UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE BIOCIÊNCIAS
CURSO DE BIOTECNOLOGIA

MARIEL BARBACHAN E SILVA

SAN-PSO: Predição da Estrutura de Proteínas utilizando um algoritmo de Otimização por Enxame de Partículas baseado em conhecimento

Monografia apresentada como requisito parcial para a obtenção do grau de Bacharel em Biotecnologia

Orientador: Prof. Dr. Márcio Dorn

Porto Alegre
2016

AGRADECIMENTOS

À minha família pelo apoio, carinho e incentivo nesta minha longa jornada pela graduação. Especialmente à Berenice Barbachan e Silva, minha mãe, por ser meu exemplo e ídolo, ao meu irmão Matheus Barbachan e Silva por ser o melhor amigo que alguém pode ter e ao Marcus Vinícius Azevedo da Silva, meu pai, por todo o carinho.

À Barbara Martinelli por ter estado comigo e me apoiado em todos os momentos desta jornada.

Ao meu orientador Dr. Márcio Dorn por todos os ensinamentos, pelo exemplo de pesquisador e pessoa que tu és e por sempre acreditar em mim e incentivar meu crescimento.

Aos colegas e amigos do SBCB pelas risadas e pela ajuda.

A todos os meus amigos, em especial ao Ricardo Albanus, à Janinne Herrlein e à Nathalie Pires que, perto ou longe, estão sempre presentes.

1 INTRODUÇÃO GERAL

As proteínas são macromoléculas que desempenham uma vasta gama de funções em organismos vivos. São polímeros formados por um conjunto de 20 diferentes monômeros, os aminoácidos. Proteínas consistem em uma ou mais cadeias de aminoácidos ligados entre si por meio de ligações peptídicas (LESK, 2002; TRAMONTANO, 2006). Os aminoácidos são compostos orgânicos que apresentam um carbono α (C_α) ligado a um grupamento amino (NH_2), um grupamento carboxil (COOH) e a cadeia lateral variável (R), a qual é responsável pelas proprieades fisico-químcias específicas de cada amino ácido(LEHNINGER; NELSON; COX, 2005) (Figura 1.1). A sequência linear de aminoácidos é chamada de estrutura primária da proteína. Segmentos da proteína tendem a se organizar de maneira regular no espaço, estes padrões de organização são chamados estrutura secundária. Cada sequência de aminoácidos resulta em um enovelamento protéico que leva a estrutura tridimensional (3D) da proteína, a qual é relacionada à sua função (LEHNINGER; NELSON; COX, 2005).

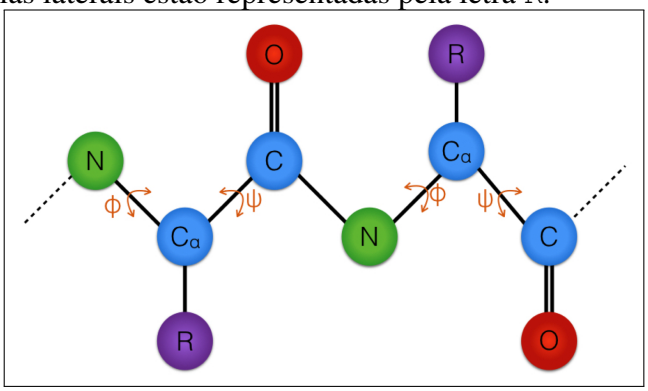
A predição da estrutura 3D de uma proteína (PEP) é um problema desafiador em diversas áreas do conhecimento, como por exemplo a ciência da computação, a matemática, a física, a biologia e a química (LANDER; WATERMAN, 1999; WOOLEY; YE, 2010; DORN et al., 2014). O desafio surge devido à explosão combinatorial de possíveis conformações que uma proteína pode assumir no espaço (LEVINTHAL, 1968), dentre estas se assume que a estrutura 3D nativa é aquela que apresenta a menor energia potencial no sistema em que está inserida.

A PEP é experimentalmente um processo árduo e caro. Por isso, diversos métodos computacionais tem sido desenvolvidas para abordar este problema. Atualmente, este é um dos principais desafios da Bioinformática Estrutural. Nas últimas décadas, diversas estratégias computacionais foram propostas para a PEP e estas podem ser estratificadas em quatro grandes grupos (DORN et al., 2014): (i) Métodos de primeiros princípios que não utilizam informação da base experimental; (ii) Métodos de primeiros princípios com informações da base de dados; (iii) Métodos de reconhecimento de enovelamento; (iv) Modelagem comparativa. O desempenho destas estratégias é avaliada bianualmente em um encontro intitulado CASP (Critical Assessment of Protein Structure Prediction) e as estratégias dos grupos ii, iii e iv têm apresentado melhores resultados(KRYSHTAFOVYCH; FIDELIS K. ANDMOULT, 2011B). Isto acontece porque o espaço de busca dos métodos do grupo i e, portanto, o custo computacional destes os torna inviáveis.

A performance dos algoritmos baseados em conhecimentos da base experimental é intimamente ligada com a qualidade da informação retirada da base e como ela é representada (DORN et al., 2014). Existem diversas maneiras de representar a estrutura 3D de uma proteina, como por exemplo: modelo *all-atom* (OSGUTHORPE, 2000), rotâmeros (SHAPOVALOV; DUNBRACK, 2011), e ângulos de torsão (CUTELLO; NARZISI; NICOSIA, 2006; DORN; BURIOL; LAMB, 2011; DORN et al., 2013; BORGUESAN et al., 2015). Neste trabalho, as proteínas são representadas por conjuntos de pares de ângulos de torção da cadeia principal: `phi` (ϕ) (`N`$-$`C`$_\alpha$) e `psi` (ψ) (`C`$_\alpha$$-$`C`) (LESK, 2002; LODISH et al., 1990), pois a rotação da cadeia polipeptídica principal, aquela que não considera as cadeias laterais, é responsável pela conformação tridimensional das proteínas (LESK, 2002; SCHEEF; FINK, 2003).

O desafio dos métodos que não utilizam informações da base experimental surge devido à explosão combinatorial de possíveis conformações que as proteínas podem assumir no espaço tridimensionl (LEVINTHAL, 1968), dentre as quais se assume que a estrutura 3D nativa é aquela que é associada com a menor energia potencial do sistema (ANFINSEN, 1973). Em 1995, Brygelson et al. introduziram a teoria do funil de enovelamento ou do panorama energético a qual se refere a possibilidade de existir mais de uma trajetória energética no processo de enovelamento proteico capaz de levar à estrutura nativa da proteína (BRYNGELSON et al., 1995). Dentre a explosão combinatorial de possíveis conformações descrita por (LEVINTHAL, 1968) sabe-se que existem combinações que são proibidas estereoquimicamente (RAMACHANDRAN; SASISEKHARAN, 1968) e que há uma preferência de pares de ângulos phi e psi que é própria de cada estrutura secundária (HOVMOLLER; OHLSON, 2002). Além disso, existe influência da vizinhança de resíduos na estrutura primária do polipeptídeo na preferência de pares de ângulos da cadeia principal de um dado resíduo de aminoácidos de acordo com a sua estrutura secundária (XIA; XIE, 2002). Baseado nestas informações, foi desenvolvida a Angle Probability List (BORGUESAN et al., 2015) que utiliza dados provenientes de estruturas obtidas experimentalmente e depositadas no Protein Data Bank para informar intervalos de ângulos de torsão mais prováveis para cada resíduo de aminoácido em determinada estrutura secundária. O NIAS-server é capaz de extrair informações do PDB e criar quatro tipos APLs: i) APL1 que é gerada sem influência dos vizinhos; ii) APL2 que é gerada com a influência de um vizinho(esquerda ou direita); iii) APL3 que é gerada com a influência completa dos vizinhos(esquerda e direita); iv) APLcentroids que considera a vizinhança completa de tamanho 5,7 ou 9 para a gerar a APL do resíduo de aminoácido

Figura 1.1: Diagrama representando uma estrutura planar de proteína. Em laranja estão evidenciados os dois ângulos de torsão (ϕ and ψ). Por simplificação,os hidrogênios foram omitidos e as cadeias laterais estão representadas pela letra R.



centróide. (BORGUESAN; INOSTROZA-PONTA; DORN, 2016)

A PEP é considerada em complexidade de algortimos como um problema NP-Completo (NGO; MARKS; KARPLUS, 2012) e muitas estratégias não-determinísticas têm sido desenvolvidas para abordar este problema de maneira não-exata. Metaheurísticas são procedimentos projetados para encontrar, gerar ou selecionar soluções aproximadas para problemas de otimização, especialmente aqueles com informações incompletas ou limitadas capacidades computacionais, para os quais muitas vezes a solução exata é impossível de ser determinada. (BIANCHI et al., 2009; GLOVER; KOCHENBERG, 2003) Buscando resolver o problema da PEP, metaheurísticas são frequentemente utilizadas por sua habilidade de encontrar soluções satisfatórias com menor custo computacional do que métodos exatos (TANTAR et al., 2007; TANTAR; MELAB; TALBI, 2008; GARZA-FABRE et al., 2016; BORGUESAN et al., 2015; CUSTÓDIO; BARBOSA; DARDENNE, 2014).

O Enxame de Partículas (EP) (EBENHART, 1995) é uma metaheurística de otimização estocástica populacional inspirada no comportamento de social de alguns animais, tais como o cardumes de peixes e bandos de pássaros. O algoritmo funciona com uma população (enxame) de soluções candidatas (partículas) que são formadas por elementos que representam os parâmetros a serem otimizados. Estas partículas são movimentadas no espaço de busca de acordo com uma função de aptidão e tem a velocidade de movimento com componentes randômicos. Os movimentos das partículas são guiadas por sua própria melhor posição no universo de busca, bem como a melhor posição de todo o enxame. Quando posições melhoradas vão sendo descobertas estas passarão, em seguida, a orientar os movimentos do enxame. O processo é iterativo e por isso espera-se que a solução satisfatória acabará por ser descoberta. EP tem sido utilizado como estretégia de

otimização para a PEP (MEISSNER; SCHNEIDER, 2007; BORGUESAN et al., 2015). Entretanto, alguns pontos fracos do algoritmo têm sido discutidos e muitas variações do OEP estão sendo desenvolvidas com o objetivo principal de melhorar a diversidade do enxame, o tempo de convergência e a taxa exploração do espaço de busca realizada pelas partículas (ZAMBRANO-BIGIARINI; CLERC; ROJAS, 2013; RINI; SHAMSUDDIN; YUHANIZ, 2011).

