modelos gerados utilizando ambos os limitantes serão submetidos a otimização utilizando uma implementação canônica de um DE. A função composta (Equação 2.6) será utilizada como função de avaliação. Concomitantemente, os modelos também serão otimizados por uma implementação multiobjetivo de um DE, onde os objetivos serão compostos por duas funções de energia, a primeira sendo a função composta, enquanto que a segunda será calculada baseada nos contatos entre aminoácidos.

3.4 Resumo do Capítulo

No presente capítulo foi realizado um levantamento teórico do atual estado da arte dos métodos propostos para a resolução do problema de PSP. Também foi explorado o cenário atual da predição de contato 3D entre aminoácidos a partir de inferências realizadas a com o resultado de alinhamentos múltiplos de sequências. Em conjunto destaca-se neste capítulo a importância da análise de acoplamento para uma melhora na acurácia dos preditores de estrutura propostos nas últimas edições do CASP. Introduziu-se neste capítulo uma visão geral de métodos de otimização estocásticos. Além de um contexto de técnicas, também foram apresentados trabalhos que propõem o emprego de meta-heurísticas para resolução do problema de PSP.

O capítulo seguinte será responsável por descrever a metodologia e análises utilizadas e proposta para a geração de modelos estruturais. É também descrito como ambos os limitantes são incorporados no processo de geração, assim como são avaliados os

parâmetros estruturais utilizados para avaliar cada modelo gerado. Os algoritmos de otimização utilizados para percorrer o espaço conformacional em busca de estruturas com menores valores energéticos são também apresentados.

4 MATERIAIS E MÉTODOS

O objetivo deste capítulo é descrever as estratégias empregadas na geração de modelos estruturais, assim como a metodologia e estruturação do método proposto para otimização de tais modelos. Semelhante com a definição de *limitante* proposta por Worth, Gong and Blundell (2009), neste trabalho utilizaremos limitantes estruturais de cunho não obrigatório. Portanto, tal limitante será constituído de uma condição composta por dois critérios de aceitação a serem avaliados para cada modelo gerado. Estes limitantes são: (*i*) o padrão conformacional de cada aminoácido sob uma estrutura secundária a partir da análise de estruturas experimentalmente determinadas e depositadas no PDB (LIGABUE-BRAUN et al., 2018); (*ii*) o contato 3D inferido a partir de análises evolutivas realizadas em alinhamento múltiplo de sequências (WEIGT et al., 2009). A união de ambos limitantes possui um grande potencial no que diz respeito à mimetização do processo de enovelamento de proteínas. Como já discutido anteriormente, estruturas secundárias adotam conformações de ângulos restritos de acordo com cada estrutura, contudo β-folhas além de conformações locais de ângulos ϕ/ψ, necessitam de proximidade 3D entre segmentos distintos de sequência peptídica (RICHARDSON, 1981). Desta maneira, a geração de modelos estruturais será guiada por dois fatores críticos: a formação de estrutura secundária e a aproximação 3D de aminoácidos que desempenham um importante papel na estabilidade da estrutura da proteína-alvo. Destaca-se que o processo de enovelamento é guiado por uma série de fatores físico-químicos e termodinâmicos (OSGUTHORPE, 2000). Em conjunto, sabe-se que o estado nativo de proteínas não corresponde a um estado único e estático (ANFINSEN, 1973). Assim, os limitantes propostos não apresentarão caráter restritivo, mas sim servirão como guias no processo de geração de modelos.

Em segunda instância, aproveitando-se da vantagem de adicionar informações previamente estabelecidas para auxiliar na eficácia de metodologias propostas ao problema de PSP, como discutidos no capítulo anterior, estes modelos gerados de acordo com os limitantes serão submetidos a processos de otimização. Primeiramente, será utilizada a implementação de um algoritmo de Evolução Diferencial canônico (STORN; PRICE, 1997). De maneira similar, os modelos serão otimizados utilizando uma estratégia multi-objetivo, a fim de aumentar a acurácia de predição, aumentando também as informações avaliadas durante o processo de otimização. Nas próximas seções serão detalhadas as implementações e estratégias utilizadas para tal.

4.1 Padrão conformacional preferencial de aminoácidos

Proteínas são capazes de assumir uma diversidade de diferentes estruturas 3D, as quais podem ser descritas através de seus ângulos de torção ϕ e ψ. A fim de utilizar a informação contida em estruturas 3D experimentalmente determinadas em métodos baseados em conhecimento, Borguesan et al. (2015) propõe a geração de lista de probabilidade de ângulos. Cada APL $H_{aa,es}$ seria constituída por [-180,+180] $\times$ [-180,+180] células para cada aminoácido (aa) sob uma estrutura secundária (es). Cada célula ij contém o número que um dado aa sob uma es específica possui a combinação de ângulos de torção ϕ/ψ ($i \leq \phi < i+1, j \leq \psi < j+1$). As ocorrências são normalizadas de acordo com a Equação 4.1. Como resultado, é obtida uma lista ordenada de acordo com a frequência de ocorrência de cada combinação de ângulos a partir de estruturas experimentalmente descritas e depositadas no PDB.

$$APL_{aa,es}(ij) = \frac{H_{aa,es}(ij)}{\sum(H_{aa,es})} \tag{4.1}$$

A ferramenta NIAS [1] (BORGUESAN; INOSTROZA-PONTA; DORN, 2017) (sigla do inglês *Neighbors Influence of Amino acids and Secondary Structures in Proteins*) está disponível gratuitamente na internet e permite que pesquisadores possam gerar APLs de diferentes tipos para a obtenção de informação de preferências conformacionais para aminoácido de interesse. Uma base de dados derivada do PDB foi gerada e utilizada pela ferramente NIAS, a fim de garantir uma alta qualidade de dados experimentais utilizados para geração de APLs. Para tal, foram selecionadas, apenas estruturas depositadas no PDB até dezembro de 2015, determinadas através de cristalografia obtidas por raio-x e que possuíam resolução menor, ou igual, a 2.5Å. O valor de resolução é responsável por indicar o nível de detalhamento o qual poderá ser determinado pelo cristal, de tal maneira que quanto menor o valor de resolução, maior a qualidade da estrutura (HOVMÖLLER; ZHOU; OHLSON, 2002). Apenas proteínas que possuem o fator R (*R-factor*) acima de 0.20 foram consideradas. Este parâmetro garante a qualidade da correspondência entre o modelo e os dados de difração dos quais o modelo foi proposto (KLEYWEGT; BRÜNGER, 1996). Para cada proteína, foram analisados os níveis de similaridade de sequência entre as sequências das demais proteínas depositadas no PDB. De tal maneira, proteínas que possuíam identidade maior do que 70% foram consideradas homólogas. Dentre os grupos formados de acordo com o critério de homologia, apenas uma estrutura foi con-

[1] <http://sbcb.inf.ufrgs.br/npas>

siderada, evitando a representação redundante de uma mesma proteína. O processo de filtragem resultou em um conjunto de dados constituído por 11.130 proteínas determinadas experimentalmente e 5.255.768 aminoácidos, após a filtragem por ocupação (*occupancy*) de valor igual a 1. A estrutura secundária de cada aminoácido foi calculada por dois descritores: STRIDE (FRISHMAN; ARGOS, 1995; HEINIG; FRISHMAN, 2004) e DSSP (KABSCH; SANDER, 1983). Os filtros utilizados para a geração do conjunto de dados utilizados pelo NIAS estão descritos na Tabela 4.1.

Tabela 4.1: Parâmetros de Filtragem do NIAS

Filtro	**Limiar**
Resolução	2.5Å
Fator R	0.2
Identidade de Sequência	70%
Ocupação	1
Total Proteínas	11.130
Total Aminoácidos	5.255.768

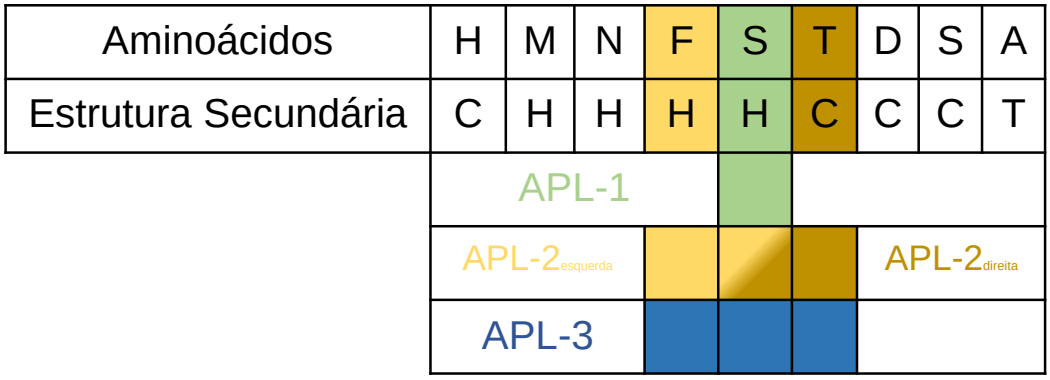
Parâmetros aplicados em estruturas depositadas no PDB para construção do conjunto de dados utilizados pelo NIAS.

Após a filtragem das estruturas que compõem o grupo de dados utilizados pelo NIAS, para cada aminoácido aa, foram calculados os ângulos de diedro ϕ e ψ nas estruturas determinadas experimentalmente e, então, foram definidas as preferências conformacionais de cada aa quando sob uma estrutura secundária específica. A partir destes dados, Correa et al. (2016) e Borguesan, Inostroza-Ponta and Dorn (2017) observaram que a preferência de conformações de cada aa dependia da estrutura secundária. De forma que um mesmo aminoácido aa possui diferentes preferências em estruturas secundárias distintas, assim como diferentes aa também possuíam padrões de combinações diferentes quando pertencentes à mesma estrutura secundária. Esta observação foi comprovada e discutida posteriormente em Ligabue-Braun et al. (2018). Desta forma, o conceito inicial da APL foi modificado de forma que se ampliasse o microambiente no qual o aminoácido de interesse está inserido (CORREA et al., 2016; BORGUESAN; INOSTROZA-PONTA; DORN, 2017).

Quando proposta, a APL calculava a ocorrência de uma dada combinação de ângulos diedrais considerando única e exclusivamente o aminoácido de interesse de acordo com a estrutura secundária na qual este se encontra, a partir de agora chamado de aminoácido de referência (aa_{ref}). Após a implementação de ampliação, não apenas o aa_{ref} co-

meça a ser considerado, mas também o microambiente no qual este está inserido, ou seja, os aminoácidos adjacentes na sequência. Desta maneira, geram-se novos tipos de APL: (*i*) **APL-1**, a qual refere-se ao conceito original da APL, considerando apenas aa_{ref} (Figura 4.1); (*ii*) **APL-2**, esta pode ser dividida em duas APLs distintas, (*a*) **APL-2**$_{direita}$ calcula a ocorrência do aa_{ref} considerando o aminoácido adjacente à direita sob uma determinada estrutura secundária, da mesma forma a (*b*) **APL-2**$_{esquerda}$ considera o vizinho à esquerda (Figura 4.1); (*iii*) **APL-3**, semelhante à **APL-2**, calcula a preferência conformacional do aa_{ref} considerando ambos vizinhos adjacentes simultaneamente. Em Borguesan, Inostroza-Ponta and Dorn (2017) ainda foram propostos 3 tipos de APL adicionais às descritas anteriormente. Estas APLs, em contraste às anteriores, consideram apenas o aa_{ref} e a estrutura secundária dos aminoácidos adjacentes. Desta maneira, há apenas a descrição de aa_{ref}, enquanto que os demais aminoácidos são desconsiderados, levando em consideração apenas a estrutura secundária. As janelas de vizinhos considerada por estas APLs adicionais são: (*I*) **APL-5**, considerando as estruturas secundárias dos dois vizinhos à direita e dois à esquerda; (*II*) **APL-7**, considerando as estruturas secundárias dos três vizinhos adjacentes tanto à direita quanto à esquerda; e (*III*) **APL-9**, abrangendo os cinco vizinhos à direita e à esquerda. Estas três APLs adicionais possuem seu nível de abstração aumentado em decorrência da diminuição de quantidade de informação disponível quando a janela de aminoácidos aumenta (BORGUESAN; INOSTROZA-PONTA; DORN, 2017).

Figura 4.1: Representação Geral da APL



Níveis de abrangência de cada tipo de APL. A APL-1 considerando apenas o aminoácido de referência. APL-2 considera individualmente cada vizinho adjacente, enquanto que a APL-e considera ambos vizinhos em conjunto.

Além das APLs, a ferramente NIAS retorna histogramas semelhantes ao mapa da

Ramachandran (RAMACHANDRAN, 1963). Esta representação gráfica das APLs possui em seu eixo x os valores dos ângulos ϕ, enquanto que em ψ está descrito no eixo y. Contudo, as diferentes combinações de valores de ângulo são provenientes de estruturas distintas, as quais possuem aa_{ref}. Exemplificados na Figura 4.2, encontram-se os casos mencionados anteriormente nesta seção. A diferença de preferência de conformações entre aminoácidos diferentes sob a mesma estrutura secundária, assim como a diferença de um mesmo aminoácido em estrutura secundária distinta, estão exemplificados na Figura 4.2a. Além disto, a diminuição de quantidade de dados à medida em que se aumenta a janela de aminoácidos avaliados está explícita nas Figuras 4.2a-4.2d. Destaca-se que as APLs são responsáveis pela geração de preferência conformacional de um *único* aminoácido, podendo ou não considerar a sua vizinhança. Portanto, ressaltamos a diferença entre o método das APLs e os utilizados pelos preditores Rosetta e QUARK. Os métodos propostos utilizando os valores provenientes da APL atribuem os valores de ângulos para cada aminoácido individualmente. Enquanto que as abordagens baseadas em fragmentos (SIMONS et al., 1997), adotado pelos métodos Rosetta, utiliza fragmentos completos de tamanho 3 e 9 aminoácidos (GRONT et al., 2011), e QUARK, calcula perfis de distância entre aminoácidos a partir de fragmentos selecionados de acordo com similaridade de domínios (OVCHINNIKOV et al., 2018).

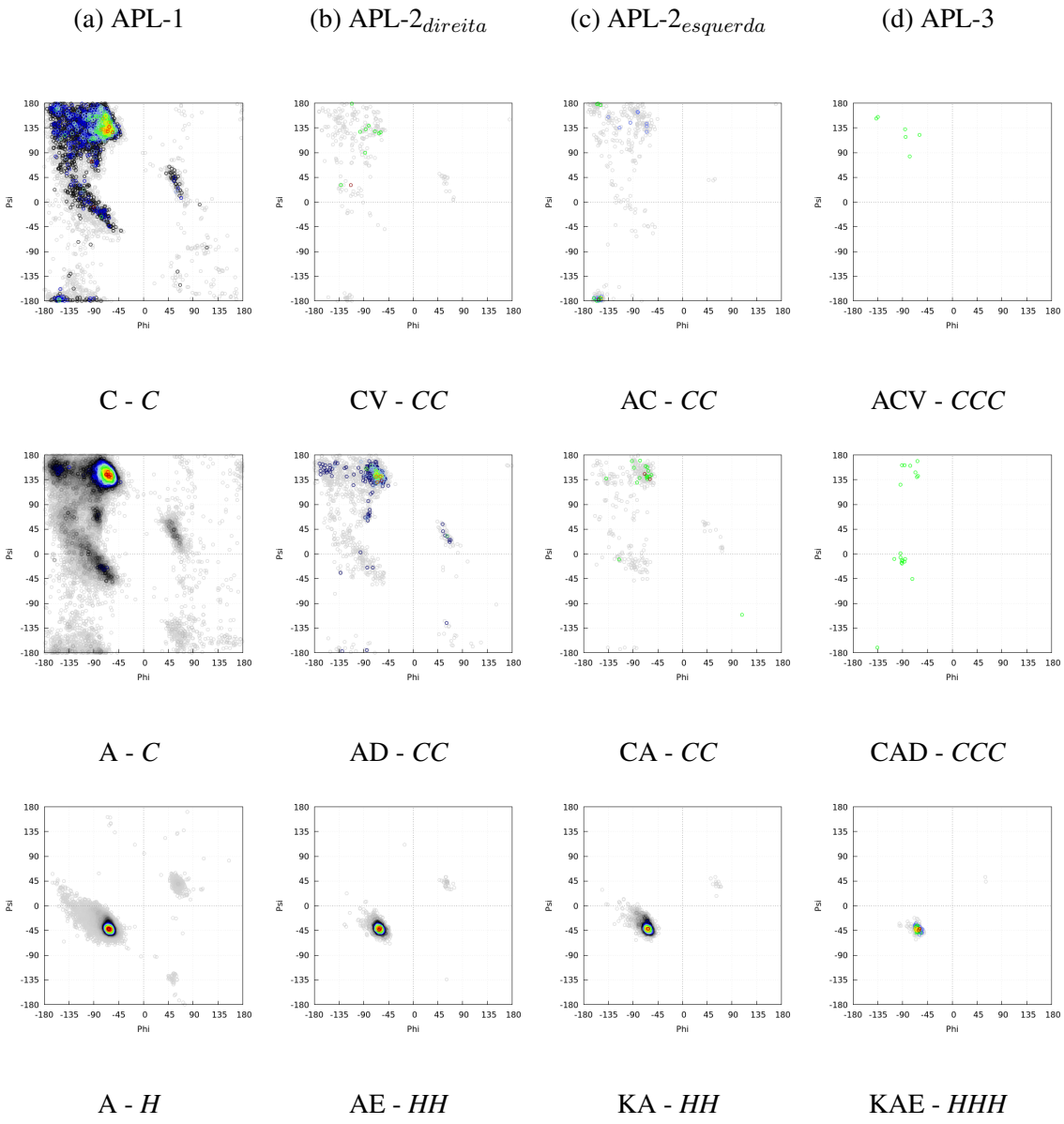
4.2 Acoplamento 3D de aminoácidos

A utilização da informação de acoplamento entre aminoácidos a partir de análises evolutivas tem sido destaque nas últimas edições do CASP, não apenas pelo aumento na precisão de predição de aminoácidos em contato direto, mas também por sua participação positiva nas metodologias propostas para a resolução do problema de PSP (SCHAARSCHMIDT et al., 2018). Como discutido na Seção 3.2, desde a implementação de um algoritmo de *message-passing* para a inferências dos parâmetros ideais da equação de Boltzmann (Equação 2.7) por Weigt et al. (2009), uma série de algoritmos e abordagens têm sido propostas ao longo dos anos de forma a aumentar a precisão das predições, assim como na diferenciação correta de contatos diretos e indiretos (Figura 2.5).

O método RaptorX-Contact [2] for escolhido para avaliar o acoplamento entre aminoácidos, o qual foi o vencedor da categoria de predição de contatos da edição CASP12 (SCHAARSCHMIDT et al., 2018). Este método difere-se dos demais métodos propostos pela

[2] <http://raptorx.uchicago.edu/ContactMap/>

Figura 4.2: Exemplificação de Preferências Conformacionais



Representação gráfica das APLs geradas através de mapas semelhantes aos de Ramachandran gerados pela ferramenta NIAS para as sequências: ACV, CAD e KAE; sob as estruturas secundárias *CCC*, *CCC* e (*HHH*), respectivamente. Onde *C* refere-se à *Coil* e *H* à α-Hélice.

implementação de um método de *Deep Learning* capaz de integrar de maneira efetiva informações provenientes de análises de coevolução e também de análises clássicas de estrutura de proteínas, permitindo uma predição de alta precisão inclusive em casos de baixa quantidade de sequências identificadas pelo MSA (WANG; SUN; XU, 2018). A partir de cada MSA são extraídos dois tipos de informações: características sequenciais e emparelhadas. As características sequenciais dizem respeito ao perfil das sequências,

assim como a estrutura secundária e área acessível ao solvente preditas pelo RaptorX-Property (WANG et al., 2016). Enquanto que dentro das informações emparelhadas estão contidas a informação mútua, potencial de contato entre aminoácidos e força co-evolutiva. O método combina duas redes neurais residuais profundas (*deep residual neural networks*), dando origem à um modelo de aprendizado profundo (*deep learning*) (WANG et al., 2017). Desta maneira, a primeira etapa consiste no aprendizado a partir das características sequenciais, enquanto que a segunda inclui informações provenientes da primeira em adição às características emparelhadas. Desta maneira, como relatado em Wang, Sun and Xu (2018), é possível informações de alta precisão acerca de contatos entre aminoácidos, independente do número de sequências homólogas disponíveis para a proteína-alvo.

4.3 Construção de modelos estruturais

A metodologia proposta para este trabalho consiste na utilização de dois limitantes de relevância biológica para construção de modelos estruturais que possuam maior grau de semelhança em relação à estrutura determinada experimentalmente. Portanto, para avaliarmos a eficácia do método, propõe-se uma maneira inédita de construção destes modelos, de maneira que sejam necessárias apenas a sequência de aminoácidos e estrutura secundária da proteína de interesse. A partir deste ponto, cada modelo estrutural será representado por um vetor de ângulos de torção, os quais descrevem a disposição dos aminoácidos que constituem o esqueleto peptídico da proteína-alvo, configurando assim a representação computacional dos modelos 3D a partir de uma sequência de aminoácidos linear, conforme descrito na seção 2.3.

A geração de indivíduos foi realizada seguindo 4 protocolos diferentes, ocasionando na criação de 4 populações distintas, cada qual contendo 10.000 indivíduos. Os modelos gerados a partir de cada metodologia foram avaliados, além da energia composta descrita pela Equação 2.6, por seus valores calculados de raio de giro (RG), área acessível ao solvente (SASA) e também avaliou-se a semelhança entre a disposição atômica do modelo gerado em comparação à estrutura determinada experimentalmente. O RG é responsável por calcular a distância quadrática média entre todos os átomos de uma dada proteína e o seu respectivo centro de massa. Tal métrica é utilizada para descrever o estado de compactação de uma proteína, portanto, quanto menor o valor calculado de RG, mais próximos estarão os átomos do centro de massas (LOBANOV; BOGATYREVA; GALZITSKAYA, 2008). O valor calculado referente ao SASA de cada proteína, refere-

se à energia calculada a partir da interação entre os átomos da proteína e os átomos do solvente (LEE; RICHARDS, 1971; CONNOLLY, 1983). Desta forma, quanto menor os valores calculados, menor o número de átomos expostos ao solvente, portanto, podemos inferir um estado de compactação maior, semelhante ao RG. Quanto maior o número de aminoácidos de proteínas globulares estiverem em contato com o solvente, menor será a compactação desta proteína, de maneira que, ao atingir um maior enovelamento ocorra a diminuição de valores de SASA (ROSE et al., 1985).

As semelhanças de distribuição 3D atômica entre modelo gerado e estrutura determinada experimentalmente foram calculadas utilizando o Desvio da Raiz Quadrada Média (RMSD, sigla do inglês *Root-mean-square Deviation*). Este cálculo mede a distância, em angstroms (Å), entre os átomos de duas estruturas distintas. Quanto mais próximo de 0 Å, maior é o nível de semelhança entre ambas estruturas. Descrito na Equação 4.2 está a fórmula utilizada para calcular o RMSD entre as estruturas, onde r_{ai} e r_{bi} correspondem ao $i$ésimo átomo no grupo n de átomos de duas proteínas. Os cálculos de RG, SASA e RMSD utilizados para avaliação das estruturas 3D de proteínas deste trabalho foram realizados utilizando as bibliotecas implementadas e oferecidas pelo PyRosetta.

$$RMSD(a,b) = \sqrt{\frac{\sum_{i=1}^{n} |r_{ai} - r_{bi}|^2}{n}} \tag{4.2}$$

Para cada proteína utilizada neste trabalho, primeiramente foram coletadas as estruturas primárias, ou seja, a sequência de aminoácidos. Então, a estrutura secundária foi descrita utilizando o algoritmo DSSP (KABSCH; SANDER, 1983). A partir destas duas informações, ambos os limitantes utilizados foram gerados. As APLs foram geradas no servidor NIAS (BORGUESAN; INOSTROZA-PONTA; DORN, 2017), enquanto que os contatos foram preditos utilizando o servidor RaptorX-Contact (WANG; SUN; XU, 2018). Para cada aminoácido compondo a sequência da proteína-alvo foram gerados a APL-1, -$2_{esquerda}$, -$2_{direita}$ e -3. Estas, foram geradas após excluir do banco de dados do NIAS as estruturas que apresentavam um grau de similaridade maior que 70% com a proteína-alvo. O contatos preditos para cada proteínas foram filtrados, segundo os parâmetros apresentados pelo CASP, onde apenas os contatos entre o aminoácido i e j fossem de média ($12 \leq |i-j| \leq 23$) ou longa ($|i-j| \geq 23$) distância (SCHAARSCHMIDT et al., 2018). Assim como foram considerados apenas os melhores $L/5$ contatos, já filtrados pela distância, foram utilizados, onde L é o tamanho da sequência da proteína de interesse.

A primeira população é constituída por indivíduos gerados aleatoriamente. Por-

tanto, cada ângulo de torção pertencente ao vetor-solução ($[\phi, \psi, \omega, \chi(z)] \times n$), onde n é o número de aminoácidos da proteína, foi gerado aleatoriamente no intervalo de $[-180.0°, 180.0°]$ sem nenhuma restrição, mantendo inclusive modelos que possuíam interferências estéricas entre átomos da cadeia principal e lateral. A segunda população foi criada utilizando apenas um dos limitantes propostos, as informações de ângulos diedrais retiradas das APLs. Para cada aminoácido da cadeia peptídica, cada APL possuía diferentes probabilidades de serem escolhidas (Tabela 4.2). Em decorrência da melhor representação do microambiente no qual o aminoácido está inserido, a APL-3 é a mais provável de ser escolhida, possuindo uma chance de 50% de escolha. Desta maneira, a segunda maior probabilidade de escolha pertence à APL-2, a qual possui 30% de chance. Uma vez que a APL-2 é escolhida, tanto a *direita* quanto a *esquerda* possuem 50% de chance de serem selecionadas. A menor probabilidade de escolha de 20% pertence à APL-1. Os aminoácidos N- e C- terminais não possuem vizinhos à esquerda e à direita, respectivamente. Portanto, estes possuem probabilidades diferentes dos demais para escolha de APL utilizada (Tabela 4.2): a APL-2 possui 60% de probabilidade, enquanto que a APL-1 possui 40% de chance. Nota-se que quanto maior a janela de aminoácidos considerada, maior o grau de restrição, portanto a quantidade de dados encontrados no PDB diminui. Assim, em casos em que não houvesse dados para computar a APL-3, as probabilidades de escolha da APL-3 eram somadas à probabilidade da APL-2. Os ângulos diedrais para cada aminoácido foram selecionados de acordo com a probabilidade calculada pela ferramenta NIAS, ou seja, as ocorrências com maior frequência possuem uma chance maior de serem selecionadas. O processo de geração do vetor de ângulos de torção para cada aminoácido é descrito pelo pseudo-código descrito pelo Algoritmo 1.

Tabela 4.2: Probabilidade de Seleção de APLs.

APL	**Probabilidade**
APL-$1_{i=1,i=n}$	40%
APL-$1_{1<i<n}$	20%
APL-$2_{i=1,i=n}$	60%
APL-$2_{1<i<n}$	30%
APL-$3_{1<i<n}$	50%

Probabilidade utilizadas para utilização de APLs, onde i corresponde ao $i$ésimo aminoácido da sequência de tamanho n. Para o primeiro ($i = 1$) e último ($i = n$) aminoácidos as probabilidades são modificadas em decorrência da falta de vizinhança à esquerda para o primeiro e o vizinho à direita do último aminoácido.

A terceira população é composta por indivíduos gerados utilizando apenas o se-

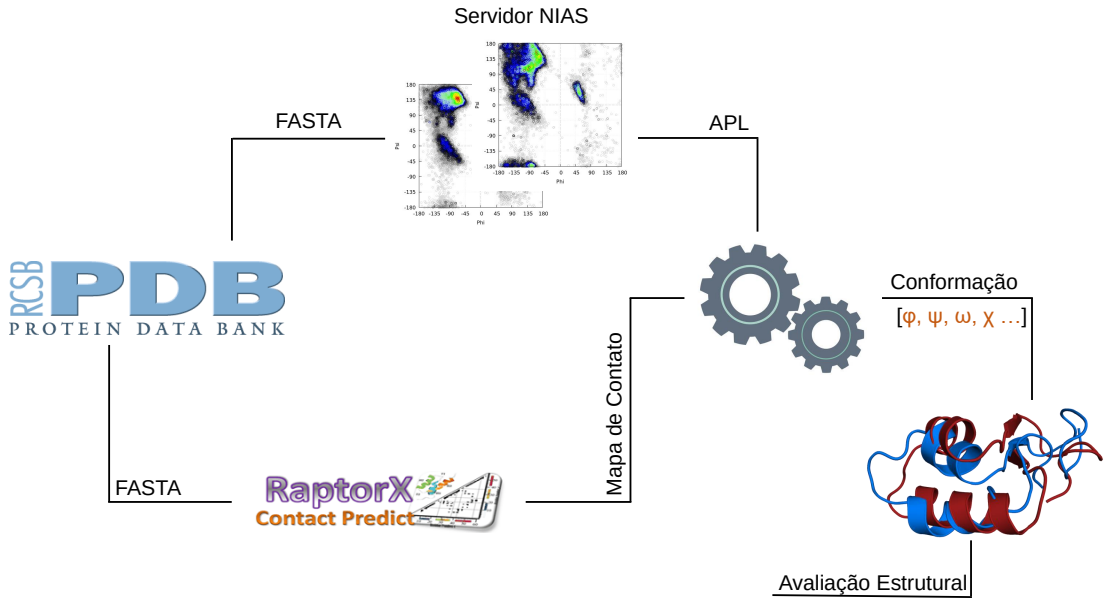
Algoritmo 1: Pseudo-código para Seleção de Ângulos da APL.

Entrada: Sequência de 3 aminoácidos, Estrutura Secundária
Saída: Vetor de ângulos diedrais para um determinado aminoácido sob uma estrutura secundária específica.

```
Seleção das APL-1, -2_esquerda, -2_direita e -3;
P ← aleatório[0,1];
se P ≥ 0.5 então
    vetorÂngulos ← selecionaÂngulos(APL-3);
senão se P ≤ 0.3 então
    nP ← aleatório[0,1];
    se nP ≥ 0.5 então
        vetorÂngulos ← selecionaÂngulos(APL-2_direita);
    senão
        vetorÂngulos ← selecionaÂngulos(APL-2_esquerda);
    fim
senão
    vetorÂngulos ← selecionaÂngulos(APL-1);
fim
retorna vetorÂngulos;
```

gundo limitante proposto por este trabalho, a aproximação 3D de aminoácidos de acordo com análises de co-evolução. Após o primeiro processo de filtragem, os contatos resultantes foram utilizados como critério de seleção para aceitação ou descarte de modelos gerados. Similar ao processo utilizado para a primeira população, os ângulos de torção utilizados para compor os modelos foram gerados aleatoriamente dentro do espaço $[-180.0°, 180.0°]$, contudo o modelo só era aceito se, e somente se, 20% dos contatos preditos estivessem em contato no modelo 3D. Para tal, consideram-se em contato 3D os pares de aminoácidos em que a distância Euclidiana entre os átomos de carbono C_β (ou C_α para Glicina) for menor do que 8 Å (SCHAARSCHMIDT et al., 2018). Nota-se que em casos onde 20% dos contatos representam um valor menor que 1, foram considerados os modelos aceitos aqueles que possuíam pelo menos um par de aminoácidos em contato. A última população, a qual deseja-se avaliar a real eficácia, constitui-se de indivíduos gerados sob ambos limitantes combinados. Então, cada indivíduo é inicialmente construído seguindo a metodologia descrita anteriormente para utilização das APLs (Algoritmo 1), contudo este é mantido se, e somente se, também estiver de acordo com o limitante de acoplamento de pares de aminoácidos descrito acima. Uma vez construídas as populações, estas foram submetidas às avaliações de parâmetros estruturais a fim de verificar a semelhança dos modelos com as estruturas experimentalmente determinadas. O processo de geração dos indivíduos está ilustrado de modo geral pela Figura 4.3.

Figura 4.3: Processo de Geração de Indivíduos



Esquema simplificado da utilização dos limitantes no processo de geração de modelos estruturais. A partir da sequência de aminoácidos são gerados as listas de preferência conformacional destes, em conjunto é gerado o mapa de contato entre pares de aminoácidos com base em análises co-evolutivas. Uma vez geradas as informações, estas são utilizadas no processo de montagem da conformação 3D de estruturas de proteínas, as quais são avaliadas de acordo com os parâmetros estruturais selecionados.

4.4 Evolução Diferencial

Segundo a teoria de complexidade computacional, o problema de PSP é classificado como um problema NP-Difícil (COOK, 1983; CRESCENZI et al., 1998). Desta forma, a utilização de algoritmos de otimização são ideais para explorar o espaço de busca conformacional a fim de encontrar as estruturas com menor valor de energia. Devido à multimodalidade apresentada pelas funções de energia geralmente aplicadas no processo de busca e otimização de estrutura de proteínas, quando um conjunto de soluções iniciais é composto por soluções ruins, este tende a convergir os resultados finais à soluções igualmente ruins. Isto decorre da ineficiência dos métodos utilizados de sair de regiões de energia mínima local (Figura 3.4). Desta forma, a utilização de algoritmos capazes de manter e gerar a diversidade de indivíduos durante o processo de otimização, tende a conseguir explorar o espaço de busca conformacional de maneira mais eficiente (DAS et al., 2011).

Borguesan et al. (2018) e Correa et al. (2016) demonstram a eficácia da utilização

das APLs na geração de indivíduos iniciais, uma vez que esta é responsável pelo aumento na precisão de predição de estrutura 3D de proteínas utilizando meta-heurísticas. Assim, a utilização dos ângulos de torção retirados do PDB juntamente com as informações evolutivas preditas a partir de MSA possuem uma grande tendência a melhorarem o processo de busca. Narloch and Dorn (2019a) demonstram a alta eficiência da utilização da APL quando constrói a população inicial de um algoritmo de Evolução Diferencial. Portanto, aproveitando-se do potencial do método proposto de gerar indivíduos com alta similaridade estrutural quando comparados à estrutura determinada experimentalmente, propôs-se a utilização destes como indivíduos pertencentes à população inicial de meta-heurísticas populacionais. Para tal, utilizou-se a implementação de um algoritmo DE canônico proposto por Storn and Price (1997). Este algoritmo pode ser segmentado em quatro estágios principais: inicialização, mutação, recombinação (*crossover*) e seleção.

Inicialização O algoritmo é um método de busca paralela que utiliza um vetor de dimensão D

$$x_{i,G}, i = 1, 2, 3, ..., N \tag{4.3}$$

como população P de tamanho N para cada geração G. Nota-se que o tamanho de N é constante durante o processo de otimização. Em sua implementação canônica, o vetor de população inicial deve ser gerado aleatoriamente. Contudo, neste trabalho também serão testadas as populações iniciais geradas sob a influência dos limitantes individualmente, assim como sob a influência de ambos simultaneamente.

Mutação Para cada indivíduo compondo a população P, um novo indivíduo v é gerado, segundo

$$v_{i,G+1} = x_{r1,G} + F \cdot (x_{r2,G} - x_{r3,G}) \tag{4.4}$$

onde $r1, r2, r3$ são indivíduos distintos entre si, assim como diferem do indivíduo-alvo i, selecionados de $\{1, 2, .., N\}$ (Algoritmo 2, linhas 4-5). Portanto, a utilização de um DE só é permitida se, e somente se $N \geq 4$. A constante F corresponde ao fator de amplificação de $(x_{r2,G} - x_{r3,G})$, o qual pode assumir qualquer valor dentro do intervalo [0,2].

