Programa Interunidades em Bioinformática

Universidade de São Paulo (USP)

**Função de Avaliação Dinâmica em Algoritmos
Genéticos Aplicados na Predição de Estruturas
Tridimensionais de Proteínas**

Luís Henrique Uchida Ishivatari

Ribeirão Preto, 2012

Luis Henrique Uchida Ishivatari

Função de Avaliação Dinâmica em Algoritmos Genéticos Aplicados
na Predição de Estruturas Tridimensionais de Proteínas

Dissertação apresentada ao Programa
Interunidades em Bioinformática da
Universidade de São Paulo para a
obtenção do título de Mestre em
Bioinformática

Área de Concentração:
Bioinformática

Orientador: Prof. Dr.
Renato Tinós

Co-Orientador: Prof. Dr.
Fernando Luis Barroso da Silva

Ribeirão Preto, 2012

Agradecimentos

Meus sinceros agradecimentos aos professores Renato Tinós e Fernando L. Barroso da Silva pela orientação, incentivo, disponibilidade, ensinamentos e discussões durante a realização deste projeto.

Aos meus pais e meu irmão, pelo apoio moral e financeiro, dando suporte às minhas decisões tomadas durante o período do mestrado.

Aos meus amigos Lariza L. de Oliveira e Helder Ken Shimo, tanto pelas discussões e sugestões relacionadas a este projeto quanto pela amizade em si, cultivada durante todos esses anos.

Aos professores Luis Paulo Scott, Otavio Henrique Thiemann e Hélio José Correa Barbosa, que compuseram a banca de qualificação, pelas sugestões e discussões para este projeto.

Por fim, agradeço à FAPESP (Projeto 2009/12931-3), pelo apoio financeiro, fundamental para o desenvolvimento deste projeto.

Resumo

O problema de predição de estruturas tridimensionais de proteínas pode ser visto computacionalmente como um problema de otimização, tal que dada a sequência de aminoácidos, deve-se encontrar a estrutura tridimensional da proteína dentre as muitas possíveis através da obtenção de mínimos de funções de energia. Vários pesquisadores têm proposto estratégias de Computação Evolutiva para a determinação de estruturas tridimensionais das proteínas, entretanto nem sempre resultados animadores têm sido alcançados visto que entre outros fatores, há um grande número de ótimos locais no espaço de busca. Geralmente as funções de *fitness* empregadas pelos algoritmos de otimização são baseadas em campos de força com diferentes termos de energia, sendo que os parâmetros destes termos são ajustados *a priori* e são mantidos estáticos ao longo do processo de otimização. Alguns pesquisadores sugerem que o uso de funções de *fitness* dinâmicas, ou seja, que mudam durante um processo de otimização evolutivo, pode aumentar a capacidade das populações fugirem de ótimos locais em problemas altamente multimodais. Neste trabalho, propõe-se que os parâmetros dos termos do campo de força utilizado sejam modificados durante o processo de otimização realizado por Algoritmos Genéticos (AGs) no problema de predição de estruturas de proteínas, sendo aumentados ou diminuídos, por exemplo, de acordo com a sua influência na formação de estruturas secundárias e no seu ajuste fino. Como a função de avaliação será modificada durante o processo de otimização, a predição de estruturas tridimensionais de proteínas torna-se um problema de otimização dinâmica, sendo que o uso de Algoritmos Genéticos específicos para tais problemas, como o AG com hipermutação e os AGs com imigrantes aleatórios são investigados aqui. É proposta uma nova métrica relacionada ao alinhamento da estrutura secundária da proteína, para auxiliar a análise dos dados obtidos e os resultados dos experimentos indicam que os algoritmos com função de avaliação dinâmica obtiveram resultados melhores que os algoritmos estáticos, o que é explicado pelo fato de as mudanças na função de *fitness* possibilitarem eventuais fugas de ótimos locais, bem como um aumento da diversidade da população.

Abstract

The protein structure prediction can be seen as an optimization problem where given an amino acid sequence, the tertiary protein structure must be found amongst many possible by obtaining energy functions minima. Many researchers have been proposing Evolutionary Computation strategies to find tridimensional structures of proteins; however results are not always satisfactory since among other factors, there are always a great number of local optima in the search space. Usually, the fitness functions used by optimization algorithms are based on force fields with different energy terms with parameters from those terms being adjusted *a priori*, kept static through the optimization process. Some researchers suggest that the use of dynamic functions, i.e., that can be changed during the evolutionary process, can help the population to escape from local optima in highly multimodal problems. In this work we propose that the force field parameters can be changed during the optimization process of Genetic Algorithms (GAs) in the protein structure prediction problem, being increased or decreased, for instance, according with its influence on formation of secondary structures and its fine tuning. Since the cost function will be changed during the optimization process, the protein tridimensional structure prediction becomes a dynamic optimization problem and specific Genetic Algorithms for this kind of problem, like the hypermutation GA and random immigrants GA are investigated. We also propose a new metric related to the protein's secondary structure alignment to help the analysis of obtained data. Results indicate that the dynamic function algorithms obtained better results than static algorithms since changes on the fitness function allow the population to escape local optima, as well as an increase on the population diversity.

Lista de Abreviaturas

AG	Algoritmo genético
CG	<i>Coarse grained</i> (Baixa granularidade)
DOP	<i>Dynamic optimization problems</i> (Problemas de otimização dinâmica)
PEP	Predição de estruturas de proteínas
RMSD	<i>Root mean squared deviation</i> (Desvio da raiz quadrada média)
MVG	<i>Modular Varying Goal</i> (Objetivos que variam modularmente)
RVG	<i>Random Varying Goal</i> (Objetivos que variam aleatoriamente)
PDB	<i>Protein data bank</i>
HG	Head-Gordon

Sumário

1	Introdução	1
1.1	Objetivos	3
1.2	Organização do Documento	3
2	Proteínas.....	4
2.1	Aminoácidos e Proteínas	4
2.2	Proteínas	5
2.3	Modelagem de Proteínas e Campos de Força.....	7
2.4	Predição de Estruturas de Tridimensionais Proteínas	9
3	Metodologia	10
3.1	Algoritmos Genéticos.....	10
3.2	Problemas de Otimização Dinâmica	12
3.3	Detalhamento da Implementação	14
3.3.1	Codificação	14
3.3.2	Operadores Genéticos	15
3.3.3	Função de Avaliação HG	15
3.3.4	Detalhamento dos termos do campo de força HG	17
3.3.5	Função de avaliação “Onuchic”	20
3.4	Algoritmos e métodos de otimização dinâmica propostos	22
3.5	Proteínas escolhidas	23
3.6	Outros parâmetros e resumo dos experimentos.....	24
3.7	Experimento com a função de avaliação “Onuchic”	27
3.8	Análise dos Resultados.....	29
3.8.1	Fitness e Energia Equivalente.....	29
3.8.2	Desvio da Raiz Média Quadrática (RMSD)	30
3.8.3	Score de Alinhamento.....	30
3.8.4	Teste estatístico Wilcoxon	31
3.8.5	Análise Visual.....	31
3.8.6	Fluxograma dos experimentos	32
4	Resultados.....	33
4.1	Tempo de execução	33
4.2	Proteína G (2GB1)	33
4.2.1	Resultados com a função de avaliação “Onuchic”	53
4.3	Fosfatase (2W4P)	55
4.4	DNA ligante (1ENH).....	62
5	Conclusões	70

Apêndice A – Experimentos Adicionais para a proteína 2GB1	72
Referências Bibliográficas	83

Lista de Figuras

Figura 1. Representação esquemática de um aminoácido.	4
Figura 2. Exemplo de uma proteína com a estrutura secundária do tipo hélice- α	6
Figura 3. Exemplo de uma proteína com a estrutura secundária do tipo folha- β	6
Figura 4. Hemaglutinina, estrutura quaternária com três subunidades idênticas.	7
Figura 5. Representação da população do AG.	10
Figura 6. Pseudo-código do AG padrão.	11
Figura 7. Esquematização de uma busca com a troca do espaço de busca.	13
Figura 8. Representação da base de dados e codificação dos indivíduos.	14
Figura 9. Representação dos ângulos de torção ψ e ϕ	14
Figura 10. Representação do ângulo de ligação formado por três átomos.	18
Figura 11. Energia do ângulo de ligação (Eq. 2).	18
Figura 12. Representação do ângulo diedral (ϕ), formado por quatro átomos consecutivos.	19
Figura 13. Energia do ângulo diedral para duas configurações dos parâmetros.	19
Figura 14. Energia de van der Waals para uma configuração de parâmetros.	20
Figura 15. <i>Pipeline</i> dos resultados para um experimento.	32
Figura 16. Energia equivalente média em relação a 10 execuções dos melhores indivíduos ao longo das gerações da proteína 2GB1 para os experimentos comuns às outras proteínas.	34
Figura 17. Energia equivalente ao longo das gerações da execução que obteve o menor RMSD da proteína 2GB1 para os experimentos comuns às outras proteínas.	34
Figura 18. Energia equivalente média em relação a 10 execuções dos melhores indivíduos ao longo das gerações da proteína 2GB1 comparando com os experimentos de mudança aleatória.	35
Figura 19. Energia equivalente ao longo das gerações da execução que obteve o menor RMSD da proteína 2GB1 comparando com os experimentos de mudança aleatória.	36
Figura 20. Energia equivalente média em relação a 10 execuções dos melhores indivíduos ao longo das gerações da proteína 2GB1 comparando com os experimentos que usam memória. .	36
Figura 21. Energia equivalente ao longo das gerações da execução que obteve o menor RMSD da proteína 2GB1 comparando com os experimentos que usam memória.	37
Figura 22. Energia equivalente média em relação a 10 execuções dos melhores indivíduos ao longo das gerações da proteína 2GB1 comparando com os experimentos que usam relaxamento de restrições.	38
Figura 23. Energia equivalente ao longo das gerações da execução que obteve o menor RMSD da proteína 2GB1 comparando com os experimentos que usam relaxamento de restrições.	38

Figura 24. Média de 10 execuções do RMSD (em Å) dos melhores indivíduos ao longo das gerações da proteína 2GB1 para os experimentos comuns às outras proteínas.	39
Figura 25. RMSD ao longo das gerações da execução onde se obteve o menor RMSD para a proteína 2GB1 para os experimentos comuns às outras proteínas.	40
Figura 26. Média de 10 execuções do RMSD (em Å) dos melhores indivíduos ao longo das gerações da proteína 2GB1 para os experimentos de mudança aleatória.	40
Figura 27. RMSD ao longo das gerações da execução onde se obteve o menor RMSD para a proteína 2GB1 para os experimentos de mudança aleatória.	41
Figura 28. Média de 10 execuções do RMSD (em Å) dos melhores indivíduos ao longo das gerações da proteína 2GB1 para os experimentos que usam memória.	41
Figura 29. RMSD ao longo das gerações da execução onde se obteve o menor RMSD para a proteína 2GB1 para os experimentos que usam memória.	42
Figura 30. Média de 10 execuções do RMSD (em Å) dos melhores indivíduos ao longo das gerações da proteína 2GB1 para os experimentos que usam relaxamento de restrições.	42
Figura 31. RMSD ao longo das gerações da execução onde se obteve o menor RMSD para a proteína 2GB1 para os experimentos que usam relaxamento de restrições.	43
Figura 32. <i>Score</i> de alinhamento da execução onde se encontrou o maior <i>score</i> ao longo das gerações para a proteína 2GB1 para os experimentos comuns às outras proteínas.	46
Figura 33. <i>Score</i> de alinhamento da execução onde se encontrou o maior <i>score</i> ao longo das gerações para a proteína 2GB1 para os experimentos de mudança aleatória.	46
Figura 34. <i>Score</i> de alinhamento da execução onde se encontrou o maior <i>score</i> ao longo das gerações para a proteína 2GB1 para os experimentos que usam memória.	47
Figura 35. <i>Score</i> de alinhamento da execução onde se encontrou o maior <i>score</i> ao longo das gerações para a proteína 2GB1 para os experimentos que usam relaxamento de restrições.	47
Figura 36. Estruturas 3D da proteína 2GB1, enfatizando as estruturas secundárias.	49
Figura 37. Estruturas 3D, enfatizando as estruturas secundárias da proteína 2GB1.	52
Figura 38. Estruturas 3D de uma execução do experimento MVG, evidenciando as estruturas secundárias.	53
Figura 39. Energia equivalente ao longo das gerações da execução que obteve o menor RMSD dos experimentos com a função de avaliação “Onuchic”.	54
Figura 40. RMSD ao longo das gerações da execução que obteve o menor RMSD dos experimentos com a função de avaliação “Onuchic”.	54
Figura 41. Estruturas 3D da proteína 2GB1, enfatizando estruturas secundárias.	55
Figura 42. Energia equivalente média em relação a 10 execuções dos melhores indivíduos ao longo das gerações da proteína 2W4P.	56

Figura 43. Energia equivalente ao longo das gerações da execução que obteve o menor RMSD da proteína 2W4P.....	56
Figura 44. Média de 10 execuções do RMSD (em Å) dos melhores indivíduos ao longo das gerações da proteína 2W4P.....	57
Figura 45. RMSD ao longo das gerações da execução onde se obteve o menor RMSD para a proteína 2W4P.....	57
Figura 46. <i>Score</i> de alinhamento da execução onde se encontrou o maior <i>score</i> ao longo das gerações para a proteína 2W4P.	59
Figura 47. Estruturas 3D da proteína 2W4P, enfatizando as estruturas secundárias.....	60
Figura 48. Estruturas 3D, enfatizando as estruturas secundárias da proteína 2W4P.....	61
Figura 49. Energia equivalente média em relação a 10 execuções dos melhores indivíduos ao longo das gerações da proteína 1ENH.	62
Figura 50. Energia equivalente ao longo das gerações da execução que obteve o menor RMSD da proteína 1ENH.....	63
Figura 51. Média de 10 execuções do RMSD (em Å) dos melhores indivíduos ao longo das gerações da proteína 1ENH.....	64
Figura 52. RMSD ao longo das gerações da execução onde se obteve o menor RMSD para a proteína 1ENH.	64
Figura 53. <i>Score</i> de alinhamento da execução onde se encontrou o maior <i>score</i> ao longo das gerações para a proteína 1ENH.	66
Figura 54. Estruturas 3D da proteína 1ENH, enfatizando as estruturas secundárias.	67
Figura 55. Estruturas 3D, enfatizando as estruturas secundárias da proteína 1ENH..	68

Lista de Tabelas

Tabela 1. Aminoácidos, suas abreviações e características das cadeias laterais [Lehninger <i>et al.</i> , 2004].	5
Tabela 2. Sumário dos experimentos realizados para a proteína 2GB1..	25
Tabela 3. Sumário dos experimentos realizados para a proteína 2W4P.....	26
Tabela 4. Sumário dos experimentos realizados para a proteína 1ENH.	27
Tabela 5. Conjunto amostral de RMSD (em Å), dos menores valores em cada execução dentre os melhores indivíduos, para a proteína 2GB1.....	44
Tabela 6. Informações estatísticas do conjunto amostral da Tabela 5.	44
Tabela 7. p-valor do teste estatístico de Wilcoxon para o RMSD, comparando os experimento estáticos SP e Est-VdW com os experimentos dinâmicos, para a proteína 2GB1.	45
Tabela 8. Informações estatísticas do <i>score</i> de alinhamento da proteína 2GB1.	48
Tabela 9. p-valor do teste estatístico de Wilcoxon para o <i>score</i> de alinhamento, comparando o experimento SP com os outros experimentos, para a proteína 2GB1.	48
Tabela 10. Informações estatísticas do RMSD da proteína 2W4P.....	58
Tabela 11. p-valor do teste estatístico de Wilcoxon para o RMSD, comparando o experimento Est-VdW com os experimentos dinâmicos, para a proteína 2W4P.....	58
Tabela 12. Informações estatísticas sobre o <i>score</i> de alinhamento da proteína 2W4P.	59
Tabela 13. p-valor do teste estatístico de Wilcoxon para o <i>score</i> de alinhamento, comparando o experimento SP com os experimentos dinâmicos, para a proteína 2W4P.	59
Tabela 14. Informações estatísticas do RMSD da proteína 1ENH.	65
Tabela 15. p-valor do teste estatístico de Wilcoxon para o RMSD, comparando o experimento Est-VdW com os experimentos dinâmicos, para a proteína 1ENH.....	65
Tabela 16. Informações estatísticas sobre o <i>score</i> de alinhamento da proteína 1ENH.	66
Tabela 17. p-valor do teste estatístico de Wilcoxon para o <i>score</i> de alinhamento, comparando o experimento SP com os experimentos dinâmicos, para a proteína 1ENH.	66

1 Introdução

O Algoritmo Genético (AG) é uma metaheurística populacional que, assim como outros Algoritmos Evolutivos, é capaz de realizar buscas em espaços de busca com alta dimensionalidade [Mitchell, 1998], procurando por ótimos locais e globais. Utilizando um conjunto de possíveis soluções que estão sujeitas a leis inspiradas na Teoria da Evolução, os Algoritmos Genéticos precisam de um contexto – um problema em específico – para que possam ser aplicados. Neste projeto, eles são aplicados em um contexto biológico, mais especificamente na predição tridimensional de estruturas de proteínas.