Melhorias na performance das metaheurísticas utilizadas na PEP tem sido feitas utilizando combinações de metaheurísticas e melhorando a interpretação dos dados retirados da base experimental (BORGUESAN; INOSTROZA-PONTA; DORN, 2016). Combinar metaheurísticas para melhorar a performance dos métodos computacionais melhorou os resultados de predição, especialmente utilizando metaheurísticas de busca local para adicionar caráter auto-adaptativo em otimizações com metaheurísticas de busca global (SUN, 1993; TANTAR; MELAB; TALBI, 2008). A otimização multimodal foca em encontrar mais de uma solução ótima para um problema de otimização. Baseada na hipótese do funil de enovelmento (BRYNGELSON et al., 1995) na qual mais de uma estrutura 3D pode estar associada com a mesma energia, a PEP pode ser abordada de maneira multimodal, tendo em vista que a otimização para PEP tem por objetivo computacional minimizar a função de energia do sistema. Estudos anteriores comprovam que abordar a PEP de forma multimodal aumentou a qualidade das estruturas finais (DAY et al., 2002; CALVO; ORTEGA; ANGUITA, 2011). Neste trabalho são apresentados dois algoritmos. O PSO, que é a implementação canônica da otimização por enxame de partículas para PEP utilizando informação proveniente de APLs, e o SAN-PSO, uma nova estratégia computacional que envolve a combinação de metaheurísticas e utilização de informação da base experimental para a PEP. Ela é composta por um EP multimodal que tem caráter auto-adaptativo utilizando informação proveniente de APLs.

2 DISCUSSÃO

Para testar a performance dos algoritmos propostos neste trabalho, utilizamos um conjunto de sete proteínas com tamanhos e composições de estrutura secundária variadas (PDB IDs: 1AB1, 1ACW, 1K43, 1WQC, 2MR9 2MTW e 2P81) e comparamos os resultados obtidos com o PSO e com o SAN-PSO. A qualidade dos resultados foi avaliada com análises estruturais e de energia. Para as análises estruturais foram utilizadas duas métricas diferentes, o RMSD que, foi medido pela distância entre os átomos de C_α entre as estruturas preditas e a experimentais otimamente sobrepostas, retirando os resíduos N- e C- terminais (Equação 2.1) e o GDT_ST, que é uma medida de porcentagem que representa a proporção da estrutura predita que equivale a nativa (equação 2.2). Para análise de energia, foi utilizada a adaptação de parâmetros *score3* (ROHL et al., 2004) do PyRosetta, uma interface em Python do Rosetta (ROHL et al., 2004) com a adição de um parâmetro de reforço positivo para a criação e manutenção de estruturas secundárias (equação 2.3). Para análise estatística dos resultados foi realizada ANOVA e considerado significativos aqueles resultados que obtiveram ($p < 0.01$).

$$\mathbf{RMSD}(a,b) = \sqrt{\left(\sum_{i=1}^{n} \|r_{ai} - r_{bi}\|^2\right)/n}, \tag{2.1}$$

onde r_{ai} e r_{bi} são vetores que representam as posições do átomo i nas estruturas a e b respectivamente.

$$\mathbf{GDT\_ST} = (GDT_{\text{P1}} + GDT_{\text{P2}} + GDT_{\text{P4}} + GDT_{\text{P8}})/4, \tag{2.2}$$

onde onde $P1$, $P2$, $P4$ e $P8$ são as porcentagens do número de resíduos alinhados com distância menor que 1Å, 2Å, 4Åe 8Å, respectivamente.

$$Fitness = E_{\text{PyRosetta}} + E_{\text{EstruturaSecundária)}} \tag{2.3}$$

A estratégia do SAN-PSO é multimodal e, portanto, resulta em uma população de soluções finalistas. Para a validação deste método foi realizada uma análise de toda a população finalista para encontrar o melhor indivíduo de cada rodada baseado no valor de RMSD em relação a estrutura experimental. E, assim, esta estrutura era adicionada ao conjunto de soluções a ser comparado com a implementação canônica do PSO.

De um ponto de vista estrutural, as duas implementações apresentaram resultados similares. No caso da análise de RMSD, a implementação canônica obteve melhor re-

sultado ao prever a proteina 2MTW (p = 0,0002) e o SAN-PSO foi melhor ao prever a 1ACW (p = 0,001), para a análise de GDT_ST os resultados foram semelhantes. As maiores variações entre as estruturas preditas e as estruturas experimentais foram observadas em regiões de estruturas irregulares principalmente nas extremidades N- e C-terminais das proteínas. Para avaliar se estas regiões irregulares influenciaram na performance das implementações, o RMSD das duas proteínas que obtiveram resultados significativamente diferentes foi recalculado eliminando estas regiões, assim como a ANOVA. Os resultados desta nova análise demonstram que a significância da diferença entre os resultados de RMSD as estruturas diminuiu em uma ordem de grandeza para a 2MTW (p = 0,02), o que significa que as regiões irregulares nas extremidades N- e C-terminais influenciaram na diferença entre essas estruturas, pois ao retirá-las as duas estruturas se tornaram mais semelhantes. No caso da 1ACW, a significância da diferença entre os resultados de RMSD das estruturas aumentou em uma ordem de grandeza (p = 0,0003) com a remoção das regiões irregulares, indicando que, ao remover o ruído causado por estas estruturas flexíveis, foi possível observar que o SAN-PSO foi melhor ainda ao gerar estruturas regulares.

De acordo com os resultados de energia, o SAN-PSO foi significativamente melhor que o PSO em 42,86% dos casos e nunca foi pior do que a implementação canônica. Estes resultados indicam que de fato a implementação multimodal e auto-adaptativa do SAN-PSO é capaz de explorar melhor o espaço de busca, podendo encontrar melhores mínimos de energia. Isto indica que a capacidade do SAN-PSO em minimizar a função de energia está aprimorada em relação ao PSO.

Os resultados indicam que a nova estratégia proposta neste trabalho é capaz de realizar o enovelamento geral das proteínas de teste, especialmente com respeito a regiões de estrutura secundária regulares tais como hélices e folhas. Na minimização da função de energia, a nova abordagem provou ser melhor do que a implementação canônica em mais de 40% dos casos e não foi, em nenhum caso, o pior. Esse resultado pode estar relacionado à introdução do caráter multimodal para a implementação, o que aumenta a exploração do espaço de busca.

REFERÊNCIAS

ANFINSEN, C. Principles that govern the folding of protein chains. **Science**, v. 181, n. 96, p. 223–230, 1973.

BIANCHI, L. et al. A survey on metaheuristics for stochastic combinatorial optimization. Kluwer Academic Publishers, v. 8, n. 2, p. 239–287, jun. 2009.

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., v. 23, 2016.

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**, v. 59, Part A, p. 142–157, 2015. ISSN 1476-9271. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S1476927115301250>.

BRYNGELSON, J. D. et al. Funnels, pathways, and the energy landscape of protein folding: a synthesis. **Proteins: Structure, Function, and Bioinformatics**, Wiley Online Library, v. 21, n. 3, p. 167–195, 1995.

CALVO, J. C.; ORTEGA, J.; ANGUITA, M. Comparison of parallel multi-objective approaches to protein structure prediction. **The Journal of Supercomputing**, Springer, v. 58, n. 2, p. 253–260, 2011.

CUSTÓDIO, F. L.; BARBOSA, H. J.; DARDENNE, L. E. A multiple minima genetic algorithm for protein structure prediction. **Applied Soft Computing**, Elsevier, v. 15, p. 88–99, 2014.

CUTELLO, V.; NARZISI, G.; NICOSIA, G. A multi-objective evolutionary approach to the protein structure prediction problem. **J. R. Soc., Interface**, v. 3, n. 6, p. 139–151, 2006.

DAY, R. O. et al. Solving the protein structure prediction problem through a multiobjective genetic algorithm. **Nanotechnology**, v. 2, p. 32–35, 2002.

DORN, M.; BURIOL, L.; LAMB, L. A hybrid genetic algorithm for the 3-d protein structure prediction problem using a path-relinking strategy. In: **IEEE Congress on Evolutionary Computation**. [S.l.: s.n.], 2011. p. 2709–2716.

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

DORN, M. et al. Three-dimensional protein structure prediction: Methods and computational strategies. **Comput. Biol. Chem.**, v. 53, Part B, p. 251 – 276, 2014.

EBENHART, R. Kennedy. particle swarm optimization. In: **Proceeding IEEE Inter Conference on Neural Networks, Perth, Australia, Piscat-away**. [S.l.: s.n.], 1995. v. 4, p. 1942–1948.

GARZA-FABRE, M. et al. Generating, maintaining and exploiting diversity in a memetic algorithm for protein structure prediction. **Evolutionary computation**, MIT Press, 2016.

GLOVER, F.; KOCHENBERG, G. Handbook of meta-heuristics. In: **International Series in Operations Research and Management Science**. [S.l.: s.n.], 2003. v. 57, p. 570.

HOVMOLLER, T.; OHLSON, T. Conformation of amino acids in protein. **Acta Crystallogr.**, v. 58, n. 5, p. 768–776, 2002.

KRYSHTAFOVYCH, A.; FIDELIS K. ANDMOULT, J. Casp9 results compared to those of previous casp experiments. **Proteins: Struct., Funct., Bioinf.**, v. 79, n. S10, p. 196–207, 2011B.