Recombinação A diversidade de indivíduos durante a execução do algoritmo é mantida e aumentada em decorrência do processo de recombinação de indivíduos. Para isso, um vetor provisório t

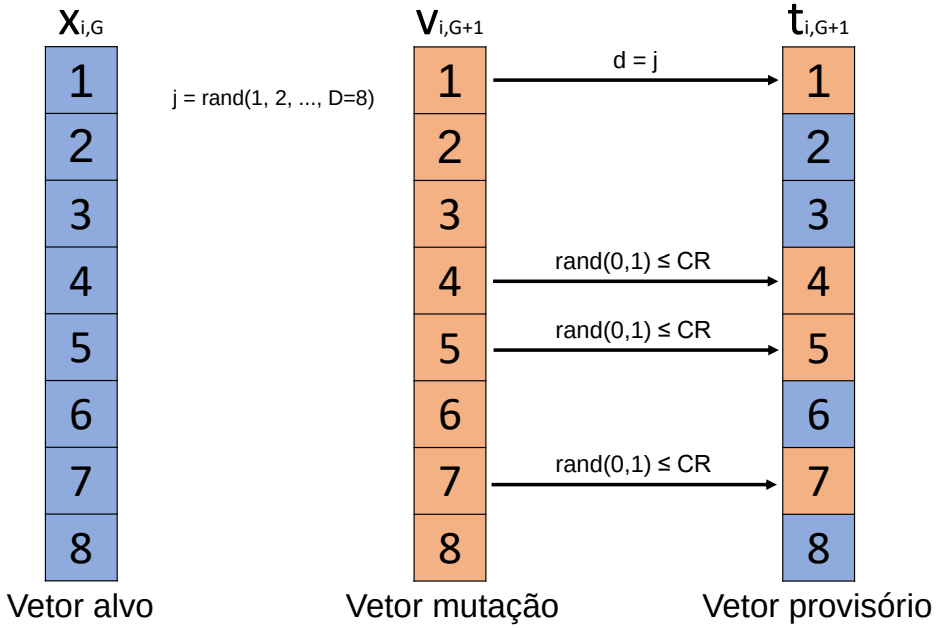
$$t_{i,G+1} = (t_{1i,G+1}, t_{2i,G+1}, ..., t_{Di,G+1}) \tag{4.5}$$

é gerado segundo

$$t_{i_d,G} = \begin{cases} v_{i_d,G} & \text{se } d = d_{rand} \text{ ou } rand[0,1] \leq CR \\ x_{i_d,G} & \text{caso contrário} \end{cases} \tag{4.6}$$

uma vez que a operação de recombinação geralmente segue um sistema de decisão binomial. Portanto, a mutação da dimensão $d \in D$ será aceita se um número escolhido aleatoriamente no intervalo [0,1] for menor do que a Constante de Recombinação (CR), onde CR $\in$ [0,1], ou se d for igual à um valor aleatório $\in \{1, 2, 3, ..., D\}$. Caso contrário, o valor de d permanece inalterado. O processo de mutação está descrito visualmente pela Figura 4.4 (Algoritmo 2, linhas 6-12).

Figura 4.4: Processo de Geração de Indivíduos



Representação do processo de recombinação. Adaptado de Storn and Price (1997).

Seleção A última etapa é o processo de seleção de vetores provisórios t que estarão presentes na geração seguinte. O vetor t é avaliado de acordo com a função $f(x)$, a qual corresponde à função utilizada para otimização. Neste trabalho será utilizada a função composta descrita pela Equação 2.6. Assim, a população de indivíduos da geração $G + 1$ será composta apenas pelos indivíduos melhore avaliados segundo a função avaliadora (Algoritmo 2, linhas 13-19). Esta operação é descrita da seguinte

maneira

$$x_{i,G+1} = \begin{cases} t_G & \text{se } f(t_G) \leq f(x_{i,G}), \\ x_{i,G} & \text{caso contrário} \end{cases} \tag{4.7}$$

Em linhas gerais, o algoritmo do DE utiliza uma população inicial, gerada aleatoriamente, para realizar um processo paralelo de busca. Então, em passos posteriores, novos vetores-solução são gerados pela diferença ponderada entre dois vetores em comparação à um terceiro vetor, todos escolhidos de maneira aleatória. Desta maneira, um quarto vetor é construído a partir de um processo de recombinação entre o vetor-solução inicial e o vetor construído a partir das diferenças. Em um último passo, os vetores são avaliados de acordo com a função de avaliação empregada para que apenas os indivíduos de melhor valor energético sejam passados para a próxima geração (STORN; PRICE, 1997). O pseudocódigo para a implementação clássica do DE é descrito pelo Algoritmo 2.

Algoritmo 2: Implementação clássica de DE.

Entrada: N: tamanho da população inicial, F: fator de amplificação, e CR: critério de recombinação

Saída: Melhor indivíduo, correspondente ao menor valor de avaliação encontrado pelo processo de otimização.

```
1  População inicial contendo N indivíduos;
2  enquanto g ≤ número total de gerações faça
3      para cada indivíduo i na população P faça
4          Seleciona aleatoriamente três indivíduos (r1,r2,r3);
5          d_rand ← dimensão aleatória a ser mutada;
6          para cada dimensão d faça
7              se d = d_rand ou random[0,1] ≤ CR então
8                  t_{i,d} ← x_{r1,G} + F · (x_{r2,G} − x_{r3,G})
9              senão
10                 t_{i,d} ← x_{i,d}
11             fim
12         fim
13         se f(t_i) ≤ f(x_i) então
14             descendente.append(t_i)
15         senão
16             descendente.append(x_i)
17         fim
18     fim
19     P ← descendente g ← g+1
20 fim
```

O problema de PSP possui uma alta multimodalidade, incapacitando a total representação do estado conformacional de uma proteína em seu estado nativo através do uso

de funções de energia (KIM et al., 2009). A função de avaliação fornecida pelo Rosetta é composta por uma série de termos diferentes (Tabela 2.2) (ALFORD et al., 2017), os quais podem incluir caráter opostos, por exemplo, termos ligantes e não-ligantes. Portanto, testamos o desempenho dos modelos gerados utilizando um algoritmo de busca Multi-objetivo (MO) (KONAK; COIT; SMITH, 2006). Durante o processo de otimização utilizando apenas a energia composta (Equação 2.6) proposto anteriormente não há garantia que a informação de contatos se mantenha ao longo da execução. A fim de reforçar o limitante de aproximação 3D entre aminoácidos, utilizamos a função de energia Lorentizana baseada em distância 3D de aminoácidos proposta por Hong et al. (2018), definida por

$$E_{contato} = \sum_{ij}^{Nc} \begin{cases} \frac{(r_{ij}-r_{cut})^2}{(r_{ij}-r_{cut})^2+\sigma^2} c(p), & \text{se } r_{ij} \geq r_{cut} \\ 0 & \text{caso contrário} \end{cases} \tag{4.8}$$

$$c(p) = ap + 1 \tag{4.9}$$

onde Nc corresponde ao número de contatos preditos com probabilidade > 0.2, os contatos utilizados neste termo correspondem aos de média e longa distância, e i e j correspondem aos aminoácidos compondo cada par predito. A distância r_{ij} é a distância Euclidiana calculada entre os carbono C_β (C_α em caso de Glicina) dos resíduos i e j. A constantes r_{cut}, σ e a, foram previamente determinadas pelos autores e estão descritas na Tabela 4.3.

Tabela 4.3: Parâmetros da Função de Energia de Contatos.

r_{cut}	σ	a
7.840362	0.084674	0.156446

Constantes utilizadas na função de energia de contatos determinados por Hong et al. (2018).

O algoritmo utilizado para realização de otimização de ambos objetivos foi a versão multi-objetiva do DE (DEMO) (ROBIČ; FILIPIČ, 2005). Nota-se que, tratando-se de um algoritmo de um único objetivo, o indivíduo da geração atual, referido como indivíduo pai, será substituído pelo vetor resultante da mutação, chamado de filho, se este for melhor avaliado do que o indivíduo pai (Algoritmo 2). Contudo, para o DEMO, o processo de avaliação se o indivíduo filho ou pai está mais apto para passar para a próxima geração não é trivial. Como descrito anteriormente, para algoritmos de um único objetivo, a comparação é direta, o indivíduo com melhor pontuação passa para a geração seguinte, enquanto que o outro é descartado. Portanto, após a avaliação se o candidato filho é dominante em

relação ao pai, o filho será mantido, caso contrário o indivíduo pai continua para próxima geração. Contudo, se ambos não exercem dominância entre si, ambos são mantidos para próxima geração (ROBIČ; FILIPIČ, 2005). Ao final do processo, gera-se uma população de tamanho variável entre N e $2N$. Esta população reduzida à N através de um processo de ranqueamento baseado em soluções não-dominadas em conjunto com uma métrica de distância de aglomeração de pontos, a fim de garantir uma maior diversidade (DEB et al., 2002). Além da implementação do algoritmo DEMO, foram utilizadas funções adaptadas retiradas do jMetalPy [3] (BENITEZ-HIDALGO et al., 2019), um conjunto de funções implementadas em Python para otimizações multi-objetivo utilizando meta-heurísticas.

4.5 Resumo do capítulo

No presente capítulo foram explicadas as metodologias implementadas para a execução do trabalho. A geração de informações utilizadas como limitantes durante a geração de modelos estruturais, tanto as informações de preferência conformacional de aminoácidos retiradas do PDB, quanto a informação de acoplamento a partir de análises evolutivas. Também foram descritos os algoritmos de meta-heurísticas escolhidos e implementados para otimização dos modelos gerados. No próximo capítulo, serão apresentados os resultados obtidos das análises descritas no presente capítulo.

[3] https://github.com/jMetal/jMetalPy

5 RESULTADOS

Os resultados obtidos através dos métodos descritos na seção anterior serão apresentados no presente capítulo. Este, será apresentado em formato de artigo, segundo as revistas pleiteadas pela editora Elsevier [1]. São divididos os resultados em um artigo principal, entitulado "*The inclusion of conserved evolutionary contact to refine validate techniques to predict protein structure*", em conjunto com um artigo de materiais suplementares ao artigo principal.

[1] <https://www.elsevier.com/pt-br>

The inclusion of conserved evolutionary contact to refine validate techniques to predict protein structure

LA Santos[a], R Ligabue-Braun[b], M Dorn[c,*]

[a] *Center of Biotechnology, Federal University of Rio Grande do Sul, Porto Alegre, Brazil*
[b] *Department of Pharmaceutical Sciences, UFCSPA, Porto Alegre, Brazil*
[c] *Institute of Informatics, Federal University of Rio Grande do Sul, Porto Alegre, Brazil*

Abstract

The disparity between the number of sequences and the number of protein structures deposited into public databases is increasing due to the high-throughput DNA sequencing technologies. Thus, the development of computational approaches to determine the three-dimensional structure of proteins is necessary since they are faster and less expensive than the experimental methods. Here we propose the incorporation of two constraints to generate structural models more similar to the native structure with potential to be improved by optimization algorithms more efficiently. The first constraint uses torsional angle information retrieved from the Protein Data Bank, while the second one takes advantage of the massive growth of sequence information available to calculate the coupling between amino acid residues within a protein family. Together, these constraints were able to generate candidates with close-to-native structural features and low free energy scores.

Keywords: Conformational Preference of Amino Acid Residues, Knowledge-based Protein Structure Prediction, Protein Structural Model, Residue Contacts, Structure Optimization.
2010 MSC: 00-01, 99-00

1. Introduction

Proteins are involved nearly in every cellular process, making the number of different protein functions virtually infinite (Nelson et al., 2008). Their function is intrinsically related to their three-dimensional (3D) structure. Hence, stable structures are essential to the well-functioning of these molecules. The determination of the 3D structure of proteins is critical for the complete elucidation of a variety of biological processes. The most common way to describe the 3D structure of proteins are the experimental methods, such as X-ray crystallography (Mazzei et al., 2017), Nuclear Magnetic Resonance (NMR) (Carneiro et al., 2015) and Cryogenic Electron Microscopy (cryo-EM) (Fan et al., 2019). These methods, although being the most commonly employed, are expensive, time-consuming and susceptible to procedural errors throughout the entire process (Edwards et al., 2000). These adversities may explain the gap between the number of known biological sequences and the number of proteins with elucidated 3D structure. The current number of structures deposited into the Protein Data Bank (PDB) (Berman et al., 2003) is equivalent to less than 1% of the number of sequences deposited in the NCBI Reference Sequence[1] (RefSeq) (Pruitt et al., 2006).

The prediction of the 3D structure of proteins is one of the milestones yet to be overcome in Structural Bioinformatics, the so-called Protein Structure Prediction (PSP) problem. The number of 3D viable

*Corresponding author
Email addresses: leonardoas95@gmail.com (LA Santos), rodrigolb@ufcspa.edu.br (R Ligabue-Braun), mdorn@inf.ufrgs.br (M Dorn)

[1] https://www.ncbi.nlm.nih.gov/refseq/

structures a single amino acid sequence can assume is humongous, classifying the PSP problem as an NP-Complete problem according to the computational complexity theory (Crescenzi et al., 1998). The problem becomes more difficult as the number of amino acids of a given protein sequence increases since the number of amino acids is directly related to the problem dimensionality (Guyeux et al., 2014). Several different approaches have been proposed in the last decades to overcome the PSP problem's complexity. These computational methodologies, share three components that must be always considered: (a) a computational way to represent the 3D structure of proteins; (b) an energy function to evaluate the structure, following the Anfinsen's thermodynamic hypothesis, which says that the lowest values of free energy also correspond to stable structures (Anfinsen, 1973); and (c) an algorithm to explore the conformational search space to find the structures with the lowest energy value (Dorn et al., 2014).

Metaheuristics are a class of optimization algorithms largely employed to obtain plausible solutions for problems with high complexity in a feasible execution time (Talbi, 2009). The optimization process is guided by several different concepts and learning processes to explore and exploit the search space in order to achieve the near-optimal solutions (Osman and Laporte, 1996). These methods rely on the optimization of a single or a group of candidate solutions (Luke, 2013). Examples of metaheuristics are Simulated Annealing (SA) (Kirkpatrick et al., 1983) and Genetic Algorithm (GA) (Goldberg and Holland, 1988). Metaheuristics are widely used to try to tackle the complexity of the PSP problem (Fonseca et al., 2010; Silva and Parpinelli, 2019; Tantar et al., 2007). The usage of biological information throughout the optimization process has already been demonstrated as a valuable addition. The efficiency of the usage of the conformational preference of amino acid torsion angles under specific secondary structures was considered during the assembly of the initial candidate solutions for two different metaheuristics (Borguesan et al., 2015). Similarly, the employment of previously established information regarding the torsion angle of amino acid residues information to build the initial candidate solutions of a Differential Evolution algorithm applied to the PSP problem (Narloch and Dorn, 2019). The results achieved by both studies show an improvement of the final solution compared to those candidates assembled without the previous information. In spite of the higher efficiency of the initial candidates optimization showed by the usage of torsional angles, this information presents higher relevance for proteins with a higher incidence of α-helix within their secondary structure due to a more restrictive allowed dihedral angle combinations. Together with the torsion angles, the β-sheets also need the spatial proximity between two independent sequence fragments (Richardson, 1981).

All sorts of relevant biological information may be used to elevate the accuracy of the prediction methods. Therefore, the usage of spatial proximity of two amino acid residues based on the analysis of multiple sequence alignment has been shown to be an effective addition when trying to improve the accuracy of the structural prediction methods (Schaarschmidt et al., 2018).

The present work propose a knowledge-based approach to generate structural models with structural features close to the experimentally determined structure, as well as the usage of these assembled models to compose the initial solutions of optimization algorithms. The generation of individuals[2] is guided by two biological constraints, first the torsion angles conformations for each amino acid residue under a specific secondary structure, retrieved from the PDB, and second, the 3D contact predicted from multiple sequence alignments (MSAs) outputs analysis. The quality of the generated individuals is measured by the structural evaluation methods: the solvent accessible surface area, the radius of gyration, and the similarity of the generated individual with the 3D structure experimentally determined for the target protein. The results obtained from the structural evaluation shows that the proposed method is able to generate individuals similar to the ones experimentally determined. It is worth to mention that these structural features are achieved without any optimization steps. Therefore, we believe these structural models generated by the merging of the two proposed biological constraints are more suitable to be used as initial candidate solutions in optimization algorithms. We expect that the optimization tends to be more efficient since these models present structural characteristics more alike to the structure experimentally determined.

[2]Here we refer to each structural model constructed by our method as an individual.

2. Materials and Methods

2.1. Biological Constraints

Given the complexity of the PSP problem, the use of critical biological constraints to reduce the search space is fundamental. Similarly to the *constraint* definition proposed by (Worth et al., 2009), we define structural constraints as a two-parts structural conformational condition to evaluate the acceptance of the folded state of a given 3D structural model. The protein folding process is guided by the combination of a series of different physical-chemical, thermodynamics and interaction processes (Osguthorpe, 2000). In addition to that, the native protein structure is not a unique static conformation (Anfinsen, 1973). Hence, the constraints do not restrict the model assembly, but rather serve as guidance to the building process and as an evaluation of the folding degree. For this work, we opted to incorporate two constraints into the structural model's generation: first the torsion angle preferences of each amino acid residue under a given secondary structure (Ligabue-Braun et al., 2018), and second, the predicted 3D contacts between pair of amino acid residues within the protein sequence (Weigt et al., 2009). The first constraint is responsible for restraining the freedom of dihedral angles within an amino acid residue under a particular secondary structure. The torsion angles adopted by those residues under α-helices, for example, tend to assume approximately PHI = -60, PSI = -60 (Richardson, 1981). The torsional preferences of β-sheets lies on the upper left quadrant of the Ramachandran's plot, which proportionate higher freedom. The dihedral angles are not the only factor working upon the β-sheets assembly, since they also require the spatial proximity of sequence fragments, not necessarily close to each other within the amino acid sequence (Richardson, 1981). The second constraint relies on the premise that throughout the evolutionary scenario, residues that are preserved or co-mutated tend to play a critical role in the protein's structural stability (De Juan et al., 2013). Thus, through analytical inspection upon MSA outputs, one can infer 3D proximity between a pair of amino acid residues. By defining these two constraints, the generation process of structural models is under two critical features, the stable assembly of secondary structures and the spatial proximity of critical amino acid residues.

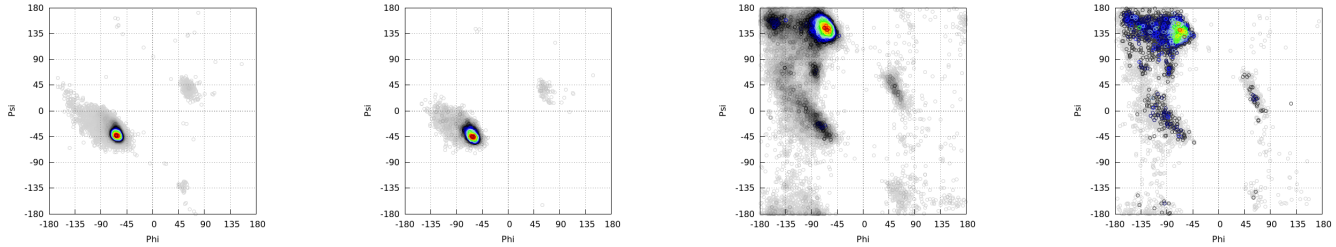
2.1.1. Angle Probability List (APL) -

Proteins can assume a vast conformation diversity, that can be described by their PHI (ϕ) and PSI (ψ) torsion angles. These angles may assume any value within the range [-180°, 180°], except those combinations culminating in steric interference between atoms from the main- and side-chain (Hovmöller et al., 2002). In addition to that, (Ligabue-Braun et al., 2018) have shown that amino acids present different conformational patterns depending on the secondary structure upon the given amino acid residue (Figure 1). The web-tool *Neighbors Influence of Amino Acids and Secondary Structures* - Server[3] (NIAS-Sever) generates the angle probability lists (APLs) (Borguesan et al., 2017). Each APL consists of a matrix $H_{aa,ss}$ of $[-180°, 180° \times -180°, 180°]$ cells, containing the relative frequency of the ϕ/ψ torsion angles combination observed in the experimental data deposited in the PDB. For each amino acid residue (aa) under a specific secondary structure (ss), the frequency of each pair of torsion angles ($i \leq \phi < i+1$, $j \leq \psi < j+1$) is calculated following Eq.1. Higher frequencies are assumed to represent those pairs of ϕ/ψ more frequently found in nature (Borguesan et al., 2015).

$$APL_{aa,ss}(i,j) = \frac{H_{aa,ss}(i,j)}{\sum(H_{aa,ss})} \tag{1}$$

The NIAS-Server utilizes a set of proteins containing 11,130 experimentally determined 3D structures deposited in the PDB. The data was filtered in a way that only those determined by X-Ray Crystallography with a 2.5Å resolution or lower, R factor less than 20% and only one structure was considered for those structures with sequence identity higher than 30%. This filtering process resulted in a 5,255,768 amino acid residues with occupancy equal to 1 to generate the APLs (Borguesan et al., 2017). The tool supports the generation of four different types of APL: (1) APL-1, only the amino acid residue under a particular

[3] http://sbcb.inf.ufrgs.br/npas



(a) ALA - α-Helix (b) TRP - α-Helix (c) ALA - Coil (d) TRP - Coil

Figure 1: `APL` generated for *Alanine* (ALA) and *Tryptophan* (TRP) under α-helix and Coil secondary structures. (a) APL for ALA under α-helix; (b) APL for TRP under α-helix; (c) APL for ALA under Coil; and (d) APL for TRP under Coil.

secondary structure; (2) APL-2, considering only one neighbor (left or right); (3) APL-3, which considers both, left and right, neighbors together. The last type (4) is $\text{APL}_{centroid}$ that analyzes the neighbors within the range five (APL-5), seven (APL-7) and nine (APL-9) regarding their secondary structure and only the central amino acid residue (Borguesan et al., 2017). Similarly to the Ramachandran plot (Ramachandran, 1963), the `APL` can be visualized by a heatmap with the ϕ and ψ angles at the axes and plotting their combination (Figure 1). The difference is that the Ramachandran-like plots generated by NIAS represent the angle combination of every ϕ/ψ pair for a given amino acid under a certain secondary structure from the tool's curated data set.

2.1.2. Protein Residue Contact -

The conservation of the 3D structure among proteins sharing the same or similar function is well established throughout the evolutionary scenario. On the other hand, the amino acid sequence of these evolutionary-related proteins is more mutable when compared to the structure (Illergård et al., 2009). Mutations on the active site, or structurally critical residues, may lead to complete protein function loss. Amino acid residue replacement upon residues that are not critical could present destabilizing effects or non-usual interactions between other amino acid residues (Zerihun and Schug, 2017). These mutations are accepted and conserved if they have positive or no impact upon the protein's stability and function. The DNA sequencing technology has been enhanced in the last few years, culminating in a massive growth of data generated by high throughput sequencing methods (Mukherjee et al., 2018). According to the latest PFAM[4] database release, there are more than 17,000 different protein families, with some of them containing thousands of sequences (El-Gebali et al., 2018).

The analysis of MSA output, where a pair of columns present conserved amino acid residues or a specific pattern of substitution, indicates the co-evolutionary relation between amino acid residues. Correlation occurs in two different ways: (1) Direct, where two amino acid residues are correlated without any additional influence; and (2) Indirect, when both residues are said to be in contact due to the influence of a third residue (Weigt et al., 2009). The correct disentanglement between direct and indirect coupling is crucial to the PSP problem, since amino acid residues said to be under direct coupling are more reliable to be in spatial proximity within the protein 3D structure (De Juan et al., 2013). Mutual information is not suitable to differentiate these contacts, since it is a local measure, not considering any other amino acid residue but the coupling pair, incapable of identifying indirect contacts, also ignoring the biochemical characteristics of amino acid residues, therefore not considering the similarity between the amino acids (De Juan et al., 2013; Zerihun and Schug, 2017).

Weigt et al. (2009) proposed the Direct Coupling Analysis (DCA), an approach that utilizes inverse statistical inferences to calculate the parameters of an Energy function, which is a simplified form of a Boltzmann's function, to disentangle direct and indirect contacts from a MSA output (Figure 2). Since then, the number of different approaches to improve the differentiation of direct coupling residues from

[4] `https://pfam.xfam.org/`

indirect is increasing not only in number, but also in accuracy as shown by the results of the Critical Assessment of Protein Structure Prediction 12th edition (CASP12), where the contact prediction accuracy was twice as high compared to CASP11 (Schaarschmidt et al., 2018). In the CASP's latest edition (CASP13), the winner's methodology included the residue contact inference into its method, together with the torsion angles information[5]. The contact prediction output can be expressed either in a list of higher to the lower probability of the possible couplings or in a $L \times L$ matrix, being L the protein sequence length (Figure 2).

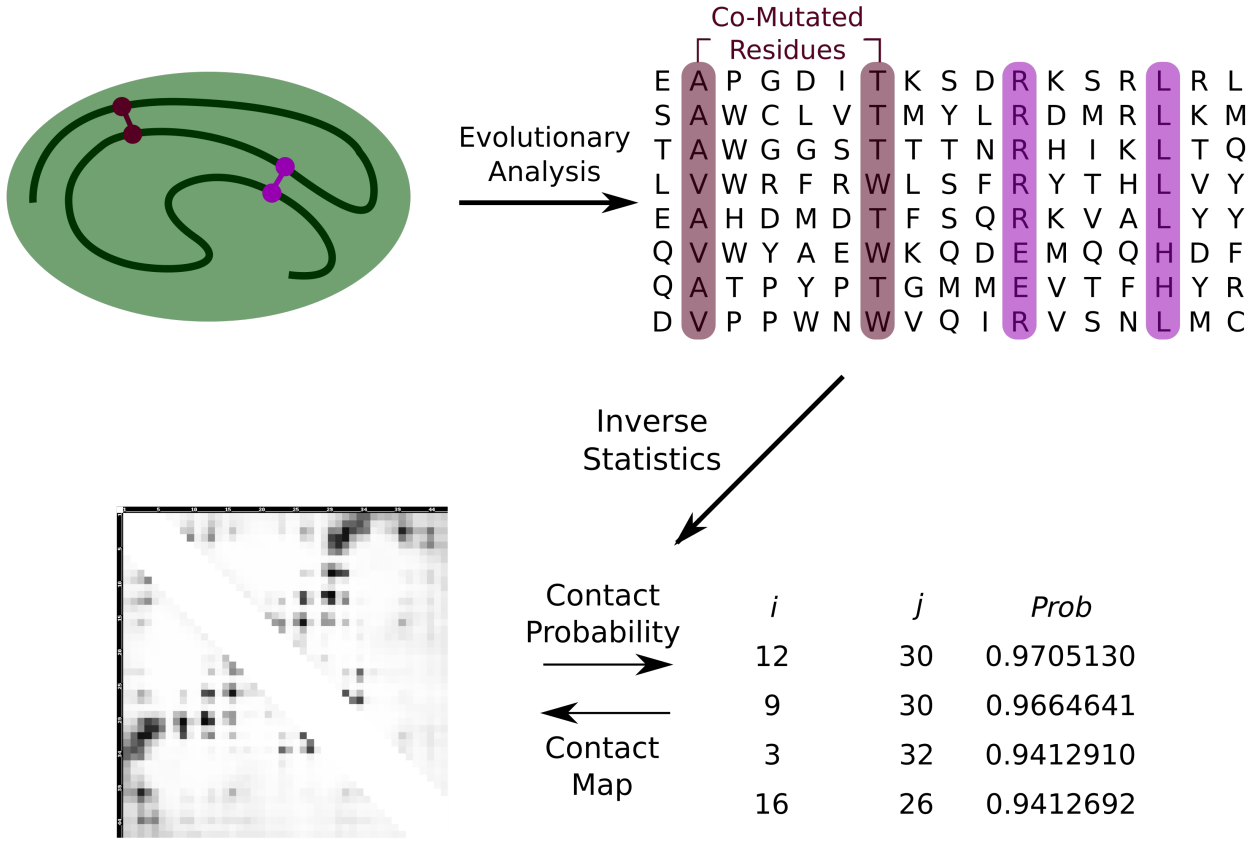


Figure 2: (Upper Left) Representation of a protein with two pair of amino acid residues. (Upper right) The MSA of the protein's family, highlighted the residue pairs shown on the protein structure. After the inverse analytic upon the MSA output the coupled residues can be visualized with a contact map (lower left) and/or a list describing the residues and the probability of being in contact.

2.2. PyRosetta's Energy Score Function

The evaluation of the structural models was performed utilizing the *PyRosetta*[6] (Chaudhury et al., 2010), a Python-based interface to the state-of-art molecular modeling software Rosetta (Rohl et al., 2004). Therefore one can assume that both energy functions, from PyRosetta and Rosetta, are equivalent. Two distinct energy functions are available: centroid and full-atom. The first treats the amino acid side-chain by its center of mass, while the second considers the side-chain with atomic-level detail. The models were then evaluated under the full-atom score energy, REF15 energy function (Alford et al., 2017). The full-atom resolution is required mostly because of the contact constraint used, which calculates the distance

[5] `http://predictioncenter.org/casp13/index.cgi`
[6] http://www.pyrosetta.org/

between the β-Carbon atoms of the coupling residues. The REF15 sums up 19 terms, including nonbonded atom interactions, hydrogen and disulfide bonds, torsional preferences, and other terms important to native structure recapitulation (Alford et al., 2017).

The attractive and repulsive forces acting upon the pair of atoms, due to Van der Waals interactions, are calculated using the Lennard-Jones (LJ) 6-12 potential (Jones and Chapman, 1924). The nonbonded electrostatics interactions of fully and partially charged atoms are based on the Coulomb's law taken and adjusted from CHARMM (Brooks et al., 1983; Park et al., 2016). The solvation approximations are estimated using a bulk water model based on Lazaridis-Karplus implicit Gaussian exclusion model (Lazaridis and Karplus, 1999). The hydrogen bonding is calculated considering both, the electrostatic term and the evaluation of high-resolution crystal structures (OMeara et al., 2015). Knowledge-based terms are also used in the Rosetta energy function. The backbone torsion angles ϕ/ψ energy is calculated based on Ramachandran maps, and side-chain conformations are calculated based on the probabilities according to the 2010 backbone-dependent rotamer library (Shapovalov and Dunbrack Jr, 2011). The total final Energy calculated by PyRosetta is the result of the sum of all weighted terms, following the standard weights for the REF2015 function (Alford et al., 2017).

Moreover, two additional terms were added to the total energy score to reinforce structural features. The first is the Solvent Accessible Surface Area (SASA) measuring of the exposure of the atoms to the solvent molecules (Connolly, 1983; Lee and Richards, 1971). The lower the SASA score, the fewer the atoms exposed to the solvent, which indicates a higher folded state of the protein structure. The SASA score was calculated using the PyRosetta library, with the atomic radius of 1.5Å. The secondary structure (SS) term is calculated as follows. For each amino acid residue within the model, a negative constant is added to the term every time that the model i_{th} residue matches the secondary structure of the i_{th} residue of the input protein. On the other hand, for each residue mismatch, a positive constant is added to the SS term. The secondary structure was assigned under the DSSP (Kabsch and Sander, 1983) algorithm implementation of PyRosetta. The final energy score used to evaluate our candidates is the sum of PyRosetta energy score and the two additional terms, shown by Equation 2.

$$E_{total} = E_{PyRosetta} + E_{SS} + E_{SASA} \tag{2}$$

It is worth to mention that the biochemical rules that govern the protein folding are only partially known; therefore there is no energy function capable of describing the potential energy of a real system. Due to these barriers, different energy functions can lead to different outputs (i.e. different final structures).

The final structure is compared to the experimentally determined structure using the Root Mean Square Deviation (RMSD). The RMSD calculates the distance, measured in angstroms (Å), between the atoms of two structures. The closer to 0Å the most similar are the two structures. In Equation 3 is shown how the RMSD is calculated, where r_{ai} and r_{bi} are the ith atom in a group of n atoms from two protein structures.

$$RMSD(a,b) = \sqrt{\frac{\sum_{i=1}^{n} |r_{ai} - r_{bi}|^2}{n}} \tag{3}$$

2.3. Differential Evolution

The PSP problem is classified into the NP-Complete category according to the computational complexity theory (Crescenzi et al., 1998). Thus, the usage of optimization algorithms are suitable to search the conformational space to find the conformation with the lower energy score of a given target protein. The *Differential Evolution* (DE) algorithm is an evolutionary algorithm, that was proposed by Storn and Price in 1997 (Storn and Price, 1997). The process implemented by this method can be separated into four main steps: initialization, mutation, crossover and selection. The algorithm is composed by a population P with N individuals, where each individual is a vector with D dimensions for each generation G.

$$x_{i,G}, i = 1, 2, ..., N \tag{4}$$

New individuals are generated by the weighted differences between two distinct individuals, through the mutation operation. The crossover operation is employed to the resulting mutated individual by mixing the parameters, creating the so-called trial vector.

Mutation. For each individual in the population, a new mutant individual v is generated following

$$v_{i,G+1} = x_{r1,G} + F \cdot (x_{r2,G} - x_{r3,G}) \tag{5}$$

where $r1, r2, r3$ are mutually different indexes randomly chosen from $\{1, 2, .., N\}$. The indexes must also be different from the target i index, therefore DE is only possible for $N \geq 4$. F is the amplification factor of $(x_{r2,G} - x_{r3,G})$, being $F \in [0, 2]$.

Crossover. This operation is responsible for the increase of the diversity throughout the execution. A trial vector $t_{i,G}$ is generated by mixing a set of dimensions from the mutated vector $v_{i,G+1}$. Usually, the crossover operation follows a binomial scheme. Thus, the dimension $d \in D$ is mutated if a randomly generated number is lower than the crossover rate (CR) constant ($CR \in [0, 1]$), or $d = d_{random}$. The operation is performed as follows

$$t_{i_d,G} = \begin{cases} v_{i_d,G} & \text{if } d = d_{rand} \text{ or } rand[0,1] \leq CR \\ x_{i_d,G} & \text{otherwise} \end{cases} \tag{6}$$

Selection. After these two operations explained above, the selection operation is performed to evaluate if the trial vector $t_{i,G}$ will be part of the population in the next generation. The individual is evaluated following a function $f(x)$, which is also the score function to be minimized. This way, the offspring is composed by the most fit individual resulting from the operations, and the population of the next generation is composed by the most fitted individual from the previous generation. The selection operator is described as

$$x_{i,G+1} = \begin{cases} t_G & \text{if } f(t_G) \leq f(x_{i,G}), \\ x_{i,G} & \text{otherwise} \end{cases} \tag{7}$$

The Algorithm 1 describes the pseudo-code for the classical implementation of a *Differential Evolution* algorithm employed to a minimization problem.

Data: N, F and CR
Result: Best individual, which correspond to the lower score achieved by the optimization process.
Initial Population generation with N individuals;
while $g \leq$ *total number of generations* **do**
  for *each individual i in population P* **do**
    Randomly select three individuals (r1,r2,r3);
    $d_{rand} \leftarrow$ random dimension to mutate;
    for *each dimension d* **do**
      if $d = d_{rand}$ **or** *random[0,1]* $\leq CR$ **then**
        $t_{i,d} \leftarrow x_{r1,G} + F \cdot (x_{r2,G} - x_{r3,G})$
      else
        $t_{i,d} \leftarrow x_{i,d}$
      end
    end
    if $f(t_i) \leq f(x_i)$ **then**
      offspring.append(t_i)
    else
      offspring.append(x_i)
    end
  end
  P $\leftarrow$ offspring g $\leftarrow$ g+1
end

Algorithm 1: Classic Differential Evolution.

Due to the multimodal characteristics of the PSP problem, it is known that energy functions are not completely able to describe the native state of the 3D conformation of a given structure (Kim et al., 2009). The Rosetta's energy function is composed by the sum of several different weighted terms (Alford et al., 2017), which are opposite to each other, for example, the bonded and non-bonded terms. Thus, to try to fully explore the advantages of using individuals generated by the proposed method, we opt for using a multi-objective algorithm (Konak et al., 2006). Throughout the optimization process the energy function presented previously (Equation 2) doesn't ensure that the contact information is preserved. Therefore, we opt to use the contact energy term proposed by (Hong et al., 2018), to reinforce the contact between coupled amino acid residues during the optimization process and not only when generating the initial population. The contact energy term is defined as follows

$$E_{contact} = \sum_{ij}^{Nc} \begin{cases} \frac{(r_{ij}-r_{cut})^2}{(r_{ij}-r_{cut})^2+\sigma^2} c(p), & \text{if } r_{ij} \geq r_{cut} \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

$$c(p) = ap + 1 \tag{9}$$

where Nc is the number of contacts predicted with a probability > 0.2, these contacts are also long and medium distance contacts, while i and j are the amino acid residues composing the predicted pair. The distance r_{ij} correspond to the Euclidian distance between the Carbon C_β (C_α for Glycine) atoms from each pair i and j. The r_{cut}, σ e a values were previously calculated by the authors and are described in Table 1.

r_{cut}	σ	a
7.840362	0.084674	0.156446

Table 1: Terms composing the Contact Energy Function previously calculated by the authors in (Hong et al., 2018).