Proteínas desempenham um importante papel biológico. Possuem várias funções, podendo catalisar processos, prover sustentação mecânica à célula, transportar metais, etc [Lehninger *et al.*, 2004]. Também apresentam grande utilização na indústria farmacêutica de um modo geral [Recktenwald *et al.*, 1992]. Proteínas são formadas por cadeias polipeptídicas, a partir de blocos básicos, conhecidos como aminoácidos; são as macromoléculas mais abundantes nas células. Proteínas também podem ser classificadas hierarquicamente, segundo quatro níveis de complexidade:

- Estrutura primária: a sequência linear de aminoácidos;
- Estrutura secundária: arranjos recorrentes, estabilizados por ligações de hidrogênio. Dentre as estruturas mais conhecidas estão a hélice- α , a folha- β e os *turns*;
- Estrutura terciária: a estrutura tridimensional da proteína. Ela é estabilizada por interações não-locais, como o núcleo hidrofóbico, pontes salinas, ligações de hidrogênio entre outras;
- Estrutura quaternária: estrutura formada por duas ou mais subunidades de proteínas, formando um complexo protéico.

A compreensão das estruturas tridimensionais é importante e fundamental para o entendimento de processos biológicos em geral, além de aplicações mais específicas, como por exemplo, o desenvolvimento de novos fármacos [Liu *et al.*, 2011].

Em uma experiência clássica, Anfinsen demonstrou que uma proteína, em seu ambiente natural, dobra ou vibra em torno de uma única estrutura tridimensional [Anfisen, 1973], aquela que minimiza sua energia livre, de modo que a estrutura tridimensional da proteína pode ser prevista, *a priori*, analisando-se somente a sequência linear de aminoácidos. Soluções computacionais têm sido estudadas para resolver o problema da determinação da estrutura tridimensional de proteínas (ver Seção 2.4).

Pode-se analisar o problema de predição de estruturas de proteínas (PEP) como um problema de otimização, no qual dada uma sequência de aminoácidos, deve-se encontrar a

estrutura da proteína, dentre as muitas possíveis, através da obtenção de mínimos de funções de energia (aqui, também chamadas de função *fitness* ou função de avaliação, etc). Este é um problema complexo, reconhecido como “NP-Completo” [Liang & Wong, 2001].

Metaheurísticas baseadas em busca estocástica, como Algoritmos Evolutivos (AEs) e Recozimento Simulado (*Simulated Annealing*), têm sido utilizadas em diversos problemas conhecidos como “NP-Completo”, como, por exemplo, o problema da PEP. Tem havido um crescente interesse por parte dos pesquisadores em aplicar tais algoritmos na determinação da estrutura tridimensional de proteínas. Por exemplo, trabalhos clássicos de algoritmos genéticos aplicados em proteínas, como [Schulze-Kremer, 1992], [Unger & Moult, 1993]. Em [Mitchell, 1998], a autora utiliza o problema de predição tridimensional de estruturas de proteínas como exemplo de problema para os algoritmos genéticos. Em [Yap & Cosic, 2001], os autores propõem um modelo híbrido Taguchi-AG na predição de proteínas com representação da proteína do tipo hidrofóbico-hidrofílico. Os autores de [Liang & Wong, 2001] mostram que o Monte Carlo evolutivo produz melhores resultados que o Monte Carlo e os AGs para modelos de proteína HP. Em [Jiang *et al.*, 2003], os autores combinam AGs com busca tabu, obtendo resultados melhores que o AG sozinho, para modelos HP. Já em [Cooper *et al.*, 2003], o AG é combinado com o Hill-Climbing porém a utilizando os ângulos diedrais como representação da proteína. [Werhli & Lemke, 2003] propõem um novo operador evolutivo, aplicado na determinação de estruturas do tipo poli-alanina. Os autores de [Liu & Tao, 2010] utilizam a técnica de recozimento simulado para minimizar a energia da proteína. Em [Oliveira, 2011], é proposto que a inserção de conhecimento de proteínas *a priori* no AG traga benefícios ao realizar a predição. Uma técnica de remoção de gêmeos no AG para o problema de PEP é proposta em [Hoque *et al.*, 2011], mostrando que esta estratégia é útil para melhorar as soluções encontradas pelo AG. Outra variante do AG, um AG diferencial é utilizado em [Jana & Sil, 2012], produzindo resultados promissores. Em [Benitez *et al.*, 2012], são combinados o AG e colônias artificiais de abelhas em um modelo HP 3D.

Neste trabalho, a proposta é investigar se a mudança da função de *fitness* durante o processo de otimização do Algoritmo Genético (AG) pode ser benéfica para o PEP. Para isso, os parâmetros do campo de força utilizado para avaliar o *fitness* de cada indivíduo (possível conformação da proteína) são modificados durante o processo de otimização realizado pelo AG. Os termos são aumentados ou diminuídos, por exemplo, de acordo com sua influência na formação de estruturas secundárias e no ajuste fino da estrutura. Este procedimento pode facilitar o processo de evolução realizado pelo AG, como indicado por [Kashtan *et al.*, 2007], que argumenta que, em alguns problemas altamente multimodais, a mudança na função de avaliação pode transformar ótimos locais ou platôs no espaço de busca antes da mudança em regiões com gradiente de *fitness* diferente de zero no novo espaço de busca, o que pode facilitar

a fuga da população de ótimos locais, problema recorrente quando AGs, e outros algoritmos de otimização, são empregados em diversos problemas complexos. Como a função de avaliação será modificada durante o processo de otimização, a predição de estruturas de proteínas torna-se assim um problema de otimização dinâmica [Branke, 2002], sendo que o uso de AGs específicos para tais problemas, como o AG com imigrantes aleatórios e o AG com hipermutação, deverá ser investigado.

1.1 Objetivos

A hipótese a ser investigada neste trabalho é a de que a mudança da função de *fitness* no problema de PEP pode facilitar a fuga de ótimos locais da população do AG, melhorando assim seu desempenho.

Assim, este trabalho tem como principal objetivo estudar a PEP computacionalmente, testando e pesquisando o uso de funções de avaliação que mudam durante o processo de otimização em AGs. Ressalta-se que este trabalho não possui o objetivo de obter um preditor para a predição de estruturas de proteínas. O objetivo é investigar se o uso de funções de avaliação dinâmicas são úteis em AGs aplicados a PEP, procedimento que pode ser utilizado em conjunto com outras técnicas poderosas utilizadas em PEP que vem sendo investigadas por diversos grupos de pesquisadores ao redor do mundo e assim contribuir para o desenvolvimento de preditores.

1.2 Organização do Documento

Este trabalho está organizado da seguinte forma: o Capítulo 2 apresenta uma breve introdução às proteínas; o Capítulo 3 descreve a metodologia empregada. O Capítulo 4 apresenta os resultados obtidos; no Capítulo 5, as conclusões; o Apêndice A possui resultados adicionais que dão suporte aos resultados discutidos no Capítulo 4. Por fim, as referências bibliográficas utilizadas neste documento

2 Proteínas

Este capítulo sobre proteínas é dividido da seguinte maneira: primeiro é feita uma breve apresentação sobre aminoácidos, ligações peptídicas e proteínas em geral; em seguida, são discutidas a modelagem de proteínas e os campos de força. Por fim, o problema de PEP é apresentado.

2.1 Aminoácidos e Proteínas

Aminoácidos são moléculas que possuem uma estrutura característica, consistindo de um átomo de carbono central – chamado C- α – ligado a um grupo amino (NH_2), uma carboxila (COOH), um hidrogênio (H) e uma cadeia com composição química variável (R), conhecida como cadeia lateral [Lehninger *et al.*, 2004]. A Figura 1 ilustra a forma geral de um aminoácido. Os seres vivos utilizam 20 aminoácidos naturais, sendo que a Tabela 1 mostra os seus nomes e abreviações.

Figura 1. Representação esquemática de um aminoácido, em sua estrutura molecular mais geral.

Tabela 1. Aminoácidos, suas abreviações e características das cadeias laterais [Lehninger *et al.*, 2004].

Nome	Abreviação (3 letras)	Abreviação (1 letra)	Polaridade	Carga (pH 7,4)
Alanina	ALA	A	Não polar	Neutra
Cisteína	CYS	C	Polar	Neutra *
Aspartato (Ácido Aspártico)	ASP	D	Polar	Negativa
Glutamato (Ácido Glutâmico)	GLU	E	Polar	Negativa
Fenilalanina	PHE	F	Não polar	Neutra
Glicina	GLY	G	Não polar	Neutra
Histidina	HIS	H	Polar	Neutra
Isoleucina	ILE	I	Não polar	Neutra
Lisina	LYS	K	Polar	Positiva
Leucina	LEU	L	Não polar	Neutra
Metionina	MET	M	Não polar	Neutra
Asparagina	ASN	N	Polar	Neutra
Prolina	PRO	P	Não polar	Neutra
Glutamina	GLN	Q	Polar	Neutra
Arginina	ARG	R	Polar	Positiva
Serina	SER	S	Polar	Neutra
Treonina	THR	T	Polar	Neutra
Valina	VAL	V	Não polar	Neutra
Triptofano	TRP	W	Não polar	Neutra
Tirosina	TYR	Y	Polar	Neutra

***Quando envolvido em ponte de enxofre.**

2.2 Proteínas

Uma ligação, conhecida como ligação peptídica, é formada quando dois aminoácidos são unidos covalentemente – o grupo amino de um aminoácido se liga ao grupo carboxila do outro [Lehninger *et al.*, 2004]. Vários aminoácidos podem fazer a ligação peptídica sequencialmente, formando os polipeptídeos, também conhecidos como proteínas. A função de uma proteína está diretamente ligada com a sua estrutura tridimensional. A alteração da estrutura primária de uma proteína pode acarretar na alteração da estrutura tridimensional e consequentemente em sua função. Mesmo a substituição de um único aminoácido por outro (mutação) pode afetar drasticamente a função biológica.

As proteínas podem ser classificadas hierarquicamente em quatro níveis. Cada um desses níveis é descrito abaixo.

A estrutura primária corresponde à sequência linear de aminoácidos. As diferentes características das cadeias laterais dos aminoácidos (Tabela 1) faz com que haja uma grande variedade nas funções desempenhadas pelas proteínas [Leach, 2001], tais como catálise de processos, sustentação mecânica, proteção, armazenamento e movimentação.

As estruturas secundárias são arranjos estáveis que as sequências de aminoácidos podem formar [Lehninger *et al.*, 2004]. A disposição dos átomos dos aminoácidos no espaço – e consequentemente os ângulos formados entre eles – ditam o tipo de arranjo formado. Dentre as estruturas secundárias, as duas mais comuns são as hélices- α e as folhas- β . As hélices- α são estruturas mais comuns pelo fato de otimizarem o uso das ligações internas de hidrogênio [Lehninger *et al.*, 2004]. A Figura 2 mostra um exemplo de uma estrutura com uma hélice- α .

Figura 2. Exemplo de uma proteína com a estrutura secundária do tipo hélice- α . Os dados estruturais desta proteína e das demais figuras foram obtidas do Protein Data Bank [Berman *et al.*, 2000]. Nesta figura, o PDB ID é 2LCM

Já as folhas- β são estruturas alinhadas, rígidas e achatadas, formadas por fitas- β e ligações de hidrogênio, podendo ser classificadas como paralela ou antiparalela, dependendo das orientações dos aminoácidos. As ligações de hidrogênio são formadas entre os aminoácidos adjacentes [Lehninger *et al.*, 2004]. A Figura 3 mostra uma estrutura com folha- β .

Figura 3. Exemplo de uma proteína com a estrutura secundária do tipo folha- β . Nesta figura, o PDB ID é 4E83

A estrutura terciária é equivalente à estrutura tridimensional da proteína. É influenciada por interações de longa distância da sequência de aminoácidos [Lehninger *et al.*, 2004], como por exemplo, interação entre aminoácidos de estruturas secundárias distintas. A estrutura terciária é estabilizada por interações mais fracas, como as interações hidrofóbicas e específicas como as ligações de hidrogênio, fazendo com que a estrutura tridimensional flutue ao longo do tempo [Lodish *et al.*, 2003].

Por fim, a estrutura quaternária é formada a partir de duas ou mais subunidades formando um complexo, importante para o regulamento da função da proteína. Podem ser compostas de subunidades idênticas (Figura 4) ou diferentes [Lodish *et al.*, 2003].

Figura 4. Hemaglutinina (PDB ID: 1QU1), estrutura quaternária com três subunidades idênticas.

2.3 Modelagem de Proteínas e Campos de Força

Para descrever o processo de enovelamento, realizar predições das estruturas de proteínas ou efetuar cálculos em qualquer outro fenômeno envolvendo proteínas, é necessário utilizar um modelo que represente a proteína. Dessa forma, foram propostos diversos tipos de representação das proteínas, cada qual com suas características e diferentes níveis de simplificações [Tozzini, 2005].

As expressões mais detalhadas aparecem nos modelos atomísticos, que assumem a localização exata de todos os átomos da proteína, fazendo com que os cálculos sejam bastante custosos pelo elevado número de interações de pares necessárias a serem determinadas. Existem vários campos de força utilizados em problemas de PEP, definidos através do uso de funções matemáticas, as quais fazem uso das propriedades dos átomos (modelos atomísticos) [Ponder & Case, 2003], [Bordner, 2012], e cada um deles possui variantes na parametrização [Hornak *et al.*, 2006], de forma que há muitas escolhas e combinações, o que pode significar uma

dificuldade para se encontrar o melhor campo de força com a melhor parametrização para se ajustar à proteína [Hao & Scheraga, 1999].

Já os modelos *coarse-grained* (CG), utilizados neste projeto, visam diminuir os graus de liberdade da representação da proteína, com o intuito de acelerar os cálculos das energias potenciais utilizados em problemas de PEP [Tozzini, 2005]. Modelos CG reduzem os graus de liberdade ao assumir que todos os átomos de um aminoácido são representados por um (ou poucos) pontos (*beads*). Os *beads* assumem então alguma propriedade do aminoácido, como atração-repulsão [Monticelli *et al.*, 2008].

No cálculo da energia potencial dos modelos utilizados em PEP, tanto os modelos atomísticos quanto os modelos CG possuem termos que são comuns, que em geral representam potenciais que são bastante conhecidos. Dentre eles, podem ser citados:

- Termo do ângulo de ligação: obtido da interação entre três átomos (ou *beads*). Este termo é calculado de maneira similar a Lei de Hook (lei das molas). Ele indica o desvio dos átomos em relação a uma geometria ideal, imitando uma ligação química. A Figura 10 mostra uma representação esquemática do ângulo de ligação
- Termo do ângulo de torção: obtido da interação entre quatro átomos (ou *beads*). O ângulo de torção consiste da localização dos planos formados pelos três primeiros átomos e dos três últimos átomos. Este termo tem um papel importante na estruturação das proteínas e outras moléculas com hidrocarbonetos. A Figura 12 mostra uma representação esquemática do ângulo torção.
- Termo de Van der Waals: baseado geralmente no potencial de Lennard-Jones, possui uma componente atrativa a médias distâncias e outra componente repulsiva, a curtas distâncias. O termo inclui forças entre dois dipolos permanentes, dipolo permanente e dipolo induzido e dois dipolos induzidos e é importante para que os átomos fiquem mais próximos, até um determinado limite, onde eles se chocam. A Figura 14 mostra um gráfico típico do termo de Van der Waals.
- Termo eletrostático: baseado na Lei de Coulomb, pode ser atrativa ou repulsiva, dependendo das suas cargas dos átomos. Indica interações de longo alcance.