LANDER, E.; WATERMAN, M. **The secrets of life: a mathematician's introduction to Molecular Biology**. Washington D. C., USA: National Academy Press, 1999. 300 p.

LEHNINGER, A.; NELSON, D.; COX, M. **Principles of Biochemistry**. 4. ed. New York, USA: W.H. Freeman, 2005. 1100 p.

LESK, A. M. **Introduction to Bioinformatics**. 1. ed. New York, USA: Oxford University Press Inc., 2002. 308 p.

LEVINTHAL, C. Are there pathways for protein folding? **J. Chim. Phys. Phys.-Chim. Biol.**, v. 65, n. 1, p. 44–45, 1968.

LODISH, H. et al. **Molecular Cell Biology**. 5. ed. New York, USA: Scientific American Books, W.H. Freeman, 1990. 970 p.

MEISSNER, M.; SCHNEIDER, G. Protein folding simulation by particle swarm optimization. **Open Struct. Biol. J**, v. 1, p. 1–6, 2007.

NGO, J. T.; MARKS, J.; KARPLUS, M. Protein structure prediction. **The Protein Folding Problem and Tertiary Structure Prediction**, Springer Science & Business Media, p. 433, 2012.

OSGUTHORPE, D. Ab initio protein folding. **Curr. Opin. Struct. Biol.**, v. 10, n. 2, p. 146–152, 2000.

RAMACHANDRAN, G.; SASISEKHARAN, V. Conformation of polypeptides and proteins. **Adv. Protein Chem.**, v. 23, p. 238–438, 1968.

RINI, D. P.; SHAMSUDDIN, S. M.; YUHANIZ, S. S. Particle swarm optimization: technique, system and challenges. **International Journal of Computer Applications**, International Journal of Computer Applications, 244 5 th Avenue,# 1526, New York, NY 10001, USA India, v. 14, n. 1, p. 19–26, 2011.

ROHL, C. A. et al. Protein structure prediction using rosetta. **Methods Enzymol.**, Elsevier, v. 383, p. 66–93, 2004.

SCHEEF, E.; FINK, J. Fundamentals of protein structure: Structural bioinformatics. In: ____. [S.l.: s.n.], 2003. chp. 2, p. 15.

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

SUN, S. Reduced representation model of protein structure prediction: Statistical potential and genetic algorithms. **Protein Science**, Cold Spring Harbor Laboratory Press, v. 2, n. 5, p. 762–785, 1993. ISSN 1469-896X. Available from Internet: <http://dx.doi.org/10.1002/pro.5560020508>.

TANTAR, A.-A.; MELAB, N.; TALBI, E.-G. A grid-based genetic algorithm combined with an adaptive simulated annealing for protein structure prediction. **Soft Computing**, v. 12, n. 12, p. 1185, 2008.

TANTAR, A.-A. et al. A parallel hybrid genetic algorithm for protein structure prediction on the computational grid. **Future Generation Computer Systems**, Elsevier, v. 23, n. 3, p. 398–409, 2007.

TRAMONTANO, A. **Protein structure prediction: concepts and applications**. 1. ed. Weinheim, Germany: John Wiley and Sons, Inc., 2006. 208 p.

WOOLEY, J.; YE, Y. A historical perspective and overview of protein structure prediction. In: ____. **Computational Methods for Protein Structure Prediction and Modeling**. [S.l.]: Springer, 2010. chp. 1, p. 1–43.

XIA, X.; XIE, Z. Protein structure, neighbor effect, and a new index of amino acid dissimilarities. **Mol. Biol. Evol.**, v. 19, n. 1, p. 58–67, 2002.

ZAMBRANO-BIGIARINI, M.; CLERC, M.; ROJAS, R. Standard particle swarm optimisation 2011 at cec-2013: A baseline for future pso improvements. In: IEEE. **2013 IEEE Congress on Evolutionary Computation**. [S.l.], 2013. p. 2337–2344.

Journal of Bioinformatics and Computational Biology

SAN-PSO: Protein Structure Prediction using a Knowledge-based Self-adaptative Multimodal Particle Swarm Optimization Algorithm

MARIEL BARBACHAN E SILVA
Federal University of Rio Grande do Sul
Porto Alegre, Rio Grande do Sul, Brasil
mariel.barbachan@ufrgs.br

MÁRCIO DORN
Federal University of Rio Grande do Sul
Porto Alegre, Rio Grande do Sul, Brasil
mdorn@inf.ufrgs.br

Proteins are a class of molecules that perform various vital functions in living organisms. Knowledge of the three-dimensional structure of proteins is crucial to the study of their function, however, this is an experimentally expensive and arduous process. Due to this constraint, many computational approaches have been developed in order to predict the three-dimensional structure of proteins. Still, based solely on the amino acid sequence, the process is arduous and computationally infeasible. Thus, several approaches attempt to circumvent the enormous computational cost involved in searching for the huge conformational search space that is characteristic of this problem using metaheuristics and knowledge of the experimental basis. The PSO is a population search metaheuristic composed of particles that travel through the search space optimizing a function. Previously used for protein structure prediction minimizing the potential energy of the system, the canonical implementation of the PSO presents flaws, mainly regarding the swarm diversity and imprisonment in local minima. In order to improve these points, we propose a metaheuristic ensemble strategy that uses information from the experimental database, the SAN-PSO. Using a set consisting of 7 test proteins and the canonical implementation of the PSO as a comparison, our results show that our strategy can predict the general folding of the proteins tested and manages to minimize the energy of the system better than the canonical implementation in more than 40% of the cases.

Keywords: protein structure prediction; swarm intelligence; multimodal particle swarm optimization; knowledge-based protein structure prediction

1. Introduction

Proteins are macromolecules that perform several critical roles in living organisms and consist of one or more chains built from a set of 20 amino acids linked together by peptide bonds [1, 2]. Amino acids consist of an α carbon ($\mathtt{C}_\alpha$) bonded to amino ($\mathtt{NH}_2$) and carboxyl (COOH) groups and a variable side chain (R), which is responsible for the specific physicochemical properties of each amino acid. The linear amino acid sequence is called the primary structure of the protein. Protein segments tend to

organize in a regular way in space, and these patterns of organization are called secondary structure. There are preferred conformations like α-helices, β-sheets, β-turns, among others [3]. In different proteins, helices and sheets are combined in many ways to create different spatial arrangements of the polypeptide chain and [4], with appropriate environmental conditions, the amino acid chain folds itself in a unique low-energy three-dimensional structure, the tertiary structure, which is strictly related to the protein function.

At the present moment, over 73 million non-redundant protein sequences have been deposited in RefSeq: NCBI Reference Sequence Database[a] and this number is expected to be fast-growing. Nonetheless, the number of experimentally determined proteins represent less than 1% of this amount, according to Protein Data Bank (PDB)[b]. This vast discrepancy between the number of sequences and structures can be explained by the arduous and expensive process involved in experimentally determine protein structures and it has motivated research toward computational methods for predicting protein structures from sequences, currently one of the cornerstones of structural bioinformatics. Protein structure prediction (PSP) from an amino acid sequence is a challenging problem in several fields of knowledge, such as Computer Science, Mathematics, Physics, Biology and Chemistry [5, 6, 7]. The challenge arises due to the combinatorial explosion of possible conformations that a protein can assume in space [8], among these it is assumed that the native 3D structure is the one associated with the lowest potential energy [9]. In 1995, Brygelson et.al. introduced the folding funnel hypothesis, which refers to the possibility of several pathways in the process of protein folding that will lead to the minimal energy structure [10].

Over the last decades, many computational strategies have been proposed for the PSP and these can be stratified into four large groups [5]: (i) First principles methods without information from experimental database; (ii) First principles methods with database information; (iii) Fold recognition methods; (iv) Comparative modeling. The performance of these methods is evaluated biannually in a meeting called CASP (Critical Assessment of Protein Structure Prediction) [c] and the strategies from groups ii, iii and iv have presented better results[11]. This outcome can be explained by the fact that, group i methods tackle the PSP by exhaustive search of the conformational space, and without any other information regarding the nature of the protein to be predicted, the number of possible conformations to be explored is many orders of magnitude larger than the computational capacity currently available, making them unfeasible [12, 13, 14].

Groups ii, iii, iv are often grouped together as Knowledge-based approaches and their performance is closely associated with the quality of the information taken from the experimental database [5].Regarding group ii methods, there are different

[a] http://www.ncbi.nlm.nih.gov/RefSeq/
[b] http://www.rcsb.org/pdb/
[c] http://predictioncenter.org

approaches to represent a polypeptide structure [5], in order to work with a reduced number of variables, in our method the polypeptide chain is represented by a set of main chain torsion angles: `phi` (ϕ) (`N-`C_α) and `psi` (ψ) (C_α`-C`) [1, 15] (see Figure 1 for clarification), since the backbone rotation of the polypeptide chain is responsible for the three dimensional shape of the protein [1, 16] .

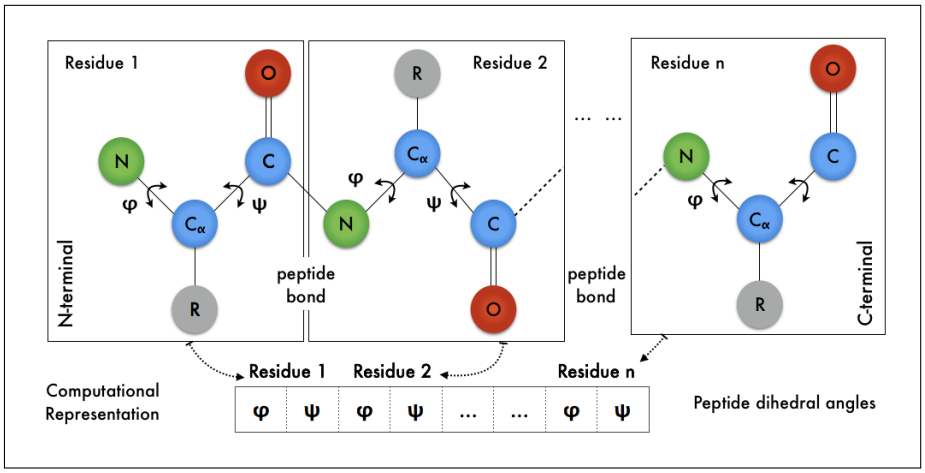


Fig. 1: Schematic diagram of a polypetide evidencing the torsion angles (ϕ and ψ) used for protein representation in our method. For simplicity, hydrogens were not represented and side chains are represented by the letter `R`.