The Differential Evolution Multi-Objective (DEMO) (Robič and Filipič, 2005) was chosen to optimize the population according to both energy functions (Equation 2 and 8). It is worth to mention that, when using a single-objective algorithm, the substitution of individuals from one generation to another is based on the score achieved by the current individuals versus the mutated one (Algorithm 1). Meanwhile, for multi-objective algorithms, the evaluation of which individual is preserved for the next generation is not simple. Thus, if the mutated individual dominates the previous individual, the new one is preserved for the next generation, otherwise the mutated is discarded and the previous is maintained within the population. However, if both individuals are non-dominant to each other, both of them are preserved for the next generation (Robič and Filipič, 2005). This way, at the end of every generation the population size vary between N and $2N$. To reduce the population size back to N a ranking based on non-dominated solutions together with a point-clustering distance method, to ensure a greater diversity, is used (Deb et al., 2002). The freely available functions implemented by jMetalPy [7] (Benitez-Hidalgo et al., 2019) were adapted to our algorithm. jMetal is a Python-implemented group of functions for multi-objective metaheuristics.

The individuals from both optimization process were evaluated according to the Global Distance Test (GDT) (Zemla et al., 1999). The GDT measures the distance between atoms, similarly to the RMSD, but it allows the atoms to be superimposed under different cutoffs (1, 2, 4 and 8 Å), which is normalized by the number of amino acid residues composing the target protein (Abriata et al., 2018a).

3. Related Works

de Lima Corrêa et al. (2018) utilized the torsion angles described by the `APL` to build the individuals to compose the populations within the proposed Memetic Algorithm. Similarly, Narloch and Dorn (2019)

[7] https://github.com/jMetal/jMetalPy

incorporated the information contained in the APL to generate the structural models, which were optimized through a DE algorithm. In Borguesan et al. (2018) the APL, together with a fragment library, was incorporated on the assembly of initial structural models and throughout the optimization process. Borguesan et al. (2015), de Lima Corrêa et al. (2018) and Narloch and Dorn (2019) used the torsion angles information retrieved from the APL to guide the creation of individuals to compose optimization algorithm's populations. Even though these works analyzed different algorithms, all of them exceeded results compared to the random generation of individuals. The works mentioned above indicate that the usage of problem-domain information helps search algorithms to achieve better results, i.e., lower energy scores and structural features that are closer to those of the experimentally determined structure.

Zhang et al. (2018) used the predicted contacts to guide the generation of structural assemblies trough simulations, as well as a score function to sort the best candidates based on the residue contacts. The function penalizes those solutions that strongly violate the predicted-to-be-in-contacts residues. The contact information was used by (Ovchinnikov et al., 2018) to identify possible domains within the protein structure. The predicted contact map was aligned against the contact maps generated from the protein structures from a non-redundant set of PDB. Those maps with highest scores were then selected as template for modeling. (Hong et al., 2018) incorporated the predicted contact information as an additional energy term, together with a secondary structure term and a solvent accessibility term. In Hopf et al. (2012) the contact information was used as a weight in the structural optimization process.

4. Results

A data set containing 21 proteins with experimentally determined structures were selected from the PDB to evaluate the influence of the biological constraints (Figure 3). The chosen proteins were selected due to their fold diversity, resolution < 2.5Å, number of amino acid residues and different determination methods (X-Ray Crystallography and NMR). In Table 2 we present the PDB ID, sequence length (i.e., number of amino acid residues) and the secondary structure composition of each protein selected to compose our data set.

PDB ID	**Length**	**SS**	**PDB ID**	**Length**	**SS**
1AB1	46	$\alpha/\beta^2/\alpha$	1UTG	70	$\alpha/\alpha/\alpha/\alpha$
1ACW	29	α/β^2	1WQC	27	α/α
1AIL	70	$\alpha/\alpha/\alpha$	2JUC	55	$\alpha/\alpha/\alpha$
1CRN	46	$\alpha/\beta^2/\alpha$	2MR9	44	$\alpha/\alpha/\alpha$
1D3Z	76	β^5/α	2P5K	63	$\alpha/\alpha/\alpha/\beta^2$
1D5Q	27	α/β^2	2P6J	52	$\alpha/\alpha/\alpha$
1DFN	30	β^3	2P81	44	$\alpha/\alpha/\alpha$
1ENH	54	$\alpha/\alpha/\alpha$	2PMR	76	$\alpha/\alpha/\alpha$
1FNA	91	β^3/β^4	3V1A	48	α/α
1Q2K	31	α/β^2	5JZR	131	β^5/α
1ROP	56	α/α	–	–	–

Table 2: Proteins composing the data set. **SS** represents the secondary structures description, whereas α refers to helices and β to sheets. The superscript number on β refers to the number of strands within the sheets.

4.1. Individual assembly

To test the structural model assembly, for each protein in our data set, four different populations were generated, each one of them composed by 10,000 individuals. The first population is the randomly-generated, i.e., each consists of a list of n (n = sequence length) torsion angles ($[\phi, \psi, \omega, \chi(z)] \times n$), with values within the range [-180°, 180°] without any restriction, including those presenting steric interference between the atoms of the residues main and side-chains.

78

The second population was created utilizing the torsion angle information retrieved from APL. For each protein the APL-1, -2 and -3 were generated using the NIAS-Server (Borguesan et al., 2017) (Figure 3). To avoid biased result, PDB entries presenting more than 70% of sequence similarity with the target protein sequence were removed from the NIAS-Server data set. The APLs were generated utilizing the secondary structure assigned by the DSSP algorithm. For each residue, each APL has a given probability of being chosen, as demonstrated in Table 3. Due to the lack of a left and right neighbor, the first and the last residues do not present APL-3, therefore the selection follows the probabilities showed by Table 3. The larger the window considered to build the APL, the fewer information is available, therefore if the NIAS-Server is not able to generate the APL-3 for a given residue, the APL-2 probability of being chosen is the sum of both APL-2 and APL-3 probabilities. If the APL-2 is selected, both RIGHT and LEFT have 50% of chance each to be chosen. The torsion angles for each residue within the protein sequence is selected from the chosen APL according to the probabilities calculated by NIAS-Server, i.e. those with more occurrences present a higher chance of being selected.

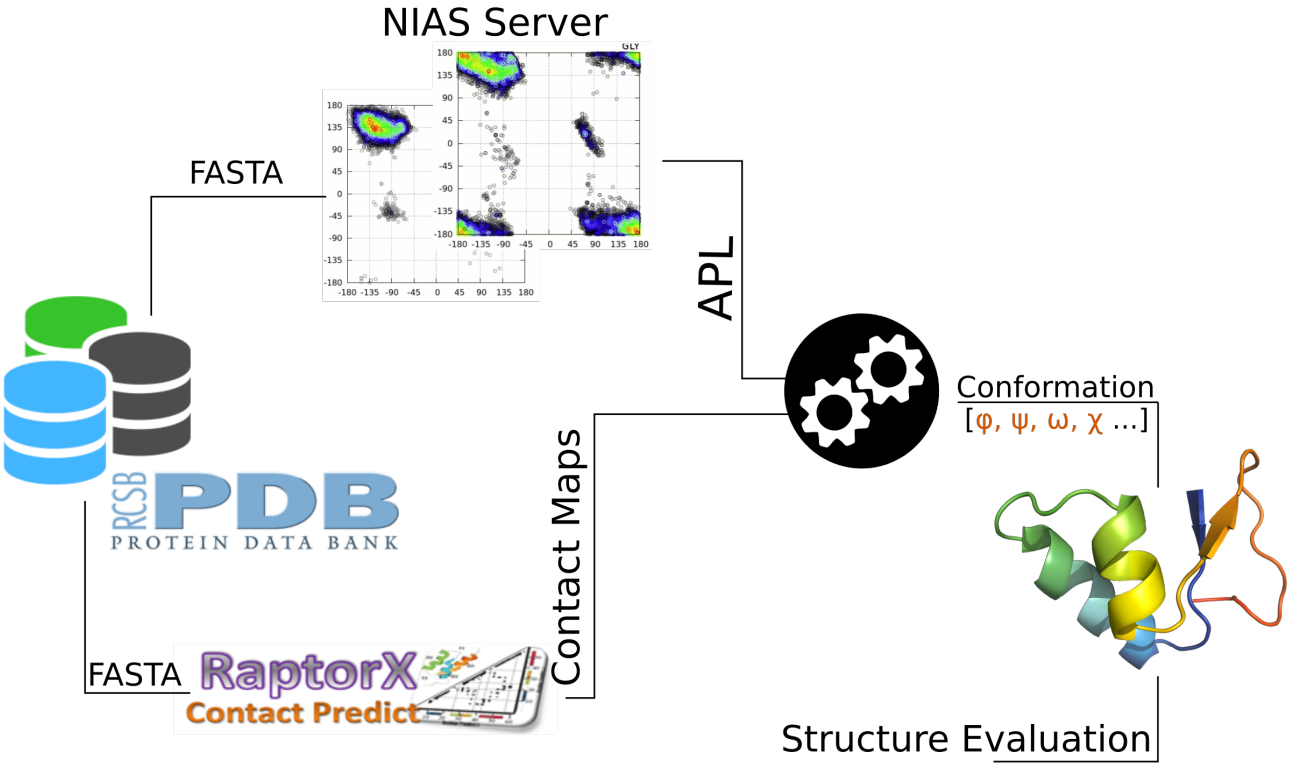


Figure 3: General graphical representation of the population generation method workflow. Starting at the selection of a protein candidate, followed by the `APL` generation and contact map prediction. Structural model assembly by merging both constraints and further structural evaluation.

The third population was generated using the information of `3D` contact between pairs of amino acid residues. The coupling between residues of each protein was predicted using the web server *RaptorX*[8] (Wang et al., 2018) (Figure 3). The residue contact list was filtered according to (Schaarschmidt et al., 2018), by considering only the top $L/5$ pairs of amino acid residues, being L the protein sequence length, separated by

[8] `http://raptorx.uchicago.edu/ContactMap/`

APL	Probability
APL-$1_{i=1,i=n}$	40%
APL-$1_{1<i<n}$	20%
APL-$2_{i=1,i=n}$	60%
APL-$2_{1<i<n}$	30%
APL-$3_{1<i<n}$	50%

Table 3: Table of APL selection probabilities, where i is the i^{th} amino acid residue within a sequence with length equals to n. For the first ($i = 1$) and last ($i = n$) residues, the probabilities are different due to the lack of a left neighbor to the first residue and a right neighbor to the last amino acid.

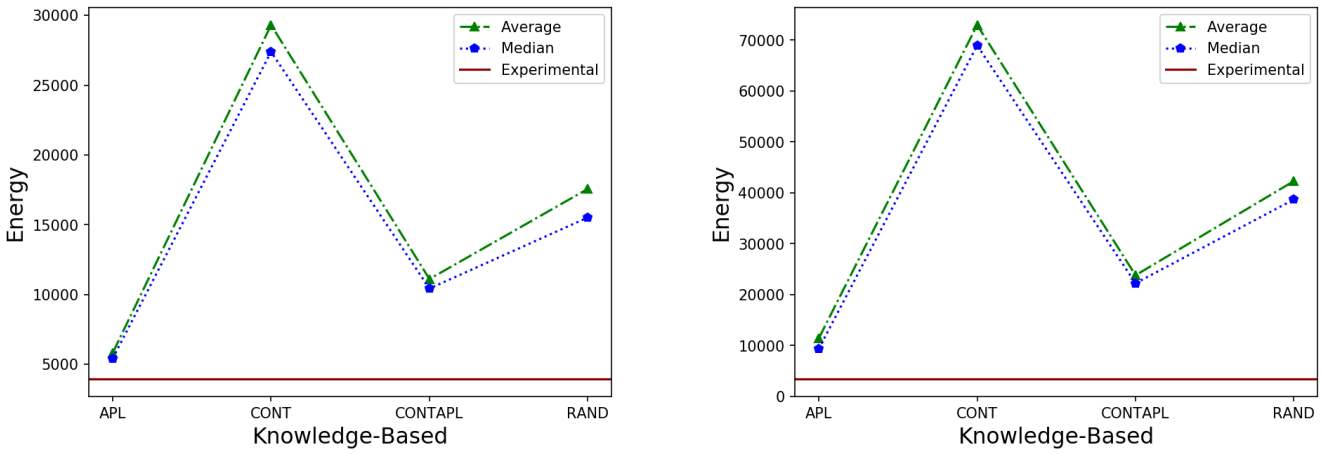
more than eleven residues within the protein sequence. Similarly to the first population, the torsion angles were randomly generated within the range [-180°, 180°], but the individual was only accounted if and only if it had at least one predicted pair of residues in contact. Here we consider two amino acids to be in contact if the Euclidean distance between their C_β carbon atoms (C_α for Glycine) is lower than 8Å (Schaarschmidt et al., 2018). The individuals of the last population were created under the merge of both constraints used to build the second and third population. The torsion angles were selected from the `APL`, following the probabilities listed in Table 3, and if and only if the contact constraint was also fulfilled then the individual was accounted for the population.

4.2. Structural Model Evaluation

4.2.1. Free Energy evaluation

The 3D structure of defensin HNP-3 experimentally determined by X-Ray Crystallography (PDB ID: 1DFN (Hill et al., 1991)) was chosen due to its secondary structure. The tertiary structure is assembled by a single anti-parallel β-Sheet, containing three strands, as well as three disulfide bonds. The quaternary structure is composed of two chains, but only one chain was considered to generate the APL and the contact maps. Although the efficiency of contacts to help the assembly of protein complexes has already been proved, shown by (Hopf et al., 2014), here we focus on the tertiary structure assemble regardless of the inter-molecular interactions. The engrailed homeodomain protein under the PDB ID 1ENH (Clarke et al., 1994) was also used to evaluate the proposed methods. Its secondary structure is composed of three α-helices.

Here we follow the thermodynamic hypothesis postulated in (Anfinsen, 1973), where the conformation of a particular protein is given by the lowest free energy achieved by the system. Therefore, the lower the energy, the more stable and adequate we consider the structural model. The α-helices are the most regular secondary structures, with the backbone torsional angles ϕ/ψ ranging around $-60°/-60°$ respectively (Richardson, 1981). The β-sheets, despite presenting a broader range than the helices, they also fall within a limited range combination of ϕ/ψ within the upper left quadrant of the Ramachandran plot (Richardson, 1981). We attribute the lower energy score achieved by all proteins within our dataset build with the APL to these observations. For both proteins the lowest mean energy was achieved by the APL (Figure 4) due to the secondary structure assembly reinforcement given by the torsion angles. The higher energy scores achieved by the contact-guided populations may be the result of superimposed atoms, since this method only guarantees a 3D approximation of residue pairs, and does not avoid atomic overlap, which results in higher energy scores. The combination of APL and residue contact resulted in individuals respecting the predicted contacts between residue pairs, based on evolutionary analysis, and the torsion angles provided by the APL. The combination of these two constraints was able to reduce the energy scores, which means that the models generated by these constraints together present higher stability combined with a higher folded state compared to the usage of each constraint separately. This was also found to be true for all proteins within the data set (Figure S1-S21).



(a) 1DFN (b) 1ENH

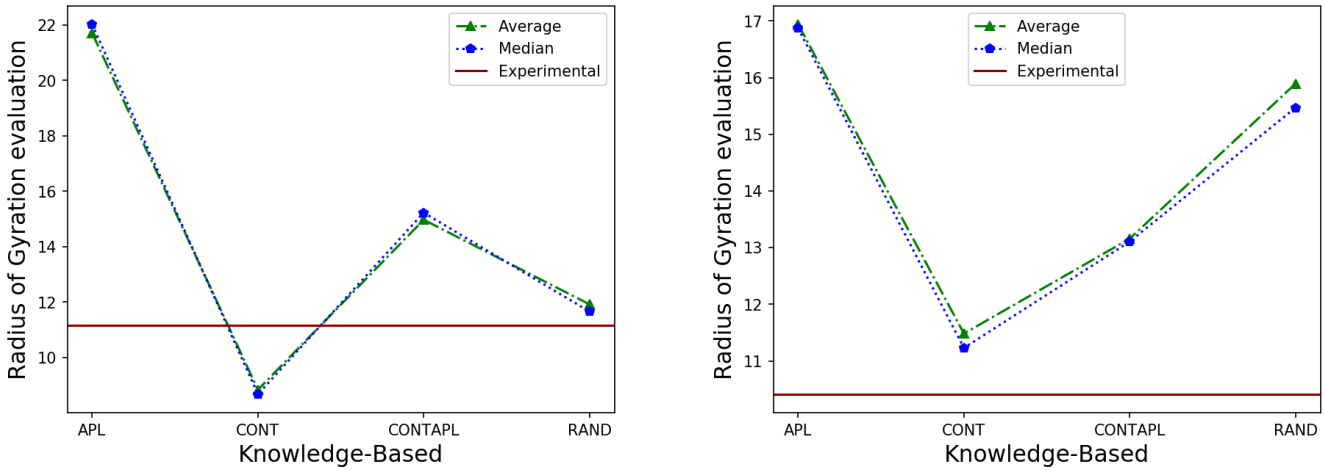
Figure 4: 1DFN and 1ENH population mean free energy score achieved by the Random (RAND), APL, Contacts only (CONT) and Contact and APL merged (CONTAPL). The dark red horizontal line represents the score calculated for the experimentally determined 3D structure.

4.2.2. Radius of Gyration evaluation

The Radius of Gyration (RG) measures the quadratic mean distance of the protein atoms to its center of mass. This metric is used to describe the folding state of a given protein, therefore the lower the RG value, the closer the atoms are to the center of mass (Lobanov et al., 2008). The RG of the experimentally determined structure of each protein was calculated using the implementation provided by PyRosetta and was taken as a reference to the folding state of the models. Here we consider that the models with similar values of RG are closer to the experimentally folded state, therefore corresponding to a structure more similar to the crystallographic one. The initial hypothesis is that the contact constraint would be responsible for forcing the molecules to fold, but not for preventing the clashing of atoms. The torsion angle information is not enough to fold the protein in a native-like conformation. Therefore the merging of APL and contact information is expected to present structural models with a lower RG and higher stability. In agreement with our first hypothesis, the population utilizing the contact constraints presents the lowest RG, for both 1DFN and 1ENH (Figure 5), as well as for the almost all other proteins within the data set (Supplementary Data). The APL population presented the highest RG values for both proteins, except for the 2MR9 protein, which the random population had the highest RG. The major β-sheet secondary structure of the 1DFN may explain the inability of the torsion angle information to fold the model to a closer native-like conformation (Figure 5a). This was expected, due to additional condition of spatial proximity between two distinct sequence fragments to the β-sheets formation (Nelson et al., 2008). The APL information is effective for the local arrangement formation, but lacks the ability to approximate distinct sequence fragments. The merge of APL and contacts was successful to approximate the individuals closer to the experimentally determined structure. As for the 1ENH, the contact information together with APL was able to lower the mean RG of the population (Figure 5b). Thus, we can infer that we were able to generate structural models with a higher folded state with lower energy scores.

4.2.3. Solvent Accessible Surface Area evaluation

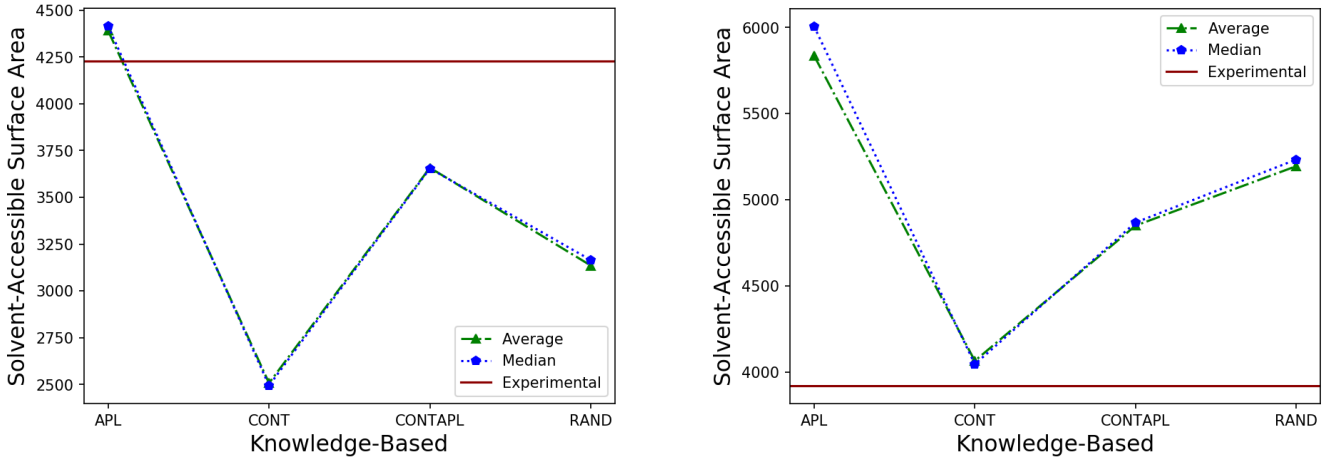
The SASA parameter measures the energy from the interactions between the atoms of the solute and the solvent; therefore here we assume that similar scores indicate a comparable 3D organization of the amino acid residues. The SASA was calculated using the PyRosetta implementation, which was also utilized to reinforce the Energy score. The good APL influence upon the SASA score for the 1DFN (Figure 6a) can be explained by the difficulties of the APL to generate individuals with closer protein segments, necessary for β-Sheets. Thus, it prevents the disordered over-approximation of segments of this small protein. Whereas for 1ENH, despite the APL being able to induce the assembly of stable helices, the folding process is more difficult,



(a) 1DFN (b) 1ENH

Figure 5: 1DFN and 1ENH population mean Radius of Gyration score achieved by the Random (RAND), APL, Contacts only (CONT) and Contact and APL merged (CONTAPL). The dark red horizontal line represents the score calculated for the experimentally determined 3D structure.

much because of the highly dense map generated for *Coil* residues, as shown by Figure 1c. The freedom provided by the APL for *Coil* residues is equivalent to the random generation. Therefore the rotation of segments within the protein structure results in the exposure of residues to the solvent, increasing the SASA score. Similar to the RG analysis, the merge of the APL and contact does improve the scores, even though they are not as close to the native structure as those individuals created only with contact information.



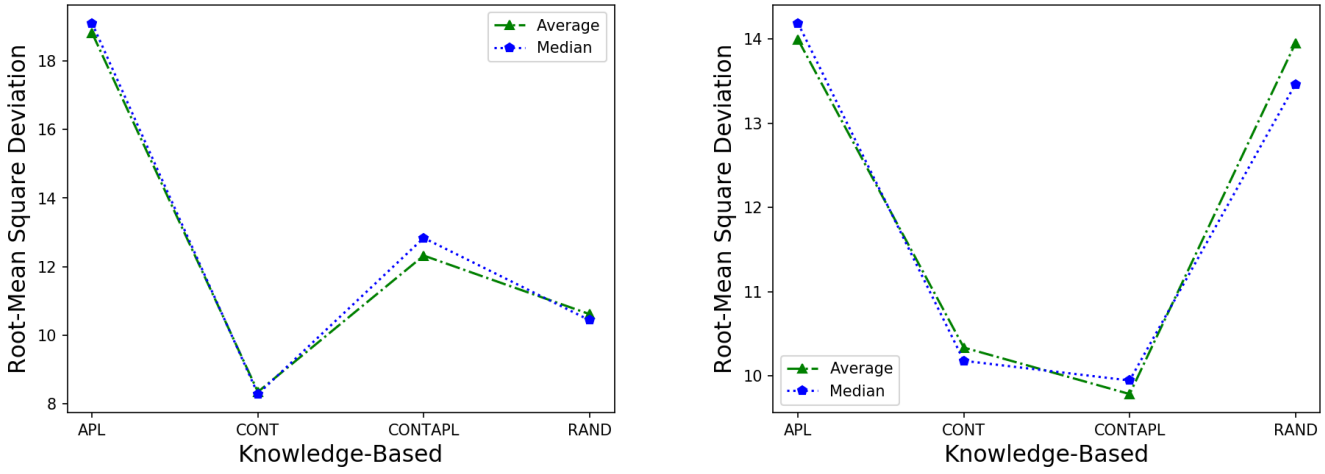
(a) 1DFN (b) 1ENH

Figure 6: 1DFN and 1ENH population mean Solvent Accessible Surface Area score achieved by the Random (RAND), APL, Contacts only (CONT) and Contact and APL merged (CONTAPL). The dark red horizontal line represents the score calculated for the experimentally determined 3D structure.

4.2.4. Root-Mean Square Deviation evaluation

The success to obtain structural parameter values, such as Free Energy, RG and SASA, close to the native structure is a useful metric to evaluate the structural candidates. The variety of 3D structures a single amino acid sequence can assume makes it possible for two distinct structures, formed by the same protein sequence, to have similar structural features, resulting in analogous RG and SASA scores between

the structures, and also low energy scores, e.g., if the 3D structure does not disrespect any physical-chemical constraint. Therefore the RMSD score was utilized to ensure that the structural models were similarly arranged when compared to the crystallographic structure. Once again, the residue contact itself presented the lowest RMSD scores for the 1DFN, much of it due to the induced folding. The RMSD of 1ENH structure, when using only contact constraint or merging both constraints, was similarly lower (Figure 7b). The 1DFN models were not as effective as the 1ENH using the two constraints, but the lower RMSD when compared to the APL highlights its usefulness (Figure 7a). The close RMSD and structural features scores to the native structure indicate that the merging of both constraints is effective.



(a) 1DFN (b) 1ENH

Figure 7: 1DFN and 1ENH population mean Root-Mean Square Deviation score achieved by the Random (RAND), APL, Contacts only (CONT) and Contact and APL merged (CONTAPL).

Both protein structures, 1DFN and 1ENH, achieved lower energy values when generating the models under the APL constraint (Figure 7a) due to the well-guided secondary structure assembly. In opposition, when the folding parameters have evaluated the use of only contact information they were enough to achieve near-native features, including the RMSD scores. These may seem contradictory, but the lower folding scores and the higher energy score indicated the existence of atom overlap. The lower structural features score and energy achieved by the merge of contact information and the torsion angle retrieved from the APL suggest that these constraints are able to guide the individual assembly with higher structural stability and similarity to the experimentally determined structure. This behavior was observed for the majority of the proteins tested in our data set[9].

It is important to highlight that these structural features alone are not capable to fully describe the 3D conformation of protein structure. Thus, we evaluated these parameters regarding the RMSD score for each model generated. The charts are all available in the Supplementary Materials. The plots show that the contact information merged with the APL torsion angles are able to generate models with higher stability and folding state more similar to the experimentally determined structure. To some cases, for example 1ROP (Figure S11) the initial population is composed by structural models with structural features almost identical to the experimental, without any optimization step. Therefore, we can assume that the merge of both constraints is responsible for the generation of good-quality structural models.

4.3. Energy Score Optimization

The models generated by the two constraints were then submitted to a optimization process. To test the improvement of structure prediction accuracy, caused by the high quality initial individuals, we set a

[9]The charts containing the evaluation for the remaining proteins within our data set are explicit in the Supplementary Materials

sub-sample containing 11 proteins. The chosen algorithm was the *Differential Evolution*, due to its previous efficiently optimization regarding the 3D structure of proteins (Narloch and Dorn, 2019). The constant parameters used by the DE algorithm are described in Table 4. For each structural model generation method to be evaluated, 30 sub-populations with 100 individuals were generated selecting 3.000 different individuals from the 10.000 individuals previously generated by every methods. The populations are mutually exclusive and different from each other, to avoid over-representation of a given individual. The number of fitness evaluation was utilized as the stop criteria, thus 1.000.000 evaluations was set, corresponding to 10.000 generations.

Parameter	Value	Description
P	100	Population size
CR	0.9	Crossover factor
F	0.5	Mutation Factor

Table 4: *Differential Evolution* algorithm parameters used during the optimization process.

Analyzing the general energy scores showed in Table 5, it is clear that the usage of the APL information to generate the initial population is responsible for the lower energy scores achieved by the optimization process. The lowest energy score achieved by all proteins were from populations generated using the APL information. Nevertheless, the RMSD calculated for those structures with the lowest energy score were not ideally close to zero, indicating a higher dissimilarity with the structures retrieved from experimental procedures (Table 6). The lower energy scores achieved by those structures initially assembled using the torsional angles information retrieved from the APL is due to the facht that α-helices present more restrictive combination of angles presented by this secondary structure (Richardson, 1981). Thus, the secondary structure folded at the very beginning, during the generation of the initial population, remained formed throughout the optimization process. In spite of that, the higher RMSD scores may be the result of the inability of the APL information to fold the tertiary structure in a more efficient way. Therefore, the presented results achieved by the individuals generated using the APLs corroborate with the initial hypothesis, however, the contact information seems to be lost during the optimization process since no constraint is considered during the execution of the algorithm. Thus, the changes made upon the angles throughout the execution may be responsible for the unfolded state of the resulting models, causing increased distance of the secondary structure assembled by the APL information.

To reinforce the contact information throughout the optimization process, we utilized the Contact Energy function proposed in (Hong et al., 2018), described by Equation 8. The multi-objective optimization performed by the DEMO algorithm has as its objectives the Free Energy and the Contact Energy. Here we chose the best individual from run of the optimization process the structure corresponding to the elbow of the Pareto frontier. Thus, we ensure the selection of the individual with the best combination of both objectives, otherwise if another structure is selected it will bias the result to one objective over the other one. The results showed by Table 7 indicates that, even with higher energy (Equation 2) when compared to the single-objective DE, the RMSD of the individuals optimized from the contact and APL population are considerably lower. The structures corresponding to the lowest energy found by the 30 runs for each protein are explicit in Table 8, as expected the secondary structures are well folded, specially those with majority of α-helices. Differently from the results showed by Table 6, the structures achieved by DEMO optimization present a higher folding state. For nine out of ten proteins, the population built under the guidance of both restraints achieved better results regarding the structure similarity according to the RMSD calculated using the experimentally determined structure (Table 7). It is worth to highlight that ten out of ten proteins achieved ideal GDT scores > 50, a threshold for good results.

The two winners from the CASP12's PSP category, Rosetta (Song et al., 2013) and QUARK [10] (Xu and Zhang, 2012), were used to compare the results obtained by the proposed method. The Rosetta method is

[10] `https://zhanglab.ccmb.med.umich.edu/QUARK/`

PDB ID	Low Energy ($\overline{x} \pm \sigma$)	Low RMSD ($\overline{x} \pm \sigma$)	High GDT ($\overline{x} \pm \sigma$)
1AB1$_{APL}$	3.68 (4.68±0.45)	**5.97** (11.33±2.0)	**46.52** (37.51±3.27)
1AB1$_{CONT}$	6.93 (7.67±0.23)	8.73 (12.74±2.89)	26.96 (19.25±2.85)
1AB1$_{CONTAPL}$	**3.61** (4.78±0.44)	8.0 (13.02±1.67)	43.48 (38.48±1.85)
1AB1$_{RAND}$	7.16 (7.68±0.24)	10.63 (13.45±1.78)	26.52 (19.19±2.06)
1ACW$_{APL}$	3.32 (3.54±0.08)	**4.99** (8.12±1.58)	**59.31** (43.56±4.62)
1ACW$_{CONT}$	4.07 (4.53±0.16)	6.49 (9.21±1.88)	35.17 (26.25±3.69)
1ACW$_{CONTAPL}$	**3.03** (3.6±0.14)	6.2 (8.33±1.25)	53.1 (43.06±3.87)
1ACW$_{RAND}$	3.97 (4.49±0.17)	6.45 (9.61±2.14)	31.72 (25.06±3.8)
1AIL$_{APL}$	**6.54** (6.8±0.09)	11.6 (18.95±3.67)	48.29 (46.69±3.2)
1AIL$_{CONT}$	12.82 (14.59±0.59)	13.8 (21.35±4.71)	15.71 (12.76±1.49)
1AIL$_{CONTAPL}$	6.79 (6.96±0.18)	**6.45** (6.82±0.75)	**50.29** (46.79±1.76)
1AIL$_{RAND}$	13.55 (14.65±0.51)	14.63 (21.8±4.8)	15.43 (12.74±1.08)
1CRN$_{APL}$	**3.11** (4.61±0.6)	**5.21** (10.97±2.82)	**50.43** (36.26±4.37)
1CRN$_{CONT}$	7.62 (8.04±0.22)	7.35 (14.06±2.78)	23.91 (19.75±2.4)
1CRN$_{CONTAPL}$	3.29 (5.01±0.54)	7.11 (11.26±2.46)	45.22 (36.36±3.55)
1CRN$_{RAND}$	7.42 (8.13±0.29)	9.26 (13.95±2.49)	27.39 (19.58±2.68)
1D5Q$_{APL}$	**2.62** (2.89±0.1)	4.72 (6.75±0.98)	63.7 (54.12±3.85)
1D5Q$_{CONT}$	3.49 (3.7±0.09)	6.12 (8.37±1.15)	34.07 (26.47±4.24)
1D5Q$_{CONTAPL}$	2.72 (2.87±0.06)	**4.22** (6.2±1.39)	**64.44** (55.31±4.54)
1D5Q$_{RAND}$	3.57 (3.71±0.08)	5.21 (7.79±1.28)	44.44 (29.63±6.57)
1DFN$_{APL}$	**4.11** (4.34±0.12)	**5.8** (10.81±2.18)	40.0 (31.31±3.55)
1DFN$_{CONT}$	4.7 (4.96±0.14)	7.68 (11.11±1.64)	34.67 (27.0±4.13)
1DFN$_{CONTAPL}$	4.25 (4.38±0.04)	7.64 (9.92±0.72)	36.0 (33.4±2.87)
1DFN$_{RAND}$	4.53 (4.91±0.19)	7.18 (11.06±2.13)	**46.0** (27.24±4.82)
1ENH$_{APL}$	**5.67** (5.93±0.12)	16.32 (16.52±0.29)	41.85 (41.59±1.03)
1ENH$_{CONT}$	10.84 (11.73±0.39)	11.59 (18.45±4.32)	20.37 (16.2±2.25)
1ENH$_{CONTAPL}$	6.14 (6.63±0.32)	7.32 (9.97±3.25)	**51.85** (50.52±2.72)
1ENH$_{RAND}$	10.45 (11.78±0.5)	**7.22** (17.89±4.21)	18.52 (15.69±2.08)
1Q2K$_{APL}$	**2.98** (3.59±0.15)	**5.64** (8.28±1.46)	**57.42** (48.9±3.23)
1Q2K$_{CONT}$	4.36 (4.56±0.09)	6.9 (10.22±1.99)	41.94 (23.96±4.44)
1Q2K$_{CONTAPL}$	3.27 (3.62±0.11)	7.67 (8.66±0.33)	48.39 (47.85±1.03)
1Q2K$_{RAND}$	4.16 (4.56±0.13)	7.15 (10.44±1.63)	33.55 (23.85±3.64)
1ROP$_{APL}$	5.23 (5.43±0.09)	11.93 (18.27±3.69)	51.43 (49.25±1.54)
1ROP$_{CONT}$	9.89 (10.71±0.34)	10.08 (17.42±3.63)	18.93 (14.31±2.26)
1ROP$_{CONTAPL}$	**5.08** (5.38±0.16)	**2.91** (17.76±8.19)	**70.36** (54.2±8.91)
1ROP$_{RAND}$	10.08 (10.78±0.36)	9.72 (16.18±3.65)	20.71 (14.31±2.77)
1UTG$_{APL}$	8.42 (8.57±0.08)	14.0 (14.35±1.13)	45.14 (42.66±1.56)
1UTG$_{CONT}$	14.97 (16.48±0.53)	11.43 (19.78±4.03)	18.0 (12.92±1.81)
1UTG$_{CONTAPL}$	**7.66** (9.66±0.43)	**7.88** (14.57±3.03)	**51.14** (37.84±3.88)
1UTG$_{RAND}$	15.16 (16.21±0.51)	13.14 (20.22±4.6)	17.71 (13.06±1.95)
1WQC$_{APL}$	**2.21** (2.61±0.18)	2.58 (4.91±1.32)	72.31 (61.23±6.25)
1WQC$_{CONT}$	4.05 (4.22±0.1)	4.8 (7.63±1.6)	42.31 (29.03±5.42)
1WQC$_{CONTAPL}$	2.27 (2.65±0.18)	**1.4** (2.9±0.61)	**90.77** (74.95±5.44)
1WQC$_{RAND}$	3.96 (4.27±0.11)	5.29 (7.53±1.43)	35.38 (27.82±4.36)

Table 5: Lowest (Mean±Std) scores for Energy (10^3), RMSD and GDT from the 30 optimization execution of DE.

available for local execution. Therefore, we were able to force the usage the same secondary structure, to generate the fragments utilized by the Rosetta method, used by our method to generate the APLs. Similarly

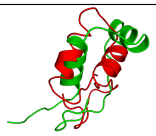
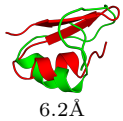
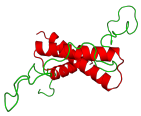
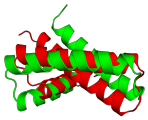
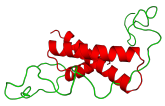
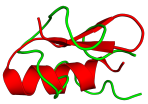
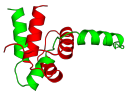
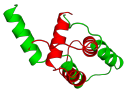
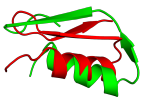
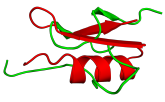
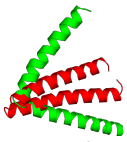
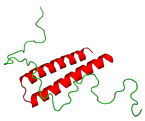
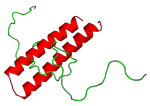
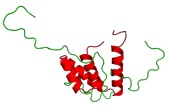
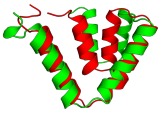
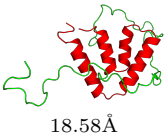
PDB	APL	CONT	CONTAPL	RAND
1AB1	 9.93Å	 12.87Å	 12.28Å	 12.94Å
1ACW	 7.72Å	 6.49Å	 6.2Å	 10.15Å
1AIL	 11.6Å	 18.75Å	 6.45Å	 20.28Å
1CRN	 5.21Å	 13.32Å	 10.31Å	 10.14Å
1D5Q	 7.81Å	 9.02Å	 4.41Å	 6.56Å
1DFN	 12.42Å	 10.84Å	 9.24Å	 7.18Å
1ENH	 16.32Å	 22.11Å	 12.04Å	 10.26Å
1Q2K	 6.36Å	 8.02Å	 9.44Å	 9.5Å
1ROP	 11.93Å	 18.5Å	 2.91Å	 17.76Å
1UTG	 15.9Å	 19.04Å	 7.88Å	 18.58Å
1WQC	5.18Å	7.37Å	2.57Å	8.42Å

Table 6: Cartoon representation of experimentally determined structures (red) and the lowest energy solution (green) found by the Differential Evolution optimization for each initial population method.