Neste projeto, utilizamos o campo de força proposto em [Yap *et al.*, 2008] como função de *fitness* dos AGs. As equações utilizadas neste projeto, bem como o detalhamento dos termos utilizados no campo de força e a parametrização do modelo estão descritos na Seção 3.3.3. Este campo de força foi escolhido para reduzir o custo computacional (em relação aos campos de força *full-atom*) e foi calibrado para chegar na solução.

2.4 Predição de Estruturas de Tridimensionais Proteínas

A determinação de estruturas tridimensionais de proteínas pode ser feita experimentalmente, através, por exemplo, dos métodos de cristalografia de Raios-X e Ressonância Nuclear Magnética [Baxevanis & Ouellette, 2001] [White & Heijne, 2004]. Entretanto essas técnicas têm certas limitações, como alto custo, necessidade de se obter um cristal – o que pode gerar artefatos ao empacotar a proteína – no caso da cristalografia ou limitações em relação ao tamanho da proteína no caso da ressonância [DePristo *et al.*, 2004] [Montelione *et al.*, 2000].

Computacionalmente, as diversas técnicas propostas podem ser classificadas em:

- Modelagem *ab initio*: nesta técnica, a modelagem assume apenas o conhecimento da estrutura primária para encontrar a estrutura tridimensional. Diversas técnicas computacionais têm sido utilizadas nesta abordagem, como Recozimento Simulado [Liu & Beveridge, 2002], Algoritmos Genéticos [Cai-Yun *et al.*, 2009], Cadeias de Markov [Hubbard & Park, 1995], [Karplus *et al.*, 1999].
- Modelagem por Homologia: baseia-se, principalmente, na similaridade estrutural, utilizando estruturas já determinadas experimentalmente de proteínas homólogas. Dentre os programas que trabalham com homologia, podem ser citados HHpred [Söding *et al.*, 2005], SWISS-MODEL [Arnold *et al.*, 2005], Modeller [Eswar *et al.*, 2006] e Rosetta [Leaver-Fay *et al.*, 2011]
- Modelagem por *Threading*: a estrutura é montada através de uma coleção de proteínas-modelo. É feita uma análise quantitativa de quão bem a sequência se encaixa no modelo. Assume-se que a modelagem reflete as propriedades físico-químicas dos aminoácidos e que as estruturas tridimensionais de proteínas possuem características semi-quantitativas previsíveis. Como programas relevantes, podem ser citados o I-TASSER [Zhang, 2008] e QUARK [Xu & Zhang, 2012].

3 Metodologia

3.1 Algoritmos Genéticos

Algoritmos Genéticos (AGs) são metaheurísticas populacionais utilizadas em processos de busca e otimização. No AG padrão, um indivíduo representa uma possível solução para o problema. Um conjunto de indivíduos, conhecido como população, é então sujeito as regras de seleção e transformação inspiradas em mecanismos da evolução natural. Uma solução $\mathbf{x}_i$, representada pelo cromossomo (indivíduo) $i=1,...,N$ (Figura 5), é avaliada segundo uma função de avaliação, ou *fitness*, $f(\mathbf{x}_i)$. Os cromossomos estão sujeitos, através de diferentes épocas, a operadores genéticos de reprodução e seleção.

Figura 5. Representação da população do AG.

Algoritmo: AG padrão**Início**

inicialize a população

avale a população inicial

repita

se critério de convergência for satisfeito

interrompa

fim se

selecione indivíduos da população corrente

aplique mutação e cruzamento nos indivíduos selecionados para gerar uma nova população

avale os indivíduos da nova população

fim repita**Fim**

Figura 6. Pseudo-código do AG padrão.

Os operadores de reprodução mais utilizados são *crossover* e mutação. No primeiro, dois indivíduos da população são escolhidos pelo operador de seleção e parte de seus cromossomos são trocados, gerando duas novas soluções. Já o operador de mutação altera aleatoriamente alguns dos genes dos indivíduos, de acordo com uma regra pré-definida. Vale ressaltar que AGs são algoritmos não-deterministas, sendo que os operadores de transformação são aplicados com probabilidades p_c e p_m , conhecidas como taxas de *crossover* e mutação, respectivamente.

Os operadores de seleção são aplicados na população para selecionar os indivíduos para serem reproduzidos. Dentre os operadores mais utilizados, podemos citar o elitismo, que sempre seleciona o indivíduo com melhor *fitness*, o método da roleta, no qual a seleção de indivíduos é proporcional ao seu *fitness*, e o torneio, no qual k indivíduos são escolhidos aleatoriamente e o melhor dentre os k segue para a próxima geração. A Figura 6 apresenta um pseudo-código do AG padrão.

As aplicações de algoritmos genéticos são diversas. Os AGs podem ser usados, entre outros, para design de asas de aeronaves [Zhao & Wang, 2010], predição de finanças e ações [Rafiei *et al.*, 2011] ou auxílio a diagnósticos médicos [Park *et al.*, 2011]. Além de aplicações, diversas novas técnicas têm sido desenvolvidas, visando a melhoria do desempenho [Mitchell, 1998].

3.2 Problemas de Otimização Dinâmica

Muitos problemas do mundo real mudam com o tempo. Fatores como falhas, mudanças no ambiente ou objetivos, variáveis humanas, entre outros, podem fazer com que o problema seja considerado dinâmico. Em problemas de otimização dinâmica (*dynamic optimization problems* - DOPs), as mesmas soluções podem ser consideradas boas ou ruins, dependendo da situação atual do problema e, desta forma, a otimização nesses cenários dinâmicos pode ser desafiadora [Cruz *et al.*, 2011].

Computacionalmente, DOPs são problemas em que a função de avaliação ou restrições mudam com o tempo. Uma das maneiras mais simples de se resolver esses problemas é considerar que cada mudança gere um novo problema de otimização, assim, podemos reiniciar o processo de busca como se fosse um novo. Esta estratégia, conhecida como *Restart* [Branke, 2002], é útil em problemas no qual o novo espaço de busca (após a mudança) não guarda semelhanças com o velho espaço de busca. Entretanto, se os espaços de busca antes e após a mudança possuem similaridades, o conhecimento adquirido durante o processo de otimização antes da mudança pode ser utilizado no novo espaço de busca, aumentando a eficiência do processo. Contudo, muitas vezes ocorre convergência das soluções para regiões promissoras no espaço antes da mudança no AG padrão, o que torna mais difícil a otimização após a mudança. Desta forma, novas estratégias como a dos imigrantes aleatórios e da hipermutação [Cobb & Grefenstette, 1993], tem sido desenvolvidas para problemas que mudam com o tempo. Outras técnicas tem sido utilizadas para resolver DOPs [Cruz *et al.*, 2011], como estratégias cooperativas [Pelta *et al.*, 2009], sistemas imunológicos artificiais [Trojanowski & Wierzbachon, 2009], busca por difusão estocástica [Meyer *et al.*, 2006] e otimização por colônia de formigas [Guntsch *et al.*, 2001].

Apesar de DOPs serem computacionalmente mais custosos, a mudança no espaço de busca pode ser benéfica. Os autores em [Kashtan *et al.*, 2007] argumentam que a alteração dos objetivos ao longo do tempo pode acelerar o processo de evolução, ou seja, encontrar boas soluções mais rapidamente, visto que a mudança pode alterar as posições dos ótimos e mudar o gradiente da superfície de *fitness*. A aceleração pode ser proporcional a complexidade do problema [Kashtan *et al.*, 2007], uma vez que problemas mais complexos tendem a ter mais ótimos locais e a mudança de ambiente pode ajudar os algoritmos a procurarem por mais áreas do espaço. As técnicas propostas em [Kashtan *et al.*, 2007] para efetuar a troca do espaço de busca foram o *Modular Varying Goal* (MVG), onde as mudanças são determinadas *a priori* e ocorrem na forma de um ciclo, e o *Random Varying Goal* (RVG), onde a escolha da mudança é feita de forma aleatória. Aplicando essas técnicas, o problema de otimização se tornará dinâmico, visto que cada vez que a função de *fitness* é alterada, os ótimos se movem e um novo processo se inicia.

A Figura 7 mostra uma esquematização de como esta mudança de ambiente pode ser benéfica. Nela, ocorre uma maximização da função de avaliação e a solução inicial está em 1. O algoritmo encontra um máximo local em 2. Em seguida há a troca do espaço de busca e a solução que estava em 2 fica na posição 3 no segundo espaço de busca. Neste segundo espaço, a solução converge para o máximo local 4. O algoritmo volta ao primeiro espaço de busca, na posição 5 e finalmente encontra o máximo global em 6. O princípio básico por trás da estratégia é a alteração do gradiente da superfície de *fitness* de tal forma a permitir uma fuga mais rápida de regiões com gradiente igual a zero (por exemplo, os ótimos locais).

Figura 7. Esquematização de uma busca, em uma dimensão, com a troca do espaço de busca. $f(x)$ é a função de avaliação.

A ideia neste projeto é induzir um DOP no problema de predição de estruturas de proteínas. Serão incorporados elementos inspirados no *folding* de proteínas, uma vez que o *folding* envolve diversos processos dinâmicos [Buchner *et al.*, 2010]. Os autores em [Voet & Voet, 2003] propõem um dos possíveis caminhos de folding: inicialmente, há uma rápida formação das estruturas secundárias como hélices- α e folhas- β , chamada “fase explosiva”. Como as proteínas possuem um núcleo hidrofóbico, ocorre então a segunda etapa, conhecida como “colapso hidrofóbico”, quando a estrutura diminui seu raio e se torna mais rígida, sendo natural admitir que a contribuição de cada interação seja mais decisiva em determinada fase do processo.

3.3 Detalhamento da Implementação

Nesta seção, serão apresentados alguns detalhes de implementação do AG para PEP, como codificação, função de avaliação e operadores genéticos.

3.3.1 Codificação

A codificação dos cromossomos do AG será a mesma empregada em [Ishivatari, 2009] e [Oliveira, 2011]. Cada gene representa os ângulos diedrais ϕ e ψ do C- α do aminoácido correspondente (ver Figura 8), de modo que o tamanho do cromossomo seja igual ao número de resíduos da proteína. Os ângulos diedrais, dos indivíduos aleatórios e aqueles gerados por dois dos três tipos de mutação, são obtidos a partir de observações experimentais [Gopalakrishnan *et al.*, 2007], formando uma base de dados que é dividida por aminoácido e estrutura secundária e ordenada pelo ângulo ϕ . A Figura 8 mostra uma representação da codificação utilizada e a Figura 9 mostra uma representação esquemática dos ângulos diedrais.

Figura 8. Representação da base de dados e codificação dos indivíduos, adaptado de [Oliveira, 2011].

Figura 9. Representação dos ângulos de torção ψ e ϕ . R é a cadeia lateral do aminoácido.

Reproduzido com autorização, [Lehninger *et al.*, 2004] ©.

3.3.2 Operadores Genéticos

Com relação aos operadores de reprodução, utilizou-se o *crossover* de dois pontos e a mutação. A mutação possui três variantes, todas com a mesma chance de ocorrer:

- Mutação com auxílio da base de dados: nesta mutação, utiliza-se o último índice para a base de dados conhecido e altera-se o índice (e, conseqüentemente, os valores de ϕ e ψ do gene) para uma posição aleatória num intervalo de $[-5, 5]$ posições do índice atual. Como a base está ordenada, espera-se que este tipo de mutação produza pequenas mudanças na estrutura da proteína.
- Mutação com auxílio de toda a base de dados: nesta, sorteia-se um novo índice no intervalo $[1, N]$, onde N é o tamanho da base de dados referente ao resíduo analisado. O intuito desta mutação é tentar promover mudanças drásticas nos ângulos dos aminoácidos.
- Mutação real: nesta mutação não há acesso à base de dados. Depois de selecionado o gene que sofrerá mutação, sorteia-se em qual ângulo ocorrerá a mudança e aplica-se uma mutação real uniforme no intervalo de $[-5^\circ, 5^\circ]$ no ângulo escolhido, preservando o valor do índice para a base de dados anterior. O objetivo desta mutação é testar outros ângulos possíveis que não estão na base de dados. Apesar do intervalo de mudança ser pequeno, há uma possibilidade de esta mutação resultar na perda de estruturas secundárias já existentes.

Os operadores de seleção utilizados foram elitismo para os dois melhores indivíduos e torneio de três participantes. O método da roleta não foi utilizado, pois em problemas de predição de estruturas de proteínas, algumas soluções do AG são muito ruins – por exemplo, quando átomos se chocam – fazendo com que o *fitness* seja muito alto, resultando em uma probabilidade próxima de zero de a solução ser escolhida.

3.3.3 Função de Avaliação HG

Algoritmos genéticos também precisam de uma função de avaliação, que é dependente do problema estudado. Neste projeto, a função de avaliação é baseada em [Yap *et al.*, 2008] – aqui designada como função de avaliação HG e propõe um modelo minimalista (*coarse-grained*), assumindo que todo o aminoácido seja representado pelo seu C- α (*bead*). Outras informações e detalhes são incorporados durante a simplificação do campo de força, que é descrito a seguir.

$$E = E_{Lig} + E_{Diedral} + E_{vdW} + E_{HB} \quad \text{Eq. 1}$$

Sendo,

$$E_{Lig} = \sum_{\text{angulos}} \frac{1}{2} k_{\theta} (\theta - \theta_0)^2 \quad \text{Eq. 2}$$

$$E_{Diedral} = \sum_{\text{diedrais}} \left[\begin{aligned} &A[1 + \cos(\phi + \phi_0)] + B[1 - \cos(\phi + \phi_0)] + \\ &+ C[1 + 3 \cos(\phi - \phi_0)] + D[1 + \cos(\phi + \phi_0 + \frac{\pi}{4})] \end{aligned} \right] \quad \text{Eq. 3}$$

$$E_{VdW} = \sum_{i,j \geq l+3} 4\epsilon_H S_1 \left[\left(\frac{\sigma}{r_{ij}} \right)^{12} - S_2 \left(\frac{\sigma}{r_{ij}} \right)^6 \right] \quad \text{Eq. 4}$$

$$E_{HB} = \sum_{ligH} \epsilon_{HB} * \exp \left[-\frac{(r_{ij} - r_{HB})^2}{\sigma_{HBdist}^2} \right] * \exp \left[|\mathbf{t}_{HB,i} \cdot \mathbf{r}_{ij}| - \frac{1}{\sigma_{HB}^2} \right] \quad \text{Eq. 5}$$

$$* \exp \left[|\mathbf{t}_{HB,j} \cdot \mathbf{r}_{ij}| - \frac{1}{\sigma_{HB}^2} \right]$$

O campo de força representado na Eq. 1 utilizado possui quatro termos, correspondentes a energia do ângulo de ligação, energia do ângulo de torção (que têm papel importante na formação de estruturas secundárias), energia de Van der Waals e energia de ligação de hidrogênio, os quais são importantes para tornar as estruturas secundárias mais próximas e rígidas. O comportamento das Eq. 2, 3 e 4 no campo de força são descritos na seção seguinte.

Na Eq. 1, θ é o ângulo de ligação definido por três C- α consecutivos, Φ é o ângulo diedral definido por quatro *beads* de C- α consecutivos e r_{ij} é a distância de separação entre os *beads* i e j . ϵ_H é a magnitude do potencial de interação de Van der Waals, θ_0 é o ângulo de ligação da estrutura nativa, A , B , C , D e Φ_0 são parâmetros que dependem do tipo diedral, k_{θ} é a constante de força do potencial harmônico, σ é a soma do raio de van der Waals e S_1 e S_2 são parâmetros que se referem a atração e repulsão dos *beads*. Esses parâmetros são definidos em [Yap *et al.*, 2008] baseado em informações estatísticas de proteínas, sendo algumas dependentes da interação atrativa-repulsiva dos aminoácidos, ou o tipo de estrutura secundária do aminoácido.

Os parâmetros da Eq. 5 são os seguintes: r_{ij} e $\mathbf{r}_{ij}$ são, respectivamente, a distância e o vetor unitário entre os *beads* i e j . A primeira exponencial é uma função Gaussiana centrada no hidrogênio ideal [Yap *et al.*, 2008], de distância r_{HB} . As duas outras exponenciais são termos direcionais de suavização e $\mathbf{t}_{HB,i}$ é o vetor unitário normal ao plano descrito pelos *beads* ($i-1$, i , $i+1$).