Among the combinatorial explosion of possible conformations described by Levintal et.al. [8] it is known that there are combinations of torsion angles that are stereochemically prohibited [17], also that secondary structures have particular preferences for a set torsion of angles in a given residue [18]. In addition, there is influence of the vicinity of residues on the primary structure of the polypeptide in the preference for pairs of angles of the main chain in a given amino acid according to its secondary structure [19]. Based on this information, Borguesan et.al developed the *Angle Probability List* (`APL`) [20] which uses data from structures obtained experimentally and deposited on `PDB` to infer the torsion angle intervals with higher probability for each amino acid in a given secondary structure.

From a computational point of view,`PSP` is considered in complexity theory as a NP-complete problem [21] and many efforts have been made to develop non exact methods that address this problem in a non-deterministic way to circumvent the complexity issue. Metaheuristics are a kind of stochastic algorithms used to discover quality solutions for tough optimization problems where often the exact solution is impossible to determine [22]. Metaheuristics can be classified by a number of properties, such as the type of strategy, the inspirational source, the number of individuals to be optimized and have been used as a strategy for solving the `PSP` [23,

24, 25, 20, 26].

Introduced by Beni and Wang in 1989, Swarm Intelligence (SI) [27] is a type of metaheuristic inspired by the collective behaviour of social insects and other animals. This field of metaheuristics is classified as a population nature-inspired global search optimization. SI focus on the self-organizing behaviour of interacting agents who follow a set of rules. SI-based algorithms have become very popular over the years due to their high efficiency and practical implementation dealing with large scale optimization problems [28, 29]. The most dominant strategies in SI are Particle Swarm Optimization (PSO) [30], Evolutionary Algortihms, such as Genetic Algorithm (GA) [31] and Ant Colony Optimization(ACO) [32] and all of those had been recurrently used to approach the PSP problem [33, 34, 35, 36, 37, 38, 39] working on minimizing the energy function to retrieve the minimal energy structure.

Improvements on the metaheuristics used in PSP have been made by combining different strategies (ensemble), improving the information extracted from the database [40] and approaching the problem in a multimodal point of view. Combining different metaheuristics to improve the overall performance of the method have been proved to better the results for PSP, specially using combined strategies of global and local search metaheuristics, conferring a self-adaptive character to the global search metaheuristic [41, 24]. Multimodal optimization focuses on finding more than one optimal solution to a given optimization problem. PSP can be considered multimodal based on the folding funnel hypothesis [10] that more than one three-dimensional conformation can be related to the same energy value. Previous studies have shown that using multimodal optimization for PSP have improved the quality of the resulting solutions [42, 43]. As an extension of EA for multimodal optimization niching methods have been used to improve and maintain diversity in population-based search methods such as GA [44], this strategy has been used in PSP [43].

PSO is a very popular SI metaheuristic due to its high efficiency and simple implementation and has been broadly used to address the PSP problem[45, 20], however it has limitations regarding the accuracy, convergence time and exploration/exploitation rates of the search space [46, 47]. This issue is often solved using PSO as part of an ensemble of metaheuristics [48, 49, 50, 51] or proposing improvements to the canonical code [52]. In this article, we propose a novel strategy to solve the protein structure prediction problem using information from APLs, a self adaptive multimodal PSO. The remainder of the paper is organized in Material and Methods, Section 2, describing the proposed implementation of PSO, as well as the canonical implementation; Section 3 presents the computational experiments and the results and Section 4 concludes the article and points out future work.

2. Material and Methods

A PSP method aims to search the protein conformational space to find the most stable conformation of a protein. The stability of a polypeptide chain is evaluated by

an energy function where the best conformation corresponds to the global minimum of its potential energy. The quality of the prediction depends on the energy scoring function for evaluating the results and an optimization algorithm to find the best conformational state. In general, there are three essential components of a PSP method: (a) a way to obtain and represent structural information from protein templates; (b) an energy function to evaluate the stability of a protein, and (c) a search procedure to find the protein structure with the lowest potential energy.

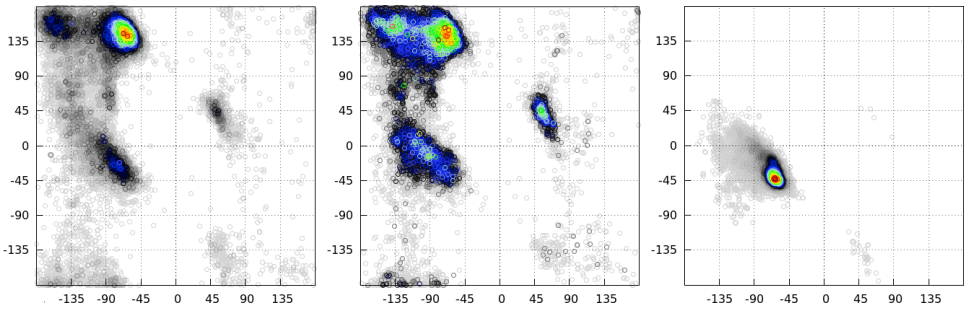
In this work we present two algorithms for PSP: i) canonical PSO: using APL information (Subsection 2.3); ii) Self-adaptive multimodal PSO using information from APL (Subsection 2.4). Both implementations were made in Python programming language in a Linux x86_64 environment. The experiments were performed using three MS Azure Standard DS5 v2 processors with 16 cores Xeon E5-2673 v3 (Haswell), 2.4 GHz, 56 GB Ram and a SSD disk with 112 GB.

2.1. *Angle Probability List:* protein structural information

Frequently PSP strategies use experimental information aiming to minimize the search space of the algorithms and thus improve their outcome. In this context, Borguesan et.al 2015 created an *Angle Probability List* approach to reduce the protein conformational search space based on the conformational preferences of amino acid in a given secondary structure and regarding its neighborhood on the primary protein structure. Briefly, the APL uses the primary and secondary structure of a query protein and PDB torsion angle information to compute conditional probabilities of torsion angles for each amino acid, e.g.$P(SS = Helix|AA = Arg)$. In a further work, Borguesan et. al. created a web-server, NIAS [40], that allows the generation of four different types of APL: APL_1 without amino acid neighbour influence [20], APL_2 with one neighbour influence (left or right), APL_3 that is completely neighbour-dependent (left and right) and $\text{APL}_{\text{centroid}}$ which takes into account a range from five to nine amino acids for the secondary structure, but only the central amino acid APL is computed. Figure 2 shows the conformational preferences of Alanine (ALA) and Arginine (ARG) in coil and Serine (SER) in α-helix secondary structure. In this paper, the conformational preferences (phi and psi) and secondary structure propensities of amino acid, obtained from experimentally determined proteins, are taken into consideration to determine the conformational flexibility of amino acids in a target sequence.

2.2. *Scoring function:* PyRosetta

Solving the PSP problem involves efficient search algorithms, which cover the relevant conformational space, and selective scoring functions, that are both efficient and effectively discriminate between native and non-native solutions. In this work, the evaluation of the candidate solutions of the metaheuristics was made using a fitness function based on the potential energy function from using *score3* parameter adaptation [53] from PyRosetta, an interactive Python-based interface of the



(a) `ALA` - Coil (b) `ARG` - Coil (c) `SER` - α-helix

Fig. 2: Conformational preferences of amino acids in proteins depends on its secondary structure. The dark red color marks the most densely occupied regions. Ramachandran plots were generated by `NIAS-server` (`www.sbcb.inf.ufrgs.br/nias`).

Rosetta [53] molecular modeling suite[d].This energy function uses a centroid mode for the amino acid side chains, which reduces significantly the computational cost for the energy calculation. Another factor was added to the fitness function in order to reinforce the maintenance of secondary structure. Using a simplified *PyRosetta* implementation of `DSSP`[54] to assign the secondary structure of the particles along the `PSO` iterations. When the ascribed secondary structure matches the one given as an input for the algorithm, it means that the perturbations made on the torsion angles did not modify the secondary structure of the protein then it gives a positive reinforcement which is added to the fitness function, on the contrary a negative reinforcement is given. Therefore, the metaheuristics are used to minimize the potential energy function from Rosetta without losing track of the secondary structure - Equation 1. It is noteworthy that choosing the energy function is a great challenge in the development of an optimization strategy for `PSP`. There is no function that perfectly describes the potential energy of a real system and it has a direct consequence in the final structural results.

$$Fitness = E_{\text{PyRosetta}} + E_{\text{SecondaryStructure)}} \quad (1)$$

The most common score function components of PyRosetta are: *van der* Waals net attractive energy and *van der* Waals net repulsive energy ; hydrogen bonds (backbone) [55]; solvation (*Lazaridis-Karplus*) [56]; *Dunbrack* rotamer probabilit [57]; statistical residue-residue pair potential; Intra-residue repulsive *van der* Waals; electrostatic potential[58]; disulfide statistical energies [59] and statistical residue-residue pair potential (centroid). The final energy value of PyRosetta scoring function ($E_{\text{PyRosetta}}$) is the sum of all considered weighted independent energy components.