PDB ID	Low Energy ($\overline{x} \pm \sigma$)	Low RMSD ($\overline{x} \pm \sigma$)	High GDT ($\overline{x} \pm \sigma$)
$1AB1_{APL}$	**12.85** (22.18±5.97)	4.41 (6.18±0.94)	52.17 (44.09±4.16)
$1AB1_{CONT}$	16.72 (29.78±6.89)	6.32 (7.83±0.82)	27.39 (22.84±2.88)
$1AB1_{CONTAPL}$	14.58 (22.79±4.95)	**3.33** (5.57±0.94)	**60.43** (47.13±5.74)
$1AB1_{RAND}$	18.27 (29.18±5.98)	6.68 (7.87±0.94)	33.91 (24.28±4.32)
$1ACW_{APL}$	8.84 (14.65±3.05)	3.99 (5.42±0.8)	56.55 (44.0±7.13)
$1ACW_{CONT}$	13.45 (19.28±3.59)	5.17 (6.48±0.71)	40.69 (28.94±6.15)
$1ACW_{CONTAPL}$	**7.79** (13.48±3.4)	**3.84** (4.98±0.71)	**60.69** (47.72±6.38)
$1ACW_{RAND}$	10.69 (18.49±4.02)	4.89 (6.71±0.85)	43.45 (28.6±6.04)
$1AIL_{APL}$	37.24 (131.16±35.93)	6.56 (10.11±1.32)	47.43 (24.65±6.19)
$1AIL_{CONT}$	50.33 (76.25±12.24)	9.47 (11.01±0.96)	17.43 (13.49±1.91)
$1AIL_{CONTAPL}$	**29.3** (79.98±47.18)	**6.39** (8.56±1.85)	**50.0** (33.15±9.68)
$1AIL_{RAND}$	61.22 (79.08±10.92)	9.06 (11.17±1.15)	16.29 (13.49±1.65)
$1CRN_{APL}$	**10.22** (24.14±6.18)	4.46 (5.77±0.76)	**61.74** (46.32±5.29)
$1CRN_{CONT}$	16.32 (29.16±6.1)	5.9 (8.35±0.82)	31.3 (23.75±3.03)
$1CRN_{CONTAPL}$	16.19 (25.81±6.29)	**4.1** (5.59±0.86)	55.22 (46.74±5.18)
$1CRN_{RAND}$	17.15 (29.67±6.6)	5.74 (8.11±1.05)	31.3 (23.67±3.04)
$1D5Q_{APL}$	6.76 (12.43±2.67)	4.33 (5.65±0.66)	59.26 (49.19±5.52)
$1D5Q_{CONT}$	9.82 (14.07±2.07)	5.09 (6.8±0.69)	42.22 (29.36±5.67)
$1D5Q_{CONTAPL}$	**5.85** (11.21±2.77)	**4.2** (5.47±0.62)	**63.7** (54.07±5.39)
$1D5Q_{RAND}$	9.61 (13.85±2.85)	4.79 (7.04±0.78)	42.96 (27.01±5.07)
$1DFN_{APL}$	10.07 (15.46±3.09)	**4.38** (6.27±0.88)	**58.67** (37.53±6.75)
$1DFN_{CONT}$	**6.41** (19.62±4.35)	4.53 (6.7±0.9)	45.33 (34.07±4.66)
$1DFN_{CONTAPL}$	10.77 (17.38±3.54)	4.94 (6.38±0.59)	52.67 (35.64±5.75)
$1DFN_{RAND}$	13.14 (20.17±4.36)	4.72 (6.4±0.82)	46.0 (34.93±4.87)
$1ENH_{APL}$	19.13 (28.69±5.61)	2.61 (3.07±0.31)	74.07 (65.54±4.08)
$1ENH_{CONT}$	36.28 (52.25±10.01)	7.44 (9.04±1.17)	22.96 (18.21±2.2)
$1ENH_{CONTAPL}$	**18.18** (27.14±4.46)	**2.17** (2.95±0.34)	**77.41** (68.95±3.57)
$1ENH_{RAND}$	30.4 (51.91±10.03)	7.24 (9.14±0.96)	22.96 (17.65±2.83)
$1Q2K_{APL}$	10.16 (15.15±3.83)	4.52 (6.02±0.74)	60.0 (47.51±5.43)
$1Q2K_{CONT}$	10.87 (18.34±3.5)	5.64 (7.37±0.78)	36.77 (27.12±4.47)
$1Q2K_{CONTAPL}$	11.71 (16.15±2.51)	**3.94** (6.05±0.84)	**62.58** (48.9±5.4)
$1Q2K_{RAND}$	**10.01** (18.15±3.96)	5.41 (7.07±0.87)	38.71 (30.15±4.9)
$1ROP_{APL}$	11.63 (16.19±2.74)	**2.34** (3.24±0.58)	73.21 (63.46±6.49)
$1ROP_{CONT}$	23.93 (35.79±7.51)	6.85 (9.3±1.18)	22.86 (15.79±2.65)
$1ROP_{CONTAPL}$	**11.36** (15.5±2.36)	2.44 (3.44±0.47)	**74.64** (62.98±4.74)
$1ROP_{RAND}$	22.35 (34.18±5.13)	6.64 (9.13±1.21)	26.07 (15.87±3.22)
$1UTG_{APL}$	**36.98** (54.77±8.76)	6.91 (7.63±0.37)	47.71 (41.95±3.37)
$1UTG_{CONT}$	49.3 (72.1±16.07)	9.86 (12.33±1.17)	17.43 (13.6±1.89)
$1UTG_{CONTAPL}$	41.08 (57.38±10.41)	**6.04** (7.64±0.53)	**55.43** (42.6±4.66)
$1UTG_{RAND}$	45.13 (64.94±12.7)	9.37 (12.44±1.54)	17.43 (13.97±1.58)
$1WQC_{APL}$	3.71 (6.0±1.56)	3.5 (4.9±0.81)	70.0 (62.56±4.84)
$1WQC_{CONT}$	8.08 (12.16±2.19)	5.9 (6.94±0.56)	41.54 (29.69±4.54)
$1WQC_{CONTAPL}$	**3.55** (5.01±0.8)	**3.43** (4.83±0.83)	**70.77** (64.23±3.14)
$1WQC_{RAND}$	7.89 (12.79±2.47)	6.08 (7.17±0.7)	42.31 (29.85±6.13)

Table 7: Lowest (Mean±Std) scores for Energy (10^3), RMSD and GDT from the 30 optimization execution of DEMO.

to the DE and DEMO optimization, for each one of the 11 proteins within the data set, 30 runs were executed to predict the final structure and the results in Table 9 are the mean and the lowest of the 30 runs. As for the

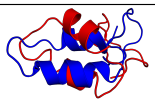
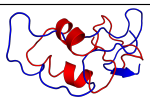
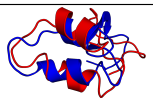
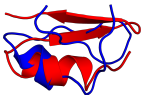
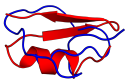
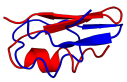
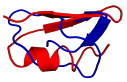
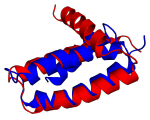
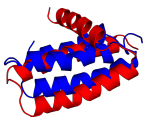
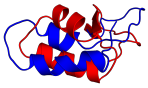
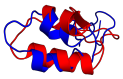
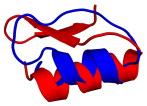
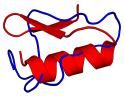
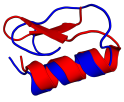
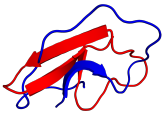
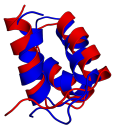
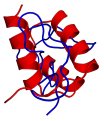
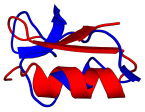
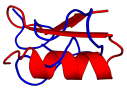
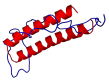
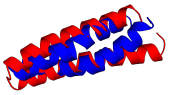
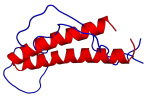
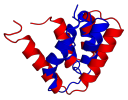
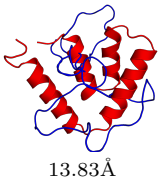
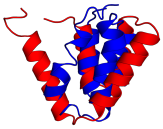
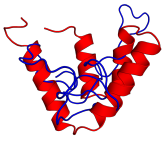
PDB	APL	CONT	CONTAPL	RAND
1AB1	 5.6Å	 7.26Å	 3.33Å	 7.75Å
1ACW	 4.49Å	 6.16Å	 4.72Å	 6.79Å
1AIL	 7.08Å	 12.69Å	 7.02Å	 11.28Å
1CRN	 6.93Å	 8.1Å	 4.5Å	 7.82Å
1D5Q	 5.58Å	 7.56Å	 4.63Å	 8.72Å
1DFN	 6.42Å	 7.7Å	 5.67Å	 6.53Å
1ENH	 2.92Å	 10.53Å	 3.49Å	 9.04Å
1Q2K	 7,07Å	 7.89Å	 5.26Å	 8.41Å
1ROP	 2.63Å	 8.36Å	 3.98Å	 9.98Å
1UTG	 7.41Å	 13.83Å	 7.33Å	 15.34Å
1WQC	4.87Å	6.74Å	5.44Å	8.34Å

Table 8: Cartoon representation of experimentally determined structures (red) and the selected solution from the Pareto Frontier (blue) found by the Multi-Objective version of Differential Evolution optimization for each initial population method.

QUARK method, it is not possible to run it local and also it is only possible to run one execution at a time. Thus, we only executed once for each of the proteins. The results obtained by the Rosetta method were the best results, as highlighted in Table 9. Nonetheless, the DEMO optimization utilizing the population generated under the guidance of contact information and APL achieved results considerably close to the state-of-the-art method Rosetta.

5. Discussion

The elucidation of the 3D structure of proteins applying exclusively computational approaches remains one of the biggest challenges in Structural Bioinformatics. The usage of significant biological constraints helps the different computational methods to achieve near-to-native structures. The results obtained in CASP12 shows that the increasing amount of data generated by the new DNA sequencing technology is helping to improve the accuracy of predicting residue-residue 3D interaction from evolutionary analysis (Schaarschmidt et al., 2018). In addition to that, the methods proposed to try to solve the PSP problem, which incorporate the residues coupling information into the workflow, were able to achieve results with higher similarity to the experimental determined structure (Abriata et al., 2018b). The usage of these evolutionary-inferred amino acid proximity was still effective in the results of the CASP13. Different usages of mutual information have already been proposed for the PSP problem; for example, Ovchinnikov et al. (2018) used the contact map to identify possible domains within the target protein while Zhang et al. (2018) used the contacts within molecular dynamics to assemble structural models, and also within a function score to prioritize those assemblies that do not violate the predicted contacts.

The proposed method to generate structural models incorporates the contact information together with the guidance of torsional angle patterns retrieved from the PDB. To test our method we generated four distinct populations: (*i*) random generation, (*ii*) torsion angle guidance, (*iii*) residue coupling, and (*iv*) torsion angle together with the contact information. As expected, the torsion angle guidance assembled the helices, increasing the protein stability and reducing the free energy scores of these models (Figures 4, Supplementary Materials). Despite that, the folding measure methods presented poor improvement (Figures 5, 6 and Supplementary Materials). The contact information improved the folding state of the models, but the higher free energy values indicated lower stability of the generated models. The merging of these two constraints achieved low energy values, indicating higher stability and also folding state near to the native structure (Figures 4,5, Supplementary Materials). It is worth to highlight that for all the proteins composing the data set, the same pattern is observed where the APL is responsible for the lower energy and the contact are responsible for the higher folding state, while the contact and APL is able to achieve a mean value for both energy and folding parameters (Supplementary Materials). The plots in Supplementary Materials shows that the contact information merged with the torsion angles retrieved from the APL generated structural models with structure very close to the experimentally determined for structures with the predominance of helices.

The usage of problem-domain information is able to guide the search process throughout the search space by optimization algorithms. The use of knowledge-based algorithm also has been proved to improve the efficiency of the search within the conformation landscape regarding the PSP problem (Borguesan et al., 2015; de Lima Corrêa et al., 2018; Narloch and Dorn, 2019). Our results shows that the incorporation of these two constraints when generating initial structural models for structural optimization improve the optimization process due to its near-to-native features already presented by the very first candidates, showed by Table 5. It is worth to mention that, even though the APL information was able to improve the search for lower energy scores (Table 5), the difficulties regarding the folding of the structural models is still a problem to solve when using the APLs, as illustrated by the high RMSD score presented in Table 6. Therefore, together with the usage of contact information at the generation of initial models, an evaluation function score based on the distance of the amino acid residues predicted to be in contact must be incorporated within the optimization process to ensure the information will not be lost during the execution of the algorithm. To do so, we proposed the usage of a second objective regarding the 3D contact between the amino acid residue pairs predicted to be in contact. The results obtained by the multi-objective optimization corroborated to the fact that the free energy functions are not capable to fully describe the native-like structure. This

PDB ID	Low Energy ($\overline{x} \pm \sigma$)	Low RMSD ($\overline{x} \pm \sigma$)	High GDT ($\overline{x} \pm \sigma$)
1AB1$_{DE}$	3.61 (4.78±0.44)	8.0 (13.02±1.67)	43.48 (38.48±1.85)
1AB1$_{DEMO}$	14.58 (22.79±4.95)	**3.33** (5.57±0.94)	60.43 (47.13±5.74)
1AB1$_{ROS}$	**2.87** (3.08±0.12)	3.87 (5.65±0.95)	**69.13** (56.57±6.17)
1AB1$_{QUA}$	2.99	4.95	60.87
1ACW$_{DE}$	3.03 (3.6±0.14)	6.2 (8.33±1.25)	53.1 (43.06±3.87)
1ACW$_{DEMO}$	7.79 (13.48±3.4)	3.84 (4.98±0.71)	60.69 (47.72±6.38)
1ACW$_{ROS}$	**2.3** (2.5±0.13)	**2.0** (3.01±0.88)	**86.9** (73.33±7.35)
1ACW$_{QUA}$	2.74	7.82	36.55
1AIL$_{DE}$	6.79 (6.96±0.18)	6.45 (6.82±0.75)	50.29 (46.79±1.76)
1AIL$_{DEMO}$	29.3 (79.98±47.18)	6.39 (8.56±1.85)	50.0 (33.15±9.68)
1AIL$_{ROS}$	**4.2** (4.49±0.19)	**3.56** (9.61±1.41)	**76.0** (37.08±8.42)
1AIL$_{QUA}$	4.6	7.31	43.43
1CRN$_{DE}$	3.29 (5.01±0.54)	7.11 (11.26±2.46)	45.22 (36.36±3.55)
1CRN$_{DEMO}$	16.19 (25.81±6.29)	4.1 (5.59±0.86)	55.22 (46.74±5.18)
1CRN$_{ROS}$	**2.8** (3.03±0.13)	**3.34** (5.45±1.05)	**75.65** (61.9±7.63)
1CRN$_{QUA}$	3.19	5.14	59.57
1D5Q$_{DE}$	2.72 (2.87±0.06)	4.22 (6.2±1.39)	64.44 (55.31±4.54)
1D5Q$_{DEMO}$	5.85 (11.21±2.77)	4.2 (5.47±0.62)	63.7 (54.07±5.39)
1D5Q$_{ROS}$	**2.05** (2.22±0.1)	**1.66** (2.49±0.61)	**85.93** (73.95±7.38)
1D5Q$_{QUA}$	2.21	4.42	66.67
1DFN$_{DE}$	4.25 (4.38±0.04)	7.64 (9.92±0.72)	36.0 (33.4±2.87)
1DFN$_{DEMO}$	10.77 (17.38±3.54)	4.94 (6.38±0.59)	52.67 (35.64±5.75)
1DFN$_{ROS}$	**2.64** (2.94±0.2)	**3.4** (6.69±1.07)	**66.67** (51.47±6.56)
1DFN$_{QUA}$	2.68	6.65	56.0
1ENH$_{DE}$	6.14 (6.63±0.32)	7.32 (9.97±3.25)	51.85 (50.52±2.72)
1ENH$_{DEMO}$	18.18 (27.14±4.46)	2.17 (2.95±0.34)	77.41 (68.95±3.57)
1ENH$_{ROS}$	**3.65** (3.96±0.18)	1.98 (3.7±1.32)	**92.96** (75.64±11.88)
1ENH$_{QUA}$	3.72	**1.78**	92.22
1Q2K$_{DE}$	3.27 (3.62±0.11)	7.67 (8.66±0.33)	48.39 (47.85±1.03)
1Q2K$_{DEMO}$	11.71 (16.15±2.51)	3.94 (6.05±0.84)	62.58 (48.9±5.4)
1Q2K$_{ROS}$	**2.36** (2.53±0.1)	**0.8** (2.27±0.97)	**98.71** (81.91±11.62)
1Q2K$_{QUA}$	2.77	4.07	61.29
1ROP$_{DE}$	5.08 (5.38±0.16)	2.91 (17.76±8.19)	70.36 (54.2±8.91)
1ROP$_{DEMO}$	11.36 (15.5±2.36)	2.44 (3.44±0.47)	74.64 (62.98±4.74)
1ROP$_{ROS}$	**3.67** (4.02±0.25)	**1.18** (6.3±3.32)	**89.64** (58.43±15.36)
1ROP$_{QUA}$	3.96	1.6	83.57
1UTG$_{DE}$	7.66 (9.66±0.43)	7.88 (14.57±3.03)	51.14 (37.84±3.88)
1UTG$_{DEMO}$	41.08 (57.38±10.41)	6.04 (7.64±0.53)	55.43 (42.6±4.66)
1UTG$_{ROS}$	**3.93** (4.4±0.28)	**4.09** (9.0±3.09)	**60.57** (42.86±9.16)
1UTG$_{QUA}$	4.12	5.04	54.0
1WQC$_{DE}$	2.27 (2.65±0.18)	**1.4** (2.9±0.61)	**90.77** (74.95±5.44)
1WQC$_{DEMO}$	3.55 (5.01±0.8)	3.43 (4.83±0.83)	70.77 (64.23±3.14)
1WQC$_{ROS}$	2.14 (2.31±0.07)	2.55 (3.04±0.35)	80.77 (71.97±3.66)
1WQC$_{QUA}$	**2.1**	2.49	78.46

Table 9: Scores achieved by the Differential Evolution (DE) and Multi-Objective Differential Evolution (DEMO) using the populations composed by the individuals constructed using the information retrieved from the APL and the contact prediction. The results were compared to the two state-of-the-art methods Rosetta (ROS) and QUARK (QUA).

is showed by the lower RMSD scores achieved by the DEMO, even with the higher energy scores. The generation method guided by both constraints achieved results near to the state-of-the-art methods, which indicates the method high potential to generate structural models.

Acknowledgements

This work was supported by grants from FAPERGS [16/2551-0000520-6], MCT/CNPq [311022/2015-4; 311611/2018-4], Alexander von Humboldt-Stiftung (AvH) [BRA 1190826 HFST CAPES-P] - Germany. This study was financed in part by the Coordenacão de Aperfeiçoamento de Pessoal de Nível Superior - Brazil (CAPES) - Finance Code 001.

References

Abriata, L. A., Kinch, L. N., Tamò, G. E., Monastyrskyy, B., Kryshtafovych, A., Dal Peraro, M., 2018a. Definition and classification of evaluation units for tertiary structure prediction in casp12 facilitated through semi-automated metrics. Proteins: Structure, Function, and Bioinformatics 86, 16–26.

Abriata, L. A., Tamò, G. E., Monastyrskyy, B., Kryshtafovych, A., Dal Peraro, M., 2018b. Assessment of hard target modeling in casp12 reveals an emerging role of alignment-based contact prediction methods. Proteins: Structure, Function, and Bioinformatics 86, 97–112.

Alford, R. F., Leaver-Fay, A., Jeliazkov, J. R., OMeara, M. J., DiMaio, F. P., Park, H., Shapovalov, M. V., Renfrew, P. D., Mulligan, V. K., Kappel, K., et al., 2017. The rosetta all-atom energy function for macromolecular modeling and design. Journal of Chemical Theory and Computation 13 (6), 3031–3048.

Anfinsen, C. B., 1973. Principles that govern the folding of protein chains. Science 181 (4096), 223–230.

Benitez-Hidalgo, A., Nebro, A. J., Garcia-Nieto, J., Oregi, I., Del Ser, J., 2019. jmetalpy: a python framework for multi-objective optimization with metaheuristics. arXiv preprint arXiv:1903.02915.

Berman, H., Henrick, K., Nakamura, H., 2003. Announcing the worldwide protein data bank. Nature Structural & Molecular Biology 10 (12), 980.

Borguesan, B., e Silva, M. B., Grisci, B., Inostroza-Ponta, M., Dorn, M., 2015. Apl: An angle probability list to improve knowledge-based metaheuristics for the three-dimensional protein structure prediction. Computational Biology and Chemistry 59, 142–157.

Borguesan, B., Inostroza-Ponta, M., Dorn, M., 2017. Nias-server: Neighbors influence of amino acids and secondary structures in proteins. Journal of Computational Biology 24 (3), 255–265.

Borguesan, B., Narloch, P. H., Inostroza-Ponta, M., Dorn, M., 2018. A genetic algorithm based on restricted tournament selection for the 3d-psp problem. In: 2018 IEEE Congress on Evolutionary Computation (CEC). IEEE, pp. 1–8.

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. a., Karplus, M., 1983. Charmm: a program for macromolecular energy, minimization, and dynamics calculations. Journal of Computational Chemistry 4 (2), 187–217.

Carneiro, M. G., Koharudin, L. M., Ban, D., Sabo, T. M., Trigo-Mourino, P., Mazur, A., Griesinger, C., Gronenborn, A. M., Lee, D., 2015. Sampling of glycan-bound conformers by the anti-hiv lectin oscillatoria agardhii agglutinin in the absence of sugar. Angewandte Chemie International Edition 54 (22), 6462–6465.

Chaudhury, S., Lyskov, S., Gray, J. J., 2010. Pyrosetta: a script-based interface for implementing molecular modeling algorithms using rosetta. Bioinformatics 26 (5), 689–691.

Clarke, N. D., Kissinger, C. R., Desjarlais, J., Gilliland, G. L., Pabo, C. O., 1994. Structural studies of the engrailed homeodomain. Protein Science 3 (10), 1779–1787.

Connolly, M. L., 1983. Solvent-accessible surfaces of proteins and nucleic acids. Science 221 (4612), 709–713.

Crescenzi, P., Goldman, D., Papadimitriou, C., Piccolboni, A., Yannakakis, M., 1998. On the complexity of protein folding. Journal of Computational Biology 5 (3), 423–465.

De Juan, D., Pazos, F., Valencia, A., 2013. Emerging methods in protein co-evolution. Nature Reviews Genetics 14 (4), 249261.

de Lima Corrêa, L., Borguesan, B., Krause, M. J., Dorn, M., 2018. Three-dimensional protein structure prediction based on memetic algorithms. Computers & Operations Research 91, 160–177.

Deb, K., Pratap, A., Agarwal, S., Meyarivan, T., 2002. A fast and elitist multiobjective genetic algorithm: Nsga-ii. IEEE Transactions on Evolutionary Computation 6 (2), 182–197.

Dorn, M., e Silva, M. B., Buriol, L. S., Lamb, L. C., 2014. Three-dimensional protein structure prediction: Methods and computational strategies. Computational Biology and Chemistry 53, 251–276.

Edwards, A. M., Arrowsmith, C. H., Christendat, D., Dharamsi, A., Friesen, J. D., Greenblatt, J. F., Vedadi, M., 2000. Protein production: feeding the crystallographers and nmr spectroscopists. Nature Structural and Molecular Biology 7 (11s), 970.

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., et al., 2018. The pfam protein families database in 2019. Nucleic Acids Research 47 (D1), D427–D432.

Fan, X., Wang, J., Zhang, X., Yang, Z., Zhang, J.-C., Zhao, L., Peng, H.-L., Lei, J., Wang, H.-W., 2019. Single particle cryo-em reconstruction of 52 kda streptavidin at 3.2 angstrom resolution. Nature Vommunications 10 (1), 2386.

Fonseca, R., Paluszewski, M., Winter, P., Jun 2010. Protein structure prediction using bee colony optimization metaheuristic. Journal of Mathematical Modelling and Algorithms 9 (2), 181–194.

Goldberg, D. E., Holland, J. H., 1988. Genetic algorithms and machine learning. Machine Learning 3 (2), 95–99.

Guyeux, C., Côté, N. M.-L., Bahi, J. M., Bienia, W., 2014. Is protein folding problem really a np-complete one? first investigations. Journal of Bioinformatics and Computational Biology 12 (01), 1350017.

Hill, C. P., Yee, J., Selsted, M. E., Eisenberg, D., 1991. Crystal structure of defensin hnp-3, an amphiphilic dimer: mechanisms of membrane permeabilization. Science 251 (5000), 1481–1485.

Hong, S. H., Joung, I., Flores-Canales, J. C., Manavalan, B., Cheng, Q., Heo, S., Kim, J. Y., Lee, S. Y., Nam, M., Joo, K., et al., 2018. Protein structure modeling and refinement by global optimization in casp12. Proteins: Structure, Function, and Bioinformatics 86, 122–135.

Hopf, T. A., Colwell, L. J., Sheridan, R., Rost, B., Sander, C., Marks, D. S., 2012. Three-dimensional structures of membrane proteins from genomic sequencing. Cell 149 (7), 1607–1621.

Hopf, T. A., Schärfe, C. P., Rodrigues, J. P., Green, A. G., Kohlbacher, O., Sander, C., Bonvin, A. M., Marks, D. S., 2014. Sequence co-evolution gives 3d contacts and structures of protein complexes. eLife 3, e03430.

Hovmöller, S., Zhou, T., Ohlson, T., 2002. Conformations of amino acids in proteins. Acta Crystallographica Section D: Biological Crystallography 58 (5), 768–776.

Illergård, K., Ardell, D. H., Elofsson, A., 2009. Structure is three to ten times more conserved than sequencea study of structural response in protein cores. Proteins: Structure, Function, and Bioinformatics 77 (3), 499–508.

Jones, J. E., Chapman, S., 1924. On the determination of molecular fields.-i. from the variation of the viscosity of a gas with temperature. Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character 106 (738), 441–462.
URL https://royalsocietypublishing.org/doi/abs/10.1098/rspa.1924.0081

Kabsch, W., Sander, C., 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers: Original Research on Biomolecules 22 (12), 2577–2637.

Kim, D. E., Blum, B., Bradley, P., Baker, D., 2009. Sampling bottlenecks in de novo protein structure prediction. Journal of Molecular Biology 393 (1), 249–260.

Kirkpatrick, S., Gelatt, C. D., Vecchi, M. P., 1983. Optimization by simulated annealing. Science 220 (4598), 671–680.

Konak, A., Coit, D. W., Smith, A. E., 2006. Multi-objective optimization using genetic algorithms: A tutorial. Reliability Engineering & System Safety 91 (9), 992–1007.

Lazaridis, T., Karplus, M., 1999. Effective energy function for proteins in solution. Proteins: Structure, Function, and Bioinformatics 35 (2), 133–152.

Lee, B., Richards, F. M., 1971. The interpretation of protein structures: estimation of static accessibility. Journal of Molecular Biology 55 (3), 379–400.

Ligabue-Braun, R., Borguesan, B., Verli, H., Krause, M. J., Dorn, M., 2018. Everyone is a protagonist: Residue conformational preferences in high-resolution protein structures. Journal of Computational Biology 25 (4), 451–465.

Lobanov, M. Y., Bogatyreva, N., Galzitskaya, O., 2008. Radius of gyration as an indicator of protein structure compactness. Molecular Biology 42 (4), 623–628.

Luke, S., 2013. Essentials of Metaheuristics, 2nd Edition. Lulu.

Mazzei, L., Cianci, M., Contaldo, U., Musiani, F., Ciurli, S., 2017. Urease inhibition in the presence of n-(n-butyl) thiophosphoric triamide, a suicide substrate: Structure and kinetics. Biochemistry 56 (40), 5391–5404.

Mukherjee, S., Stamatis, D., Bertsch, J., Ovchinnikova, G., Katta, H. Y., Mojica, A., Chen, I.-M. A., Kyrpides, N. C., Reddy, T., 2018. Genomes online database (gold) v. 7: updates and new features. Nucleic Acids Research 47 (D1), D649–D659.

Narloch, P. H., Dorn, M., 2019. A knowledge based differential evolution algorithm for protein structure prediction. In: International Conference on the Applications of Evolutionary Computation (Part of EvoStar). Springer, pp. 343–359.

Nelson, D. L., Lehninger, A. L., Cox, M. M., 2008. Lehninger principles of biochemistry. Macmillan.

Osguthorpe, D. J., 2000. Ab initio protein folding. Current Opinion in Structural Biology 10 (2), 146–152.

Osman, I. H., Laporte, G., 1996. Metaheuristics: A bibliography.

Ovchinnikov, S., Park, H., Kim, D. E., DiMaio, F., Baker, D., 2018. Protein structure prediction using rosetta in casp12. Proteins: Structure, Function, and Bioinformatics 86, 113–121.

OMeara, M. J., Leaver-Fay, A., Tyka, M. D., Stein, A., Houlihan, K., DiMaio, F., Bradley, P., Kortemme, T., Baker, D., Snoeyink, J., et al., 2015. Combined covalent-electrostatic model of hydrogen bonding improves structure prediction with rosetta. Journal of Chemical Theory and Computation 11 (2), 609–622.

Park, H., Bradley, P., Greisen Jr, P., Liu, Y., Mulligan, V. K., Kim, D. E., Baker, D., DiMaio, F., 2016. Simultaneous optimization of biomolecular energy functions on features from small molecules and macromolecules. Journal of Chemical Theory and Computation 12 (12), 6201–6212.

Pruitt, K. D., Tatusova, T., Maglott, D. R., 2006. Ncbi reference sequences (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Research 35 (suppl_1), D61–D65.

Ramachandran, G. N., 1963. Stereochemistry of polypeptide chain configurations. Journal of Molecular Biology 7, 95–99.

Richardson, J. S., 1981. The anatomy and taxonomy of protein structure. In: Advances in protein chemistry. Vol. 34. Elsevier, pp. 167–339.

Robič, T., Filipič, B., 2005. Differential evolution for multiobjective optimization. In: International Conference on Evolutionary Multi-Criterion Optimization. Springer, pp. 520–533.

Rohl, C. A., Strauss, C. E., Misura, K. M., Baker, D., 2004. Protein structure prediction using rosetta. In: Numerical Computer Methods, Part D. Vol. 383 of Methods in Enzymology. Academic Press, pp. 66 – 93.
URL http://www.sciencedirect.com/science/article/pii/S0076687904830040

Schaarschmidt, J., Monastyrskyy, B., Kryshtafovych, A., Bonvin, A. M., 2018. Assessment of contact predictions in casp12: Co-evolution and deep learning coming of age. Proteins: Structure, Function, and Bioinformatics 86, 51–66.

Shapovalov, M. V., Dunbrack Jr, R. L., 2011. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. Structure 19 (6), 844–858.

Silva, R. S., Parpinelli, R. S., 2019. A self-adaptive differential evolution with fragment insertion for the protein structure prediction problem. In: International Workshop on Hybrid Metaheuristics. Springer, pp. 136–149.

Song, Y., DiMaio, F., Wang, R. Y.-R., Kim, D., Miles, C., Brunette, T., Thompson, J., Baker, D., 2013. High-resolution comparative modeling with rosettacm. Structure 21 (10), 1735–1742.

Storn, R., Price, K., 1997. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. Journal of Global Optimization 11 (4), 341–359.

Talbi, E.-G., 2009. Metaheuristics: from design to implementation. Vol. 74. John Wiley & Sons.

Tantar, A.-A., Melab, N., Talbi, E.-G., Parent, B., Horvath, D., 2007. A parallel hybrid genetic algorithm for protein structure prediction on the computational grid. Future Generation Computer Systems 23 (3), 398 – 409.

Wang, S., Sun, S., Xu, J., 2018. Analysis of deep learning methods for blind protein contact prediction in casp12. Proteins: Structure, Function, and Bioinformatics 86, 67–77.

Weigt, M., White, R. A., Szurmant, H., Hoch, J. A., Hwa, T., 2009. Identification of direct residue contacts in protein–protein interaction by message passing. Proceedings of the National Academy of Sciences 106 (1), 67–72.

Worth, C. L., Gong, S., Blundell, T. L., 2009. Structural and functional constraints in the evolution of protein families. Nature Reviews Molecular Cell Biology 10 (10), 709.

Xu, D., Zhang, Y., 2012. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins: Structure, Function, and Bioinformatics 80 (7), 1715–1735.

Zemla, A., Venclovas, Č., Moult, J., Fidelis, K., 1999. Processing and analysis of casp3 protein structure predictions. Proteins: Structure, Function, and Bioinformatics 37 (S3), 22–29.

Zerihun, M. B., Schug, A., 2017. Biomolecular coevolution and its applications: Going from structure prediction toward signaling, epistasis, and function. Biochemical Society Transactions 45 (6), 1253–1261.
URL http://www.biochemsoctrans.org/content/45/6/1253

Zhang, C., Mortuza, S., He, B., Wang, Y., Zhang, Y., 2018. Template-based and free modeling of i-tasser and quark pipelines using predicted contact maps in casp12. Proteins: Structure, Function, and Bioinformatics 86, 136–151.