A função de *fitness* utilizada no AG neste trabalho é descrita pela Eq. 6, incorporando a dinamicidade do problema. Na Eq. 6, $w_{vdW}(t)$ e $w_{HB}(t)$ são respectivamente os pesos dos termos E_{vdW} e E_{HB} na geração t do AG

$$f(\mathbf{x}) = E_{Lig} + E_{Diedral} + w_{vdW}(t)E_{vdW} + w_{HB}(t)E_{HB} \quad \text{Eq. 6}$$

A justificativa de apenas os termos de Van der Waals e ligação de hidrogênio possuírem pesos é que o objetivo é tentar simular as etapas de “fase explosiva” e “colapso hidrofóbico” do *folding* de proteínas [Voet & Voet, 2003]. Os padrões de mudança para os pesos (Eq. 6), que neste projeto, são valores reais maiores que 1,0, estão descritos na Seção 3.4. É interessante notar que a análise de tais pesos pode ajudar na elucidação de quais contribuições são mais significativas em diferentes fases do dobramento protéico.

É importante ressaltar também que as informações relativas à estrutura secundária de cada resíduo das proteínas utilizadas nos experimentos são extraídas da base de dados do PDB. Entretanto mesmo em uma situação mais realista de predição de proteínas onde não se conhecem as estruturas secundárias é possível utilizar um preditor de estrutura secundária. Nos experimentos realizados em [Oliveira, 2011] com o mesmo campo de força e codificação do AG, não houve diferença significativa ao usar as informações do PDB ou aqueles obtidos pelo preditor de estrutura secundária.

3.3.4 Detalhamento dos termos do campo de força HG

Nesta seção, os termos do campo de força HG serão detalhados, focando no impacto de cada termo no resultado final.

Termo do ângulo de ligação

O ângulo de ligação é obtido da interação de três átomos (ou *beads*) subsequentes, como pode ser visto na Figura 10. Alterações no ângulo de ligação podem afetar a conformação da estrutura, bem como a formação de estruturas secundárias. Mudanças nos ângulos são mais fáceis de serem aplicadas quando comparado com mudanças no comprimento de ligação, uma vez que as constantes de força são proporcionalmente menores. O gráfico da energia (Eq. 2) é mostrado na Figura 11.

Figura 10. Representação do ângulo de ligação formado por três átomos, a , b e c . θ é o ângulo de ligação.

Figura 11. Energia do ângulo de ligação (Eq. 2). A unidade da energia é uma unidade reduzida, ϵ_H .

Termo do ângulo de torção

O ângulo diedral – ou ângulo de torção – é definido por quatro pontos no espaço. Com esses quatro pontos são formados dois planos, um que utiliza os três primeiros pontos e outro que utiliza os três últimos. O ângulo diedral é dado pelos vetores normais aos dois planos, como é exemplificado na Figura 12.

Potenciais de torção podem ser expressos por uma expansão de série de cossenos (Eq. 3) [Leach, 2001] e, apesar de serem menos flexíveis que os ângulos de ligação, os ângulos diedrais são fundamentais para a formação de estruturas secundárias. A Figura 13 mostra o gráfico da energia do ângulo diedral, obtida da Eq. 3.

Figura 12. Representação do ângulo diedral (ϕ), formado por quatro átomos consecutivos.

Figura 13. Energia do ângulo diedral para duas configurações dos parâmetros. A unidade da energia é uma unidade reduzida, ϵ_H .

Termo de van der Waals

O termo van der Waals é uma interação relativamente fraca, formada por dipolos permanentes e/ou induzidos, sendo geralmente modelada pelo potencial de Lennard-Jones 6-12. É responsável pela aproximação ou repulsão dos átomos dependendo se eles estiverem distantes ou muito próximos respectivamente. A Figura 14 mostra o gráfico da energia de van der Waals (Eq. 4).

Figura 14. Energia de van der Waals para uma configuração de parâmetros. A unidade da energia é uma unidade reduzida, ϵ_H .

Termo da ligação de hidrogênio

O termo da ligação ou ponte de hidrogênio é usado para modelar interação entre o hidrogênio doador e o heteroátomo receptor. O termo da ligação de hidrogênio é utilizado para aumentar a acurácia em que a geometria de sistemas de ligação de hidrogênio é predita [Leach., 2001]. Este termo é baseado no modelo de água Mercedes Benz [Marcus & Ben-Naim, 1985].

O potencial de interação de hidrogênio é avaliado para todos os *beads* i - j tais que os pares sejam capazes de formar ligações de hidrogênio. Dependendo da sua propensão diedral, cada *bead* assume um dos três possíveis tipos: hélice (A), folha (B) e nenhum (C). Para A, o potencial de hidrogênio é avaliado se o terceiro vizinho de A também for A. Já para B, o potencial é avaliado entre ele e todos os outros *beads* B que estejam a uma distância de três unidades de comprimento [Yap *et al.*, 2008].

3.3.5 Função de avaliação “Onuchic”

Esta função de avaliação é baseada em um campo de força proposto pelo grupo liderado por Onuchic [Clementi *et al.*, 2000], [Lammert *et al.*, 2009]. Similarmente à função de avaliação HG descrita acima, este campo de força utiliza um modelo C- α *coarse-grained* para representar a proteína. Para calcular o valor da energia – cuja unidade é equivalente a kcal/mol – é necessário fornecer tanto a estrutura a ser avaliada quanto a estrutura desejada ou nativa. Desta forma, já que se necessita da estrutura nativa para o cálculo da aptidão da solução, esta função não é utilizada em PEP, sendo apropriada para o estudo de *folding* de proteínas. Neste trabalho,

esta função é utilizada para validar o método proposto em uma função diferente daquela apresentada em [Yap *et al.*, 2008]. O campo de força é dado por:

$$E(\Gamma, \Gamma_0) = E_{lig} + E_{ang} + E_{diedral} + E_{NL} \quad \text{Eq. 7}$$

Sendo,

$$\begin{aligned} E(\Gamma, \Gamma_0) = & \sum_{\text{ligações}} K_r (r - r_0)^2 \\ & + \sum_{\text{ângulos}} K_\theta (\theta - \theta_0) \\ & + \sum_{\text{diedral}} K_\phi^{(n)} [1 + \cos(n \times (\phi - \phi_0))] \\ & + \sum_{i < j-3} \left\{ \varepsilon(i, j) \left[5 \left(\frac{\sigma_{i,j}}{r_{i,j}} \right)^{12} - 6 \left(\frac{\sigma_{i,j}}{r_{i,j}} \right)^{10} \right] + \varepsilon_2(i, j) \left(\frac{\sigma_{i,j}}{r_{i,j}} \right)^{12} \right\} \end{aligned} \quad \text{Eq. 8}$$

Tal que a variável sem índice refere-se a estrutura avaliada enquanto a variável com índice 0 é referente a estrutura nativa. Dessa forma:

- Γ e Γ_0 são as estruturas tridimensionais das proteína;
- r e r_0 são as distâncias entre dois resíduos subsequentes;
- θ e θ_0 são os ângulos formados por três resíduos subsequentes;
- ϕ e ϕ_0 são os ângulos diedrais formados por quatro resíduos subsequentes;
- n assume dois valores, $n=1$ e $n=3$. O somatório então é calculado para os dois valores de n para cada termo do somatório.
- $\varepsilon(i, j) = 1$ e $\varepsilon_2(i, j) = 0$ se i - j faz contato;
- $\varepsilon(i, j) = 0$ e $\varepsilon_2(i, j) = 1$ se i - j não faz contato;
- Um par i - j faz contato se a distância entre os C- α for até 20% da distância nativa
- $\sigma_{i,j} = 4 \text{ \AA}$ para interações repulsivas e $\sigma_{i,j}$ é a distância de i até j no estado nativo da proteína;
- $K_r = 100\varepsilon$; $K_\theta = 20\varepsilon$; $K_\phi^{(1)} = \varepsilon$ e $K_\phi^{(3)} = 0.5\varepsilon$

A função de avaliação para o AG neste caso é:

$$f(\mathbf{x}) = E_{lig} + E_{ang} + E_{diedral} + w_{NL} E_{NL} \quad \text{Eq. 9}$$

Sendo w_{NL} o peso referente ao termo da energia dos não-ligantes.

3.4 Algoritmos e métodos de otimização dinâmica propostos

Diversos experimentos foram realizados para verificar se a mudança do espaço de busca é benéfica para o problema de PEP. Com base nas conclusões em [Kashtan *et al.*, 2007], foram propostos os seguintes experimentos dinâmicos, utilizando o AG Padrão:

- *Modular Varying Goal* (MVG): os pesos da Eq. 6 são alterados de maneira determinista;
- *Random Varying Goal 1* (RVG1): os pesos da Eq. 6 são alterados para valores aleatórios. Neste experimento, apenas o peso da Energia de Van der Waals é modificado;
- *Random Varying Goal 2* (RVG 2): similar ao RVG1, exceto que o peso da Energia da Ligação de Hidrogênio também pode ser alterada.

Os pesos utilizados na Eq. 6 para cada enfoque são mostrados na Seção 3.6. Com o objetivo de comparação, foram realizados experimentos em que a função de *fitness* não muda:

- Sem pesos (SP): nenhuma alteração nos pesos;
- Estático Van der Waals (Est-VdW): apenas o peso para a Energia de Van der Waals na Eq. 6 possui um valor maior que 1 e este valor é fixo durante toda a execução do experimento. A justificativa deste experimento é para confirmar se o melhor desempenho é devido a presença de dinamicidade ou ao aumento da contribuição de van der Waals.

Também foram propostos experimentos com AGs desenvolvidos para DOP, especialmente algoritmos que inserem diversidade na população. Os algoritmos utilizados foram os AGs com:

- Hipermutação: A cada troca do espaço de busca, a taxa de mutação é aumentada em três vezes, por um tempo equivalente a 2% do número de gerações que o AG fica em um determinado espaço de busca;
- Imigrantes Aleatórios: A cada geração no AG, uma quantidade pré-determinada de indivíduos aleatórios (neste projeto, 6%) é inserida na população;
- Restart Parcial: A cada troca do espaço de busca, exceto pelos dois melhores indivíduos, toda a população é substituída por uma nova, aleatória.
- Restart Full: Similar ao anterior, com a diferença que neste experimento não há exceções: até os melhores indivíduos são substituídos.

- Relaxamento de Restrições: Diminui-se a contribuição de um termo (neste caso o termo de van der Waals) para que soluções consideradas ruins (como átomos se chocando) possam ser aceitas. Ao trocar de espaço (por exemplo, voltando o peso de van der Waals para 1), as soluções ruins se reajustam e eventualmente se tornam boas soluções.
- Memória 1: A cada troca de espaço de busca, o melhor indivíduo de um ciclo passado correspondente, por exemplo, o indivíduo do ciclo SP é reinserido na população no próximo ciclo SP.
- Memória População: Uma subpopulação é salva a cada troca de espaço de busca. A subpopulação é formada pelos dois melhores indivíduos de cada ciclo e logo após ocorrer a mudança do espaço de busca, dois indivíduos são selecionados aleatoriamente desta subpopulação. Estes dois indivíduos podem realizar *crossover* em pontos específicos: nas extremidades das estruturas secundárias.
- Memória População Antes/Depois: Similar ao anterior, exceto que os indivíduos são reinseridos na população antes e depois da troca do espaço de busca. A ideia neste experimento é aproveitar-se do fato de que as trocas são artificiais, ou seja, ocorrem em intervalos conhecidos. Inserir indivíduos algumas gerações antes da troca pode ajudar o algoritmo a se preparar para a mudança do espaço de busca.

3.5 Proteínas escolhidas

Nos experimentos propostos foram utilizadas três proteínas. A proteína G (PDB ID: 2GB1), uma proteína variante da fosfatase (PDB ID: 2W4P) e uma proteína DNA ligante (PDB ID: 1ENH).

A proteína G é uma proteína ligante de imunoglobulina do grupo G Streptococcus, determinada experimentalmente por NMR. É uma conformação altamente estável, devido a rede de ligações de hidrogênio e um núcleo hidrofóbico compacto [Gronenborn *et al.*, 1991]. Ela foi escolhida por possuir estruturas secundárias Hélice- α e Folha- β , já ter sido utilizada em trabalhos relacionados anteriores [Oliveira, 2011] e [Yap *et al.*, 2008] e por ser razoavelmente pequena (56 resíduos) para que as simulações sejam concluídas em um tempo aceitável. A partir daqui, esta proteína será chamada pelo seu ID da base de dados PDB, 2GB1.

A proteína de ID 2W4P é uma variante de acilfosfatase presente nos seres humanos, com uma alta rigidez em sua conformação, fazendo com que suas atividades enzimáticas sejam reduzidas em temperaturas baixas [Lam *et al.*, 2011]. Ela foi escolhida por também possuir tanto estruturas Hélice- α quanto Folha- β . Por ser maior (94 resíduos), quando comparada com a 2GB1, a proposta é verificar o comportamento dos algoritmos na predição de estruturas mais

complexas. A partir daqui, esta proteína será chamada pelo seu ID da base de dados PDB, 2W4P.

A proteína de ID 1ENH, é uma proteína do tipo DNA Ligante da espécie *Drosophila melanogaster*, contém uma extensa rede de pontes salinas e interações dos átomos de oxigênio das carbonilas, garantindo estabilidade à conformação nativa [Clarke *et al.*, 1994]. Possui 54 resíduos, aproximadamente o mesmo número da 2GB1 e foi escolhida pelo fato de que sua estrutura secundária ser composta apenas de Hélices- α . As Hélices- α são reconhecidamente mais fáceis de serem preditas, de modo que se deseja verificar se os resultados obtidos são melhores que para a 2GB1. A partir daqui, esta proteína será chamada pelo seu ID da base de dados PDB, 1ENH.

3.6 Outros parâmetros e resumo dos experimentos

Dentre os parâmetros do AG, as taxas de mutação e de *crossover* foram, respectivamente, 15% e 60%, valores determinados em experimentos anteriores [Oliveira, 2011]. Quanto ao número de gerações, foram realizados alguns testes e verificou-se que a partir de 50 gerações por resíduo os valores de *fitness* geralmente apresentaram pouca variação. Dessa forma, a duração do ciclo para a troca do espaço de busca τ , foi de 2500 gerações para as proteínas 1ENH e 2GB1 e 5000 para a proteína 2W4P. Os experimentos foram executados por 23 ciclos para as duas proteínas menores e 15 ciclos para a 2W4P, totalizando respectivamente 57500 e 75000 gerações para cada experimento, de forma que proteínas maiores são processadas durante mais tempo. Já que a população inicial é aleatória, cada experimento foi executado 10 vezes com diferentes sementes aleatórias.

Os valores dos pesos estão no intervalo $[1, 2N]$, onde N é o número de resíduos da proteína. Esses valores foram escolhidos com o intuito de mudar drasticamente o espaço de busca. As Tabelas 2, 3 e 4 resumizam os experimentos realizados com as proteínas 2GB1, 2W4P e 1ENH, respectivamente.

Da Tabela 2, verifica-se que o número de experimentos realizados para a proteína 2GB1 é maior que para as outras proteínas. Isso ocorre porque a proteína 2GB1 foi a primeira a ser testada, além do fato de que ela foi utilizada para investigação do processo de enovelamento protéico com este campo de força [Yap *et al.*, 2008]. Para as outras proteínas, foram testados apenas os algoritmos que obtiveram os melhores resultados nos testes com a 2GB1.

Tabela 2. Sumário dos experimentos realizados para a proteína 2GB1. N na coluna pesos corresponde ao número de resíduos da proteína (56 para a 2GB1). A sequência de ciclos é repetida até o fim do experimento. Continua na próxima página.