[d]`https://www.rosettacommons.org`

2.3. *Particle Swarm Optimization for the PSP problem: the canonical algorithm*

Metaheuristics designate a class of approximate computational methods that optimize a problem by an iterative generation process which guides a subordinate heuristic by intelligently combining different concepts for exploring and exploiting the search space. PSO is a SI-based metaheuristic with a socio-cognitive approach used to find a solution for an optimization problem. The algorithm is modelled by a population (*swarm*) of candidate solutions (*particles*) that consist of elements representing the parameters to be optimized. These particles move in the search space according to a fitness function and have their velocity governed by components with random factors. The movement of the particles is guided by their own best position in the search space - cognitive component - as well as the best position of the whole swarm - social component. The global best solution is constantly updated and informed to the swarm, guiding the swarm movement towards the best-known solution. The search procedure is stochastic, and so it is expected that the satisfactory solution will be discovered after a certain number of iterations of the algorithm. For each iteration of the PSO, the jth amino acid of particle i, the velocity v (Eq. 2) and the position x are updated (Eq. 3), considering y as the particle's best position and g as the global best position.

$$v_{\mathrm{i,j}}(t+1) = w * v_{\mathrm{i,j}}(t) + c_1 r_{1,\mathrm{j}}(t)(y_{\mathrm{i,j}}(t) - x_{\mathrm{i,j}}(t)) + c_2 r_{2,\mathrm{j}}(t)(g_{\mathrm{i,j}}(t) - x_{\mathrm{i,j}}(t)) \quad (2)$$

$$x_{\mathrm{i}}(t+1) = x_{\mathrm{i}}(t) + v_{\mathrm{i}}(t+1) \quad (3)$$

where w is the inertia weight, $c1$ and $c2$ are the social cognitive and social acceleration rates, respectively, and $r_{1,j}(t)$, $r_{2,j}(t)$ $U(0,1)$.

Meissner [45] built an PSO algorithm to optimize backbone geometries of proteins considering secondary structure information in the optimization process. In [51], a distributed parallel particle swarm optimization algorithm was developed for protein structure prediction problem. [60] presents a hybrid PSO-GA algorithm to search for the native structure of a protein molecule in a hydrophobic-hydrophilic lattice model representation. Other works have already been done using PSO implementations for the PSP problem. In 2005, Liu [61] used the canonical implementation of the PSO in a hydrophobic-hydrophilic lattice model and concluded that this implementation would need improvements. In 2009, Lin [60] introduced a PSO ensemble with GA to the PSP using a hydrophobic-hydrophilic lattice model. Kondov [51] presented a parallel distributed model for PSP. In 2015, Borguesan[20] used the canonical implementation of the PSO using APL information.

To apply the PSO to the PSP, each particle was formed by a set of torsion angle pairs (ϕ and ψ) for each amino acid of the protein (see Figure 1), its current fitness value, velocity, inertia value and the set of angles for its individual best fitness value (best individual fitness memory). APL_1, APL_2 and APL_3 were used to generate a

swarm of size 1000 particles (Algorithm 1, line 1). These particles move in the search space being guided by the score function described on Subsection 2.2, in order to minimize this value. A global best topology has been used for optimization, which means that all particles are able to see the whole swarm at all times, and best of all particles will guide the entire swarm. In each iteration, there is the possibility of finding a better individual solution (tested on Algorithm 1, line 7) and a better global solution (tested on Algorithm 1, line 10), if this occurs, these values will be updated and will guide the movement of the particle, and the swarm in the next iteration. Based on the convergence time of the algorithms and the possibility to execute the codes in several architectures guaranteeing the same computational effort, 10^6 energy evaluations were used as a stop criterium.

Algorithm 1: Canonical PSO for the 3-D PSP Problem

```
   Data: Swarm created using the APL.
   Result: The best particle
 1 particle ← Generate the particles using APL;
 2 setOfinitialFitness ← Fitness(Particles);
 3 globalBest ← min(setOfinitialFitness);
 4 c1 = 4.0; c2 = 2.0; w = 0.6; fitnessCalculations = 10^6;
 5 for i in particles do
 6     i.actualFitness ← Fitness(i);
 7     if i.actualFitness < i.fitness then
 8         i.fitness ← actualFitness;
 9         i.fitnessPosition ← i.actualPosition;
10     end
11     if i.actualFitness < globalBest.fitness then
12         globalBest ← i;
13         cognitive ← i.fitnessPosition − i.actualPosition;
14         social ← globalBest.position − .actualPosition;
15         term_1 ← (c1 ∗ rand ∗ cognitive);
16         term_2 ← (c2 ∗ rand ∗ social);
17         v_i ← w ∗ v_{i−1} + term_1 + term_2;
18         for angle in i.actualPosition do
19             position_i = angle + v_i;
20             i.actualPosition[angle] ← position_i;
21         end
22     end
23     return globalBest
24 end
```

2.4. *SAN-PSO: Self Adaptative Niching PSO*

The canonical implementation of the PSO described in Section 2.3 presents some negative points regarding its performance, such as premature convergence and entrapment in local minima due to the lack of diversity of the swarm due to the social propagation of the global best solution which result in a low rate of exploration of the immense search space that is characteristic of the PSP problem. As discussed, several approaches have been used to address these failures. Based on the folding funnel hypothesis, which proposes more than one possible structure with minimum energy, in this work we propose a multimodal approach for the PSO, aiming to increment the search space exploitation rate by increasing the diversity of the particles in order to find, in the end, more than one finalist structure. Our multimodal approach is based on niching, so we chose an unsupervised learning technique, hierarchical clustering, to divide the swarm into groups without having to know the number of clusters *a priori*. To insert a self-adaptive character in our proposal, we added a single-solution local search, simulated annealing (SA) [62, 63], after PSO execution, to improve each of the finalist structures.

Figure 3 represents a diagram of the process involved in the method proposed in this work - the SAN-PSO. Receiving as input the amino acid sequence of the proteins and their secondary structure, APL is used to create an initial swarm composed by 1000 particles torsion angles information. Our proposal is multimodal PSO based on an iterative process that intercalates niching, made with hierarchical clustering based on structural similarity of the particles, and optimization of the fitness function using PSO in each cluster separately. After the stop criterium is reached, the finalist solutions are submitted to a single-solution optimization local search using Simulated Annealing in order to improve in a self-adaptive way the solutions found by the niching PSO.From obtaining the torsion angles data from APL and creation of the initial swarm, to a schematic description of the iterative process of hierarchical clustering and PSO that characterizes the method together with the application of Simulated Annealling, conferring the self-adaptive character and improving the finalist structures.

There are many studies regarding the influence of the inertia, social and cognitive terms on the convergence time and trajectory of the particles[64, 65, 66]. We propose an adaptive version of the inertia weight calculation proposed by Ismail and Engelbrecht[66] for convergence reasons as follows:

$$w_{\mathrm{i}}(t) = w_{\mathrm{i}}(t-1) + (w_{\mathrm{b}} - w_{\mathrm{a}})\frac{CurrentIteration}{MaxIterations} \tag{4}$$

where w_a and w_b are the limits of the inertia interval.

We implemented an adaptation of the niching PSO introduced by Brits et.al. [67] using hierarchical clustering from SciPy package [68] (Algorithm 2, line 6) to create niches based on the structural distance of each particle and the entire swarm using

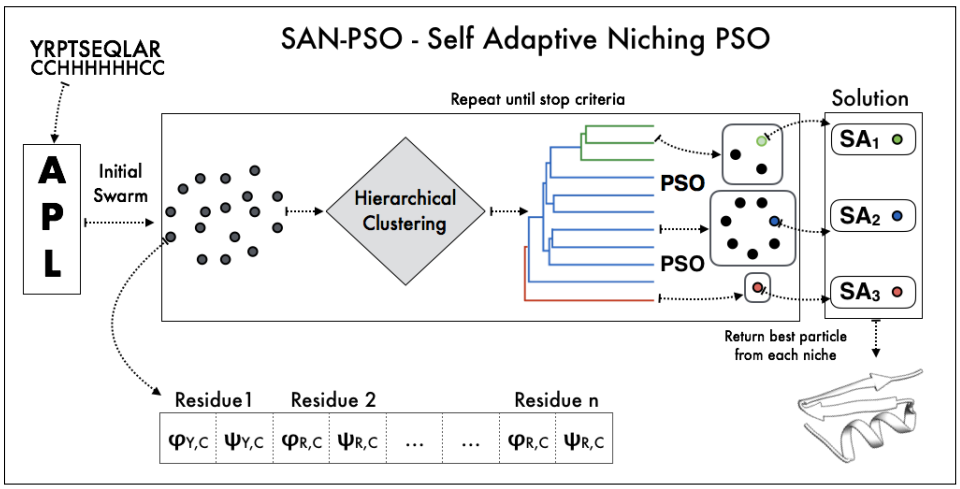


Fig. 3: Graphical scheme of the proposed method.

the root mean square deviation (RMSD)(see Equation 5). Hierarchical clustering uses a condensed distance matrix and a cutoff to cluster with no *a priori* information of the number o clusters needed. Clustering cutoff criterium was used as described in equation 6 [69]. We applied the PSO with adaptive weight inertia in each cluster (Algorithm 2, lines 7 - 28). After each iteration, the entire swarm was subjected to hierarchical clustering once again, allowing particles to move from one cluster to another, aiming to increase the diversity of the swarm. After the stop criterium was reached, SA was applied to improve the resulting structures (Algorithm 2, line 33). SA is a metaheuristic inspired by the process of heating and slow cooling in metallurgy, it is a local search algorithm that focus on retrieving the approximate global optimization for a single solution.

$$\mathbf{RMSD}(a,b) = \sqrt{\left(\sum_{i=1}^{n} \|r_{ai} - r_{bi}\|^2\right)/n}, \tag{5}$$

where r_{ai} and r_{bi} are vectors representing the positions of the same atom i in each of two structures, a and b respectively, and where the structures a and b are optimally superimposed.

$$\text{cutoff} = (\text{length of sequence})^{\frac{1}{3}} \tag{6}$$

3. Computational Experiments and Results

To test the predictive performance of our proposed self-adaptive niching PSO (SAN-PSO), described in Subsection 2.4, comparing it with the canonical implemen-