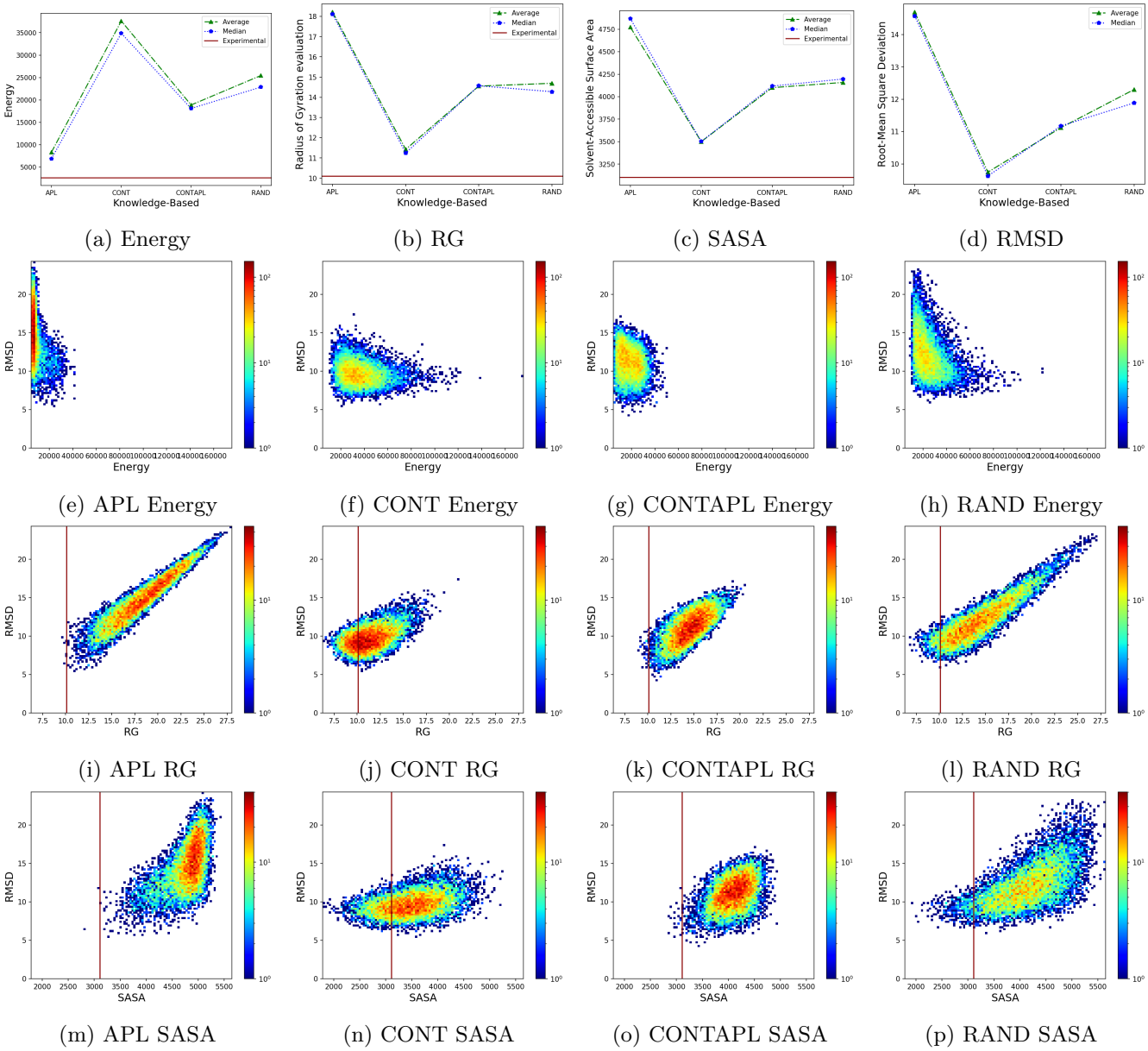
The inclusion of conserved evolutionary contact to refine validate techniques to predict protein structure Supplementary Materials

LA Santos[a], PH Narloch[c], R Ligabue-Braun[b], M Dorn[c,*]

[a] *Center of Biotechnology, Federal University of Rio Grande do Sul, Porto Alegre, Brazil*
[b] *Department of Pharmaceutical Sciences, UFCSPA, Porto Alegre, Brazil*
[c] *Institute of Informatics, Federal University of Rio Grande do Sul, Porto Alegre, Brazil*

The Figures S1-S21, items (a)-(d), show the average scores achieved by each population: Random (RAND), APL, Contacts only (CONT) and Contact and APL merged (CONTAPL); generated for each one of the 21 proteins within our data set. The structural features evaluated here are Energy, Radius of Gyration (RG), Solvent-Accessible Surface Area (SASA) and Root-Mean Square Deviation (RMSD). While the items (e)-(p) shows the heatmap generated based on the density of structural models sharing the combination between the RMSD and the other structural features for each protein and population. The dark red horizontal line represents the score calculated for the experimentally determined 3D structure.

*Corresponding author
Email addresses: leonardoas95@gmail.com (LA Santos), phnarloch@inf.ufrgs.br (PH Narloch), rodrigolb@ufcspa.edu.br (R Ligabue-Braun), mdorn@inf.ufrgs.br (M Dorn)



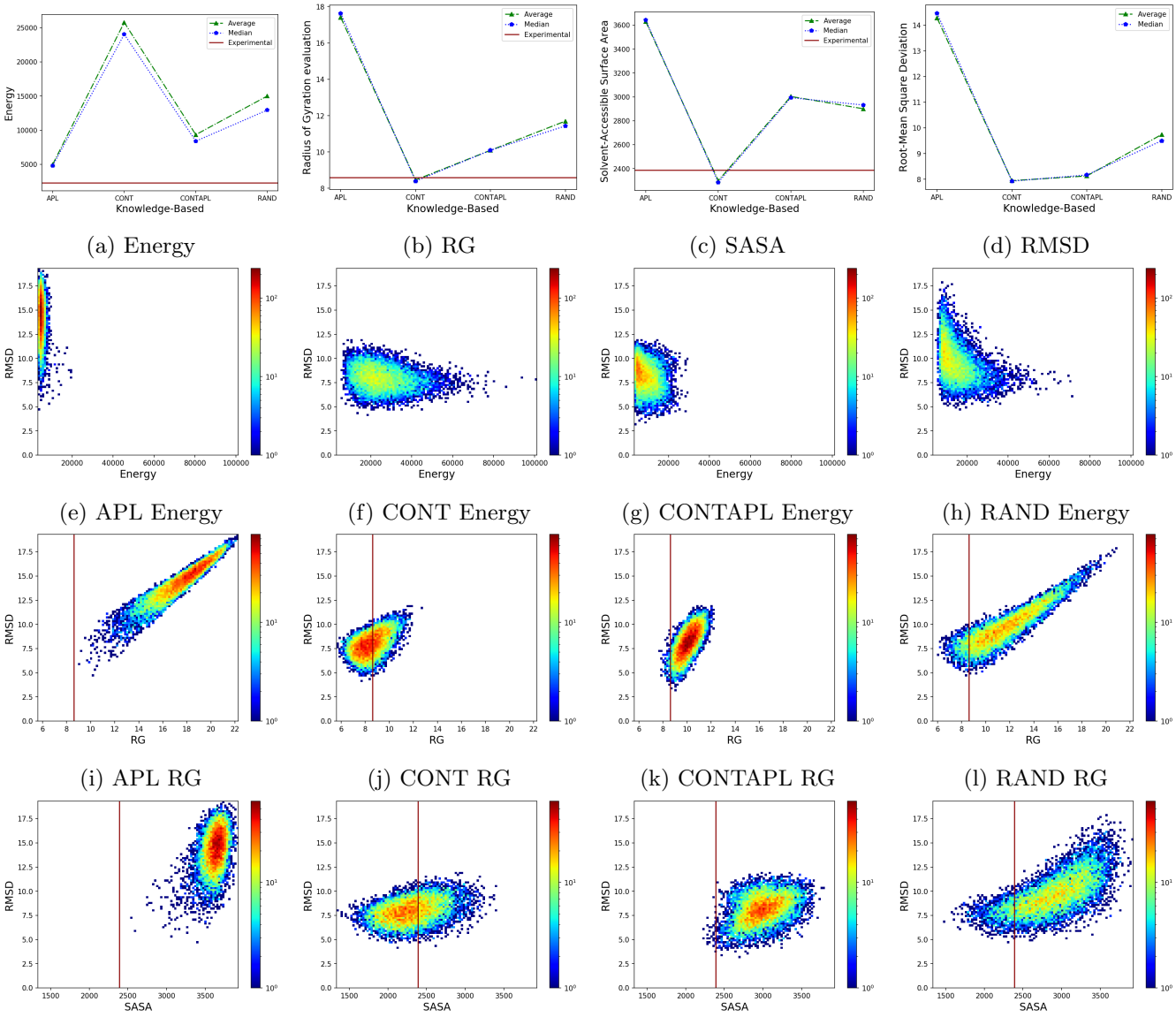
(a) Energy (b) RG (c) SASA (d) RMSD

(e) APL Energy (f) CONT Energy (g) CONTAPL Energy (h) RAND Energy

(i) APL RG (j) CONT RG (k) CONTAPL RG (l) RAND RG

(m) APL SASA (n) CONT SASA (o) CONTAPL SASA (p) RAND SASA

Figure S1: 1AB1 population's mean of the structural features score achieved by the Random (RAND), APL, Contacts only (CONT) and Contact and APL merged (CONTAPL), for each one of the proteins within our data set. The dark red horizontal line represents the score calculated for the experimentally determined 3D structure.



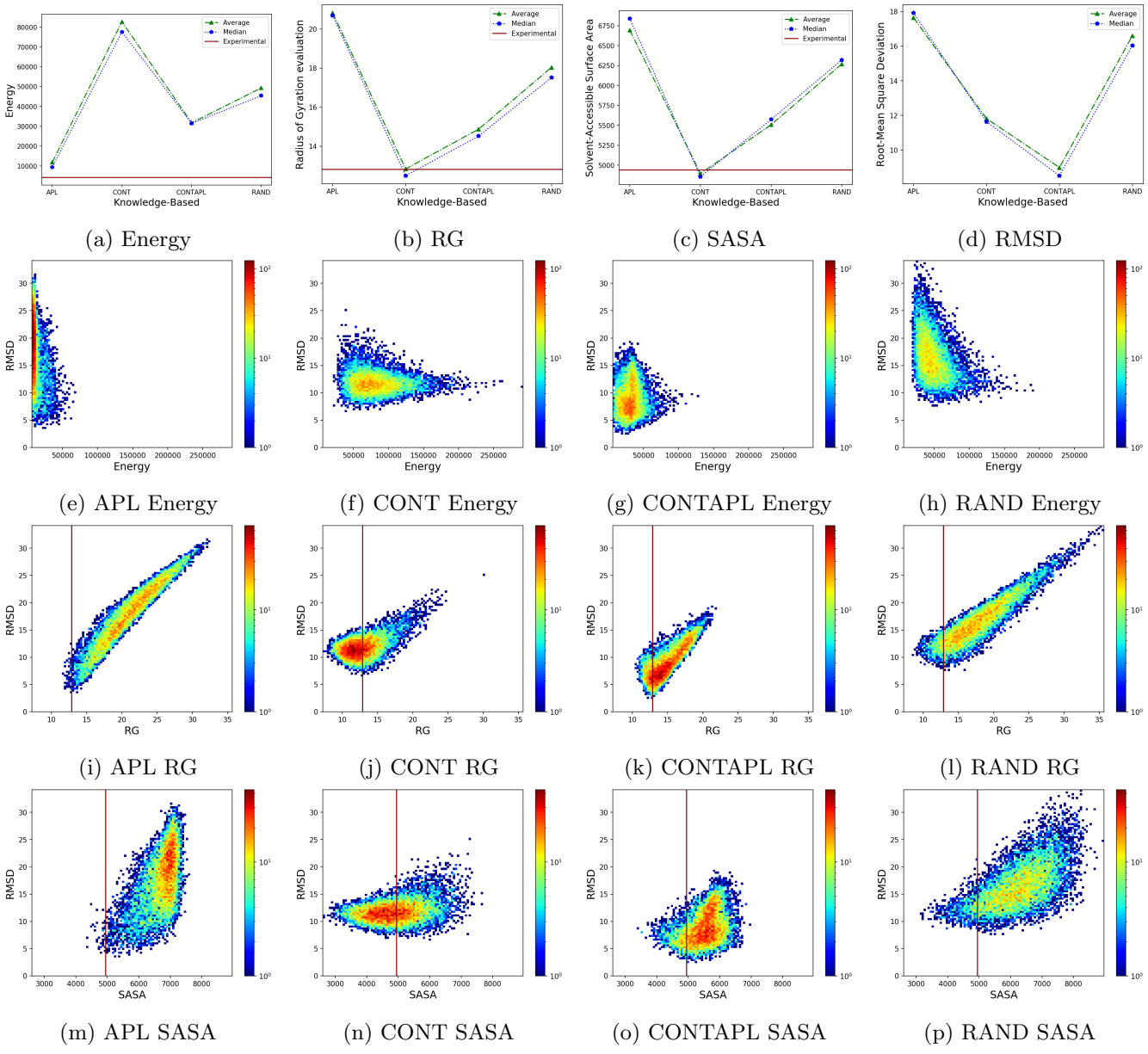
(a) Energy (b) RG (c) SASA (d) RMSD

(e) APL Energy (f) CONT Energy (g) CONTAPL Energy (h) RAND Energy

(i) APL RG (j) CONT RG (k) CONTAPL RG (l) RAND RG

(m) APL SASA (n) CONT SASA (o) CONTAPL SASA (p) RAND SASA

Figure S2: 1ACW populations' mean of the structural features score achieved by the Random (RAND), APL, Contacts only (CONT) and Contact and APL merged (CONTAPL), for each one of the proteins within our data set. Each heatmap was generated based on the density of combination between the RMSD and the other structural features. The dark red horizontal line represents the score calculated for the experimentally determined 3D structure.



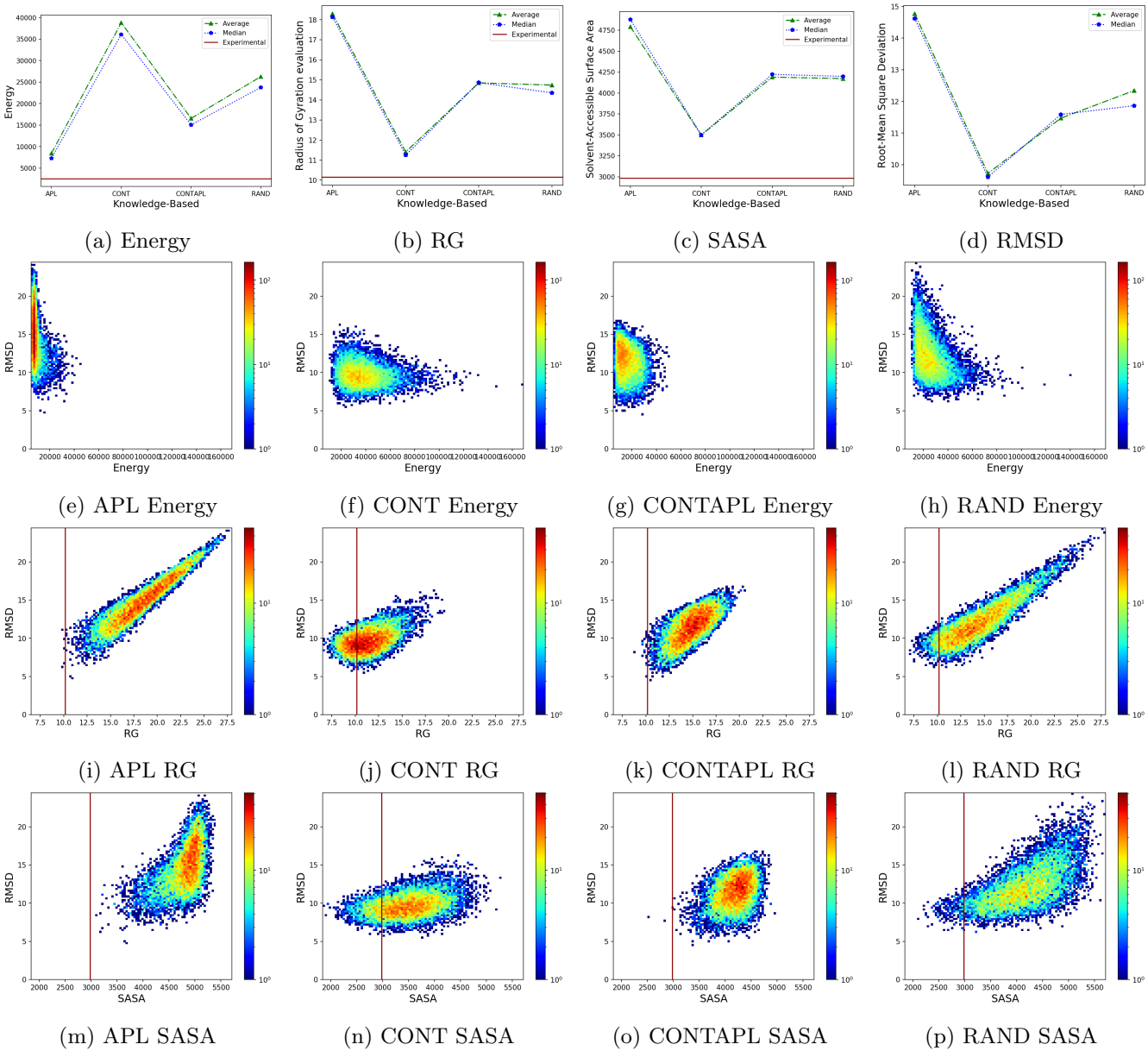
(a) Energy (b) RG (c) SASA (d) RMSD

(e) APL Energy (f) CONT Energy (g) CONTAPL Energy (h) RAND Energy

(i) APL RG (j) CONT RG (k) CONTAPL RG (l) RAND RG

(m) APL SASA (n) CONT SASA (o) CONTAPL SASA (p) RAND SASA

Figure S3: 1AIL populations' mean of the structural features score achieved by the Random (RAND), APL, Contacts only (CONT) and Contact and APL merged (CONTAPL), for each one of the proteins within our data set. Each heatmap was generated based on the density of combination between the RMSD and the other structural features. The dark red horizontal line represents the score calculated for the experimentally determined 3D structure.



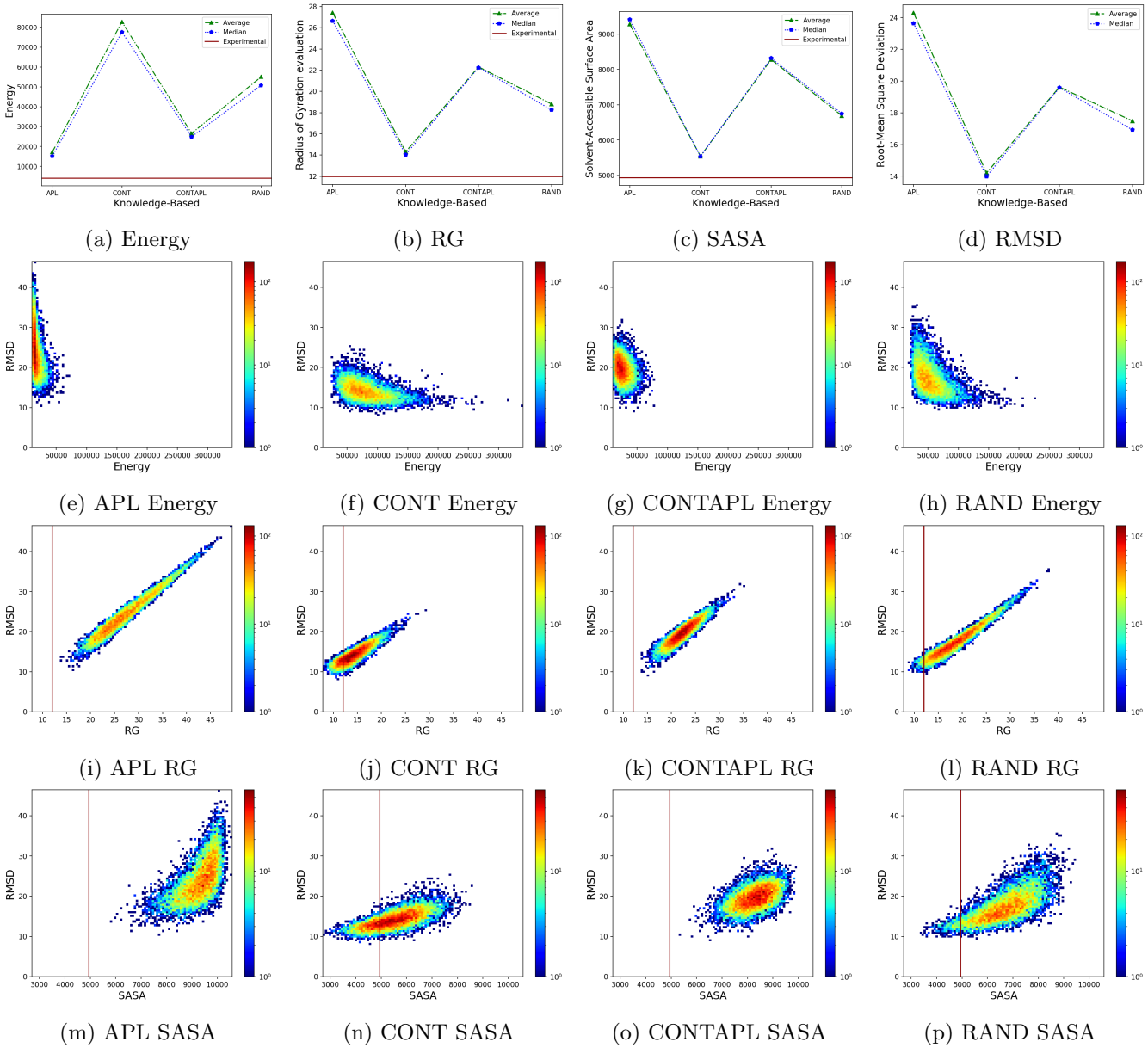
(a) Energy (b) RG (c) SASA (d) RMSD

(e) APL Energy (f) CONT Energy (g) CONTAPL Energy (h) RAND Energy

(i) APL RG (j) CONT RG (k) CONTAPL RG (l) RAND RG

(m) APL SASA (n) CONT SASA (o) CONTAPL SASA (p) RAND SASA

Figure S4: 1CRN populations' mean of the structural features score achieved by the Random (RAND), APL, Contacts only (CONT) and Contact and APL merged (CONTAPL), for each one of the proteins within our data set. Each heatmap was generated based on the density of combination between the RMSD and the other structural features. The dark red horizontal line represents the score calculated for the experimentally determined 3D structure.



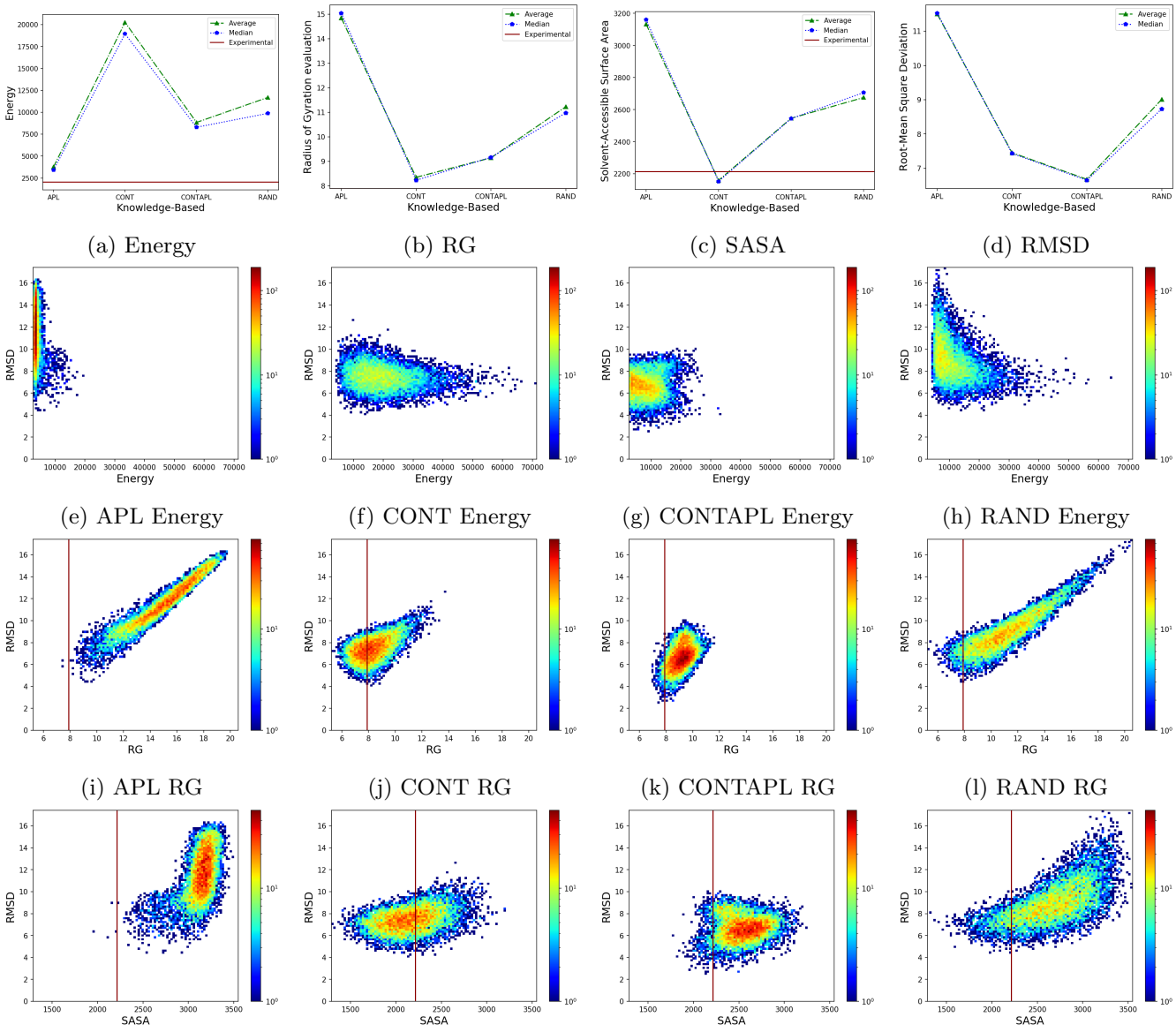
(a) Energy (b) RG (c) SASA (d) RMSD

(e) APL Energy (f) CONT Energy (g) CONTAPL Energy (h) RAND Energy

(i) APL RG (j) CONT RG (k) CONTAPL RG (l) RAND RG

(m) APL SASA (n) CONT SASA (o) CONTAPL SASA (p) RAND SASA

Figure S5: 1D3Z populations' mean of the structural features score achieved by the Random (RAND), APL, Contacts only (CONT) and Contact and APL merged (CONTAPL), for each one of the proteins within our data set. Each heatmap was generated based on the density of combination between the RMSD and the other structural features. The dark red horizontal line represents the score calculated for the experimentally determined 3D structure.



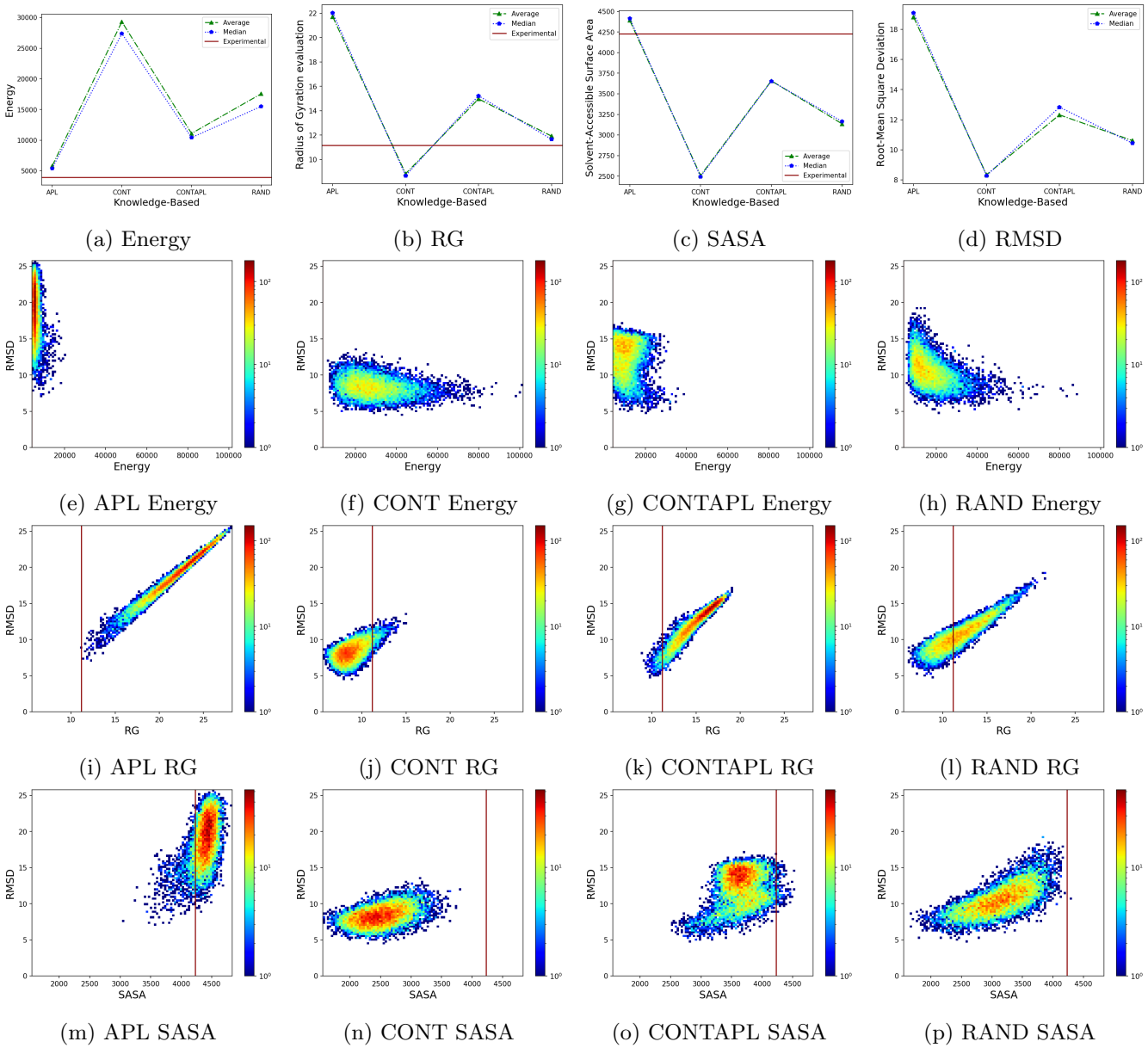
(a) Energy (b) RG (c) SASA (d) RMSD

(e) APL Energy (f) CONT Energy (g) CONTAPL Energy (h) RAND Energy

(i) APL RG (j) CONT RG (k) CONTAPL RG (l) RAND RG

(m) APL SASA (n) CONT SASA (o) CONTAPL SASA (p) RAND SASA

Figure S6: 1D5Q populations' mean of the structural features score achieved by the Random (RAND), APL, Contacts only (CONT) and Contact and APL merged (CONTAPL), for each one of the proteins within our data set. Each heatmap was generated based on the density of combination between the RMSD and the other structural features. The dark red horizontal line represents the score calculated for the experimentally determined 3D structure.



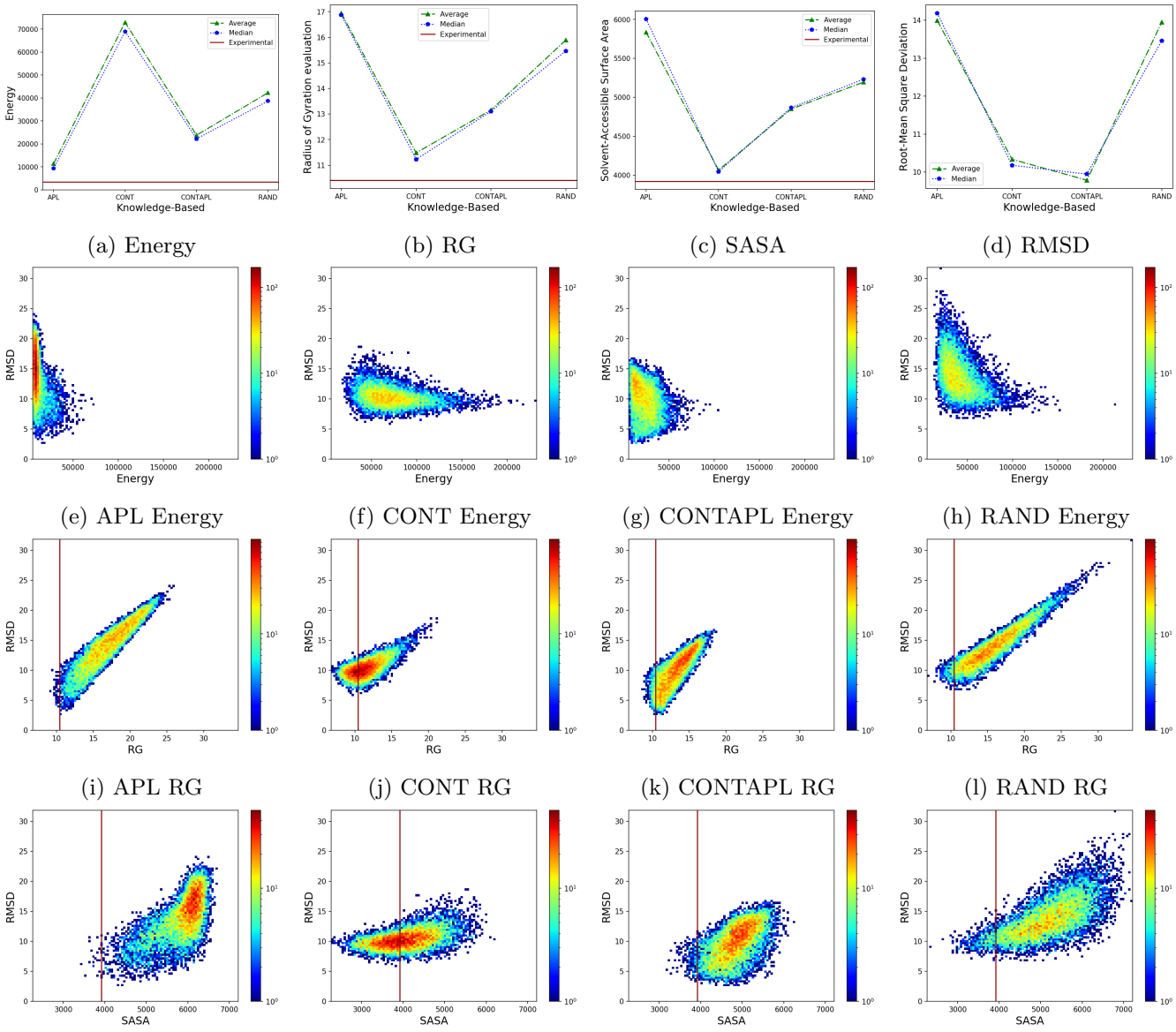
(a) Energy (b) RG (c) SASA (d) RMSD

(e) APL Energy (f) CONT Energy (g) CONTAPL Energy (h) RAND Energy

(i) APL RG (j) CONT RG (k) CONTAPL RG (l) RAND RG

(m) APL SASA (n) CONT SASA (o) CONTAPL SASA (p) RAND SASA

Figure S7: 1DFN populations' mean of the structural features score achieved by the Random (RAND), APL, Contacts only (CONT) and Contact and APL merged (CONTAPL), for each one of the proteins within our data set. Each heatmap was generated based on the density of combination between the RMSD and the other structural features. The dark red horizontal line represents the score calculated for the experimentally determined 3D structure.