Experimento	τ – tempo do ciclo (gerações)	Técnicas para inserção de diversidade na população	Pesos (ver Eq. 6)
SP	Não se aplica	Não se aplica	$w_{VdW} = w_{PH} = 1$ para todos os ciclos
Est-VdW	Não se aplica	Não se aplica	$w_{VdW} = N/2$ e $w_{PH} = 1$ para todos os ciclos
MVG	2500	Não se aplica	$w_{VdW} = w_{PH} = 1$ (ciclo 1); $w_{VdW} = N/2$ e $w_{PH} = 1$ (ciclo 2); $w_{VdW} = 1$ e $w_{PH} = N$ (ciclo 3)
RVG 1	2500	Não se aplica	$w_{PH} = 1$ e aleatório entre 1 e N para w_{VdW}
RVG 2	2500	Não se aplica	Aleatório entre 1 e N para w_{PH} ou w_{VdW}
MVG-Hip	2500	Hipermutação	Mesmo que MVG
MVG-IA6	2500	6% Imigrantes Aleatórios	Mesmo que MVG
MVG-Rst	2500	Restart	Mesmo que MVG
MVG-RstAll	2500	Restart Full	Mesmo que MVG
MVG-Relax0.1	2500	Relaxamento de restrições	$w_{VdW} = 0,1$ e $w_{PH} = 1$ (ciclo 1); $w_{VdW} = w_{PH} = 1$ (ciclo 2); $w_{VdW} = N/2$ e $w_{PH} = N$ (ciclo 3)

Tabela 2. Continuação da tabela da página anterior. Sumário dos experimentos realizados para a proteína 2GB1. N na coluna pesos corresponde ao número de resíduos da proteína (56 para a 2GB1). A sequência de ciclos é repetida até o fim do experimento.

Experimento	τ – tempo do ciclo (gerações)	Técnicas para inserção de diversidade na população	Pesos (ver Eq. 6)
MVG- Relax0.01	2500	Relaxamento de restrições	$w_{VdW} = 0,01$ e $w_{PH} = 1$ (ciclo 1); $w_{VdW} = w_{PH} = 1$ (ciclo 2); $w_{VdW} = N/2$ e $w_{PH} = N$ (ciclo 3)
MVG-Mem1	2500	Memória 1	Mesmo que MVG
MVG-MemPop	2500	Memória População	Mesmo que MVG
MVG- MemPopAD	2500	Memória População Antes/Depois	Mesmo que MVG

Tabela 3. Sumário dos experimentos realizados para a proteína 2W4P. N na coluna pesos corresponde ao número de resíduos da proteína (94 para a 2W4P).

Experimento	τ – tempo do ciclo (gerações)	Técnicas para inserção de diversidade na população	Pesos (ver Eq. 6)
SP	Não se aplica	Não se aplica	$w_{VdW} = w_{PH} = 1$ para todos os ciclos
Est-VdW	Não se aplica	Não se aplica	$w_{VdW} = N/2$ e $w_{PH} = 1$ para todos os ciclos
MVG	5000	Não se aplica	$w_{VdW} = w_{PH} = 1$ (ciclo 1); $w_{VdW} = N/2$ e $w_{PH} = 1$ (ciclo 2); $w_{VdW} = 1$ e $w_{PH} = N$ (ciclo 3)
MVG-Hip	5000	Hipermutação	Mesmo que MVG
MVG-IA6	5000	6% Imigrantes Aleatórios	Mesmo que MVG
MVG-Rst	5000	Restart	Mesmo que MVG

Tabela 4. Sumário dos experimentos realizados para a proteína 1ENH. N na coluna pesos corresponde ao número de resíduos da proteína (54 para a 1ENH).

Experimento	τ – tempo do ciclo (gerações)	Técnicas para inserção de diversidade na população	Pesos (ver Eq. 6)
SP	Não se aplica	Não se aplica	$w_{VdW} = w_{PH} = 1$ para todos os ciclos
Est-VdW	Não se aplica	Não se aplica	$w_{VdW} = N/2$ e $w_{PH} = 1$ para todos os ciclos
MVG	2500	Não se aplica	$w_{VdW} = w_{PH} = 1$ (ciclo 1); $w_{VdW} = N/2$ e $w_{PH} = 1$ (ciclo 2); $w_{VdW} = 1$ e $w_{PH} = N$ (ciclo 3)
MVG-Hip	2500	Hipermutação	Mesmo que MVG
MVG-IA6	2500	6% Imigrantes Aleatórios	Mesmo que MVG
MVG-Rst	2500	Restart	Mesmo que MVG

No experimento sem pesos (SP), não há ciclos, ou seja, a Eq. 1 é utilizada ao longo de todas as gerações. No MVG, a alteração segue a sequência de ciclos da Tabela 2, e ao terminar os três ciclos de pesos, a sequência é executada novamente, até atingir o número total de gerações. No experimento RVG 1, o peso do termo de Van der Waals, é alterado para um novo peso aleatório a cada troca de ciclo. O RVG 2 é similar ao RVG 1, exceto que o peso escolhido pode ser w_{VdW} ou w_{PH} , com 50% de chance cada. Por fim, o experimento Est-VdW é similar ao experimento realizado em [Ishivatari, 2009], onde a função de *fitness* é estática e um peso definido previamente é atribuído a componente de van der Waals por todas as gerações da execução.

De acordo com [Kashtan *et al.*, 2007], o modelo que mais produziu aceleração ao trocar de espaço de busca foi o MVG. Dessa forma, se estabeleceu o modelo MVG como sendo o padrão utilizado nos algoritmos de inserção de diversidade da população. A proteína G (2GB1) possui testes adicionais, quando comparados aos testes das proteínas 2W4P e 1ENH por ser do tipo $\alpha+\beta$ e relativamente pequena, como comentado anteriormente.

3.7 Experimento com a função de avaliação “Onuchic”

Três experimentos foram realizados com a função de avaliação “Onuchic” e são descritos a seguir:

- Experimento SPO (Sem Pesos Onuchic)

Este experimento foi executado a título de comparação. Nele, os parâmetros do AG utilizados foram:

- AG Estático Padrão
- 150.000 gerações
- População: 200 indivíduos
- Taxa de mutação: 10%
- Taxa de crossover: 40%

- Experimento MVGO (MVG Onuchic)

Por sua vez, o experimento MVGO foi executado com as seguintes configurações:

- AG Dinâmico MVG
- 150.000 gerações
- População: 200 indivíduos
- Taxa de mutação: 10%
- Taxa de crossover: 40%
- Inserção de diversidade: 6% Imigrantes aleatórios e relaxamento de restrições no termo de energia dos não-ligantes

- Experimento MVGO-PopMelhor (MVG Onuchic com troca da população baseada no melhor indivíduo)

Por fim, este experimento foi realizado com base nos resultados de experimentos anteriores (Seção 3.6) e foi dividido em três etapas. As configurações comuns às três etapas foram:

- AG Dinâmico MVG
- 150.000 gerações (total de gerações das três etapas)
- População: 200 indivíduos
- Inserção de diversidade: 6% Imigrantes aleatórios e relaxamento de restrições no termo de energia dos não-ligantes

A primeira etapa foi executada com os seguintes parâmetros:

- 50.000 gerações
- Taxa de mutação: 20%
- Taxa de crossover: 60%

A segunda etapa foi executada com os seguintes parâmetros:

- População inicial: 200 indivíduos próximos da melhor solução obtida na primeira etapa
- 50.000 gerações
- Taxa de mutação: 5%
- Taxa de crossover: 40%

A terceira e última etapa foi utilizada com os seguintes parâmetros:

- População inicial: 200 indivíduos próximos da melhor solução obtida na segunda etapa
- 50.000 gerações
- Taxa de mutação: 5%
- Taxa de crossover: 40%

3.8 Análise dos Resultados

Por ser um trabalho que abrange diversas áreas, diferentes tipos de informações podem ser extraídas dos experimentos, que se traduzem em diferentes resultados. Nas subseções a seguir é feito um breve comentário do significado e da justificativa de cada tipo de resultado.

3.8.1 Fitness e Energia Equivalente

O *fitness* é a função que guia o processo de evolução do AG, de forma que utilizá-lo para a análise dos resultados é fundamental. Através dele é possível dizer se um determinado algoritmo obtém melhores soluções que outro.

Todavia, como este trabalho tem como objetivo comparar resultados de experimentos estáticos e experimentos dinâmicos, não é possível fazer uma comparação direta utilizando o valor do *fitness* dos algoritmos, uma vez que nos experimentos dinâmicos a função de *fitness* varia com o tempo. Assim, apresentaremos os valores de *fitness* equivalentes àqueles calculados para o caso estático, i.e. dados pela Eq. 1 mesmo sabendo que nem sempre esta função é diretamente otimizada em um dado momento. Chamaremos este valor de *energia equivalente*.

Dessa forma, uma comparação entre os algoritmos poderá ser feita, ainda que a análise seja mais complexa, visto que a energia equivalente apenas deixa os resultados na mesma escala, mas ainda não resolve o problema de que os algoritmos estão otimizando diferentes funções de *fitness* ao longo do tempo. Contudo, será mostrado que, mesmo sem otimizar por todo o tempo a função de energia equivalente, alguns resultados para os casos dinâmicos são melhores (quando a energia equivalente é analisada) indicando que tornar as funções dinâmicas

permite às populações escapar eventualmente dos ótimos locais. Vale ressaltar que os valores das energias de todos os resultados estão em unidades reduzidas, ϵ_H , como no modelo proposto por [Yap *et al.*, 2008].

3.8.2 Desvio da Raiz Média Quadrática (RMSD)

O desvio da raiz quadrática média (RMSD) é uma métrica amplamente utilizada em problemas de predição de estruturas tridimensionais de proteínas. Ela diz a proximidade (geralmente em Å) da estrutura predita e a estrutura nativa, alinhando os *backbones*, ou seja, os carbonos- α das estruturas. Os valores de RMSD obtidos neste trabalho foram calculados utilizando o pacote Bio3D [Grant *et al.*, 2006] da ferramenta estatística R [Ihaka & Gentleman, 1996].

3.8.3 Score de Alinhamento

O RMSD é bastante utilizado em problemas de PEP, porém sabe-se que ele possui limitações, como dificuldade de interpretação ou falta de normalização [Baber *et al.*, 2009]. Além disso, o RMSD não leva em conta a estrutura secundária da proteína, de forma que seja possível prever estruturas com um valor de RMSD baixo (menor que 5 Å [Unger, 2004]), mas que seja completamente distinta da estrutura nativa, quanto as suas estruturas secundárias [Capriles, 2011].

Dessa forma, neste trabalho é proposta uma nova métrica, que leva em consideração o alinhamento das estruturas secundárias, similar ao que é feito no BLAST [Altschul *et al.*, 1990]. A ideia é aumentar a pontuação de uma estrutura predita caso a estrutura do aminoácido seja igual à estrutura do aminoácido da nativa. A pontuação deve ser mantida se as estruturas forem diferentes, para que seja possível visualizar – em forma de porcentagem – quão boa foi a predição. Para padronizar as estruturas secundárias, utilizou-se o programa e a base de dados DSSP [Kabsch & Sander, 1983] [Joosten *et al.*, 2010]. A fórmula utilizada para calcular o *score* é dada pelas Eq. 10 e 11.

$$hit(i) = \begin{cases} +1 & \text{se } p(i) = n(i) \\ 0 & \text{se } p(i) \neq n(i) \end{cases} \quad \text{Eq. 10}$$

$$score = \frac{\sum_{i=1}^N hit(i)}{N} \quad \text{Eq. 11}$$

Tal que $p(i)$ e $n(i)$ são os tipos de estrutura secundária, definidas pelo DSSP, das proteínas preditas e nativas, respectivamente e N é o número de resíduos da proteína. Dessa maneira o resultado é dado na forma de porcentagem e pode ser comparada entre diferentes proteínas.

3.8.4 Teste estatístico Wilcoxon

Foram realizados também testes estatísticos para verificar a significância dos resultados. Optou-se pelo teste não-paramétrico *Wilcoxon signed-rank test* [Wilcoxon, 1945], uma vez que não é possível assumir que os resultados – RMSD ou *score* de alinhamento – possuam uma distribuição normal. Os conjuntos amostrais correspondem ao menor RMSD obtido de cada execução, dentre os melhores indivíduos ao longo das gerações. Analogamente, para o *score* de alinhamento, os conjuntos amostrais são formados pelos maiores *scores*.

É importante ressaltar que o teste de Wilcoxon diz se dois conjuntos amostrais são diferentes ou não. De forma que se eles forem estatisticamente diferentes, a mediana entre os dois conjuntos deve ser comparada para poder afirmar se o resultado de um algoritmo é superior a outro.

3.8.5 Análise Visual

Além do RMSD, em problemas de PEP é possível analisar as estruturas 3D geradas. Apesar de não ser exatamente uma métrica, já que a análise visual é subjetiva, é possível verificar a estrutura em determinado instante da execução. Isso é interessante para AGs com *fitness* dinâmico, pois é possível ver o comportamento da estrutura ao trocar o espaço de busca. As imagens das estruturas apresentadas neste trabalho foram geradas com a ferramenta Jmol [Herráez, 2006].

3.8.6 Fluxograma dos experimentos

A Figura 15 apresenta um resumo do fluxo dos experimentos.

Figura 15. *Pipeline* dos resultados para um experimento.

4 Resultados

Os resultados dos experimentos são apresentados e discutidos neste capítulo. Ele está organizado da seguinte maneira: a primeira parte mostra um comparativo em relação ao tempo de execução; em seguida os resultados da proteína 2GB1 são apresentados e discutidos, incluindo testes com a função de avaliação “Onuchic”. Por fim, os resultados para as proteínas 2W4P e 1ENH são discutidos.

4.1 Tempo de execução

Uma das vantagens de se trabalhar com um campo de força baseado em um modelo *coarse-grained* é o fato de o custo computacional ser reduzido quando comparado a um campo de força *full-atom*.

O teste foi realizado em um Intel® Xeon® 2,00 GHz e 5GB de RAM, com os mesmos parâmetros discutidos na Seção 3.6, exceto que o número total de gerações foi 1000. O algoritmo utilizado foi o AG Padrão Sem Pesos (SP), alterando apenas a função de *fitness*. O campo de força CG foi o mesmo utilizado neste projeto [Yap *et al.*, 2008] e o campo de força *full-atom* utilizado foi o CHARMM 22.

O tempo de execução foi de aproximadamente 18 minutos para o modelo CG, enquanto que o modelo *full-atom* levou 24 horas para terminar. Percebe-se que o ganho em termos de tempo é bastante expressivo, passando da ordem de horas para minutos. Os experimentos descritos na Seção 3.6 levaram em média 17 horas por execução. Assumindo um processamento serial, o processamento das 10 execuções demora aproximadamente 1 semana. Nessas mesmas condições, o modelo *full-atom* levaria aproximadamente 1 ano e meio.

4.2 Proteína G (2GB1)

Os principais resultados da proteína 2GB1 estão reunidos nesta seção (para experimentos adicionais desta proteína, ver Apêndice A), divididos pelo tipo de experimento, para facilitar a visualização. Os experimentos utilizados como comparação foram: SP, Est-VdW e MVG; de forma que seus dados aparecerão em várias figuras. As Figuras 16 e 17 apresentam os gráficos referentes às energias equivalentes para os diferentes algoritmos.

Figura 16. Energia equivalente média em relação a 10 execuções dos melhores indivíduos ao longo das gerações da proteína 2GB1 para os experimentos comuns às outras proteínas.

Figura 17. Energia equivalente ao longo das gerações da execução que obteve o menor RMSD da proteína 2GB1 para os experimentos comuns às outras proteínas.

Pela Figura 16 verifica-se que o experimento SP possui um decaimento contínuo. A queda na energia é rápida inicialmente e se torna lenta ao final das gerações, evidenciando que a população pode estar presa em ótimos locais. O experimento Est-VdW possui valores de energia que fica estagnado em aproximadamente 95 ϵ_H , o que sugere que manter a contribuição do termo de van der Waals alta durante todo o processo evolutivo não é benéfico para o processo de otimização.

Já os experimentos MVG e suas variantes apresentam um comportamento típico de experimentos dinâmicos. O *fitness* decai até o momento onde ocorre a troca no espaço de busca. O melhor indivíduo do espaço anterior já pode não ser o melhor no novo espaço e com isso ocorre o aumento brusco no *fitness*. Em seguida o processo de otimização continua no novo espaço e o *fitness* continua a diminuir. Durante a repetição deste processo, nota-se também que os vales e picos dos experimentos MVG tendem a diminuir na média, aproximando dos resultados do SP.

Já pela Figura 17, observa-se que os comportamentos descritos para a média das energias (Figura 16) também são visíveis nas execuções separadamente. Características como o aumento da energia dos experimentos dinâmicos MVG são mantidas, entretanto novas informações podem ser extraídas da Figura 17. Percebe-se que alguns vales dos experimentos MVG conseguem atingir valores menores que o SP, evidenciando que o algoritmo consegue eventualmente fugir de ótimos locais através da mudança do espaço do *fitness*, similarmente ao que ocorre nos experimentos apresentados em [Kashtan *et al.*, 2007].