Algorithm 2: Self-adaptative Niching PSO for the 3-D PSP Problem

Data: Swarm created using the APL.
Result: The best particle of each cluster

particle ← Generate the particles using APL;
setOfinitialFitness ← Fitness(Particles);
globalBest ← min(setOfinitialFitness);
c1 = 4.0; c2 = 2.0; wBeginning = 0.4; wEnd = 0.9, w = 0; fitnessCalculations = 10^6;
while *energyCalculations*$\leq 10^6$ **do**
  clusters ← *hierarchicalClustering(Swarm)* **for** *c in cluster* **do**
    for *particle in cluster* **do**
      particle.actualFitness ← *Fitness(particle);*
      if *particle.actualFitness* < *particle.fitness* **then**
        particle.fitness ← *actualFitness;*
        particle.fitnessPosition ← *particle.actualPosition;*
      end
      if *particle.actualFitness* < *globalBest.fitness* **then**
        globalBest ← *particle;*
        for *aa in sequence* **do**
          $cognitive \leftarrow particle.fitnessPosition - particle.actualPosition;$
          $social \leftarrow globalBest.position - particle.actualPosition;$
          $term_1 \leftarrow (c1 * rand * cognitive);$
          $term_2 \leftarrow (c2 * rand * social);$
          $w_t \leftarrow w_{t-1} + (w_b - w_a) * (CurrentIteration/MaxIterations)$
          $v_{aa}(t) \leftarrow w * v_{aa}(t-1) + term_1 + term_2;$
          for *angle in particle[aa].actualPosition* **do**
            $position_{aa}$ = *angle* + v_{aa};
            $particle.actualPosition[angle] \leftarrow position_i;$
          end
        end
      end
    end
  end
  return *setOfBestParticleForEachCluster*
end
for *bestParticle in setOfBestParticleForEachCluster* **do**
  optimezedParticle ← *simulatedAnnealling*(*bestParticle*)
  return *optimezedParticle*
end

tation described in Subsection 2.3, we selected a set o seven proteins with diverse size and secondary structure components from PDB.Table 1 presents details of the target protein sequences. Column 1 shows the PDB identification number, column 2 shows the reference of protein structure, column 3 shows its size and column 4 describe the secondary structure content of each protein. The secondary structure

content was assigned using STRIDE[70] and along with the amino acid sequence was the input for creating the initial swarm from APL in both implementations. For each protein, we ran the algorithms 30 times with a stop criterium of 10^6 energy function calculations. The results were analysed in terms of structural quality and energy values.

Table 1: Target protein sequences. The size of the amino acid sequences varies from 14-46 amino acids.

PDB ID	Reference	Sequence Size	SS Component
1AB1	Yamano et al. [71]	46	2 helices 1 sheet
1ACW	Blanc et al. [72]	29	1 helix 1 sheet
1K43	Pastor et al. [73]	14	1 sheet
1WQC	Chagot et al. [74]	26	2 helices
2MR9	Nowicka et al. [75]	44	3 helices
2MTW	Cifuentes et al. [76]	20	1 helix
2P81	Religa et al. [77]	44	2 helices

For quality analysis, three parameters were evaluated: RMSD2, Global Distance Total Score Test (GDT_ST) and Energy. Of these, both the RMSD and the GDT_ST are structural parameters while the energy evaluates the stereochemical quality of the solutions. As previously described, RMSD is a distance metric between the atoms, in this case, the distances between the C_α of the predicted and the experimental structures were evaluated. The two last amino acids in the extremity of the structure were not considered due their high flexibility. GDT_ST measures the percentage of atoms of the predicted structure that are in agreement with those of the experimental structure, this metric has been repeatedly used as qualtity parameter in CASP. The energy analysis was made by calculating the fitness function value as described in Equation 1.

Concerning the overall performance of SAN-PSO, Table2 describes the structural and energy results. On column 2 mean and standard deviation RMSD. In the column 3 are described the values of mean and standard deviation of GDT_ST for for the best particles from the 30 runs for PSO and the lowest GDT_ST particle among the finalists from the 30 runs of SAN-PSO. Column 4 represent the mean and standard deviation energy values for the best particles from the 30 runs for PSO and the best particle among the finalists from the 30 runs of SAN-PSO. An Analysis of Variance (ANOVA) was performed to compare these results and those that obtained $p < 0.01$ are highlighted in bold on Table 2, meaning they are significantly different.

Figure 4 shows the best results for PSO and SAN-PSO algorithms compared with the experimental structure in green, blue and red, respectively. Visually, it is possible to conclude that the two implementations are able to predict very reliably and consistently the regular structures, such as helices and sheets, of the test proteins.

Table 2: Structural and Energy Results

PDB ID	RMSD ($\AA$)	GDT_ST (%)	Energy
1AB1 (PSO)	7.60 ± 1.63	46.45±4.44	-121.10±23.86
1AB1 (SAN-PSO)	7.52 ± 1.98	48.86±5.43	-111.75±22.61
1ACW (PSO)	5.14 ± 1.67	56.06±8.17	-118.08±42.49
1ACW (SAN-PSO)	**3.90 ± 1.07**	**61.93±4.64**	**-147.50±15.24**
1K43 (PSO)	2.71 ± 0.63	77.86±4.31	-42.03±1.80
1K43 (SAN-PSO)	2.77 ± 0.65	76.96±5.05	-41.20±1.71
1WQC (PSO)	3.95 ± 0.80	64.81±5.03	-142.02±6.42
1WQC (SAN-PSO)	3.92 ± 1.10	65.38±6.83	**-145.66±3.41**
2MR9 (PSO)	6.22 ± 2.08	53.79±8.88	-275.95±7.33
2MR9 (SAN-PSO)	5.53 ± 2.00	55.87±8.13	-274.91±5.36
2MTW (PSO)	**4.09 ± 0.86**	**62.21±6.08**	-94.05 ±2.26
2MTW (SAN-PSO)	4.99 ± 0.75	58.20±4.81	**-95.46±1.97**
2P81 (PSO)	8.04 ± 1.5	33.54±1.87	-248.88±5.39
2P81 (SAN-PSO)	7.39 ± 1.22	33.73±2.22	-246.95±3.48

The same can not be said about non-regular regions, such as coils and turns. This is because these are very flexible structures that have a greater diversity of possible conformations. The fitness function (Equation 1) used for optimization is to prioritize the packing of proteins, which will lead to the formation of more globular structures, which does not always occur in the native structure.

From a structural point of view, the results of the two implementations are similar except for two cases in which each approach was better. Regarding RMSD values, the PSO was better when predicting 2MTW ($p = 0.0002$) and the SAN-PSO results were better with 1ACW ($p = 0.001$), the same behaviour was observed in the GDT_ST results. As can be observed in Figure 4, the variation of the predicted structures with the experimental one is more evident in the irregular regions such as coils and turns in N- and C- terminal regions of the proteins. To evaluate whether the results could be explained by this assumption, we calculated the RMSD of the two proteins using the same data from Table 2 analysis, however, the flexible N and C terminal regions were removed and, as was done previously, the results were analyzed with ANOVA. The results showed that the significance of the difference between the structures decreased by an order of magnitude in the case of 2MTW ($p = 0.02$), which means that the irregular regions at the ends of the protein influenced the difference between the structures, These regions have become less structurally different. In the case of 1ACW, the significance of the difference between the structures increased by an order of magnitude ($p = 0.0003$) with the removal of the coil regions in N- and C- terminal regions of the proteins, indicating that, by removing the noise from the flexible ends, the SAN-PSO performance for this protein was even

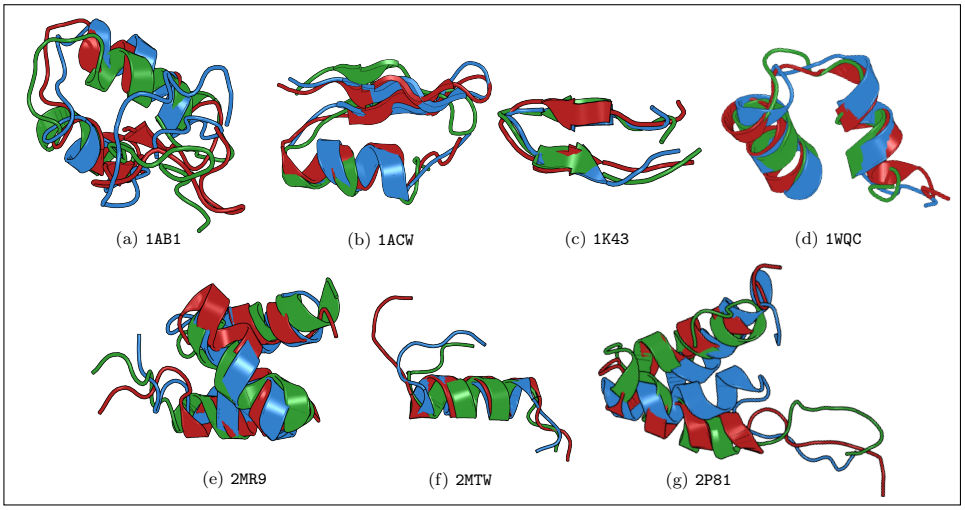


Fig. 4: Ribbon representation of the experimental (red), lowest RMSD from SAN-PSO (blue) and lowest RMSD from Canonical PSO (green). The C_α of the experimental and the predicted 3-D structure are fitted. Amino acid side chains are not shown for clarity. Graphic representation was prepared with PyMOL.

better than the PSO in generating regular structures.

According to the energy results, the SAN-PSO was significantly better than the PSO in 42.86% of the cases and was never worse in relation to the PSO. This result indicates that in fact the multimodal and self adaptive approach is able to better explore the search space, finding better energy minima. This indicates that the capacity of optimization of the energy function is improved in relation to the canonical implementation of PSO.