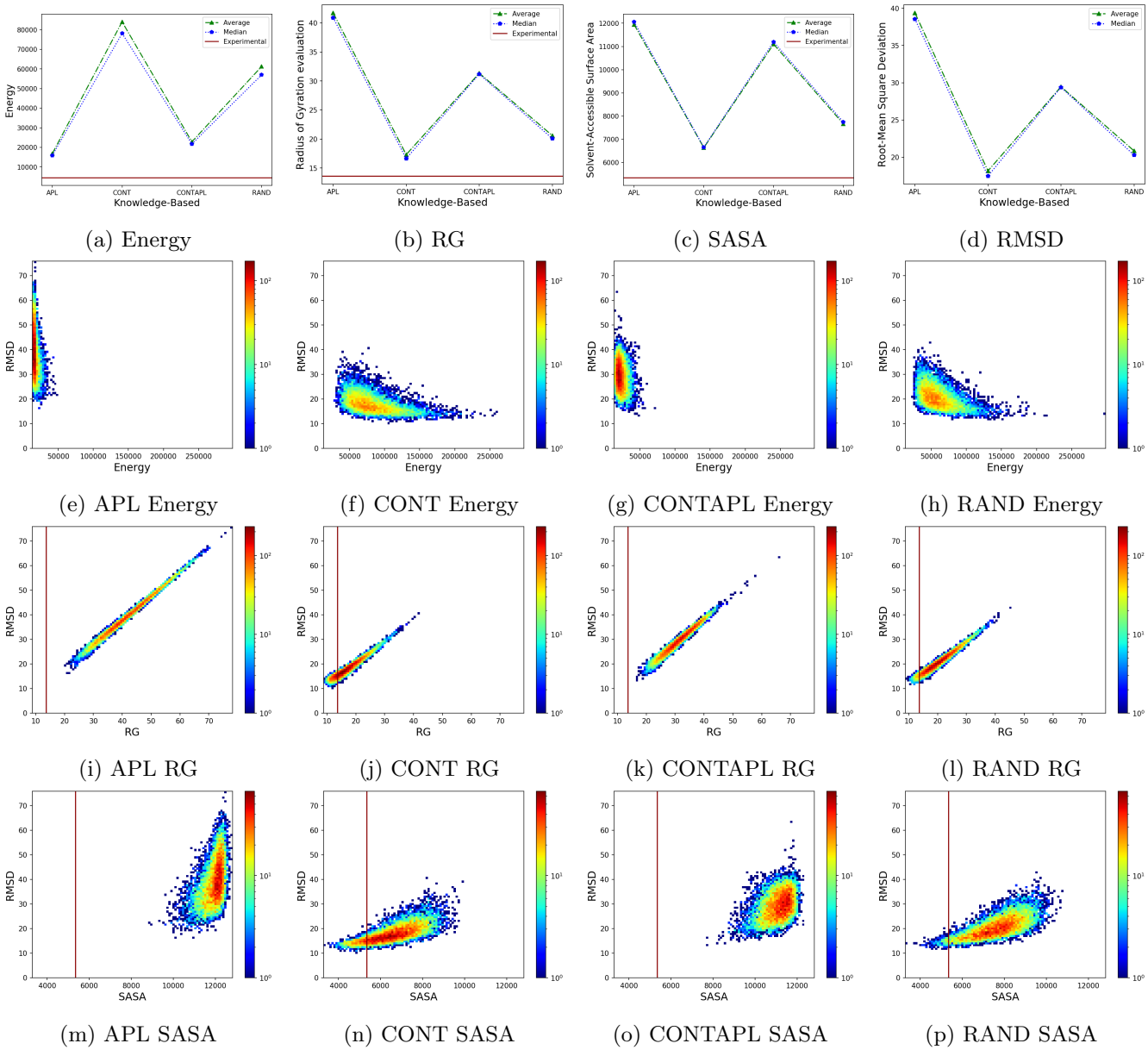
(a) Energy (b) RG (c) SASA (d) RMSD

(e) APL Energy (f) CONT Energy (g) CONTAPL Energy (h) RAND Energy

(i) APL RG (j) CONT RG (k) CONTAPL RG (l) RAND RG

(m) APL SASA (n) CONT SASA (o) CONTAPL SASA (p) RAND SASA

Figure S8: 1ENH populations' mean of the structural features score achieved by the Random (RAND), APL, Contacts only (CONT) and Contact and APL merged (CONTAPL), for each one of the proteins within our data set. Each heatmap was generated based on the density of combination between the RMSD and the other structural features. The dark red horizontal line represents the score calculated for the experimentally determined 3D structure.



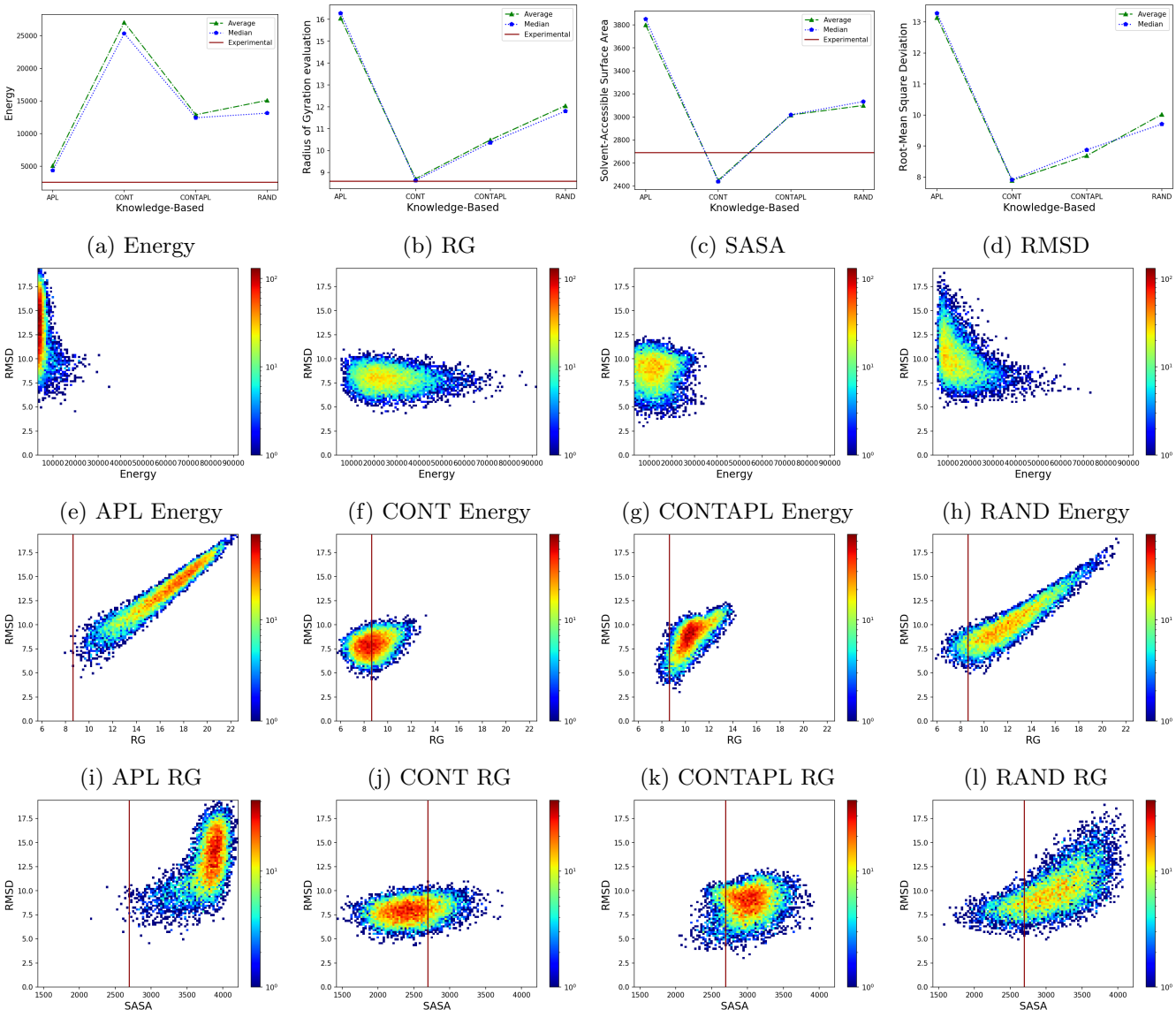
(a) Energy (b) RG (c) SASA (d) RMSD

(e) APL Energy (f) CONT Energy (g) CONTAPL Energy (h) RAND Energy

(i) APL RG (j) CONT RG (k) CONTAPL RG (l) RAND RG

(m) APL SASA (n) CONT SASA (o) CONTAPL SASA (p) RAND SASA

Figure S9: 1FNA populations' mean of the structural features score achieved by the Random (RAND), APL, Contacts only (CONT) and Contact and APL merged (CONTAPL), for each one of the proteins within our data set. Each heatmap was generated based on the density of combination between the RMSD and the other structural features. The dark red horizontal line represents the score calculated for the experimentally determined 3D structure.



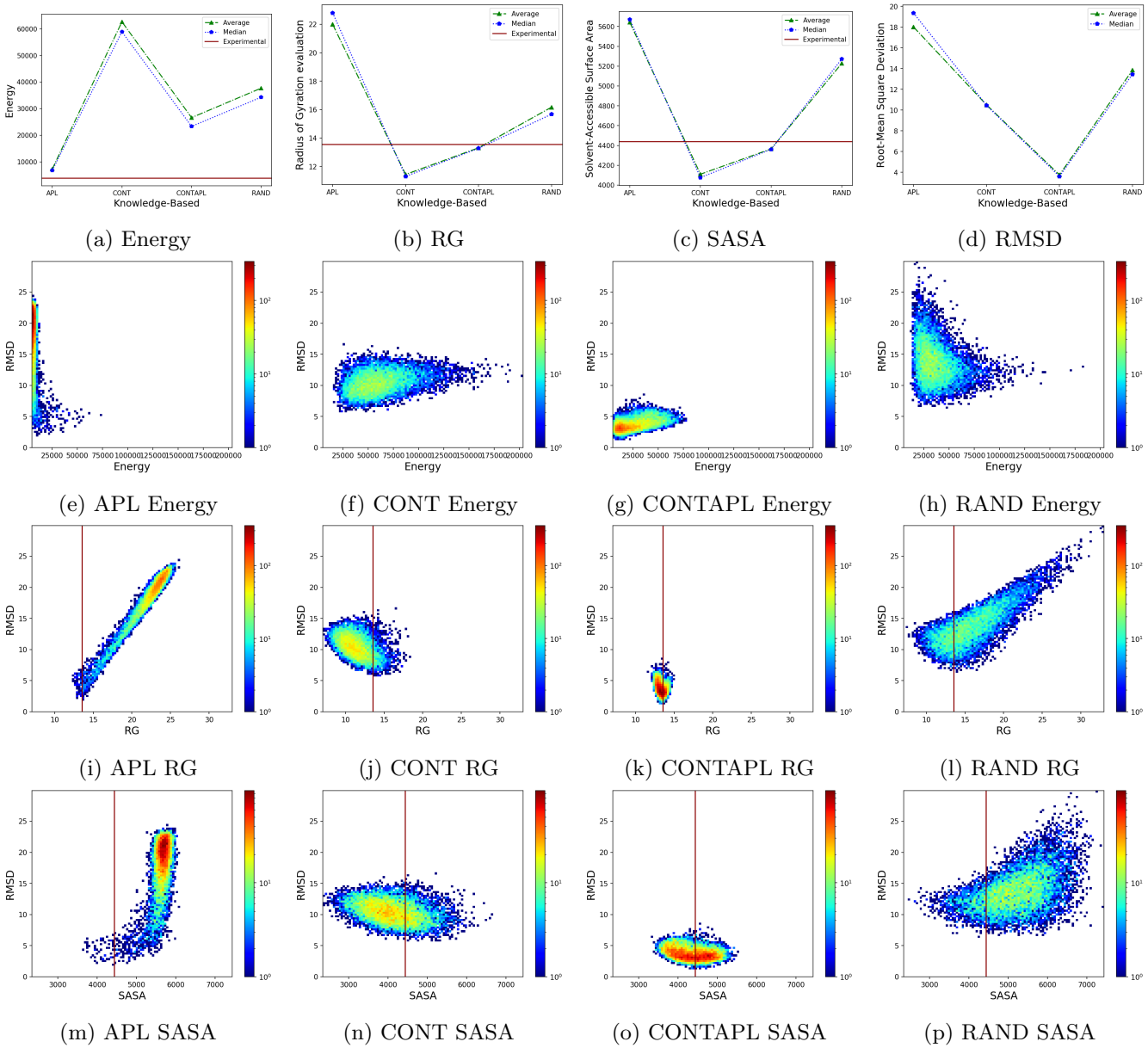
(a) Energy (b) RG (c) SASA (d) RMSD

(e) APL Energy (f) CONT Energy (g) CONTAPL Energy (h) RAND Energy

(i) APL RG (j) CONT RG (k) CONTAPL RG (l) RAND RG

(m) APL SASA (n) CONT SASA (o) CONTAPL SASA (p) RAND SASA

Figure S10: 1Q2K populations' mean of the structural features score achieved by the Random (RAND), APL, Contacts only (CONT) and Contact and APL merged (CONTAPL), for each one of the proteins within our data set. Each heatmap was generated based on the density of combination between the RMSD and the other structural features. The dark red horizontal line represents the score calculated for the experimentally determined 3D structure.



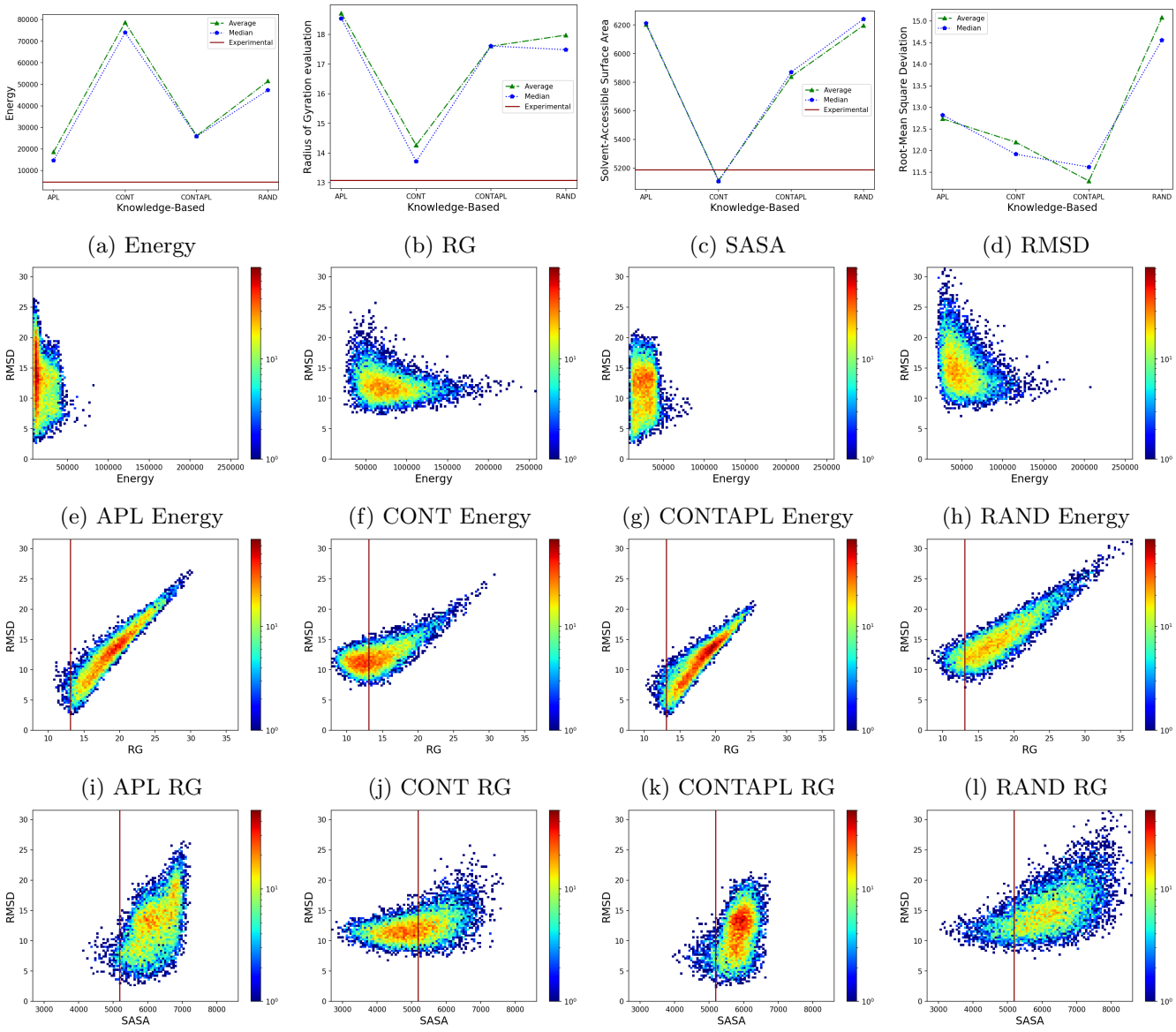
(a) Energy (b) RG (c) SASA (d) RMSD

(e) APL Energy (f) CONT Energy (g) CONTAPL Energy (h) RAND Energy

(i) APL RG (j) CONT RG (k) CONTAPL RG (l) RAND RG

(m) APL SASA (n) CONT SASA (o) CONTAPL SASA (p) RAND SASA

Figure S11: 1ROP populations' mean of the structural features score achieved by the Random (RAND), APL, Contacts only (CONT) and Contact and APL merged (CONTAPL), for each one of the proteins within our data set. Each heatmap was generated based on the density of combination between the RMSD and the other structural features. The dark red horizontal line represents the score calculated for the experimentally determined 3D structure.



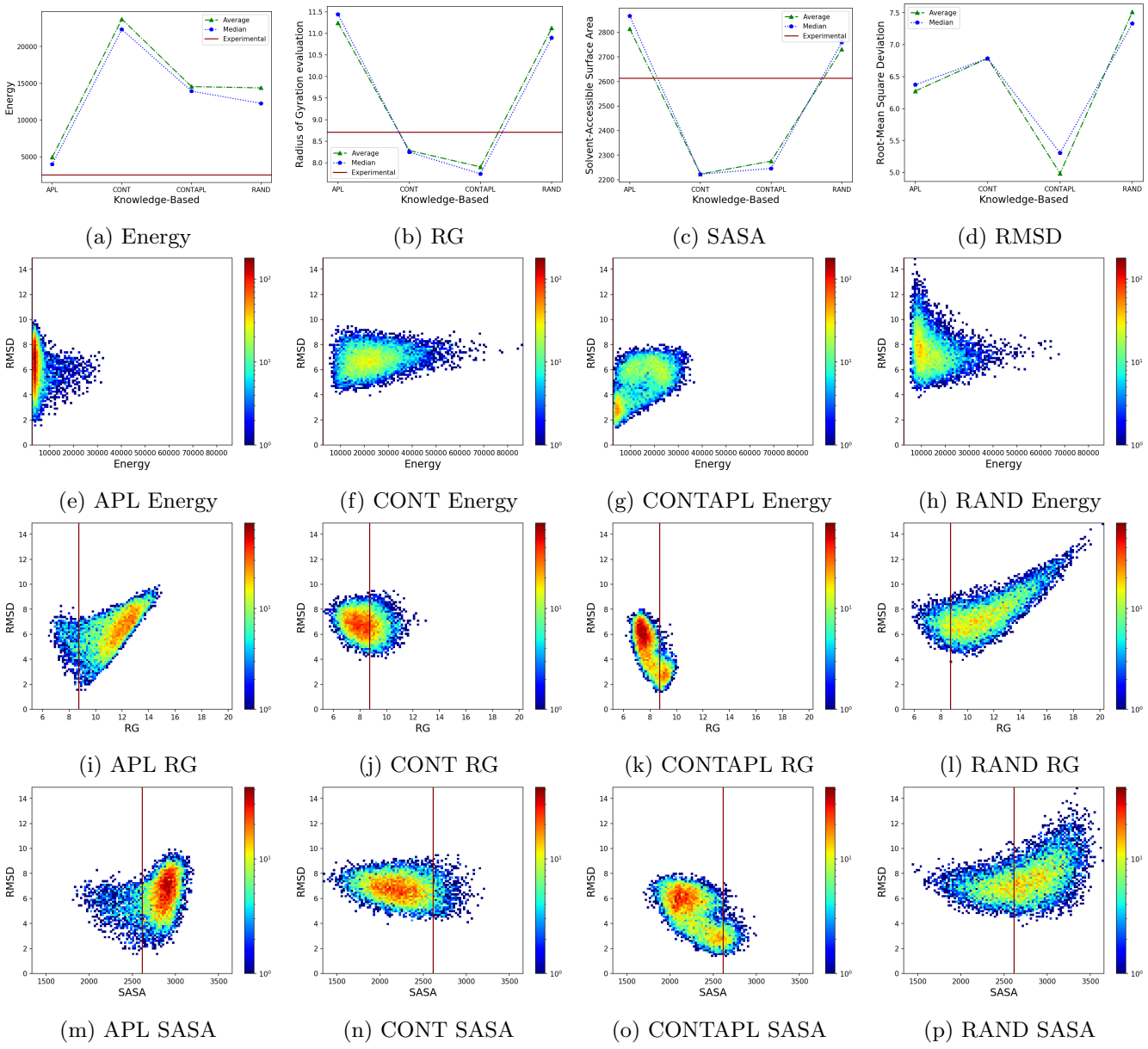
(a) Energy (b) RG (c) SASA (d) RMSD

(e) APL Energy (f) CONT Energy (g) CONTAPL Energy (h) RAND Energy

(i) APL RG (j) CONT RG (k) CONTAPL RG (l) RAND RG

(m) APL SASA (n) CONT SASA (o) CONTAPL SASA (p) RAND SASA

Figure S12: 1UTG populations' mean of the structural features score achieved by the Random (RAND), APL, Contacts only (CONT) and Contact and APL merged (CONTAPL), for each one of the proteins within our data set. Each heatmap was generated based on the density of combination between the RMSD and the other structural features. The dark red horizontal line represents the score calculated for the experimentally determined 3D structure.



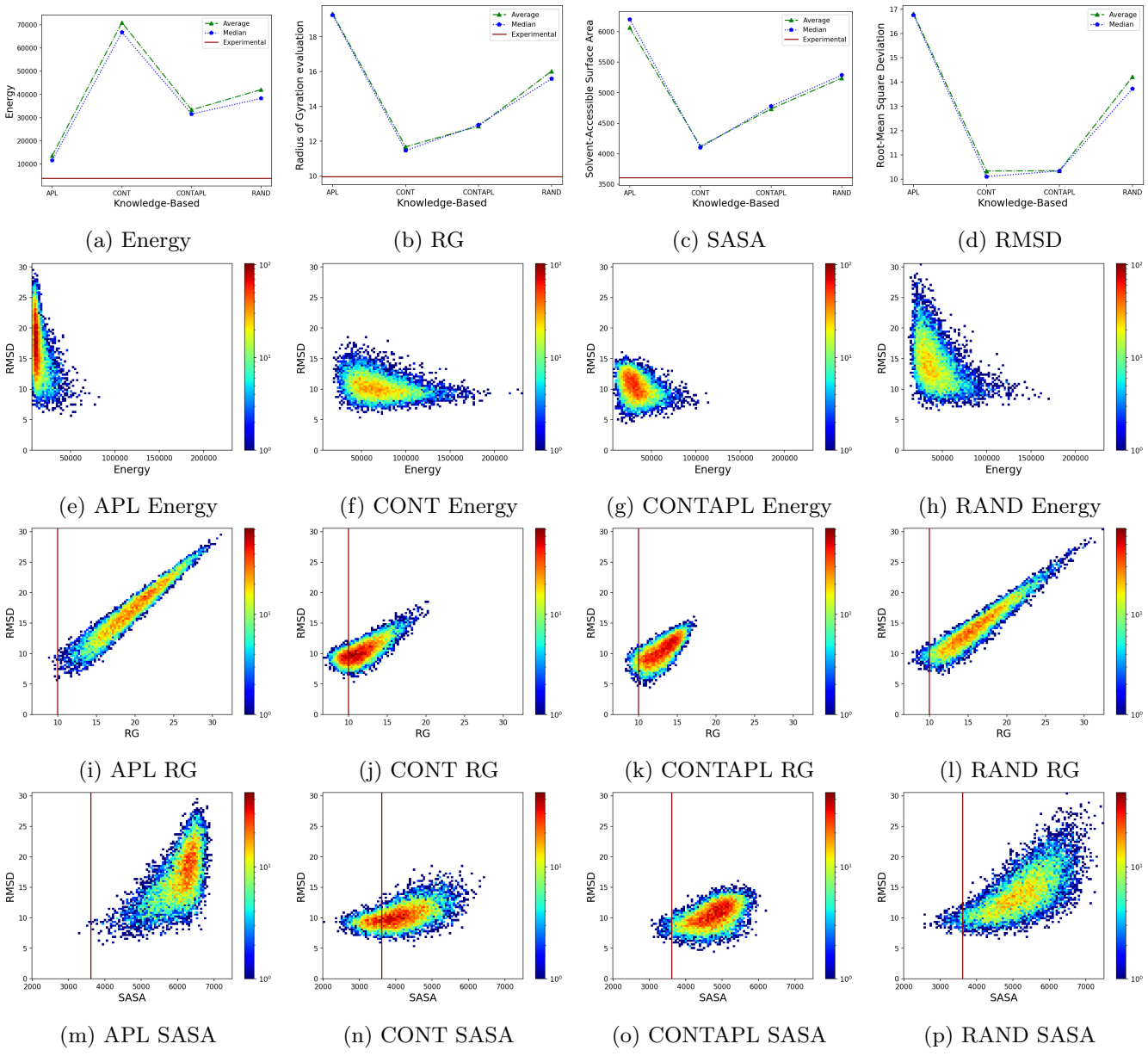
(a) Energy (b) RG (c) SASA (d) RMSD

(e) APL Energy (f) CONT Energy (g) CONTAPL Energy (h) RAND Energy

(i) APL RG (j) CONT RG (k) CONTAPL RG (l) RAND RG

(m) APL SASA (n) CONT SASA (o) CONTAPL SASA (p) RAND SASA

Figure S13: 1WQC populations' mean of the structural features score achieved by the Random (RAND), APL, Contacts only (CONT) and Contact and APL merged (CONTAPL), for each one of the proteins within our data set. Each heatmap was generated based on the density of combination between the RMSD and the other structural features. The dark red horizontal line represents the score calculated for the experimentally determined 3D structure.



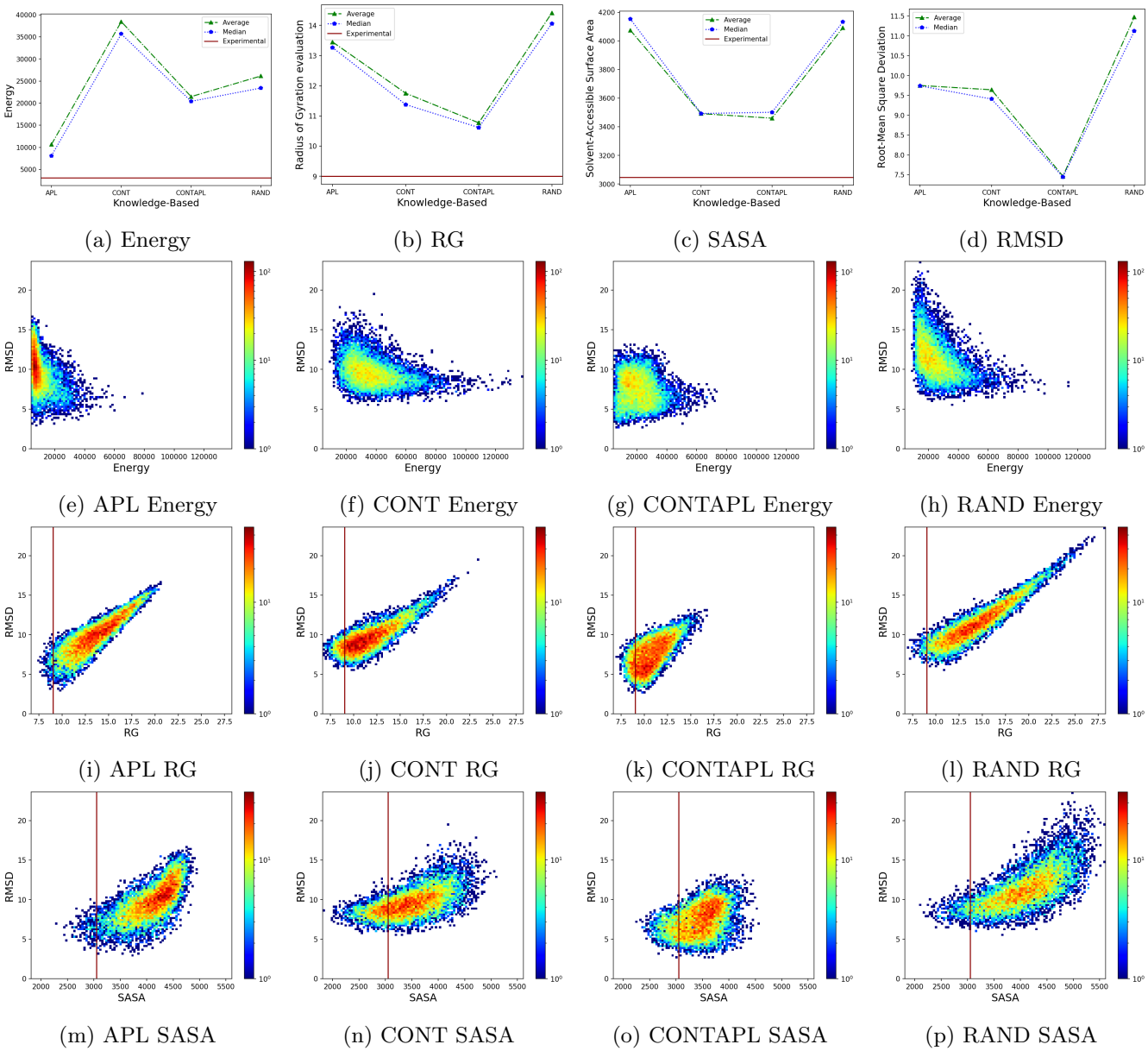
(a) Energy (b) RG (c) SASA (d) RMSD

(e) APL Energy (f) CONT Energy (g) CONTAPL Energy (h) RAND Energy

(i) APL RG (j) CONT RG (k) CONTAPL RG (l) RAND RG

(m) APL SASA (n) CONT SASA (o) CONTAPL SASA (p) RAND SASA

Figure S14: 2JUC populations' mean of the structural features score achieved by the Random (RAND), APL, Contacts only (CONT) and Contact and APL merged (CONTAPL), for each one of the proteins within our data set. Each heatmap was generated based on the density of combination between the RMSD and the other structural features. The dark red horizontal line represents the score calculated for the experimentally determined 3D structure.



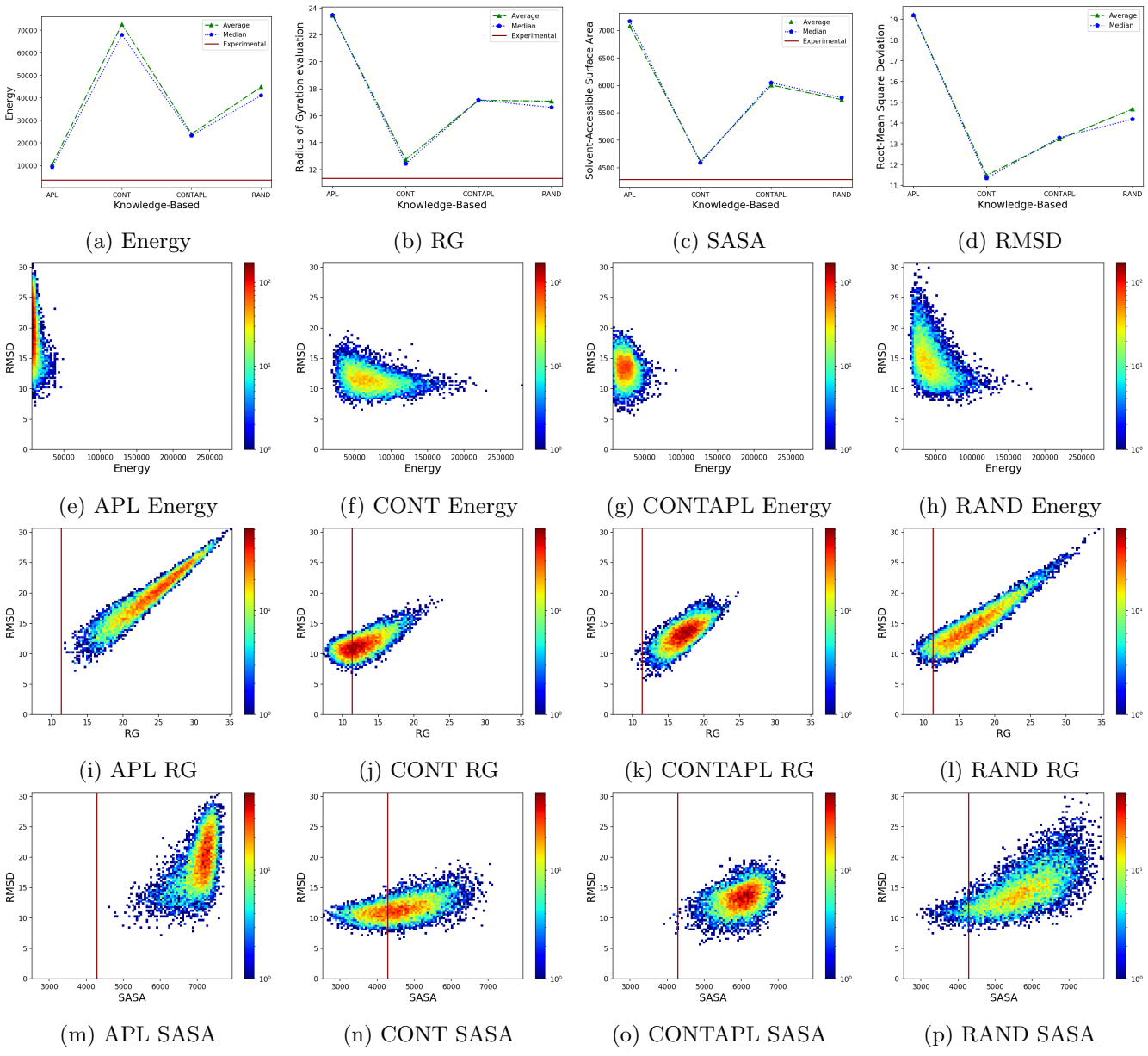
(a) Energy (b) RG (c) SASA (d) RMSD

(e) APL Energy (f) CONT Energy (g) CONTAPL Energy (h) RAND Energy

(i) APL RG (j) CONT RG (k) CONTAPL RG (l) RAND RG

(m) APL SASA (n) CONT SASA (o) CONTAPL SASA (p) RAND SASA

Figure S15: 2MR9 populations' mean of the structural features score achieved by the Random (RAND), APL, Contacts only (CONT) and Contact and APL merged (CONTAPL), for each one of the proteins within our data set. Each heatmap was generated based on the density of combination between the RMSD and the other structural features. The dark red horizontal line represents the score calculated for the experimentally determined 3D structure.



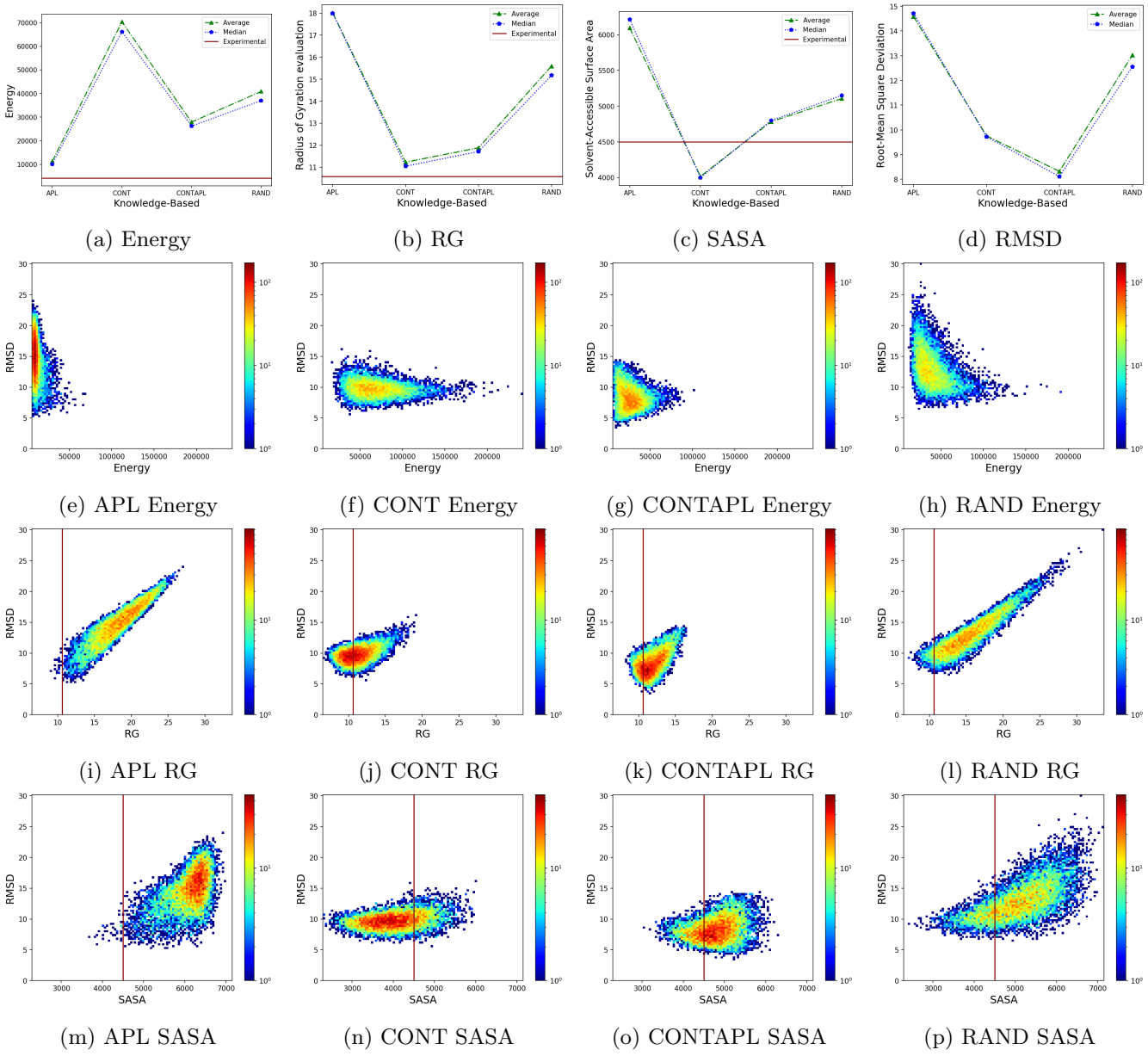
(a) Energy (b) RG (c) SASA (d) RMSD

(e) APL Energy (f) CONT Energy (g) CONTAPL Energy (h) RAND Energy

(i) APL RG (j) CONT RG (k) CONTAPL RG (l) RAND RG

(m) APL SASA (n) CONT SASA (o) CONTAPL SASA (p) RAND SASA

Figure S16: 2P5K populations' mean of the structural features score achieved by the Random (RAND), APL, Contacts only (CONT) and Contact and APL merged (CONTAPL), for each one of the proteins within our data set. Each heatmap was generated based on the density of combination between the RMSD and the other structural features. The dark red horizontal line represents the score calculated for the experimentally determined 3D structure.