Figura 18. Energia equivalente média em relação a 10 execuções dos melhores indivíduos ao longo das gerações da proteína 2GB1 comparando com os experimentos de mudança aleatória.

Figura 19. Energia equivalente ao longo das gerações da execução que obteve o menor RMSD da proteína 2GB1 comparando com os experimentos de mudança aleatória.

As figuras 18 e 19 mostram as energias dos experimentos de mudança aleatória. A partir delas, observa-se que os experimentos RVG apresentam a característica de mudança brusca no *fitness*, similar ao experimento MVG. Os valores de energia, por outro lado, ficam mais próximos dos valores do experimento Est-VdW, devido à forma como os pesos são alterados (Tabela 2).

Figura 20. Energia equivalente média em relação a 10 execuções dos melhores indivíduos ao longo das gerações da proteína 2GB1 comparando com os experimentos que usam memória.

Figura 21. Energia equivalente ao longo das gerações da execução que obteve o menor RMSD da proteína 2GB1 comparando com os experimentos que usam memória.

As figuras 20 e 21 mostram os resultados para os experimentos que utilizam memória de indivíduos de ciclos passados. Observa-se que os experimentos com memória possuem uma característica em comum: poucas melhoras no *fitness* quando há uma troca no espaço de busca. Isso se deve ao fato da memória recuperar os melhores indivíduos do ciclo anterior, o que pode acarretar em um domínio deste indivíduo sob toda a população. Entretanto, deve-se destacar também que o experimento MVG-MemPop conseguiu obter valores de energia menores que o experimento SP, fato que não havia ocorrido até então.

Figura 22. Energia equivalente média em relação a 10 execuções dos melhores indivíduos ao longo das gerações da proteína 2GB1 comparando com os experimentos que usam relaxamento de restrições.

Figura 23. Energia equivalente ao longo das gerações da execução que obteve o menor RMSD da proteína 2GB1 comparando com os experimentos que usam relaxamento de restrições.

As figuras 22 e 23 mostram as energias dos experimentos com relaxamento de restrições. Observa-se que a adição do relaxamento de restrições faz com que a função de *fitness* tenha picos maiores, uma vez que com o relaxamento, o algoritmo aceita soluções que não poderiam ser aceitas normalmente. Além disso, o relaxamento de restrições (experimento MVG-Relax0.1) também conseguiu obter um vale menor que o experimento SP, indicando que esta técnica pode melhorar os resultados em relação ao SP e ao MVG.

Figura 24. Média de 10 execuções do RMSD (em Å) dos melhores indivíduos ao longo das gerações da proteína 2GB1 para os experimentos comuns às outras proteínas.

A Figura 24 apresenta a média de dez execuções do RMSD e observa-se que para os experimentos dinâmicos MVG o mesmo comportamento da energia também é observado no RMSD. A diferença mais significativa é que os valores de RMSD para o experimento Est-VdW é menor que o experimento SP, devido ao fato de que o aumento da contribuição do termo de van der Waals faz com que os átomos fiquem mais próximos uns dos outros o que favorece a formação de estruturas mais compactas e com menor RMSD. Entretanto, assim como na energia, a média no RMSD para várias execuções pode ocultar informações. Assim, a Figura 25 apresenta os valores de RMSD da execução que se obteve o menor RMSD.

Figura 25. RMSD ao longo das gerações da execução onde se obteve o menor RMSD para a proteína 2GB1 para os experimentos comuns às outras proteínas.

A Figura 25 mostra que na melhor execução todos os experimentos dinâmicos conseguiram obter pelo menos um vale com RMSD menor que o RMSD dos experimentos estáticos.

Figura 26. Média de 10 execuções do RMSD (em Å) dos melhores indivíduos ao longo das gerações da proteína 2GB1 para os experimentos de mudança aleatória.

Figura 27. RMSD ao longo das gerações da execução onde se obteve o menor RMSD para a proteína 2GB1 para os experimentos de mudança aleatória.

Das Figuras 26 e 27, verifica-se que os experimentos RVG obtiveram valores de RMSD bastante próximos ou menores que os valores do experimento Est-VdW, mesmo com valores de energia (Figuras 18 e 19) parecidos, o que indica que os experimentos RVG estão em ótimos locais diferentes.

Figura 28. Média de 10 execuções do RMSD (em Å) dos melhores indivíduos ao longo das gerações da proteína 2GB1 para os experimentos que usam memória.

Figura 29. RMSD ao longo das gerações da execução onde se obteve o menor RMSD para a proteína 2GB1 para os experimentos que usam memória.

Os experimentos com memória obtiveram, em alguns casos, valores de energia menores que os experimentos estáticos (Figuras 20 e 21), entretanto, apesar de apresentarem a mesma característica no RMSD (Figuras 28 e 29) – pouca ou nenhuma melhoria no *fitness* quando há troca no espaço de busca – observa-se que a adição da memória ao Algoritmo Genético não melhorou o RMSD em nenhum dos casos.

Figura 30. Média de 10 execuções do RMSD (em Å) dos melhores indivíduos ao longo das gerações da proteína 2GB1 para os experimentos que usam relaxamento de restrições.

Figura 31. RMSD ao longo das gerações da execução onde se obteve o menor RMSD para a proteína 2GB1 para os experimentos que usam relaxamento de restrições.

As Figuras 30 e 31 mostram os resultados comparativos dos experimentos com relaxamento de restrições. Constatou-se que assim como nas energias (Figuras 22 e 23), os picos dos valores de RMSD são mais altos que os outros experimentos. Para a média das dez execuções não obtiveram nenhum ganho de RMSD, entretanto para a melhor execução obteve-se um valor que foi menor que os outros experimentos, reforçando a ideia de que o relaxamento de restrições pode ser benéfico para problemas de PEP.

Para auxiliar o entendimento dos gráficos acima, as Tabelas 5, 6 e 7 apresentam informações estatísticas e o teste de Wilcoxon para os menores RMSDs de cada execução.

Tabela 5. Conjunto amostral de RMSD (em Å), dos menores valores em cada execução dentre os melhores indivíduos, para a proteína 2GB1. Os valores em negrito são os menores valores de cada experimento.

Execução	1	2	3	4	5	6	7	8	9	10
SP	17,106	12,593	11,104	11,699	18,616	15,026	16,720	17,580	20,102	13,454
MVG	8,105	8,805	7,754	10,256	9,905	10,783	10,739	10,938	9,174	11,271
Est-VdW	10,393	10,240	9,017	13,115	13,563	11,867	10,219	9,241	8,452	10,506
RVG 1	10,379	7,480	13,616	9,805	10,532	11,049	12,139	10,244	11,038	8,866
RVG 2	11,885	9,276	12,781	9,509	12,230	11,124	11,364	11,636	11,917	9,810
MVG-Hip	10,617	8,451	10,684	9,560	9,280	10,529	7,514	9,636	7,929	9,458
MVG-IA6	8,753	10,069	10,405	8,173	7,044	7,454	7,826	8,475	10,905	8,141
MVG-Rst	9,194	8,065	8,078	9,389	7,757	7,992	8,248	11,273	8,634	9,477
MVG-RstAll	9,803	9,178	11,477	8,209	9,837	9,531	10,332	9,162	8,982	10,897
MVG-Mem1	11,338	9,767	11,220	9,727	8,953	9,495	11,279	10,647	9,127	11,424
MVG-MemPop	10,877	9,196	11,225	13,383	9,784	11,801	10,540	12,514	11,002	11,919
MVG-MemPopAD	10,877	10,718	10,531	11,283	8,972	11,801	10,622	12,064	11,494	11,916
MVG-Relax0.1	8,655	10,437	9,811	7,285	11,470	11,322	8,924	10,840	10,957	10,206
MVG-Relax0.01	10,831	8,823	9,740	9,803	10,732	8,842	10,739	11,308	11,330	10,572

Tabela 6. Informações estatísticas do conjunto amostral da Tabela 5. A entrada com asterisco indica o experimento onde se encontrou o menor RMSD.

	Média ± Desvio Padrão	Mediana	Mínimo	Máximo
SP	15,390 ± 3,103	15,873	11,004	20,102
MVG	9,755 ± 1,279	10,081	7,574	11,271
Est-VdW	10,661 ± 1,699	10,317	8,452	13,563
RVG 1	10,515 ± 1,678	10,456	7,480	13,616
RVG 2	11,153 ± 1,213	11,500	9,276	12,781
MVG-Hip	9,366 ± 1,111	9,509	7,514	10,684
MVG-IA6*	8,725 ± 1,306	8,324	7,044	10,905
MVG-Rst	8,811 ± 1,065	8,441	7,757	11,273
MVG-RstAll	9,741 ± 0,962	9,667	8,209	11,477
MVG-Mem1	10,298 ± 0,984	10,207	8,953	11,424
MVG-MemPop	11,224 ± 1,246	11,114	9,196	13,383
MVG-MemPopAD	11,028 ± 0,913	11,080	8,972	12,064
MVG-Relax0.1	9,991 ± 1,339	10,322	7,285	11,470
MVG-Relax0.01	10,272 ± 0,923	10,652	8,823	11,330

Tabela 7. p-valor do teste estatístico de Wilcoxon para o RMSD, comparando os experimento estáticos SP e Est-VdW com os experimentos dinâmicos, para a proteína 2GB1. Os sinais + e – entre parêntesis indicam respectivamente que a mediana do experimento da linha é menor ou maior que o experimento estático. A letra ‘s’ indica que o resultado é estatisticamente significativo.

	SP	Est-VdW
MVG	0,002 (s+)	0,160 (+)
RVG 1	0,006 (s+)	0,770 (-)
RVG 2	0,004 (s+)	0,492 (-)
MVG-Hip	0,002 (s+)	0,064 (+)
MVG-IA6	0,004 (s+)	0,084 (+)
MVG-Rst	0,002 (s+)	0,037 (s+)
MVG-RstAll	0,004 (s+)	0,322 (+)
MVG-Mem1	0,004 (s+)	1,000 (+)
MVG-MemPop	0,014 (s+)	0,275 (-)
MVG-MemPopAD	0,002 (s+)	0,375 (-)
MVG-Relax0.1	0,002 (s+)	0,492 (-)
MVG-Relax0.01	0,002 (s+)	0,770 (-)

Da Tabela 7 verifica-se que todos os experimentos dinâmicos são estatisticamente superiores ao experimento estático SP. Já com relação ao experimento Est-VdW, destaque para o experimento MVG-Rst, que obteve um conjunto amostral superior ao Est-VdW com 95% de confiança. Nenhum experimentos com memória e relaxamento obtiveram medianas de RMSD menores que o experimento Est-VdW.

Os gráficos abaixo apresentam os resultados do *score* de alinhamento para a proteína 2GB1, lembrando que nesta medida, quanto maior o valor, melhor o resultado do experimento.

Figura 32. *Score* de alinhamento da execução onde se encontrou o maior *score* ao longo das gerações para a proteína 2GB1 para os experimentos comuns às outras proteínas.

Figura 33. *Score* de alinhamento da execução onde se encontrou o maior *score* ao longo das gerações para a proteína 2GB1 para os experimentos de mudança aleatória.

Figura 34. *Score* de alinhamento da execução onde se encontrou o maior *score* ao longo das gerações para a proteína 2GB1 para os experimentos que usam memória.

Figura 35. *Score* de alinhamento da execução onde se encontrou o maior *score* ao longo das gerações para a proteína 2GB1 para os experimentos que usam relaxamento de restrições.

Pelas Figuras 32, 33, 34 e 35, verifica-se que apesar do experimento Est-VdW ter obtido menores valores de RMSD quando comparado com o experimento SP, o alinhamento das estruturas do teste Est-VdW foi pior que o SP. No geral, observa-se que os melhores *scores* foram de 50%, devido à presença das folhas- β na estrutura, que não foram formadas nas estruturas preditas. As Tabelas 8 e 9 apresentam as informações estatísticas do *score* de alinhamento. O conjunto de dados corresponde aos maiores *scores* de cada execução, dentre os melhores indivíduos.

Tabela 8. Informações estatísticas do *score* de alinhamento da proteína 2GB1.

	Média ± Desvio Padrão	Mediana	Mínimo	Máximo
SP	0,414 ± 0,052	0,402	0,339	0,482
Est-VdW	0,454 ± 0,035	0,455	0,393	0,500
MVG	0,382 ± 0,044	0,366	0,339	0,464
RVG 1	0,418 ± 0,040	0,420	0,357	0,464
RVG 2	0,441 ± 0,028	0,446	0,411	0,500
MVG-Hip	0,445 ± 0,023	0,446	0,411	0,482
MVG-IA6	0,427 ± 0,021	0,429	0,393	0,446
MVG-Rst	0,429 ± 0,038	0,429	0,357	0,482
MVG-RstAll	0,438 ± 0,015	0,438	0,411	0,464
MVG-Mem1	0,421 ± 0,031	0,411	0,357	0,464
MVG-MemPop	0,409 ± 0,031	0,402	0,375	0,446
MVG-MemPopAD	0,411 ± 0,037	0,411	0,357	0,464
MVG-Relax0.1	0,455 ± 0,024	0,446	0,429	0,518
MVG-Relax0.01	0,438 ± 0,009	0,438	0,429	0,446

Tabela 9. p-valor do teste estatístico de Wilcoxon para o *score* de alinhamento, comparando o experimento SP com os outros experimentos, para a proteína 2GB1. Os sinais +, – e ~ entre parêntesis indicam respectivamente que a mediana do experimento da linha é maior, menor ou igual que o SP.

	SP
Est-VdW	0,060 (+)
MVG	0,114 (-)
RVG1	0,753 (+)
RVG2	0,059 (+)
MVG-Hip	0,183 (+)
MVG-IA6	0,507 (+)
MVG-Rst	0,260 (+)
MVG-RstAll	0,097 (+)
MVG-Mem1	0,674 (+)
MVG-MemPop	0,918 (~)
MVG-MemPopAD	0,838 (+)
MVG-Relax0,1	0,085 (+)
MVG-Relax0,01	0,191 (+)

Pela Tabela 9 verifica-se que nenhum experimento é estatisticamente diferente do experimento SP, considerando um grau de confiança de 95%. Por outro lado, se abrandarmos o grau de confiança para 90%, os experimentos Est-VdW, RVG2, MVG-RstAll e MVG-Relax0.1 possuem conjuntos amostrais estatisticamente superiores ao conjunto amostral do experimento

SP. A Figura 36 abaixo apresenta um comparativo da estrutura 3D nativa com a estrutura de menor RMSD e menor energia.

Figura 36. Estruturas 3D da proteína 2GB1, enfatizando as estruturas secundárias. (a) Estrutura nativa, (b) Estrutura de menor RMSD do experimento MVG e (c) Estrutura de menor energia do experimento MVG.

Observa-se pela Figura 36 que há discrepância entre a estrutura com menor RMSD e a estrutura com menor energia. Este problema é devido à função de energia utilizada e será

discutido com mais detalhes no Apêndice A. A Figura 37 mostra as estruturas 3D de menor RMSD dos experimentos realizados.

Energia nativa: 160,56 ϵ_H

(a)

Energia: 52,286 ϵ_H

RMSD: 11,004 Å

(b)

Energia: 116,908 ϵ_H

RMSD: 8,452 Å

(c)

Energia: 108,339 ϵ_H

RMSD: 7,48 Å

(d)

Energia: 102,039 ϵ_H

RMSD: 9,276 Å

(e)

Energia: 82,753 ϵ_H

RMSD: 7,574 Å

(f)

Figura 37. (a) Estrutura nativa, (b) SP, (c) Est-VdW, (d) RVG 1, (e) RVG 2, (f) MVG

Energia: 92,48 ϵ_H

RMSD: 7,514 Å

(g)

Energia: 97,714 ϵ_H

RMSD: 7,044 Å

(h)

Energia: 75,502 ϵ_H

RMSD: 7,757 Å

(i)

Energia: 97,677 ϵ_H

RMSD: 8,209 Å

(j)

Energia: 84,523 ϵ_H

RMSD: 8,953 Å

(k)

Energia: 81,314 ϵ_H

RMSD: 9,196 Å

(l)

Figura 37. (g) MVG-Hip, (h) MVG-IA6, (i) MVG-Rst, (j) MVG-RstAll, (k) MVG-Mem1, (l) MVG-MemPop

Energia: 81,115 ϵ_H

RMSD: 8,972 Å

(m)

Energia: 81,993 ϵ_H

RMSD: 7,285 Å

(n)

Energia: 78,539 ϵ_H

RMSD: 8,823 Å

(o)

Figura 37. Estruturas 3D, enfatizando as estruturas secundárias da proteína 2GB1. (a) Estrutura nativa, (b) SP, (c) Est-VdW, (d) RVG 1, (e) RVG 2, (f) MVG, (g) MVG-Hip, (h) MVG-IA6, (i) MVG-Rst, (j) MVG-RstAll, (k) MVG-Mem1, (l) MVG-MemPop, (m) MVG-MemPopAD, (n) MVG-Relax0.1, (o) MVG-Relax0.01. Para as siglas dos experimentos, ver Tabela 2.