To evaluate the distribution data of the two implementations, Figure 5 shows a box plot of the best particles of each run for the two implementations taking into account their RMSD values to the experimental structure. By visual inspection, it is possible to interpret that, overall, the distribution present smaller medians and interquartile distances although these differences have not been shown significantly ($p < 0.01$), except for 1ACW and 2MTW. Corroborating with the data of Table 2, 1ACW presents a more concise distrubution, with an interquartile distance smaller than that of the canonical implementation, and with median clearly smaller. Still in agreement with Table 2, 2MTW presents greater median and greater interquartile distance.

4. Conclusion

In this work we propose a novel ensemble of methods for the PSP using information from *Angle Probability Lists*. Combining particle swarm optimization with niching

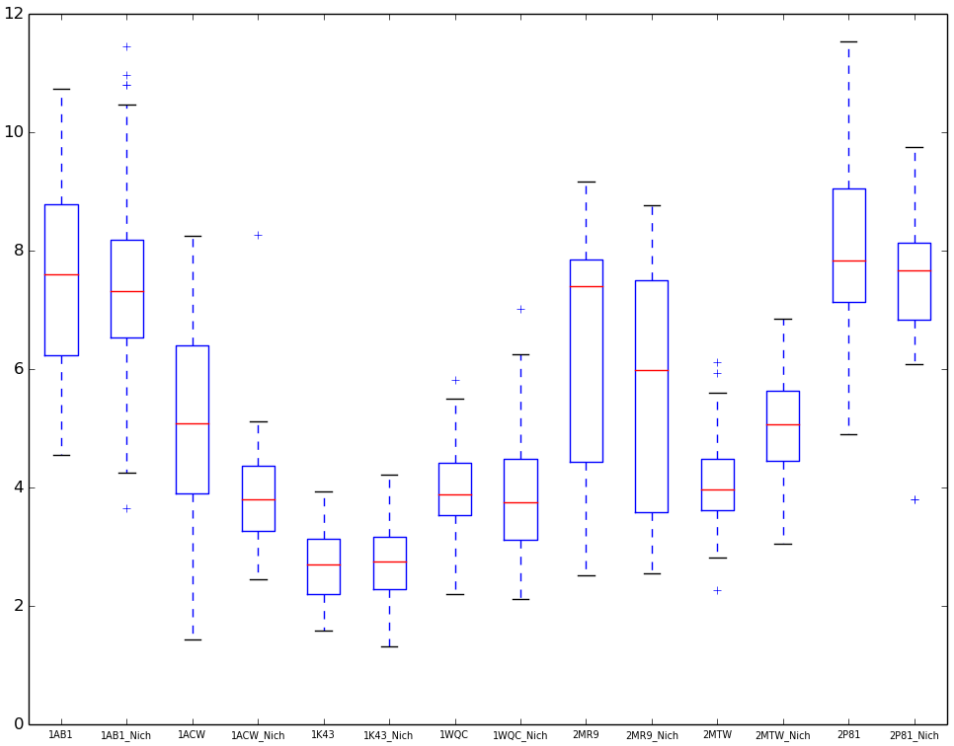


Fig. 5: Box Plot representing `RMSD` (Å) data from the best particles in each of the 30 runs for `PSO` and `SAN-PSO`.

by unsupervised learning, we seek to increase the exploration rate of the search space while minimizing the fitness function, and to improve the finalist solutions, we chose to use local search with simulated annealing. For comparison, the canonical version of the particle swarm optimization was also implemented.

The results indicate that the new strategy proposed in this work is capable of performing the general folding of test proteins, especially with respect to regular secondary structure regions such as helices and sheets. On the minimization of the fitness function, the new approach proved to be better than canonical implementation in more than 40% of the cases and was not worse in any test case. This results might be related to the introduction of the multimodal character to the implementation, which increases the search space exploration. From this point of view, a future work will be carried out with the `SAN-PSO` optimizing other energy functions, with the objective of obtaining structural results that accompany the optimization. This becomes very challenging, since the choice of the energy function is not trivial, because there is no function capable of representing the real system in a completely reliable way.

ACKNOWLEDGEMENTS

This research received funding by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (MD); the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) (MD); and the Fundação de Amparo a Pesquisa do Estado do Rio Grande do Sul (FAPERGS) (MD). This Research is supported by Microsoft under a Microsoft Azure for Research Award (MD).

References

1. A. M. Lesk. *Introduction to Bioinformatics*. Oxford University Press Inc., New York, USA, 1 edition, 2002.
2. A. Tramontano. *Protein structure prediction: concepts and applications*. John Wiley and Sons, Inc., Weinheim, Germany, 1 edition, 2006.
3. A.L. Lehninger, D.L. Nelson, and M.M. Cox. *Principles of Biochemistry*. W.H. Freeman, New York, USA, 4 edition, 2005.
4. A. M. Lesk. *Introduction to Protein Sci.* Oxford University Press, New York, 2 edition, 2010.
5. M. Dorn, M. Barbachan e Silva, L. S. Buriol, and L. C. Lamb. Three-dimensional protein structure prediction: Methods and computational strategies. *Comput. Biol. Chem.*, 53, Part B:251 – 276, 2014.
6. E.S. Lander and M.S. Waterman. *The secrets of life: a mathematician's introduction to Molecular Biology*. National Academy Press, Washington D. C., USA, 1999.
7. J.C. Wooley and Y. Ye. *A historical perspective and overview of protein structure prediction*, chapter 1, pages 1–43. Springer, 2010.
8. C. Levinthal. Are there pathways for protein folding? *J. Chim. Phys. Phys.-Chim. Biol.*, 65(1):44–45, 1968.
9. C.B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181(96):223–230, 1973.
10. Joseph D Bryngelson, Jose Nelson Onuchic, Nicholas D Socci, and Peter G Wolynes. Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins: Structure, Function, and Bioinformatics*, 21(3):167–195, 1995.
11. Andriy Kryshtafovych, Alessandro Barbato, Krzysztof Fidelis, Bohdan Monastyrskyy, Torsten Schwede, and Anna Tramontano. Assessment of the assessment: evaluation of the model quality estimates in casp10. *Proteins: Structure, Function, and Bioinformatics*, 82(S2):112–126, 2014.
12. J Thomas Ngo, Joe Marks, and Martin Karplus. Computational complexity, protein structure prediction, and the levinthal paradox. In *The protein folding problem and tertiary structure prediction*, pages 433–506. Springer, 1994.
13. Pierluigi Crescenzi, Deborah Goldman, Christos Papadimitriou, Antonio Piccolboni, and Mihalis Yannakakis. On the complexity of protein folding. *Journal of computational biology*, 5(3):423–465, 1998.

14. Christophe Guyeux, Nathalie M-L Côté, Jacques M Bahi, and Wojciech Bienia. Is protein folding problem really a np-complete one? first investigations. *Journal of bioinformatics and computational biology*, 12(01):1350017, 2014.
15. H. Lodish, A. Berk, P. Matsudaira, C. A. Kaiser, M. Krieger, and M.P. Scott. *Molecular Cell Biology*. Scientific American Books, W.H. Freeman, New York, USA, 5 edition, 1990.
16. E.D. Scheef and J.L. Fink. *Fundamentals of protein structure: Structural Bioinformatics*, chapter 2, page 15. 2003.
17. G.N. Ramachandran and V. Sasisekharan. Conformation of polypeptides and proteins. *Adv. Protein Chem.*, 23:238–438, 1968.
18. T.Z. Hovmoller and T. Ohlson. Conformation of amino acids in protein. *Acta Crystallogr.*, 58(5):768–776, 2002.
19. X. Xia and Z. Xie. Protein structure, neighbor effect, and a new index of amino acid dissimilarities. *Mol. Biol. Evol.*, 19(1):58–67, 2002.
20. Bruno Borguesan, Mariel Barbachan e Silva, Bruno Grisci, Mario Inostroza-Ponta, and Mrcio Dorn. APL: An angle probability list to improve knowledge-based metaheuristics for the three-dimensional protein structure prediction. *Computational Biology and Chemistry*, 59, Part A:142–157, 2015.
21. J Thomas Ngo, Joe Marks, and Martin Karplus. Protein structure prediction. *The Protein Folding Problem and Tertiary Structure Prediction*, page 433, 2012.
22. Xin-She Yang and Luniver Press. Nature-inspired metaheuristic algorithms second edition. 2010.
23. A-A Tantar, Nouredine Melab, E-G Talbi, Benjamin Parent, and Dragos Horvath. A parallel hybrid genetic algorithm for protein structure prediction on the computational grid. *Future Generation Computer Systems*, 23(3):398–409, 2007.
24. Alexandru-Adrian Tantar, Nouredine Melab, and El-Ghazali Talbi. A grid-based genetic algorithm combined with an adaptive simulated annealing for protein structure prediction. *Soft Computing*, 12(12):1185, 2008.
25. Mario Garza-Fabre, Shaun M Kandathil, Julia Handl, Joshua Knowles, and Simon C Lovell. Generating, maintaining and exploiting diversity in a memetic algorithm for protein structure prediction. *Evolutionary computation*, 2016.
26. Fábio Lima Custódio, Helio JC Barbosa, and Laurent Emmanuel Dardenne. A multiple minima genetic algorithm for protein structure prediction. *Applied Soft Computing*, 15:88–99, 2014.
27. G. Beni and J. Wang. Swarm intelligence in cellular robotic systems. In *NATO Advanced Workshop on Robotics and Biological Systems*, June 1989.
28. Amit Garg and Pawan Gill. An insight into swarm intelligence. 2009.
29. Iztok Fister Jr, Xin-She Yang, Iztok Fister, Janez Brest, and Dušan Fister. A brief review of nature-inspired algorithms for optimization. *arXiv preprint arXiv:1307.4186*, 2013.
30. Russel Ebenhart. Kennedy. particle swarm optimization. In *Proceeding IEEE Inter Conference on Neural Networks, Perth, Australia, Piscat-away*, volume 4,