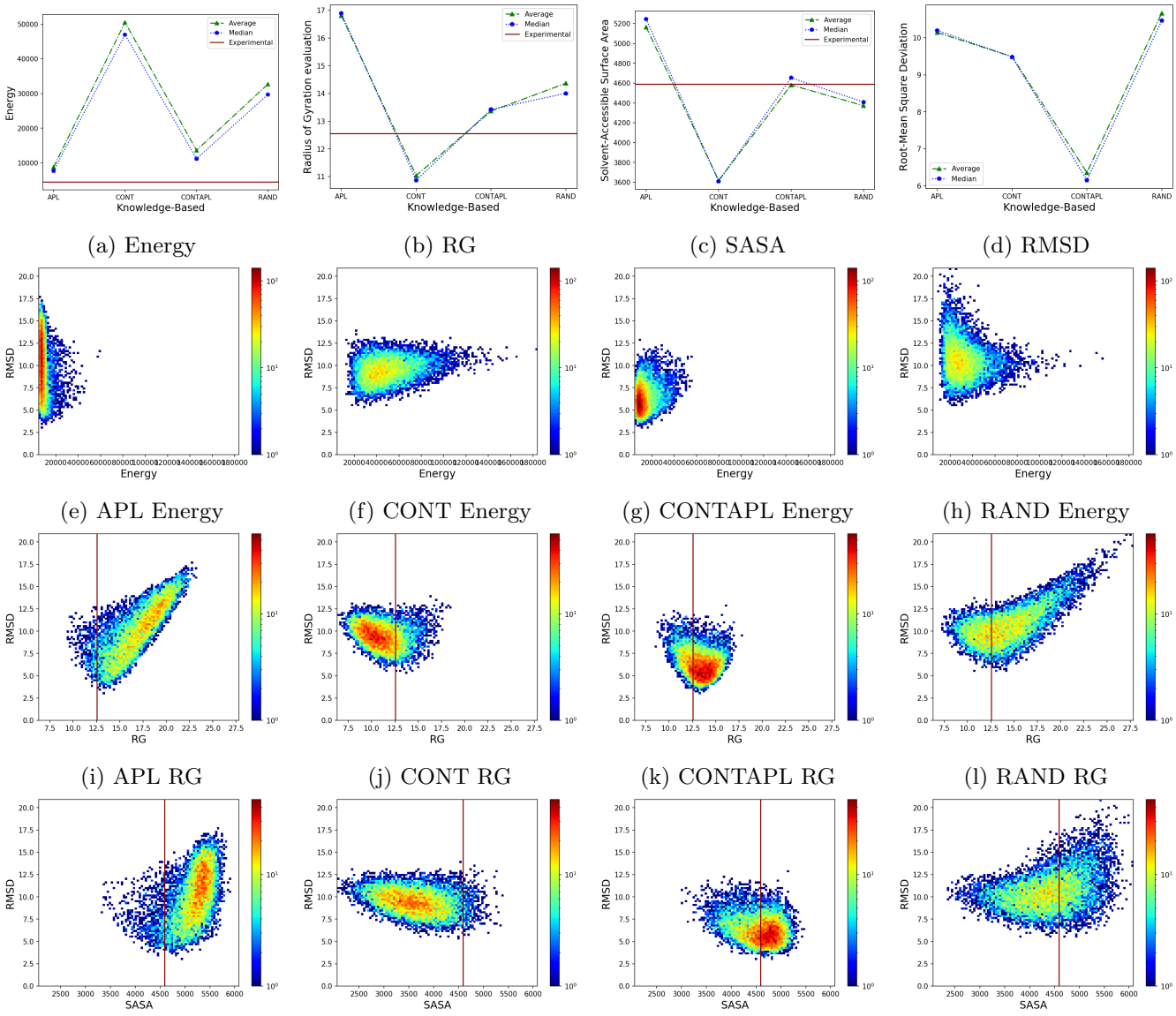
(a) Energy (b) RG (c) SASA (d) RMSD

(e) APL Energy (f) CONT Energy (g) CONTAPL Energy (h) RAND Energy

(i) APL RG (j) CONT RG (k) CONTAPL RG (l) RAND RG

(m) APL SASA (n) CONT SASA (o) CONTAPL SASA (p) RAND SASA

Figure S17: 2P6J populations' mean of the structural features score achieved by the Random (RAND), APL, Contacts only (CONT) and Contact and APL merged (CONTAPL), for each one of the proteins within our data set. Each heatmap was generated based on the density of combination between the RMSD and the other structural features. The dark red horizontal line represents the score calculated for the experimentally determined 3D structure.



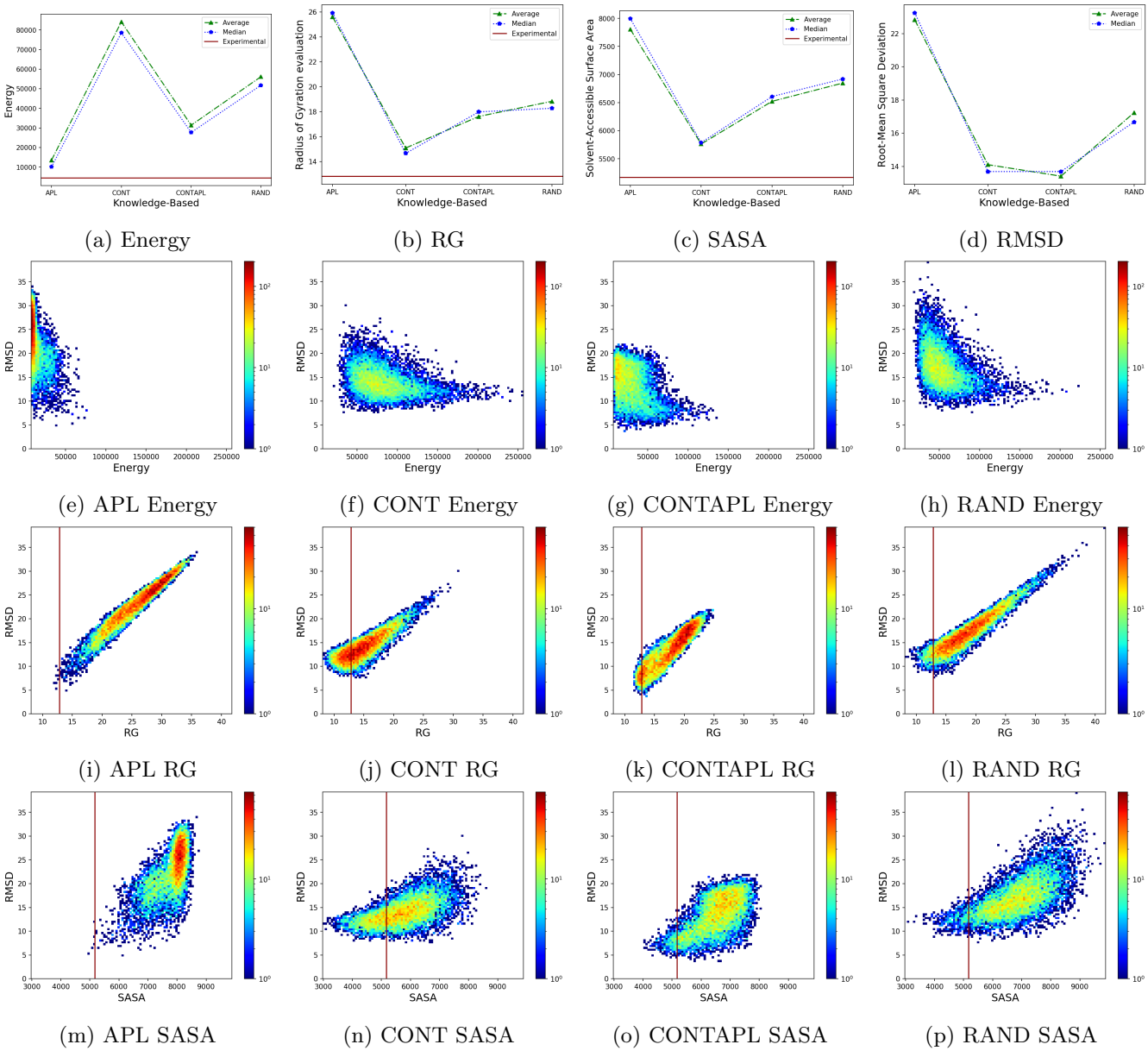
(a) Energy (b) RG (c) SASA (d) RMSD

(e) APL Energy (f) CONT Energy (g) CONTAPL Energy (h) RAND Energy

(i) APL RG (j) CONT RG (k) CONTAPL RG (l) RAND RG

(m) APL SASA (n) CONT SASA (o) CONTAPL SASA (p) RAND SASA

Figure S18: 2P81 populations' mean of the structural features score achieved by the Random (RAND), APL, Contacts only (CONT) and Contact and APL merged (CONTAPL), for each one of the proteins within our data set. Each heatmap was generated based on the density of combination between the RMSD and the other structural features. The dark red horizontal line represents the score calculated for the experimentally determined 3D structure.



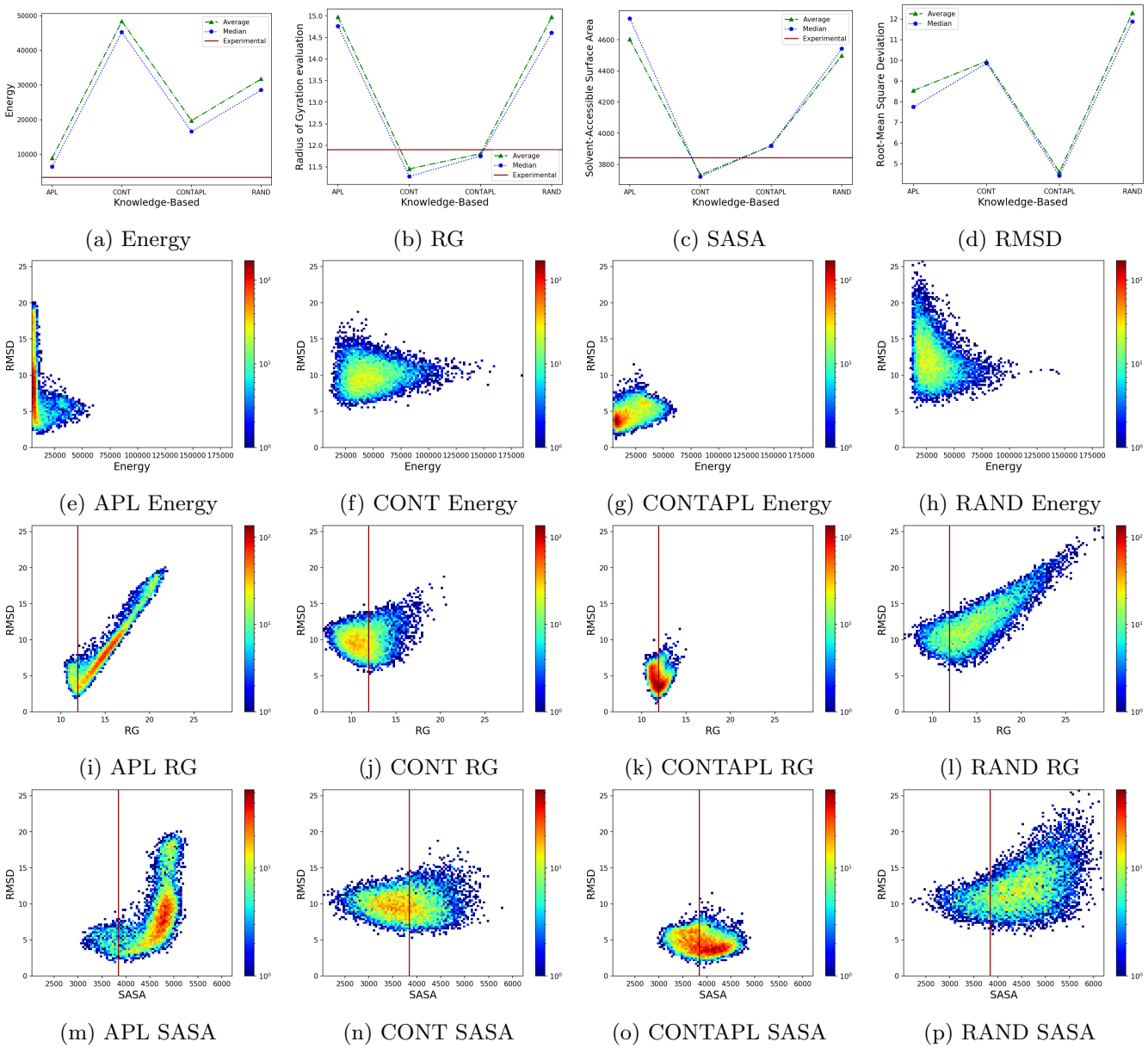
(a) Energy (b) RG (c) SASA (d) RMSD

(e) APL Energy (f) CONT Energy (g) CONTAPL Energy (h) RAND Energy

(i) APL RG (j) CONT RG (k) CONTAPL RG (l) RAND RG

(m) APL SASA (n) CONT SASA (o) CONTAPL SASA (p) RAND SASA

Figure S19: 2PMR populations' mean of the structural features score achieved by the Random (RAND), APL, Contacts only (CONT) and Contact and APL merged (CONTAPL), for each one of the proteins within our data set. Each heatmap was generated based on the density of combination between the RMSD and the other structural features. The dark red horizontal line represents the score calculated for the experimentally determined 3D structure.



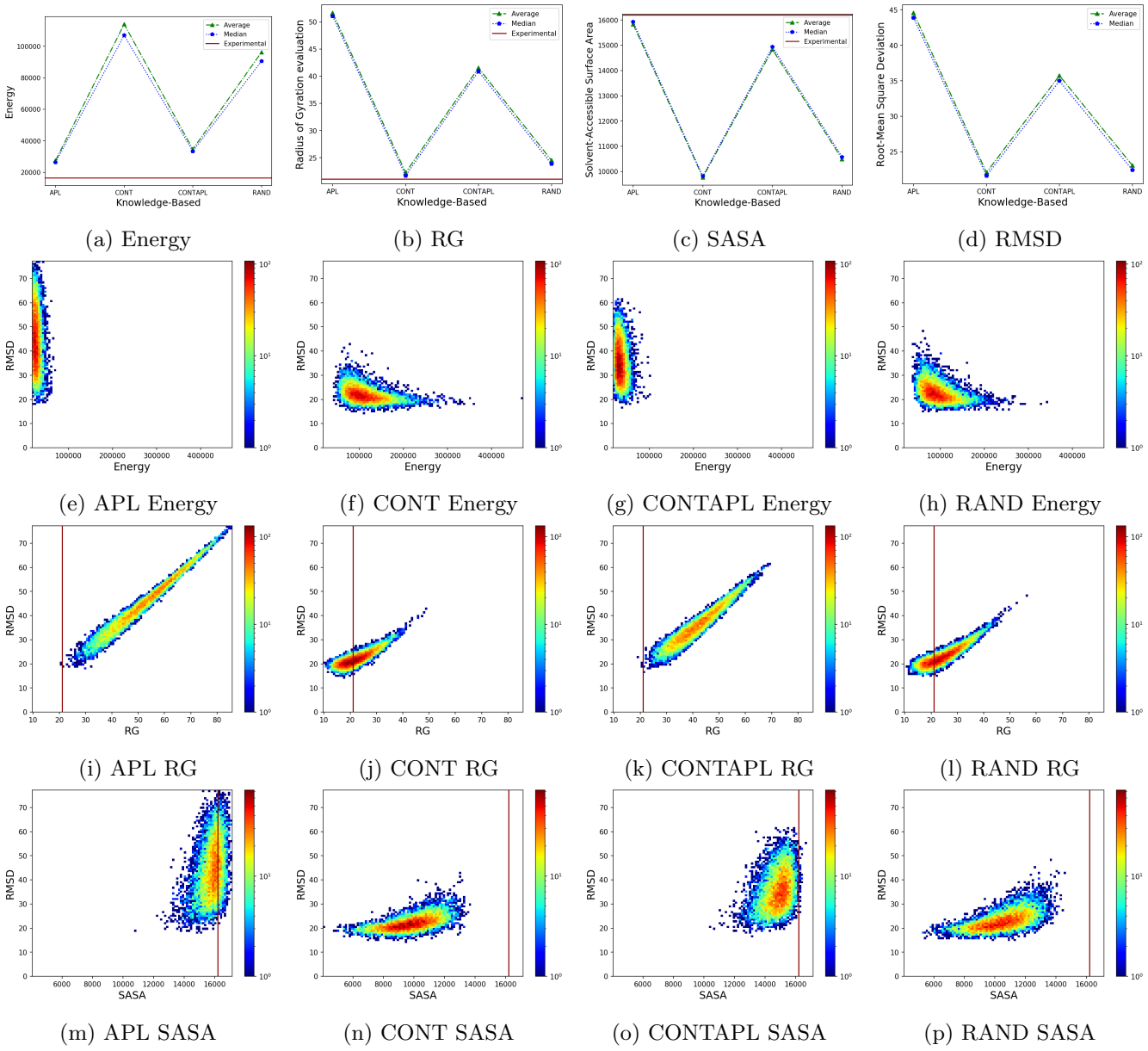
(a) Energy (b) RG (c) SASA (d) RMSD

(e) APL Energy (f) CONT Energy (g) CONTAPL Energy (h) RAND Energy

(i) APL RG (j) CONT RG (k) CONTAPL RG (l) RAND RG

(m) APL SASA (n) CONT SASA (o) CONTAPL SASA (p) RAND SASA

Figure S20: 3V1A populations' mean of the structural features score achieved by the Random (RAND), APL, Contacts only (CONT) and Contact and APL merged (CONTAPL), for each one of the proteins within our data set. Each heatmap was generated based on the density of combination between the RMSD and the other structural features. The dark red horizontal line represents the score calculated for the experimentally determined 3D structure.



(a) Energy (b) RG (c) SASA (d) RMSD

(e) APL Energy (f) CONT Energy (g) CONTAPL Energy (h) RAND Energy

(i) APL RG (j) CONT RG (k) CONTAPL RG (l) RAND RG

(m) APL SASA (n) CONT SASA (o) CONTAPL SASA (p) RAND SASA

Figure S21: 5JZR populations' mean of the structural features score achieved by the Random (RAND), APL, Contacts only (CONT) and Contact and APL merged (CONTAPL), for each one of the proteins within our data set. Each heatmap was generated based on the density of combination between the RMSD and the other structural features. The dark red horizontal line represents the score calculated for the experimentally determined 3D structure.

6 DISCUSSÃO GERAL

A quantidade de dados biológicos gerados tem aumentado consideravelmente ao longo dos últimos anos, principalmente no que diz respeito ao sequenciamento de DNA, RNA e sequência de aminoácidos. Este aumento na geração de dados é atribuído à diminuição de custos atrelados a metodologias de sequenciamento, assim como no surgimento de métodos de larga escala (*high throughput*) (MUIR et al., 2016). Contudo, ao comparar o número de sequências depositadas em bancos de dados de referência, por exemplo, RefSeq (PRUITT; TATUSOVA; MAGLOTT, 2006), e o número de estrutura 3D depositados no RCSB PDB, nota-se que o número de estruturas correspondem a 1% do número de sequências. Esta discrepância pode ser explicada pela metodologia mais comumente empregada para determinação destas estruturas. As metodologias experimentais, tais como Cristalografia obtida por Raio-X (JOHANNSSON; NEUMANN; FICNER, 2018) e Microscopia Eletrônica (ME) (GE et al., 2015), apesar de amplamente adotadas, possuem alguns limitantes, como o alto custo de execução, assim como o tempo necessário e possíveis erros ao longo da execução destes protocolos experimentais (VERLI, 2014). Tais limitantes e dificuldades justificam os grandes esforços de pesquisadores das mais diversas áreas do conhecimento em buscar metodologias alternativas para obtenção de informações estruturais.

A elucidação computacional de estrutura de proteínas é o principal problema ainda a ser solucionado pela Bioinformática Estrutural. O problema de predição de estrutura de proteínas, apesar de já ser objeto de estudos há mais de 20 anos, ainda permanece sem uma resolução ótima (DILL; MACCALLUM, 2012). Este problema é categorizado como um problema NP-difícil segundo a teoria de complexidade computacional (UNGER; MOULT, 1993), permitindo que ao longo destes anos, uma grande variedade de abordagens diferentes fossem propostas para resolução de tal. Idealmente, estes métodos seriam capazes de predizer a conformação 3D correta de uma determinada proteína utilizando única e exclusivamente a sequência de aminoácidos constituintes da proteína-alvo, os chamados métodos *ab initio* (OSGUTHORPE, 2000). Contudo, estes métodos baseiam-se no postulado de que a conformação que apresente a menor energia livre corresponde ao estado nativos destas proteínas (ANFINSEN, 1973). Porém, ainda não existe um entendimento total das leis físico-químicas que regem o processo de enovelamento, para o desenvolvimento de uma função de energia fidedigna. Estes métodos também são limitados pelo aumento de complexidade do problema pela alta dimensionalidade dos

mínimos energéticos.

Uma maneira de contornar estes limitantes é a utilização de dados previamente determinados, a fim de diminuir a complexidade do problema. Estes métodos são capazes de gerar resultados de alta qualidade, porém dependem diretamente da quantidade e qualidade de dados disponibilizados em bancos de dados (DORN et al., 2014). Contudo, são descritas metodologias as quais se beneficiam de informações previamente depositadas em bancos de dados, mas que também utilizam processos de minimizações energéticas a fim de aumentar a acurácia de predição (FLOUDAS et al., 2006). Tais métodos utilizam fragmentos provenientes de bancos de dados estruturais, seguidos de etapas de otimização de estruturas através de minimização energética (SIMONS et al., 1997). Estes métodos são destacados por serem os métodos utilizados pelos vencedores das últimas edições do descritor do estado-da-arte de PSP (MOULT et al., 2016).

Os resultados obtidos na edição CASP12 demonstram que a enorme quantidade de dados gerados em decorrência das novas metodologias disponíveis para sequenciamento de DNA são responsáveis pelo aumento significativo na precisão dos resultados de preditores de contrato 3D entre pares de aminoácidos a partir de análises evolutivas (SCHAARSCHMIDT et al., 2018). O aumento de precisão de predição de contatos proporcionou um aumento na eficiência de metodologias propostas para resolução do problema de PSP, as quais incorporam esta informação ao longo de sua metodologia (ABRIATA et al., 2018b). Este aumento na precisão de tais métodos ainda se manteve verdadeiro para a edição CASP13, reforçando seu alto grau de importância. A utilização destas informações evolutivas são implementadas das mais diversas maneiras. Por exemplo, Ovchinnikov et al. (2018) utilizam as informações contidas nos mapas de contato para identificação de possíveis domínios presentes na proteína-alvo, enquanto que Zhang et al. (2018) utiliza a informação de contatos para construção de modelos estruturais através de dinâmicas moleculares, assim como propõem a utilização destes contato em uma função para avaliação de qualidade de modelos gerados embasada na não-violação destes contatos preditos.

Durante a execução deste trabalho foram utilizadas informações de alta relevância biológica, a fim de gerar modelos estruturais com características semelhantes às observadas nas estruturas experimentalmente determinadas. A primeira informação utilizada foi baseada na preferência conformacional de aminoácidos específicos de acordo com a estrutura secundária na qual este está inserido (BORGUESAN et al., 2015). A adição deste limitante foi escolhida, primeiramente, pela vasta descrição de preferência de combinações de ângulos diedrais de acordo com a estrutura secundária. Em Richardson (1981)

são descritas as preferências conformacionais de hélices, as quais possuem as regiões de preferência ϕ = ($-70.0°$,$-60.0°$) e ψ = ($-45.0°$,$-39.0°$), assim como as β-folhas ϕ = ($-139.0°$,$-119.0°$) e ψ = ($-135.0°$,$-113.0°$). Porém, ao analisar de maneira mais refinada, é possível notar que estes padrões conformacionais são dependentes não apenas da estrutura secundária, mas também dependem do tipo de aminoácido. Descritos por Ligabue-Braun et al. (2018), cada aminoácido possui diferentes combinações de ângulos de torção dependendo do tipo de estrutura secundária sob o qual este está inserido, os quais não necessariamente se assemelham aos padrões comumente associados às estruturas secundárias regulares.

Diante disto, foi testada a eficiência de utilização destas preferências conformacionais de aminoácidos sob estrutura secundária específica, assim como a utilização de contatos preditos entre pares de aminoácidos. Para tal foram gerados modelos estruturais utilizando diferentes combinações destas informações: (*i*) apenas APL, (*ii*) apenas contatos, (*iii*) informações da APL em conjunto com contatos, além de (*iv*) modelos gerados aleatoriamente. As conformações de preferência provenientes das APLs foram responsáveis pela geração de modelos estruturais com os menores valores energéticos em comparação aos modelos gerados utilizando a geração aleatória de ângulos de torção. Desta maneira, é possível inferir que estas estruturas estão mais próximas aos estados nativos das proteínas. Contudo, no que diz respeito a proteínas, diversas conformações estruturais possuem valores energéticos próximos ou idênticos (Figura 3.4). Portanto, ao analisarmos as semelhanças de estrutura 3D entre a estrutura experimentalmente determinada e os modelos gerados apenas com a utilização dos ângulos retirados da APL, é possível notar que os valores de RMSD estão elevados (Capítulo 5 - Material suplementar). Este fato provavelmente é o resultado da dificuldade da APL de aproximação entre segmentos distintos da cadeia polipeptídica, fator crucial para formação de β-Folhas (NELSON; LEHNINGER; COX, 2008). As combinações de SASA e RMSD, assim como RG e RMSD, reforçam ainda mais os indícios de que a APL, apesar de gerar estruturas de alta estabilidade, por exemplo hélices, não é capaz de aproximar segmentos das cadeias da proteínas, de forma a mimetizar o enovelamento desta proteína. A Figura 4.2a demonstra a grande liberdade de rotação de aminoácidos sob estrutura secundária *Coil*, portanto nota-se que isto pode ocasionar na possível torção errônea de segmentos que deveriam estar próximos no espaço 3D.

Os indivíduos gerados a partir de informação de acoplamento evolutivo entre pares de aminoácidos (WEIGT et al., 2009), em oposição aos gerados a partir das APL,

possuíam altos valores energéticos, porém com níveis de empacotamento mais similares aos descritos das estruturas depositadas no PDB (Capítulo 5 - Material suplementar). Os valores energéticos elevados podem ser resultados de possíveis sobreposições atômicas causada pela aproximação de segmentos distintos da proteína, uma vez que este limitante não possui nenhuma restrição que impeça tal sobreposição. Diferente da APL, pois esta permite a formação de α-hélices, por exemplo, permitindo a distribuição correta e não sobreposto dos átomos da proteína-alvo. Apesar disto, ao avaliarmos os descritores que indicam a aproximação entre as cadeias do peptídeo, é possível constatar que os contatos foram responsáveis por aproximar os indivíduos da estrutura experimentalmente descrita. Portanto, ao avaliarmos as populações de modelos estruturais construídos sob a influência de ambos os parâmetros é possível constatar valores energéticos mais baixos, em decorrência da APL, assim como parâmetros indicadores de enovelamento mais próximos ao experimental, por influência dos contatos (Capítulo 5 - Material suplementar).

A fim de testar a eficácia e a utilidade dos modelos gerados para construção de populações iniciais, as quais possuam a capacidade de conferir vantagens no processo de busca de algoritmos de otimização, foi proposta a utilização de dois algoritmos populacionais. O primeiro deles é a versão canônica do algoritmo Evolução Diferencial, proposto por Storn and Price (1997), otimizado para minimização do valor de energia calculado a partir da equação de energia composta descrita previamente (Equação 2.6). A escolha deste algoritmo decorre da sua eficiência previamente demonstrada por Narloch and Dorn (2019a), onde também foram utilizadas as informações provenientes das APLs para geração de indivíduos iniciais. Contudo, não utilizaram-se as informações de contato entre aminoácidos. O segundo algoritmo proposto é a versão multiobjetivo do algoritmo de Evolução Diferencial, o qual avalia os indivíduos não apenas utilizando a equação composta, mas também uma função de energia baseada em contatos (Equação 4.8) (HONG et al., 2018).

Os resultado obtidos a partir da otimização utilizando um único objetivo (Capítulo 5 - *Table 5*) demonstram que, em termos energéticos, as populações geradas a partir das APLs são as detentoras dos melhores resultados, independente da utilização de contatos. Tal resultado corrobora com os achados de outros autores que utilizam esta informação para geração de estruturas iniciais a serem otimizadas (CORRÊA et al., 2018; BORGUESAN et al., 2015; NARLOCH; DORN, 2019a). Contudo, ao avaliarmos os valores de RMSD mínimos encontrados, salvo para proteína 1WQC, observamos que não são idealmente próximos às estruturas determinadas experimentalmente. Isto confirma as

hipóteses iniciais de que os baixos valores de energia obtidos pelas populações construídas utilizando os ângulos de torção retirados das APLs eram resultados da formação de hélices, como é possível notar pelas estruturas de menor valor energético encontradas pela otimização (Capítulo 5 - *Table 6*). Estes afastamentos de contatos podem ser ocasionados pela falta de uma métrica que penalize as estruturas que violem esta informação, tal qual utilizada em Zhang et al. (2018).

Os objetivos utilizados na versão multiobjetivo do algoritmo de Evolução Diferencial foram determinados de maneira a garantir a minimização de energia livre, a fim de encontrar estruturas de maior estabilidade (ANFINSEN, 1973), assim como garantir que a informação de contatos não fosse perdida ao longo da otimização. Os indivíduos selecionados de cada execução correspondem ao modelo cujo valores de objetivo não podem ser minimizados sem que o resultado do outro objetivo seja prejudicado. Os resultados obtidos para esta otimização possuem valores energéticos maiores, quando comparados aos resultados da otimização de um único objetivo (Capítulo 5 - *Table 7*). Contudo, os valores de RMSD são consideravelmente mais baixos, demonstrando a eficácia do método proposto. Em conjunto, os resultados obtidos pela análise de GDT demonstram a maior precisão do método quando utilizados ambos limitantes propostos para geração da população inicial. Dentro do conjunto de teste, composto por 10 proteínas, 9 proteínas obtiveram os maiores scores de GTD para populações geradas utilizando o método proposto.

Os desempenhos das otimizações, de um único objetivo e multiobjetivos, utilizando as populações geradas a partir de ambos limitantes, foram comparadas as predições realizadas pelos dois vencedores da edição CASP12, Rosetta (SONG et al., 2013) e QUARK (XU; ZHANG, 2012). Os resultados descritos na Capítulo 5 - *Table 9* demonstram que, apesar de os valores energéticos, segundo a Equação 2.6, não serem idealmente baixos, ao comparar os valores de RMSD obtidos para o método de otimização utilizando ambos objetivos propostos, são próximos aos obtidos pelo método do Rosetta. Destaca-se o valor de RMSD mais baixo encontrado para proteína 1AB1 utilizando o método proposto por esta dissertação. De tal maneira é correto afirmar que o método proposto é relevante no que diz respeito à geração de modelos estruturais com características semelhantes às descritas a partir de estruturas determinadas experimentalmente. Assim como a utilização destes modelos influenciam positivamente no desempenho de algoritmos de otimização, aproximando os resultados aos resultados obtidos pelos métodos referentes ao estado-da-arte.

7 CONCLUSÕES E PERSPECTIVAS

Os avanços científicos dentro da Bioinformática nas últimas décadas pode ser constatado a partir da análise do crescimento no número de dados depositados em bancos de dados. Contudo, apesar deste crescimento, uma grande discrepância pode ser observado ao comparar a quantidade exorbitante de dados de sequência em relação aos dados estruturais. Desta forma, é possível atestar a alta necessidade do desenvolvimento de metodologias capazes de elucidar a estrutura 3D de proteínas de maneira eficaz e com maior precisão. O problema de PSP ainda permanece sem uma solução ótima. Portanto, ao longo dos últimos anos, pesquisadores das mais diversas áreas voltam seu esforços ao desenvolvimento de novas metodologias para resolução de tal problema.

Os resultados obtidos nas últimas edições do CASP demonstram que a utilização de informações previamente descritas é de extrema importância para obtenção de estruturas de proteínas com maior similaridade às depositadas no PDB. Dentre as informações de maior relevância utilizadas pelos preditores vencedores da edição CASP12, estão as informações de contatos 3D inferidos a partir de análises evolutivas. Além disto, demonstrando a eficácia de métodos de otimização de estrutura, ambas as metodologias vencedoras implementam processos de otimização de modelos ao longo de sua execução.

Neste trabalho foi proposto, em primeira instância, uma metodologia de geração de modelos estruturais guiados por informações biológicas de alta relevância. Posteriormente, foi proposta a utilização destes modelos como indivíduos pertencentes às populações iniciais de dois algoritmos de otimização baseados em população. Sendo um dos algoritmos de otimização de um único indivíduo, enquanto que o outro pertence à categoria de algoritmos multiobjetivos.

A primeira etapa deste trabalho propôs a construção de 4 populações distintas, cada qual constituída por 10.000 indivíduos. Cada população foi construída utilizando diferentes metodologias: (*i*) geração aleatória de indivíduos, onde os valores de ângulos poderiam assumir qualquer valor dentro do intervalo estipulado; (*ii*) construção a partir dos ângulos extraídos das APLs; (*iii*) geração de indivíduos onde os contatos preditos estavam dentro de uma distância 3D de 8 Å; e (*iv*) indivíduos gerados a partir de ângulos retirados da APL que mantinham os contatos preditos. Para cada população foram avaliados as características estruturais segundo sua energia livre, SASA e RG, assim como a semelhança entre os modelos gerados e a estrutura determinada experimentalmente. Desta etapa, conclui-se que:

- Em decorrência da maior restrição ao que se refere às conformações angulares permitidas para estruturas secundárias específicas, em especial para hélices, as APLs possuem uma grande importância para geração de indivíduos de baixa energia livre. Contudo, devido à grande liberdade conformacional apresentada por regiões *Coil*, assim como a dificuldade de aproximação das fitas que compõem β-Folhas, estas populações não foram capazes de atingir parâmetros referentes a RG e SASA próximos aos calculados para as estruturas referência.
- As populações geradas apenas utilizando os limitantes de contato entre pares de aminoácidos possuíram altos valores energéticos, em decorrência de aproximação de átomos de aminoácidos. Os valores de SASA e RG indicam que este limitante é capaz de gerar um aumento no grau de empacotamento apresentado pelos modelos estruturais.
- A junção de ambos limitantes propostos foi capaz de gerar indivíduos com valores energéticos menores pela influência das APLs, quando comparados aos modelos gerados apenas guiados pela informação de contato. Interessantemente, estes indivíduos também preservaram, superando ambas populações anteriores, os bons resultados de RG e SASA obtidos pela influência dos contatos. Destacam-se populações com valores de RMSD próximos ao ideal quando comparados com a estrutura depositada no PDB.

A partir disto, é possível concluir que a utilização conjunta de ambos os limitantes propostos possui o potencial de gerar modelos estruturas com níveis de estabilidade elevados e contendo características estruturais, tais como nível de enovelamento, próximos aos descritos a partir de estruturas experimentalmente determinadas.