Pela Figura 37, observa-se que as estruturas possuem RMSDs parecidos, porém as energias estão com valores bastante diferentes. As estruturas dos experimentos Est-VdW, RVG 1 e RVG 2 estão com o segmento da Hélice- α mais curvado, devido ao fato do termo de van der Waals estar com uma contribuição alta durante todo o processo, como discutido anteriormente. Percebe-se que em algumas estruturas, as fitas- β estão quase paralelas, ainda que o software DSSP não as aceite como folhas- β . A Figura 38 apresenta o que ocorre com a estrutura durante a troca do espaço de busca para uma execução do experimento MVG.

Figura 38. Estruturas 3D de uma execução do experimento MVG, evidenciando as estruturas secundárias. (a) geração 38000, (b) geração 39000, (c) 40000, (d) geração 41000, (e) geração 42000 e (f) estrutura nativa.

Verifica-se da Figura 38 que durante a evolução da estrutura no experimento MVG, o algoritmo desenovela a estrutura ao trocar de espaço de busca. Isto pode parecer, a princípio, mais custoso, porém este processo de desenovelamento pode ser importante para a PEP, visto que neste processo as ligações são rompidas e novas ligações, eventualmente mais promissoras, podem ser formadas. Como dito anteriormente, sabe-se que o problema de PEP é multimodal e complexo, de modo que isto pode ser benéfico para retirar a população de ótimos locais em que se encontram.

4.2.1 Resultados com a função de avaliação “Onuchic”

Esta apresenta os resultados do melhor experimento utilizando a função de avaliação “Onuchic”. A Figura 39 mostra os valores de energia para os três experimentos.

Figura 39. Energia equivalente ao longo das gerações da execução que obteve o menor RMSD dos experimentos com a função de avaliação “Onuchic”.

O valor de energia da estrutura nativa é de -1400 kJ/mol. Comparando com a Figura 39, verifica-se que os experimentos estabilizaram em valores de energia próximos a -800 kJ/mol. Destaca-se também que o experimento dinâmico MVGO não possui a curva característica dos experimentos dinâmicos vistos nas figuras acima e o experimento “MVGO-PopMelhor” possui os saltos devido a troca de população durante a execução.

Figura 40. RMSD ao longo das gerações da execução que obteve o menor RMSD dos experimentos com a função de avaliação “Onuchic”.

A Figura 40 apresenta os valores de RMSD dos experimentos. Constatase que o experimento SPO conseguiu obter valores de RMSD de 4,0 Å, menores que os valores do experimento dinâmico MVGO, de aproximadamente 6,5 Å. Entretanto o experimento com troca da população durante a execução obteve valores de RMSD de 2,5 Å com uma estrutura próxima da nativa, como pode ser visto na Figura 41. Vale ressaltar também que a própria estrutura nativa ao passar pela mudança de coordenadas cartesianas para ângulos diedrais realizada pelo AG sofre um desvio de 1,6 Å, portanto as estruturas estão mais próximas do que os valores de RMSD apresentam.

Figura 41. Estruturas 3D da proteína 2GB1, enfatizando estruturas secundárias. Estrutura nativa (azul) sobreposta com a estrutura de menor RMSD (vermelha). O valor de RMSD é de 2,577 Å e a energia da estrutura é -845,989 kcal/mol.

A Figura 41 mostra que a estrutura predita consegue se alinhar bem com a estrutura nativa, tanto na hélice- α quanto nas quatro fitas- β . Não houve formação das estruturas de folhas- β pois o visualizador PyMOL [Schrodinger, 2010] utiliza outro algoritmo para padronizar os tipos de estruturas das proteínas.

4.3 Fosfatase (2W4P)

Os resultados para a proteína Fosfatase 2W4P estão reunidos nesta seção. A apresentação dos resultados seguirá o modelo da seção anterior. As Figuras 42 e 43 apresentam gráficos das energias equivalentes ao longo das gerações.

Figura 42. Energia equivalente média em relação a 10 execuções dos melhores indivíduos ao longo das gerações da proteína 2W4P.

Figura 43. Energia equivalente ao longo das gerações da execução que obteve o menor RMSD da proteína 2W4P.

Com base na Figura 42, verifica-se o mesmo comportamento observado nas energias dos experimentos para a proteína 2GB1 (Figura 16). Os experimentos estáticos possuem um decaimento rápido inicial e em seguida um decaimento lento, enquanto que os quatro experimentos dinâmicos MVG apresentaram o comportamento característico de troca de espaço

de busca. Já a Figura 43 nos mostra que o experimento MVG-IA6 obteve energias menores que o SP no primeiro ciclo (5000 gerações), entretanto nas gerações subsequentes o MVG-IA6 não obteve bons resultados. O MVG-Rst por outro lado, conseguiu obter dois vales menores que o SP.

Figura 44. Média de 10 execuções do RMSD (em Å) dos melhores indivíduos ao longo das gerações da proteína 2W4P.

Figura 45. RMSD ao longo das gerações da execução onde se obteve o menor RMSD para a proteína 2W4P.

A Figura 16 apresenta a média dos RMSDs e pode-se constatar que os quatro experimentos dinâmicos conseguiram obter vales menores que o Est-VdW na média e na execução onde se obteve o RMSD mínimo (Figura 45). Os valores dos RMSDs são bem maiores que os valores da proteína 2GB1, como esperado, pois esta proteína possui quase o dobro do número de resíduos que a 2GB1. As Tabelas 10 e 11 apresentam as informações estatísticas do RMSD.

Tabela 10. Informações estatísticas do RMSD da proteína 2W4P. Os dados foram gerados de maneira similar a proteína 2GB1 (Tabela 5).

	<i>Média ± Desvio Padrão</i>	<i>Mediana</i>	<i>Mínimo</i>	<i>Máximo</i>
SP	24,514 ± 5,013	24,573	17,319	32,430
Est-VdW	17,184 ± 2,680	17,680	13,739	21,439
MVG	14,872 ± 0,981	15,150	12,344	15,920
MVG-Hip	15,438 ± 1,328	15,395	12,453	17,361
MVG-IA6	15,132 ± 1,024	15,097	13,651	17,049
MVG-Rst	15,038 ± 1,427	15,003	12,917	16,860

Tabela 11. p-valor do teste estatístico de Wilcoxon para o RMSD, comparando o experimento Est-VdW com os experimentos dinâmicos, para a proteína 2W4P. Os sinais + e – entre parêntesis indicam respectivamente que a mediana do experimento da coluna é menor ou maior que o Est-VdW. A letra ‘s’ indica que o resultado é estatisticamente significativo.

	MVG	MVG-Hip	MVG-IA6	MVG-Rst
SP	0,002 (s+)	0,002 (s+)	0,002 (s+)	0,002 (s+)
Est-VdW	0,084 (+)	0,131 (+)	0,106 (+)	0,131 (+)

Pelas Tabelas 10 e 11, observa-se que os quatro experimentos dinâmicos obtiveram resultados superiores (com nível de confiança de 99%) ao experimento SP. Os experimentos dinâmicos também obtiveram medianas menores que o Est-VdW, porém não são significativamente menores, assumindo um nível de confiança de 95%. Por outro lado, se o intervalo de confiança for ampliado para 85%, todos os experimentos dinâmicos podem ser considerados melhores que o Est-VdW. Os valores de RMSD, como comentado anteriormente, estão acima dos 5 Å discutidos em [Unger, 2004].

Figura 46. *Score* de alinhamento da execução onde se encontrou o maior *score* ao longo das gerações para a proteína 2W4P.

Tabela 12. Informações estatísticas sobre o *score* de alinhamento da proteína 2W4P.

	Média	Mediana	Mínimo	Máximo
SP	0,324 ± 0,019	0,328	0,292	0,354
Est-VdW	0,330 ± 0,035	0,323	0,281	0,385
MVG	0,362 ± 0,024	0,354	0,313	0,396
MVG-Hip	0,359 ± 0,014	0,354	0,333	0,385
MVG-IA6	0,357 ± 0,016	0,354	0,333	0,385
MVG-Rst	0,349 ± 0,017	0,354	0,313	0,365

Tabela 13. p-valor do teste estatístico de Wilcoxon para o *score* de alinhamento, comparando o experimento SP com os experimentos dinâmicos, para a proteína 2W4P. Os sinais + e – entre parêntesis indicam respectivamente que a mediana do experimento da coluna é maior ou menor que o SP. A letra ‘s’ indica que o resultado é estatisticamente significativo.

	Est-VdW	MVG	MVG-Hip	MVG-IA6	MVG-Rst
SP	0,918 (+)	0,008 (s+)	0,006 (s+)	0,009 (s+)	0,015 (s+)

Pela Figura 46 e Tabelas 12 e 13, verifica-se que os resultados do alinhamento também são piores quando comparados com os resultados da proteína 2GB1. O melhor alinhamento não passou dos 39%, enquanto que para a 2GB1 foi possível obter alinhamentos de até 50%. Por outro lado, os experimentos MVG-Hip, MVG-IA6 e MVG-Rst conseguiram conjuntos amostrais significativamente superiores ao conjunto amostral do experimento SP.

Figura 47. Estruturas 3D da proteína 2W4P, enfatizando as estruturas secundárias. (a) Estrutura nativa, (b) Estrutura de menor RMSD do experimento MVG e (c) Estrutura de menor energia do experimento MVG.

Pela Figura 47 é possível notar que assim como na proteína 2GB1 (Figura 36), há uma diferença entre as estruturas de menor energia e menor RMSD. Há a formação de Hélices- α e fitas β nos dois casos, uma vez que o modelo utilizado assume informações da estrutura secundária *a priori*, porém as estruturas preditas – principalmente a de menor energia – não se dobraram o suficiente.

Energia: 142,952 ϵ_H

RMSD: 17,319 Å

(a)

Energia: 179,341 ϵ_H

RMSD: 13,739 Å

(b)

Energia: 162,334 ϵ_H

RMSD: 12,344 Å

(c)

Energia: 159,43 ϵ_H

RMSD: 12,453 Å

(d)

Energia: 174,993 ϵ_H

RMSD: 13,651 Å

(e)

Energia: 175,391 ϵ_H

RMSD: 12,917 Å

(f)

Energia nativa: 204 ϵ_H

(g)

Figura 48. Estruturas 3D, enfatizando as estruturas secundárias da proteína 2W4P. (a) SP, (b) Est-VdW, (c) MVG, (d) MVG-Hip, (e) MVG-IA6, (f) MVG-Rst, (g) Estrutura nativa.

Da Figura 48, observa-se que novamente o experimento Est-VdW está com as estruturas secundárias mais curvadas do que as estruturas dos outros experimentos. Em alguns casos – SP e MVG-Rst – houve a formação de pequenas Hélices- α em locais não esperados, o que pode acarretar em uma piora no *score* de alinhamento, uma vez que esses resíduos poderiam ser de folhas- β .

4.4 DNA ligante (1ENH)

Os resultados da proteína 1ENH são apresentados nesta seção. As Figuras 49 e 50 mostram os gráficos das energias equivalentes médias e mínimas, respectivamente.

Figura 49. Energia equivalente média em relação a 10 execuções dos melhores indivíduos ao longo das gerações da proteína 1ENH.

Figura 50. Energia equivalente ao longo das gerações da execução que obteve o menor RMSD da proteína 1ENH.

Pela Figura 49 verifica-se que o experimento MVG-Rst conseguiu obter vales menores que os valores do experimento SP. Nota-se também que os picos dos experimentos dinâmicos são menores que das outras duas proteínas, o que pode traduzir em um menor número de gerações necessárias para diminuir o *fitness* ao trocar o espaço de busca.

Já da Figura 50, observa-se que mesmo fazendo a análise apenas sob os melhores indivíduos, a energia do experimento Est-VdW tem um aumento na geração 45000, o que não ocorre em experimentos estático. A justificativa deste aumento da energia é a presença do peso de van der Waals, causa uma maior compactação da proteína, o que acaba acarretando em um aumento da energia equivalente.

Figura 51. Média de 10 execuções do RMSD (em Å) dos melhores indivíduos ao longo das gerações da proteína 1ENH.

Figura 52. RMSD ao longo das gerações da execução onde se obteve o menor RMSD para a proteína 1ENH.

Pela Figura 51, não é observado o comportamento visto até então para os experimentos dinâmicos, onde há picos quando o espaço de busca é alterado. No entanto, verifica-se que na

média os resultados dos experimentos dinâmicos são melhores que os dois experimentos estáticos, ao longo de quase todas as gerações.

Baseado na Figura 52, alguns picos se tornam mais evidentes, como no MVG-Hip. Assim como nas médias, a execução onde se obteve o mínimo também tem, no geral, valores de RMSD menores que os dois experimentos estáticos. Verifica-se também que os valores de RMSD do SP e do Est-VdW estão mais próximos para esta proteína, algo que não havia ocorrido nos experimentos das proteínas anteriores. As Tabelas 14 e 15 apresentam as informações estatísticas sobre o RMSD.

Tabela 14. Informações estatísticas do RMSD da proteína 1ENH. Os dados foram gerados de maneira similar a proteína 2GB1 (Tabela 5).

	<i>Média ± Desvio Padrão</i>	<i>Mediana</i>	<i>Mínimo</i>	<i>Máximo</i>
SP	11,252 ± 2,227	11,320	8,553	15,255
Est-VdW	10,052 ± 1,005	10,279	7,993	11,732
MVG	8,327 ± 1,223	8,576	5,822	9,988
MVG-Hip	8,276 ± 1,180	8,225	6,779	10,362
MVG-IA6	8,482 ± 1,1377	8,998	6,356	10,129
MVG-Rst	8,105 ± 1,146	8,123	6,483	10,237

Tabela 15. p-valor do teste estatístico de Wilcoxon para o RMSD, comparando o experimento Est-VdW com os experimentos dinâmicos, para a proteína 1ENH. Os sinais + e – entre parêntesis indicam respectivamente que a mediana do experimento da coluna é menor ou maior que o Est-VdW. A letra ‘s’ indica que o resultado é estatisticamente significativo.

	MVG	MVG-Hip	MVG-IA6	MVG-Rst
SP	0,010 (s+)	0,004 (s+)	0,010 (s+)	0,004 (s+)
Est-VdW	0,002 (s+)	0,01 (s+)	0,01 (s+)	0,002 (s+)

Das Tabelas 14 e 15, observa-se que as medianas dos quatro experimentos dinâmicos são significativamente menores que o dois experimentos estáticos, SP e Est-VdW, com um nível de confiança de 99%. Nota-se também que o mínimo obtido foi de 5,822 Å, valor que está bem próximo dos 5 Å discutidos em [Unger, 2004]. A Figura 53 e as Tabelas 16 e 17 a seguir apresentam os resultados referentes ao *score* de alinhamento.

Figura 53. Score de alinhamento da execução onde se encontrou o maior score ao longo das gerações para a proteína 1ENH.

Tabela 16. Informações estatísticas sobre o score de alinhamento da proteína 1ENH.

	Média $\pm$ Desvio Padrão	Mediana	Mínimo	Máximo
SP	0,737 $\pm$ 0,051	0,741	0,630	0,815
Est-VdW	0,574 $\pm$ 0,122	0,528	0,426	0,778
MVG	0,728 $\pm$ 0,040	0,732	0,667	0,815
MVG-Hip	0,724 $\pm$ 0,059	0,732	0,630	0,852
MVG-IA6	0,682 $\pm$ 0,048	0,676	0,611	0,759
MVG-Rst	0,722 $\pm$ 0,045	0,732	0,667	0,815

Tabela 17. p-valor do teste estatístico de Wilcoxon para o score de alinhamento, comparando o experimento SP com os experimentos dinâmicos, para a proteína 1ENH. Os sinais + e – entre parêntesis indicam respectivamente que a mediana do experimento da coluna é maior ou menor que o SP. A letra ‘s’ indica que o resultado é estatisticamente significativo.