pages 1942–1948, 1995.
31. M. Mitchell. *An Introduction to Genetic Algorithms*. MI2005T Press, Cambridge, USA, 5 edition, 1999.
32. Marco Dorigo, Mauro Birattari, and Thomas Stutzle. Ant colony optimization. *IEEE computational intelligence magazine*, 1(4):28–39, 2006.
33. Mehul M. Khimasia and Peter V. Coveney. Protein structure prediction as a hard optimization problem: The genetic algorithm approach. *Molecular Simulation*, 19(4):205–226, 1997.
34. R. Das, B. Qian, S. Raman, R. Vernon, J. Thompson, P. Bradley, S. Khare, M.D.D. Tyka, D. Bhat, D. Chivian, D.E.E. Kim, W.H.H. Sheffler, L. Malmstroem, A.M.M. Wollacott, C. Wang, I. Andre, and D. Baker. Structure prediction for casp7 targets using extensive all-atom refinement with rosetta@home. *Proteins*, 68(S8):118–128, 2007.
35. Alena Shmygelska and Holger H Hoos. An ant colony optimisation algorithm for the 2d and 3d hydrophobic polar protein folding problem. *BMC bioinformatics*, 6(1):1, 2005.
36. J.T. Pedersen and J. Moult. Genetic algorithms for protein structure prediction. *Curr. Opin. Struct. Biol.*, 6(2):227–231, 1996.
37. Natalio Krasnogor, William E Hart, Jim Smith, and David A Pelta. Protein structure prediction with evolutionary algorithms. In *Proceedings of the 1st Annual Conference on Genetic and Evolutionary Computation-Volume 2*, pages 1596–1601. Morgan Kaufmann Publishers Inc., 1999.
38. Andrei Băutu and Henri Luchian. Protein structure prediction in lattice models with particle swarm optimization. In *International Conference on Swarm Intelligence*, pages 512–519. Springer, 2010.
39. Heitor Silvério Lopes. Evolutionary algorithms for the protein folding problem: A review and current trends. In *Computational intelligence in biomedicine and bioinformatics*, pages 297–315. Springer, 2008.
40. Bruno Borguesan, Mario Inostroza-Ponta, and Márcio Dorn. Nias-server: Neighbors influence of amino acids and secondary structures in proteins. *Journal of Computational Biology*, 23, 2016.
41. Shaojian Sun. Reduced representation model of protein structure prediction: Statistical potential and genetic algorithms. *Protein Science*, 2(5):762–785, 1993.
42. Richard O Day, Jesse B Zydallis, Gary B Lamont, and R Pachter. Solving the protein structure prediction problem through a multiobjective genetic algorithm. *Nanotechnology*, 2:32–35, 2002.
43. José C Calvo, Julio Ortega, and Mancia Anguita. Comparison of parallel multiobjective approaches to protein structure prediction. *The Journal of Supercomputing*, 58(2):253–260, 2011.
44. Alain Pétrowski. A clearing procedure as a niching method for genetic algorithms. In *Evolutionary Computation, 1996., Proceedings of IEEE International Conference on*, pages 798–803. IEEE, 1996.

45. M. Meissner and G. Schneider. Protein folding simulation by particle swarm optimization. *Open Struct. Biol. J*, 1:1–6, 2007.
46. Mauricio Zambrano-Bigiarini, Maurice Clerc, and Rodrigo Rojas. Standard particle swarm optimisation 2011 at cec-2013: A baseline for future pso improvements. In *2013 IEEE Congress on Evolutionary Computation*, pages 2337–2344. IEEE, 2013.
47. Dian Palupi Rini, Siti Mariyam Shamsuddin, and Siti Sophiyati Yuhaniz. Particle swarm optimization: technique, system and challenges. *International Journal of Computer Applications*, 14(1):19–26, 2011.
48. Juan Lin, Jing Ning, Qing-Liang Du, and Yi-Wen Zhong. Multi-agent simulated annealing algorithm based on particle swarm optimization algorithm for protein structure prediction. *Journal of Bionanoscience*, 7(1):84–91, 2013.
49. Tianyu Hu, Mandong Hu, Ling Lv, and Changjun Zhou. Improved genetic algorithm-particle swarm optimization based on multiple populations for 3d protein structure prediction. *Journal of Computational and Theoretical Nanoscience*, 12(7):1414–1419, 2015.
50. Xiaoli Lin, Fengli Zhou, and Huayong Yang. Effective protein structure prediction with the improved lapso algorithm in the ab off-lattice model. In *International Conference on Intelligent Computing*, pages 448–454. Springer, 2016.
51. I. Kondov. Protein structure prediction using distributed parallel particle swarm optimization. *Nat. Comput.*, 12(1):29–41, 2013.
52. Christopher W Cleghorn and Andries P Engelbrecht. Particle swarm variants: standardized convergence analysis. *Swarm Intelligence*, 9(2-3):177–203, 2015.
53. Carol A Rohl, Charlie EM Strauss, Kira MS Misura, and David Baker. Protein structure prediction using rosetta. *Methods Enzymol.*, 383:66–93, 2004.
54. W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–637, 1983.
55. T. Kortemme, A.V. Morozov, and D. Baker. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein–protein complexes. *J. Mol. Biol.*, 326(4):1239–1259, 2003.
56. Themis Lazaridis and Martin Karplus. Effective energy functions for protein structure prediction. *Curr. Opin. Struct. Biol.*, 10(2):139–145, 2000.
57. Roland L Dunbrack and Martin Karplus. Backbone-dependent rotamer library for proteins application to side-chain prediction. *J. Mol. Biol.*, 230(2):543–574, 1993.
58. Brian Kuhlman and David Baker. Native protein sequences are close to optimal for their structures. *Proc. Natl. Acad. Sci. U.S.A.*, 97(19):10383–10388, 2000.
59. Carolyn S Sevier and Chris A Kaiser. Formation and transfer of disulphide bonds in living cells. *Nature reviews Molecular cell biology*, 3(11):836–847, 2002.
60. C.J. Lin and M.H. Hsieh. An efficient hybrid taguchi-genetic algorithm for protein folding simulation. *Expert Syst. with Applic.*, 36(10):12446–12453, 2009.
61. Juan Liu, Longhui Wang, Lianlian He, and Feng Shi. Analysis of toy model

for protein folding based on particle swarm optimization algorithm. In *International Conference on Natural Computation*, pages 636–645. Springer, 2005.
62. S. Kirkpatrick, C.D. Gelatt, and M.P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671, 1983.
63. V. Granville, M. Krivanek, and J.-P. Rasson. Simulated annealing: A proof of convergence. *IEEE T. Pattern Anal.*, 16(6):652, 1994.
64. Idel Montalvo, Joaqun Izquierdo, Rafael Prez-Garca, and Manuel Herrera. Improved performance of pso with self-adaptive parameters for computing the optimal design of water supply systems. *Engineering Applications of Artificial Intelligence*, 23(5):727 – 735, 2010.
65. A Rezaee Jordehi and Jasronita Jasni. Parameter selection in particle swarm optimisation: a survey. *Journal of Experimental & Theoretical Artificial Intelligence*, 25(4):527–542, 2013.
66. Adiel Ismail and Andries P Engelbrecht. Self-adaptive particle swarm optimization. In *Asia-Pacific Conference on Simulated Evolution and Learning*, pages 228–237. Springer, 2012.
67. Riaan Brits, Andries P Engelbrecht, and F Van den Bergh. A niching particle swarm optimizer. In *Proceedings of the 4th Asia-Pacific conference on simulated evolution and learning.*, pages 692–696, 2002.
68. Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001–. [Online; accessed 2016-11-20].
69. Vladimir N Maiorov and Gordon M Crippen. Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins. *Journal of molecular biology*, 235(2):625–634, 1994.
70. D. Frishman and P. Argos. Knowledge-based protein secondary structure assignment. *Proteins*, 23(4):566–579, 1995.
71. A. Yamano, N.H. Heo, and M.M. Teeter. Crystal structure of Ser-22/Ile-25 form crambin confirms solvent, side chain substate correlations. *J. Biol. Chem.*, 272(15):9597–9600, Apr 1997.
72. E. Blanc, V. Fremont, P. Sizun, S. Meunier, J. Van Rietschoten, A. Thevand, J.M. Bernassau, and H. Darbon. Solution structure of P01, a natural scorpion peptide structurally analogous to scorpion toxins specific for apamin-sensitive potassium channel. *Proteins: Struct., Funct., Bioinf.*, 24(3):359–369, Mar 1996.
73. M.T. Pastor, M. Lpez de la Paz, E. Lacroix, L. Serrano, and E. Prez-Pay. Combinatorial approaches: A new tool to search for highly structured beta-hairpin peptides. *Proc. Natl. Acad. Sci.*, 99(2):614–619, 2002.
74. B. Chagot, C. Pimentel, L. Dai, J. Pil, J. Tytgat, T. Nakajima, G. Corzo, H. Darbon, and G. Ferrat. An unusual fold for potassium channel blockers: Nmr structure of three toxins from the scorpion opisthacanthus madagascariensis. *Biochem. J.*, 388:263–271, 2005.
75. U. Nowicka, D. Zhang, O. Walker, D. Krutauz, C.A. Castaneda, A. Chaturvedi, T.Y. Chen, N. Reis, M.H. Glickman, and D. Fushman. DNA-damage-inducible 1 protein (Ddi1) contains an uncharacteristic ubiquitin-like domain that binds

ubiquitin. *Structure*, 23(3):542–557, Mar 2015.

76. G. Cifuentes, L.M. Salazar, L.E. Vargas, C.A. Parra, M. Vanegas, J. Cortes, and M.E. Patarroyo. Evidence supporting the hypothesis that specifically modifying a malaria peptide to fit into HLA-DRbeta1*03 molecules induces antibody production and protection. *Vaccine*, 23(13):1579–1587, Feb 2005.
77. T.L. Religa, C.M. Johnson, D.M. Vu, S.H. Brewer, R.B. Dyer, and A.R. Fersht. The helix-turn-helix motif as an ultrafast independently folding domain: the pathway of folding of Engrailed homeodomain. *Proc. Natl. Acad. Sci. U.S.A.*, 104(22):9272–9277, May 2007.