Para testar a utilidade destes modelos gerados, foi proposta uma etapa de otimização de estrutura utilizando algoritmos de busca. Para tal, foi proposta a utilização de dois algoritmos, DE e DEMO, sendo o primeiro de um único objetivo, enquanto que o segundo pertence à classe de algoritmos multiobjetivos. Os modelos foram otimizados a partir de uma função de energia livre de resolução atômica, enquanto que para o DEMO, o segundo objetivo era composto por uma função de energia baseada na preservação de contatos preditos nas estruturas. A fim de avaliar os resultados obtidos em relação ao estado-da-arte de PSP, foi realizada a comparação com os resultados obtidos pelos dois vencedores do CASP12, Rosetta e QUARK. Observou-se, então:

- Os resultados obtidos a partir da otimização pela minimização de energia livre dos

modelos gerados demonstram que a inicialização de população utilizando informações provenientes da APL são capazes de auxiliar no processo de busca. Os resultados indicam que as informações provenientes dos contatos 3D entre aminoácidos foram perdidas durante o processo, pois não foi utilizada nenhuma métrica para penalização de possíveis violações.

- A otimização realizada pelo algoritmo DEMO foi capaz de gerar estruturas com valores adequados tanto de RMSD quanto de GDT para as populações inicializadas com os indivíduos construídos a partir de ambos limitantes. Demonstrando que ambos os objetivos, tanto de energia livre e energia de contatos, são importantes para geração de modelos mais próximos aos encontrados no PDB.
- Quando comparados os valores energéticos obtidos pelos indivíduos otimizados por DE e por DEMO, é possível constatar um aumento significativo para os resultantes do DEMO. Apesar disto, as similaridades encontradas destes para com as estruturas do PDB indicam que apesar do valor energético, estas estão mais próximas dos modelos ideais.
- Ao comparar os resultados obtidos dos preditores pertencentes ao estado-da-arte, é possível observar a grande similaridade de valores mínimos de RMSD encontrados por eles em comparação aos obtidos pela otimização multiobjetivo proposta.

Os resultados obtidos das otimizações, principalmente multiobjetivo, demonstram o grande potencial do método de geração de modelos estruturais proposto por esta dissertação. É evidente que melhoras necessitam ser realizadas, assim como uma maior quantidade de testes para avaliar a robustez do método em relação ao estado-da-arte. De tal maneira, a conclusão deste trabalho abre espaço para futuras investigações a afim de explorar de maneira exaustiva a utilização de contatos e preferências conformacionais a fim de propor uma resolução ao problema de PSP, assim como utilização de diferentes informações para construção de indivíduos de populações para otimização. Destaca-se também a exploração de diferentes métricas a serem utilizadas como função de avaliação ao longo do processo de otimização de estruturas, de modo que sejam avaliados diversos parâmetros relevantes ao que concerne o processo correto de enovelamento de proteínas. Como trabalhos futuros no que tange o trabalho descrito nesta dissertação, propõe-se:

1. Aumentar o número de proteínas consideradas na base de dados para geração de APLs, a fim de abranger uma gama maior de conformações. De tal maneira, à medida que o número de estruturas aumenta, também espera-se uma maior quantidade

de dados compondo APLs de janelas maiores.

2. Aumentar o número de proteínas otimizadas, a fim de explorar de maneira mais ampla a capacidade de utilização dos indivíduos gerados a partir do método proposto.
3. Testar diferentes tipos de funções de energia baseadas em contatos preditos. Desta maneira, espera-se que a informação contida nas análises evolutivas possa ser utilizada de maneira mais eficaz ao longo do processo de otimização.
4. Nota-se que a exploração das soluções geradas através de algoritmos multiobjetivos é realizada a partir de um conjuntos de soluções ótimas de acordo com os valores de objetivos encontrados. Portanto, a análise de forma mais refinada, priorizando os resultados encontrados com energias menores, a fim de diminuir sobreposição atômica, por exemplo, pode ser explorada de maneira mais minuciosa.
5. Utilizar proteínas de tamanho maiores e topologias mais complexas. De acordo com descrições na literatura, as informações contidas nas análises evolutivas são capazes inclusive de inferir contato intermoleculares. Portanto, possivelmente será permitida a análise de complexos proteicos.
6. Explorar a utilização dos modelos gerados em diferentes técnicas. Por exemplo, utilizando dinâmica molecular para minimização dos modelos gerados, sem a necessidade de utilização de algoritmos de otimização. Utilizar a metodologia proposta para geração de fragmentos, segundo ambos limitantes, os quais podem ser utilizados para construção de modelos complexos.

REFERÊNCIAS

ABRIATA, L. A. et al. Definition and classification of evaluation units for tertiary structure prediction in casp12 facilitated through semi-automated metrics. **Proteins: Structure, Function, and Bioinformatics**, Wiley Online Library, v. 86, p. 16–26, 2018.

ABRIATA, L. A. et al. Assessment of hard target modeling in casp12 reveals an emerging role of alignment-based contact prediction methods. **Proteins: Structure, Function, and Bioinformatics**, Wiley Online Library, v. 86, p. 97–112, 2018.

ACKLEY, D. H.; HINTON, G. E.; SEJNOWSKI, T. J. A learning algorithm for boltzmann machines. **Cognitive Science**, Elsevier, v. 9, n. 1, p. 147–169, 1985.

ALFORD, R. F. et al. The rosetta all-atom energy function for macromolecular modeling and design. **Journal of Chemical Theory and Computation**, ACS Publications, v. 13, n. 6, p. 3031–3048, 2017.

ALTSCHUL, S. F. et al. Basic local alignment search tool. **Journal of molecular biology**, Elsevier, v. 215, n. 3, p. 403–410, 1990.

ANFINSEN, C. B. Principles that govern the folding of protein chains. **Science**, JSTOR, v. 181, n. 4096, p. 223–230, 1973.

ANFINSEN, C. B. et al. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. **Proceedings of the National Academy of Sciences of the United States of America**, National Academy of Sciences, v. 47, n. 9, p. 1309, 1961.

BACK, T. **Evolutionary algorithms in theory and practice: evolution strategies, evolutionary programming, genetic algorithms**. [S.l.]: Oxford University Press, 1996.

BALDASSI, C. et al. Fast and accurate multivariate gaussian modeling of protein families: predicting residue contacts and protein-interaction partners. **PloS One**, Public Library of Science, v. 9, n. 3, p. e92721, 2014.

BAXEVANIS, A. D.; OUELLETTE, B. F. **Bioinformatics: a practical guide to the analysis of genes and proteins**. [S.l.]: John Wiley & Sons, 2004.

BELDA, I. et al. Evolutionary computation and multimodal search: A good combination to tackle molecular diversity in the field of peptide design. **Molecular Diversity**, Springer, v. 11, n. 1, p. 7–21, 2007.

BENITEZ-HIDALGO, A. et al. jmetalpy: a python framework for multi-objective optimization with metaheuristics. **arXiv preprint arXiv:1903.02915**, 2019.

BERKHOLZ, D. S. et al. Nonplanar peptide bonds in proteins are common and conserved but not biased toward active sites. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 109, n. 2, p. 449–453, 2012.

BERMAN, H. M. et al. The protein data bank. **Nucleic Acids Research**, Oxford University Press, v. 28, n. 1, p. 235–242, 2000.

BIRATTARI, M. et al. Classification of metaheuristics and design of experiments for the analysis of components tech. rep. aida-01-05. 2001.

BLUNDELL, T. et al. Knowledge-based prediction of protein structures and the design of novel molecules. **Nature**, Nature Publishing Group, v. 326, n. 6111, p. 347, 1987.

BOHR, J. et al. Protein structures from distance inequalities. **Journal of Molecular Biology**, Elsevier, v. 231, n. 3, p. 861–869, 1993.

BORGUESAN, B.; INOSTROZA-PONTA, M.; DORN, M. Nias-server: Neighbors influence of amino acids and secondary structures in proteins. **Journal of Computational Biology**, Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA, v. 24, n. 3, p. 255–265, 2017.

BORGUESAN, B. et al. A genetic algorithm based on restricted tournament selection for the 3d-psp problem. In: IEEE. **2018 IEEE Congress on Evolutionary Computation (CEC)**. [S.l.], 2018. p. 1–8.

BORGUESAN, B. et al. Apl: An angle probability list to improve knowledge-based metaheuristics for the three-dimensional protein structure prediction. **Computational Biology and Chemistry**, Elsevier, v. 59, p. 142–157, 2015.

BOUSSAÏD, I.; LEPAGNOT, J.; SIARRY, P. A survey on optimization metaheuristics. **Information Sciences**, Elsevier, v. 237, p. 82–117, 2013.

BOWIE, J. U.; LUTHY, R.; EISENBERG, D. A method to identify protein sequences that fold into a known three-dimensional structure. **Science**, American Association for the Advancement of Science, v. 253, n. 5016, p. 164–170, 1991.

BRANDEN, C. I.; TOOZE, J. **Introduction to protein structure**. [S.l.]: Garland Science, 2012.

BRASIL, C. R. S.; DELBEM, A. C. B.; SILVA, F. L. B. da. Multiobjective evolutionary algorithm with many tables for purely ab initio protein structure prediction. **Journal of Computational Chemistry**, v. 34, n. 20, p. 1719–1734, 2013. Available from Internet: <https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.23315>.

BROOKS, B. R. et al. Charmm: a program for macromolecular energy, minimization, and dynamics calculations. **Journal of Computational Chemistry**, Wiley Online Library, v. 4, n. 2, p. 187–217, 1983.

BURGER, L.; NIMWEGEN, E. V. Disentangling direct from indirect co-evolution of residues in protein alignments. **PLoS Computational Biology**, Public Library of Science, v. 6, n. 1, p. e1000633, 2010.

CAMACHO, C. et al. Blast+: architecture and applications. **BMC bioinformatics**, BioMed Central, v. 10, n. 1, p. 421, 2009.

CASINO, P.; RUBIO, V.; MARINA, A. Structural insight into partner specificity and phosphoryl transfer in two-component signal transduction. **Cell**, Elsevier, v. 139, n. 2, p. 325–336, 2009.

CHAUDHURY, S.; LYSKOV, S.; GRAY, J. J. Pyrosetta: a script-based interface for implementing molecular modeling algorithms using rosetta. **Bioinformatics**, Oxford University Press, v. 26, n. 5, p. 689–691, 2010.

CHENNA, R. et al. Multiple sequence alignment with the clustal series of programs. **Nucleic acids research**, Oxford University Press, v. 31, n. 13, p. 3497–3500, 2003.

CHIVIAN, D. et al. Ab initio methods. **Methods of biochemical analysis**, Wiley Online Library, v. 44, p. 547–558, 2003.

CHOU, K.-C. Structural bioinformatics and its impact to biomedical science. **Current Medicinal Chemistry**, Bentham Science Publishers, v. 11, n. 16, p. 2105–2134, 2004.

CHOU, K.-C.; ZHANG, C.-T. Prediction of protein structural classes. **Critical Reviews in Biochemistry and Molecular Biology**, Taylor & Francis, v. 30, n. 4, p. 275–349, 1995.

COCCO, S. et al. Inverse statistical physics of protein sequences: a key issues review. **Reports on Progress in Physics**, IOP Publishing, v. 81, n. 3, p. 032601, 2018.

CONNOLLY, M. L. Solvent-accessible surfaces of proteins and nucleic acids. **Science**, American Association for the Advancement of Science, v. 221, n. 4612, p. 709–713, 1983.

COOK, S. A. An overview of computational complexity. **Communications of the ACM**, ACM, v. 26, n. 6, p. 400–408, 1983.

CORREA, L. et al. A memetic algorithm for 3d protein structure prediction problem. **IEEE/ACM Transactions on Computational Biology and Bioinformatics**, IEEE, v. 15, n. 3, p. 690–704, 2016.

CORRÊA, L. de L. et al. Three-dimensional protein structure prediction based on memetic algorithms. **Computers & Operations Research**, Elsevier, v. 91, p. 160–177, 2018.

CORRÊA, L. de L.; DORN, M. Multi-agent systems in three-dimensional protein structure prediction. In: **Multi-Agent-Based Simulations Applied to Biological and Environmental Systems**. [S.l.]: IGI Global, 2017. p. 241–278.

CRESCENZI, P. et al. On the complexity of protein folding. **Journal of computational biology**, v. 5, n. 3, p. 423–465, 1998.

CUTELLO, V.; NARZISI, G.; NICOSIA, G. A multi-objective evolutionary approach to the protein structure prediction problem. **Journal of The Royal Society Interface**, The Royal Society London, v. 3, n. 6, p. 139–151, 2005.

DAS, S. et al. Real-parameter evolutionary multimodal optimization—a survey of the state-of-the-art. **Swarm and Evolutionary Computation**, Elsevier, v. 1, n. 2, p. 71–88, 2011.

DAWSON, N. L. et al. Cath: an expanded resource to predict protein function through structure and sequence. **Nucleic Acids Research**, Oxford University Press, v. 45, n. D1, p. D289–D295, 2016.

DEB, K. et al. A fast and elitist multiobjective genetic algorithm: Nsga-ii. **IEEE Transactions on Evolutionary Computation**, IEEE, v. 6, n. 2, p. 182–197, 2002.

DESJARLAIS, J. R.; CLARKE, N. D. Computer search algorithms in protein modification and design. **Current Opinion in Structural Biology**, Elsevier, v. 8, n. 4, p. 471–475, 1998.

DILL, K. A.; MACCALLUM, J. L. The protein-folding problem, 50 years on. **Science**, American Association for the Advancement of Science, v. 338, n. 6110, p. 1042–1046, 2012.

DORN, M. et al. A knowledge-based genetic algorithm to predict three-dimensional structures of polypeptides. In: IEEE. **2013 IEEE Congress on Evolutionary Computation**. [S.l.], 2013. p. 1233–1240.

DORN, M. et al. Three-dimensional protein structure prediction: Methods and computational strategies. **Computational Biology and Chemistry**, Elsevier, v. 53, p. 251–276, 2014.

DUNKER, A. K. et al. Intrinsically disordered protein. **Journal of Molecular Graphics and Modelling**, Elsevier, v. 19, n. 1, p. 26–59, 2001.

EDWARDS, A. M. et al. Protein production: feeding the crystallographers and nmr spectroscopists. **Nature Structural and Molecular Biology**, Nature Publishing Group, v. 7, n. 11s, p. 970, 2000.

ELOFSSON, A.; GRAND, S. M. L.; EISENBERG, D. Local moves: An efficient algorithm for simulation of protein folding. **Proteins: Structure, Function, and Bioinformatics**, Wiley Online Library, v. 23, n. 1, p. 73–82, 1995.

FARAGGI, E.; KLOCZKOWSKI, A. A global machine learning based scoring function for protein structure prediction. **Proteins: Structure, Function, and Bioinformatics**, Wiley Online Library, v. 82, n. 5, p. 752–759, 2014.

FATTORUSSO, R. et al. Nmr structure of the human oncofoetal fibronectin ed-b domain, a specific marker for angiogenesis. **Structure**, Elsevier, v. 7, n. 4, p. 381–390, 1999.

FLOUDAS, C. et al. Advances in protein structure prediction and de novo protein design: A review. **Chemical Engineering Science**, Elsevier, v. 61, n. 3, p. 966–988, 2006.

FOX, N. K.; BRENNER, S. E.; CHANDONIA, J.-M. The value of protein structure classification information—surveying the scientific literature. **Proteins: Structure, Function, and Bioinformatics**, Wiley Online Library, v. 83, n. 11, p. 2025–2038, 2015.

FRISHMAN, D.; ARGOS, P. Knowledge-based protein secondary structure assignment. **Proteins: Structure, Function, and Bioinformatics**, Wiley Online Library, v. 23, n. 4, p. 566–579, 1995.

GE, J. et al. Architecture of the mammalian mechanosensitive piezo1 channel. **Nature**, Nature Publishing Group, v. 527, n. 7576, p. 64, 2015.

GLIBOVETS, N.; GULAYEVA, N. A review of niching genetic algorithms for multimodal function optimization. **Cybernetics and Systems Analysis**, Springer, v. 49, n. 6, p. 815–820, 2013.

GRONT, D. et al. Generalized fragment picking in rosetta: design, protocols and applications. **PloS One**, Public Library of Science, v. 6, n. 8, p. e23294, 2011.

GUNASEKARAN, K. et al. Extended disordered proteins: targeting function with less scaffold. **Trends in Biochemical Sciences**, Elsevier, v. 28, n. 2, p. 81–85, 2003.

HAO, M.-H.; SCHERAGAT, H. A. Designing potential energy functions for protein folding. **Current Opinion in Structural Biology**, Elsevier, v. 9, n. 2, p. 184–188, 1999.

HE, B. et al. Nebcon: protein contact map prediction using neural network training coupled with naïve bayes classifiers. **Bioinformatics**, Oxford University Press, v. 33, n. 15, p. 2296–2306, 2017.

HEINIG, M.; FRISHMAN, D. Stride: a web server for secondary structure assignment from known atomic coordinates of proteins. **Nucleic acids research**, Oxford University Press, v. 32, n. suppl_2, p. W500–W502, 2004.

HONG, S. H. et al. Protein structure modeling and refinement by global optimization in casp12. **Proteins: Structure, Function, and Bioinformatics**, Wiley Online Library, v. 86, p. 122–135, 2018.

HOU, X.-F. et al. Enzymology of anthraquinone-γ-pyrone ring formation in complex aromatic polyketide biosynthesis. **Angewandte Chemie International Edition**, Wiley Online Library, v. 57, n. 41, p. 13475–13479, 2018.

HOVMÖLLER, S.; ZHOU, T.; OHLSON, T. Conformations of amino acids in proteins. **Acta Crystallographica Section D: Biological Crystallography**, International Union of Crystallography, v. 58, n. 5, p. 768–776, 2002.

ILLERGÅRD, K.; ARDELL, D. H.; ELOFSSON, A. Structure is three to ten times more conserved than sequence—a study of structural response in protein cores. **Proteins: Structure, Function, and Bioinformatics**, Wiley Online Library, v. 77, n. 3, p. 499–508, 2009.

JOHANNSSON, S.; NEUMANN, P.; FICNER, R. Crystal structure of the human trna guanine transglycosylase catalytic subunit qtrt1. **Biomolecules**, Multidisciplinary Digital Publishing Institute, v. 8, n. 3, p. 81, 2018.

JONES, J. E. On the determination of molecular fields.—ii. from the equation of state of a gas. **Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character**, The Royal Society London, v. 106, n. 738, p. 463–477, 1924.

JONES, J. E.; CHAPMAN, S. On the determination of molecular fields.-i. from the variation of the viscosity of a gas with temperature. **Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character**, v. 106, n. 738, p. 441–462, 1924. Available from Internet: <https://royalsocietypublishing.org/doi/abs/10.1098/rspa.1924.0081>.

JR, A. D. M. Empirical force fields for biological macromolecules: overview and issues. **Journal of Computational Chemistry**, Wiley Online Library, v. 25, n. 13, p. 1584–1604, 2004.

JR, R. L. D.; COHEN, F. E. Bayesian statistical analysis of protein side-chain rotamer preferences. **Protein Science**, Wiley Online Library, v. 6, n. 8, p. 1661–1681, 1997.

JUAN, D. D.; PAZOS, F.; VALENCIA, A. Emerging methods in protein co-evolution. **Nature Reviews Genetics**, Nature Publishing Group, v. 14, n. 4, p. 249–261, 2013.

KABSCH, W.; SANDER, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. **Biopolymers: Original Research on Biomolecules**, Wiley Online Library, v. 22, n. 12, p. 2577–2637, 1983.

KALLEN, J. et al. X-ray structures and analysis of 11 cyclosporin derivatives complexed with cyclophilin a1. **Journal of Molecular Biology**, Elsevier, v. 283, n. 2, p. 435–449, 1998.

KARPLUS, M.; MCCAMMON, J. A. Molecular dynamics simulations of biomolecules. **Nature Structural and Molecular Biology**, Nature Publishing Group, v. 9, n. 9, p. 646, 2002.

KESSEL, A.; BEN-TAL, N. **Introduction to proteins: structure, function, and motion**. [S.l.]: CRC Press, 2010.

KIM, D. E. et al. Sampling bottlenecks in de novo protein structure prediction. **Journal of Molecular Biology**, Elsevier, v. 393, n. 1, p. 249–260, 2009.

KIM, D. E.; CHIVIAN, D.; BAKER, D. Protein structure prediction and analysis using the robetta server. **Nucleic acids research**, Oxford University Press, v. 32, n. suppl_2, p. W526–W531, 2004.

KITCHEN, D. B. et al. Docking and scoring in virtual screening for drug discovery: methods and applications. **Nature Reviews Drug Discovery**, Nature Publishing Group, v. 3, n. 11, p. 935, 2004.

KLEYWEGT, G. J.; BRÜNGER, A. T. Checking your imagination: applications of the free r value. **Structure**, Elsevier, v. 4, n. 8, p. 897–904, 1996.

KONAK, A.; COIT, D. W.; SMITH, A. E. Multi-objective optimization using genetic algorithms: A tutorial. **Reliability Engineering & System Safety**, Elsevier, v. 91, n. 9, p. 992–1007, 2006.

KORTEMME, T.; MOROZOV, A. V.; BAKER, D. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein–protein complexes. **Journal of Molecular Biology**, Elsevier, v. 326, n. 4, p. 1239–1259, 2003.

KRYSHTAFOVYCH, A.; FIDELIS, K.; MOULT, J. Casp10 results compared to those of previous casp experiments. **Proteins: Structure, Function, and Bioinformatics**, Wiley Online Library, v. 82, p. 164–174, 2014.

KRYSHTAFOVYCH, A. et al. Evaluation of the template-based modeling in casp12. **Proteins: Structure, Function, and Bioinformatics**, Wiley Online Library, v. 86, p. 321–334, 2018.

KUHLMAN, B.; BAKER, D. Native protein sequences are close to optimal for their structures. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 97, n. 19, p. 10383–10388, 2000.

LASKOWSKI, R. A.; WATSON, J. D.; THORNTON, J. M. Profunc: a server for predicting protein function from 3d structure. **Nucleic Acids Research**, Oxford University Press, v. 33, n. suppl_2, p. W89–W93, 2005.

LAZARIDIS, T.; KARPLUS, M. Effective energy function for proteins in solution. **Proteins: Structure, Function, and Bioinformatics**, Wiley Online Library, v. 35, n. 2, p. 133–152, 1999.

LAZARIDIS, T.; KARPLUS, M. Effective energy functions for protein structure prediction. **Current Opinion in Structural Biology**, Elsevier, v. 10, n. 2, p. 139–145, 2000.

LEAVER-FAY, A. et al. Scientific benchmarks for guiding macromolecular energy function improvement. In: **Methods in Enzymology**. [S.l.]: Elsevier, 2013. v. 523, p. 109–143.

LEE, B.; RICHARDS, F. M. The interpretation of protein structures: estimation of static accessibility. **Journal of Molecular Biology**, Elsevier, v. 55, n. 3, p. 379–400, 1971.

LESK, A. **Introduction to protein science: architecture, function, and genomics**. [S.l.]: Oxford University Press, 2010.

LESK, A. M. Casp2: report on ab initio predictions. **Proteins: Structure, Function, and Bioinformatics**, Wiley Online Library, v. 29, n. S1, p. 151–166, 1997.

LESK, A. M.; CHOTHIA, C. How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. **Journal of Molecular Biology**, Elsevier, v. 136, n. 3, p. 225–270, 1980.

LEUNG, Y.-W.; WANG, Y. An orthogonal genetic algorithm with quantization for global numerical optimization. **IEEE Transactions on Evolutionary computation**, IEEE, v. 5, n. 1, p. 41–53, 2001.

LIGABUE-BRAUN, R. et al. 3-to-1: unraveling structural transitions in ureases. **Naturwissenschaften**, Springer, v. 100, n. 5, p. 459–467, 2013.

LIGABUE-BRAUN, R. et al. Everyone is a protagonist: residue conformational preferences in high-resolution protein structures. **Journal of Computational Biology**, Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA, v. 25, n. 4, p. 451–465, 2018.

LOBANOV, M. Y.; BOGATYREVA, N.; GALZITSKAYA, O. Radius of gyration as an indicator of protein structure compactness. **Molecular Biology**, Springer, v. 42, n. 4, p. 623–628, 2008.

MARTÍ-RENOM, M. A. et al. Comparative protein structure modeling of genes and genomes. **Annual Review of Biophysics and Biomolecular Structure**, Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA, v. 29, n. 1, p. 291–325, 2000.

MCGUFFIN, L. J.; BRYSON, K.; JONES, D. T. The psipred protein structure prediction server. **Bioinformatics**, Oxford University Press, v. 16, n. 4, p. 404–405, 2000.

MÉZARD, M.; MORA, T. Constraint satisfaction problems and neural networks: A statistical physics perspective. **Journal of Physiology-Paris**, Elsevier, v. 103, n. 1-2, p. 107–113, 2009.

MIRNY, L.; SHAKHNOVICH, E. Protein folding theory: from lattice to all-atom models. **Annual Review of Biophysics and Biomolecular Structure**, Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA, v. 30, n. 1, p. 361–396, 2001.

MOULT, J. et al. Critical assessment of methods of protein structure prediction: Progress and new directions in round xi. **Proteins: Structure, Function, and Bioinformatics**, Wiley Online Library, v. 84, p. 4–14, 2016.

MOULT, J. et al. Critical assessment of methods of protein structure prediction (casp)—round xii. **Proteins: Structure, Function, and Bioinformatics**, Wiley Online Library, v. 86, p. 7–15, 2018.

MOULT, J. et al. A large-scale experiment to assess protein structure prediction methods. **Proteins: Structure, Function, and Bioinformatics**, Wiley Online Library, v. 23, n. 3, p. ii–iv, 1995.

MUIR, P. et al. The real cost of sequencing: scaling computation to keep pace with data generation. **Genome Biology**, BioMed Central, v. 17, n. 1, p. 53, 2016.

MUKHERJEE, S. et al. Genomes online database (gold) v. 7: updates and new features. **Nucleic Acids Research**, Oxford University Press, v. 47, n. D1, p. D649–D659, 2018.

MURZIN, A. G. et al. Scop: a structural classification of proteins database for the investigation of sequences and structures. **Journal of Molecular Biology**, Elsevier, v. 247, n. 4, p. 536–540, 1995.

NARLOCH, P. H.; DORN, M. A knowledge based differential evolution algorithm for protein structure prediction. In: SPRINGER. **International Conference on the Applications of Evolutionary Computation (Part of EvoStar)**. [S.l.], 2019. p. 343–359.

NARLOCH, P. H.; DORN, M. A knowledge based self-adaptive differential evolution algorithm for protein structure prediction. In: SPRINGER. **International Conference on Computational Science**. [S.l.], 2019. p. 87–100.

NEEDLEMAN, S. B.; WUNSCH, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. **Journal of Molecular Biology**, Elsevier, v. 48, n. 3, p. 443–453, 1970.

NELSON, D. L.; LEHNINGER, A. L.; COX, M. M. **Lehninger principles of biochemistry**. [S.l.]: Macmillan, 2008.

NERIA, E.; FISCHER, S.; KARPLUS, M. Simulation of activation free energies in molecular systems. **The Journal of Chemical Physics**, AIP, v. 105, n. 5, p. 1902–1921, 1996.

NEUMAIER, A. Molecular modeling of proteins and mathematical prediction of protein structure. **SIAM review**, SIAM, v. 39, n. 3, p. 407–460, 1997.

OSGUTHORPE, D. J. Ab initio protein folding. **Current Opinion in Structural Biology**, Elsevier, v. 10, n. 2, p. 146–152, 2000.

OSMAN, I. H.; LAPORTE, G. **Metaheuristics: A bibliography**. [S.l.]: Springer, 1996.

OVCHINNIKOV, S. et al. Protein structure prediction using rosetta in casp12. **Proteins: Structure, Function, and Bioinformatics**, Wiley Online Library, v. 86, p. 113–121, 2018.

OVCHINNIKOV, S. et al. Protein structure determination using metagenome sequence data. **Science**, American Association for the Advancement of Science, v. 355, n. 6322, p. 294–298, 2017.

O'MEARA, M. J. et al. Combined covalent-electrostatic model of hydrogen bonding improves structure prediction with rosetta. **Journal of Chemical Theory and Computation**, ACS Publications, v. 11, n. 2, p. 609–622, 2015.

PARK, H. et al. Simultaneous optimization of biomolecular energy functions on features from small molecules and macromolecules. **Journal of Chemical Theory and Computation**, ACS Publications, v. 12, n. 12, p. 6201–6212, 2016.

PAULING, L.; COREY, R. B. The pleated sheet, a new layer configuration of polypeptide chains. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 37, n. 5, p. 251–256, 1951.

PAULING, L.; COREY, R. B.; BRANSON, H. R. The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 37, n. 4, p. 205–211, 1951.

PRUITT, K. D.; TATUSOVA, T.; MAGLOTT, D. R. Ncbi reference sequences (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. **Nucleic Acids Research**, Oxford University Press, v. 35, n. suppl_1, p. D61–D65, 2006.

RAMACHANDRAN, G. N. Stereochemistry of polypeptide chain configurations. **Journal of Molecular Biology**, v. 7, p. 95–99, 1963.

RAMACHANDRAN, G. T.; SASISEKHARAN, V. Conformation of polypeptides and proteins. In: **Advances in Protein Chemistry**. [S.l.]: Elsevier, 1968. v. 23, p. 283–437.

RAMAN, S. et al. Structure prediction for casp8 with all-atom refinement using rosetta. **Proteins: Structure, Function, and Bioinformatics**, Wiley Online Library, v. 77, n. S9, p. 89–99, 2009.

RICHARDSON, J. S. The anatomy and taxonomy of protein structure. In: **Advances in protein chemistry**. [S.l.]: Elsevier, 1981. v. 34, p. 167–339.

ROBIČ, T.; FILIPIČ, B. Differential evolution for multiobjective optimization. In: SPRINGER. **International Conference on Evolutionary Multi-Criterion Optimization**. [S.l.], 2005. p. 520–533.

ROHL, C. A. et al. Protein structure prediction using rosetta. In: **Numerical Computer Methods, Part D**. Academic Press, 2004, (Methods in Enzymology, v. 383). p. 66 – 93. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0076687904830040>.

ROSE, G. D. et al. Hydrophobicity of amino acid residues in globular proteins. **Science**, American Association for the Advancement of Science, v. 229, n. 4716, p. 834–838, 1985.

ROST, B. Twilight zone of protein sequence alignments. **Protein Engineering**, Oxford University Press, v. 12, n. 2, p. 85–94, 1999.

SCHAARSCHMIDT, J. et al. Assessment of contact predictions in casp12: Co-evolution and deep learning coming of age. **Proteins: Structure, Function, and Bioinformatics**, Wiley Online Library, v. 86, p. 51–66, 2018.

SHAPOVALOV, M. V.; JR, R. L. D. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. **Structure**, Elsevier, v. 19, n. 6, p. 844–858, 2011.

SHI, Y. et al. Particle swarm optimization: developments, applications and resources. In: IEEE. **Proceedings of the 2001 Congress on Evolutionary Computation (IEEE Cat. No. 01TH8546)**. [S.l.], 2001. v. 1, p. 81–86.

SIKOSEK, T.; CHAN, H. S. Biophysics of protein evolution and evolutionary protein biophysics. **Journal of The Royal Society Interface**, v. 11, n. 100, p. 20140419, 2014. Available from Internet: <https://royalsocietypublishing.org/doi/abs/10.1098/rsif.2014.0419>.

SIMONS, K. T. et al. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. **Journal of Molecular Biology**, Elsevier, v. 268, n. 1, p. 209–225, 1997.

SIMONS, K. T. et al. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. **Proteins: Structure, Function, and Bioinformatics**, Wiley Online Library, v. 34, n. 1, p. 82–95, 1999.

SKOLNICK, J.; KOLINSKI, A.; ORTIZ, A. R. Monsster: a method for folding globular proteins with a small number of distance restraints. **Journal of Molecular Biology**, Elsevier, v. 265, n. 2, p. 217–241, 1997.

SONG, Y. et al. High-resolution comparative modeling with rosettacm. **Structure**, Elsevier, v. 21, n. 10, p. 1735–1742, 2013.

STORN, R.; PRICE, K. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. **Journal of Global Optimization**, Springer, v. 11, n. 4, p. 341–359, 1997.

TALBI, E. Common concepts for metaheuristics. **Metaheuristics: from design to implementation, John Wiley & Sons, Inc., Hoboken, New Jersey**, p. 23–25, 2009.

TALBI, E.-G. **Metaheuristics: from design to implementation**. [S.l.]: John Wiley & Sons, 2009.

TAYLOR, T. J. et al. Assessment of casp10 contact-assisted predictions. **Proteins: Structure, Function, and Bioinformatics**, Wiley Online Library, v. 82, p. 84–97, 2014.

UNGER, R.; MOULT, J. Finding the lowest free energy conformation of a protein is an np-hard problem: proof and implications. **Bulletin of Mathematical Biology**, Springer, v. 55, n. 6, p. 1183–1198, 1993.

VERLI, H. Bioinformática: da biologia à flexibilidade molecular. Sociedade Brasileira de Bioquímica e Biologia Molecular, 2014.

VOET, D.; VOET, J. G. **Biochemistry, 4th Edition**. [S.l.]: Wiley, 2010.

WANG, S. et al. Raptorx-property: a web server for protein structure property prediction. **Nucleic Acids Research**, Oxford University Press, v. 44, n. W1, p. W430–W435, 2016.

WANG, S. et al. Accurate de novo prediction of protein contact map by ultra-deep learning model. **PLoS Computational Biology**, Public Library of Science, v. 13, n. 1, p. e1005324, 2017.

WANG, S.; SUN, S.; XU, J. Analysis of deep learning methods for blind protein contact prediction in casp12. **Proteins: Structure, Function, and Bioinformatics**, Wiley Online Library, v. 86, p. 67–77, 2018.

WANG, Z.-X. A re-estimation for the total numbers of protein folds and superfamilies. **Protein Engineering**, v. 11, n. 8, p. 621–626, 1998.

WEIGT, M. et al. Identification of direct residue contacts in protein–protein interaction by message passing. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 106, n. 1, p. 67–72, 2009.

WORTH, C. L.; GONG, S.; BLUNDELL, T. L. Structural and functional constraints in the evolution of protein families. **Nature Reviews Molecular Cell Biology**, Nature Publishing Group, v. 10, n. 10, p. 709, 2009.

XU, D.; ZHANG, Y. Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. **Biophysical Journal**, Elsevier, v. 101, n. 10, p. 2525–2534, 2011.

XU, D.; ZHANG, Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. **Proteins: Structure, Function, and Bioinformatics**, Wiley Online Library, v. 80, n. 7, p. 1715–1735, 2012.

XU, D.; ZHANG, Y. Toward optimal fragment generations for ab initio protein structure assembly. **Proteins: Structure, Function, and Bioinformatics**, Wiley Online Library, v. 81, n. 2, p. 229–239, 2013.

XUE, Z. et al. Threadom: extracting protein domain boundary information from multiple threading alignments. **Bioinformatics**, Oxford University Press, v. 29, n. 13, p. i247–i256, 2013.

YANG, H. et al. 4.4 å resolution cryo-em structure of human mtor complex 1. **Protein & Cell**, Springer, v. 7, n. 12, p. 878–887, 2016.

YANOVER, C.; BRADLEY, P. Extensive protein and dna backbone sampling improves structure-based specificity prediction for c2h2 zinc fingers. **Nucleic Acids Research**, Oxford University Press, v. 39, n. 11, p. 4564–4576, 2011.

ZERIHUN, M. B.; SCHUG, A. Biomolecular coevolution and its applications: Going from structure prediction toward signaling, epistasis, and function. **Biochemical Society Transactions**, Portland Press, v. 45, n. 6, p. 1253–1261, 2017.

ZHANG, C. et al. Template-based and free modeling of i-tasser and quark pipelines using predicted contact maps in casp12. **Proteins: Structure, Function, and Bioinformatics**, Wiley Online Library, v. 86, p. 136–151, 2018.

ZHANG, J.; LIANG, Y.; ZHANG, Y. Atomic-level protein structure refinement using fragment-guided molecular dynamics conformation sampling. **Structure**, Elsevier, v. 19, n. 12, p. 1784–1795, 2011.

ZHANG, Y.; SKOLNICK, J. Spicker: a clustering approach to identify near-native protein folds. **Journal of Computational Chemistry**, Wiley Online Library, v. 25, n. 6, p. 865–871, 2004.

ZHOU, A. et al. Multiobjective evolutionary algorithms: A survey of the state of the art. **Swarm and Evolutionary Computation**, Elsevier, v. 1, n. 1, p. 32–49, 2011.