	Est-VdW	MVG	MVG-Hip	MVG-IA6	MVG-Rst
SP	0,012 (-)	0,271 (-)	0,209 (-)	0,022 (s-)	0,271 (s-)

Pela Figura 53 e Tabelas 16 e 17, verifica-se que a porcentagem de alinhamento foi maior que da 2GB1, como esperado, devido à ausência de folhas- β . Em particular, o experimento MVG-Hip obteve 85% de alinhamento, enquanto que o máximo na proteína 2GB1

não passou dos 50%. Comprova-se pelo gráfico e pelas tabelas também que o experimento SP obteve medianas melhores que todos os outros experimentos.

Figura 54. Estruturas 3D da proteína 1ENH, enfatizando as estruturas secundárias. (a) Estrutura nativa, (b) Estrutura de menor RMSD do experimento MVG e (c) Estrutura de menor energia do experimento MVG.

Comparando a estrutura de menor RMSD com a estrutura nativa pela Figura 54, observa-se que algumas dobras não foram previstas deixando a estrutura mais compacta que a nativa. Já a estrutura de menor energia teve alguns locais onde a Hélice- α foi dividida, a estrutura está menos compacta e há uma formação de uma pequena fita β , provavelmente uma interpretação incorreta do visualizador, visto que esta fita aparece em outras predições, como pode ser visto na Figura 55.

Energia: 95,077 ϵ_H

RMSD: 8,553 Å

(a)

Energia: 109,927 ϵ_H

RMSD: 7,993 Å

(b)

Energia: 102,552 ϵ_H

RMSD: 5,822 Å

(c)

Energia: 99,057 ϵ_H

RMSD: 6,794 Å

(d)

Energia: 96,92 ϵ_H

RMSD: 6,356 Å

(e)

Energia: 95,176 ϵ_H

RMSD: 6,483 Å

(f)

Energia nativa: 108,20 ϵ_H

(g)

Figura 55. Estruturas 3D, enfatizando as estruturas secundárias da proteína 1ENH. (a) SP, (b) Est-VdW, (c) MVG, (d) MVG-Hip, (e) MVG-IA6, (f) MVG-Rst, (g) Estrutura nativa.

Com base na Figura 55, verifica-se que houve a formação não esperada da fita β em três dos seis experimentos. Observa-se também que o experimento Est-VdW foi o que mais perdeu informações sobre as Hélices- α , confirmando o resultado do baixo valor da mediana no *score* de alinhamento (Tabela 16).

5 Conclusões

De uma maneira geral, pode-se dizer que os resultados utilizando AGs com função de avaliação dinâmica para o problema de predição de estruturas foram positivos, com base nos experimentos realizados para as três proteínas. Foi possível obter valores de energia equivalente próximos – e em alguns casos até melhores – que as energias dos experimentos estáticos, ainda que, como discutido anteriormente, a comparação não seja justa para os algoritmos de *fitness* dinâmico, visto estarmos comparando parametrizações distintas do campo de força.

Com relação aos experimentos realizados para a proteína G (2GB1), observou-se que a introdução da dinamicidade ao problema foi benéfica. Os experimentos dinâmicos obtiveram algumas estruturas de menor energia, menor RMSD e maior *score* de alinhamento que os experimentos estáticos. Quando comparado com o experimento Sem Pesos, os experimentos dinâmicos obtiveram também melhores resultados com significância estatística de 99% de confiança. Destaque também para os experimentos com relaxamento de restrição, que conseguiram melhorar o processo de otimização, aceitando estruturas que não seriam aceitas normalmente e eventualmente encontrando boas soluções a partir delas.

Ao comparar os dois campos de força testados, o campo de força “Onuchic” [Clementi *et al.*, 2000] produziu melhores resultados considerando o RMSD e o fato da estrutura estar no ótimo global, porém ressalta-se que este campo de força requer o conhecimento prévio da estrutura tridimensional da proteína, ou seja, não seria viável para uma aplicação real em predição de proteínas. Por outro lado, o campo de força HG [Yap *et al.*, 2008], apesar da estrutura nativa não estar no ótimo global da função (o que dificulta o processo de otimização e a localização da estrutura nativa), é genérico e pode ser aplicado em problemas de PEP.

Já para os experimentos da proteína fosfatase (2W4P), há um destaque para a questão da escalabilidade do sistema. Os resultados não foram tão bons quanto os resultados da proteína G, principalmente com relação aos valores de RMSD. Para esta proteína (e outras com mais resíduos) deve-se levar em consideração também o tempo de execução e o aumento no número de gerações necessárias para se obter um resultado satisfatório. Ainda assim, os experimentos dinâmicos tiveram um forte impacto na melhoria das soluções visto que quanto maior a complexidade do problema, maior será a aceleração para encontrar melhores soluções ao trocar o espaço de busca [Kashtan *et al.*, 2007].

Para a proteína DNA ligante (1ENH), notou-se que a otimização dos AGs dinâmicos gerou resultados que são estatisticamente melhores que os resultados dos dois experimentos estáticos, fato que não havia ocorrido para as outras duas proteínas. Observou-se também que os métodos de introdução de diversidade na população não foram tão benéficos para esta proteína – os valores de RMSD são maiores que o experimento MVG. Isso se justifica pelo fato da

proteína 1ENH ser mais simples que as outras duas. Assim é mais interessante explorar os o espaço de busca já conhecido, do que explorar outros locais do espaço.

Deve-se ficar atento aos possíveis artefatos que o aumento ou diminuição de um dos termos pode trazer, como as estruturas mais curvadas, discutido na seção anterior. A utilização dos pesos de forma dinâmica mostrou-se útil, pois foi possível aumentar a contribuição dos termos por um número limitado de gerações. Com isso, foram obtidas estruturas com RMSD menor que o experimento SP e ainda assim não tão deformadas quanto as estruturas do experimento Est-VdW.

Como conclusão geral, os experimentos dinâmicos foram melhores que os experimentos estáticos – principalmente o AG padrão, corroborando com a hipótese proposta em [Kashtan *et al.*, 2007]: a mudança no espaço de busca causa um aumento na diversidade da população, evitando a convergência prematura do AG. Dentre os algoritmos dinâmicos, destaque para o algoritmo modular (MVG) que obteve melhores resultados que os algoritmos aleatórios (RVG). Por fim, com relação às técnicas de inserção de diversidade não é possível afirmar com clareza qual técnica foi melhor, porém constatou-se que a utilização de várias dessas técnicas na mesma execução mostrou-se bastante eficaz.

E para trabalhos futuros, uma das possibilidades é a integração deste método com outros preditores reconhecidos e premiados com o ROSETTA [Leaver-Fay *et al.*, 2011] e I-TASSER [Zhang, 2008]. A integração poderá ser feita realizando a predição através desses preditores e em seguida utilizar a estrutura predita como entrada na população do AG dinâmico. Dessa forma, ao invés do AG iniciar a busca a partir de estruturas aleatórias, ele poderia otimizar tendo como ponto de partida uma estrutura mais estável, realizando a busca por outros ângulos e possivelmente melhorar as soluções obtidas dos preditores.

Além disso, outro possível trabalho futuro seria a utilização deste método dinâmico em campos de força *full-atom*, com a possível integração com o campo de força CG, efetuando tanto uma busca geral quanto um ajuste fino das estruturas.

Apêndice A – Experimentos Adicionais para a proteína 2GB1

1. Diversidade da População

Um dos problemas do AG padrão, como discutido anteriormente, é que após um determinado número de gerações o ganho no *fitness* é muito baixo em relação ao tempo de execução. Isso se deve ao fato da população convergir prematuramente para ótimos locais [Mitchell, 1998], sendo que, nestes casos, os indivíduos da população ficam próximos, o que pode ser verificado pela baixa diversidade da população. Para comparar a diversidade de experimentos estáticos e dinâmicos, as seguintes equações foram utilizadas:

$$d_{i,k} = \sqrt{\sum_j ([\phi_{i,j} - \phi_{k,j}]^2 + [\psi_{i,j} - \psi_{k,j}]^2)} \quad \text{Eq. 12}$$

$$d_m = \frac{1}{N} \sum_{i, k>i} d_i, \quad \text{Eq. 13}$$

Tal que $d_{i,k}$ é a distância euclidiana entre os indivíduos i e k da população, j refere-se aos resíduos de cada indivíduo, $\phi_{i,j}$, $\phi_{k,j}$, $\psi_{i,j}$, $\psi_{k,j}$ referem-se aos ângulos (genes) do resíduo j dos indivíduos i e k . O d_m é a soma das distâncias calculadas par a par e divididas pelo número de vezes que foram calculadas, N o número de comparações feitas. Este d_m foi registrado a cada 5 gerações para o experimento estático SP e o experimento dinâmico RVG 1 e os resultados são apresentados na Figura A-1.

Figura A-1. Distância euclidiana média da população Eq. 12 e 13 a cada (a) 5 gerações e (b) 200 gerações

Observa-se da Figura A-1 que o experimento dinâmico possui uma distância euclidiana maior que o experimento estático. Isso faz com que a população do experimento dinâmico seja

mais diversa, o que facilita a fuga de ótimos locais, ocasionando numa exploração melhor do espaço de busca e, eventualmente, melhorando o desempenho.

2. População Inicial do AG Próxima da Estrutura Nativa

Com base nos resultados discutidos na seção 4.2, notou-se que as estruturas de um modo geral não conseguiram obter valores de RMSD próximos da estrutura nativa. Além disso, várias estruturas preditas possuem energia menor que a da estrutura nativa. O campo de força proposto em [Yap *et al.*, 2008] foi desenvolvida para convergir nas proximidades do estado enovelado, sendo que para experimentos que começam longe da estrutura nativa, o algoritmo pode convergir para outros pontos ótimos de superfícies com energia menor. O objetivo dos testes nesta subseção é investigar esta hipótese.

Um teste inicial verificou a transformação de ângulos diedrais para coordenadas cartesianas utilizado no AG. Neste teste, o AG possui uma população de 1 indivíduo, com taxa de mutação e *crossover* iguais a zero. Por fim, calculou-se o RMSD entre o indivíduo que sofreu a transformação e a estrutura nativa, obtendo um valor de aproximadamente 1,6 Å. Este resultado confirma que há uma perda de informação ao transformar de ângulos diedrais para coordenadas cartesianas.

Optou-se por realizar um teste adicional, para verificar se, ao realizar a otimização de uma população próxima da estrutura nativa, haveria divergência ou convergência nos resultados, isto é, se as estruturas ainda ficariam próximas da estrutura nativa ou não.

Para tanto, seis testes foram realizados, com as seguintes configurações:

- Número de gerações: 2000
- Taxa de mutação, taxa de *crossover* e número de indivíduos na população foram mantidos em 15%, 60% e 100 respectivamente (mesmos valores que os experimentos anteriores, Seção 3.6 para mais detalhes)
- População inicial do AG foi dividida em dois tipos
 - Idêntica: indivíduos da população são todos idênticos a estrutura nativa
 - Aproximada: indivíduos da população são parecidos com a estrutura nativa. Neste caso, em cada ângulo de torção ϕ e ψ foi aplicado uma perturbação com distribuição normal com média centrada no próprio ângulo e desvio padrão de 1,0°.
- Os tipos de mutação também foram divididos em dois:
 - Os três tipos de mutação utilizados durante o processo de otimização (mesmo que os experimentos propostos na Seção 3.6)

- Apenas uma mutação foi utilizada durante o processo. A mutação real, onde não há acesso a base de dados de ângulos. Neste tipo de mutação, diferentemente dos outros dois, os saltos no espaço de busca sempre são pequenos, o que não permite, em muitos casos, a fuga de ótimos locais

A Tabela A-1 apresenta um resumo dos testes realizados:

Tabela A-1. Resumo dos testes com a população inicial próxima da estrutura com estado nativo.

Nome do Teste	População Inicial	Tipos de Mutação
3 mut., Idêntico	Idêntica	Três mutações
1 mut., Idêntico	Idêntica	Apenas mutação real
3 mut., Aprox.	Aproximada	Três mutações
1 mut., Aprox.	Aproximada	Apenas mutação real

A Figura A-2 apresenta um gráfico com as energias equivalentes dos quatro testes realizados. Observa-se que as energias equivalentes dos quatro testes estão relativamente próximas umas das outras, entretanto evidencia-se que os testes com apenas a mutação real se estabilizam em valores acima dos valores obtidos pelos experimentos com as três mutações. A Figura A-3 apresenta os resultados referentes aos RMSDs dos testes e a Figura A-4 a diversidade da população dos quatro testes.

Figura A-2. Energias equivalentes para os quatro testes com a população inicial próxima da estrutura com estado nativo.

Pela Figura A-3 observa-se a distinção entre os RMSDs para os testes com as três mutações quando comparado com os valores de RMSD dos testes com apenas a mutação real. Os testes com três mutações apresentam comportamentos similares aos vistos na Seção 4.2, com valores de RMSD entre 10,0 e 25,0 Å. Por outro lado, os dois testes com apenas a mutação real atingiram RMSDs entre 2,0 e 4,0 Å, ou seja, bem próximos da estrutura nativa e dos valores obtidos em [Yap *et al.*, 2008].

A Figura A-4 reforça a ideia de que as populações dos experimentos com uma mutação ficaram presas em um ótimo local. Verifica-se que os experimentos com uma mutação possuem uma diversidade baixa quando comparada com a diversidade dos experimentos com três mutações.

Por fim, com base nas Figuras Figura A-2, Figura A-3 e Figura A-4, verifica-se que quando a população inicial possui soluções próximas ou idênticas à estrutura nativa, o AG com uma mutação consegue manter-se próximo da estrutura nativa, pois a mutação real possibilita apenas pequenos saltos, que não são suficientes para atravessar uma barreira energética maior. Nota-se também que utilizando este campo de força, as soluções com menor RMSD não são as soluções com menor energia. Enfatiza-se novamente que esses testes foram feitos com uma população próxima da nativa e, portanto, os valores baixos de RMSD. Nos experimentos com população inicial aleatória as três mutações são importantes, para que o AG possa explorar melhor o espaço de busca.

Figura A-3. RMSDs para os quatro testes com a população inicial próxima da estrutura com estado nativo.

Figura A-4. Distância euclidiana da população Eq. 12 e 13 dos quatro testes com a população inicial próxima da estrutura com estado nativo.

3. Função de fitness específica para a proteína 2GB1

Alguns parâmetros da função de *fitness* Eq. 1 foram ajustados e calibrados para serem exatamente como os da estrutura nativa da proteína 2GB1. Estes parâmetros são os valores dos ângulos de ligação e a distância entre os carbonos- α . Deseja-se verificar se, ao calibrar a função de *fitness* especificamente para a proteína 2GB1, há uma melhor correlação entre baixo *fitness* e baixo RMSD para o campo de força utilizado. Foram propostos dois tipos de teste: com a população inicial próxima a conformação da estrutura nativa (similar aos testes da Seção A-2) e com a população inicial aleatória (similar aos experimentos da Seção 3.6). A Tabela A-2 apresenta um resumo dos testes realizados.

Tabela A-2. Resumo dos testes com a função de *fitness* específica para a proteína 2GB1. Os outros parâmetros são os mesmos descritos na Seção 3.6.

Nome do Teste	População Inicial	Tipos de Mutação	Número de Gerações
3 mut., Idêntico, Espec.	Aproximada	Três mutações	2000
1 mut., Idêntico, Espec.	Aproximada	Apenas mutação real	2000
SP, Espec.	Aleatória	Três mutações	57500
Est-VdW, Espec.	Aleatória	Três mutações	57500
MVG, Espec.	Aleatória	Três mutações	57500

Figura A-5. Energias equivalentes para os testes com a população inicial próxima da estrutura com estado nativo.

As Figuras Figura A-5 e Figura A-6 mostram as energias equivalentes e seus termos separadamente, para os testes com a população inicial do AG idêntica ou próxima da estrutura de estado nativo. Observa-se que o uso da função de *fitness* específica para a proteína 2GB1 não resultou em uma melhora na energia equivalente como um todo. Contudo, ao analisar a Figura A-6a, verifica-se que a energia do ângulo de ligação do teste “1 mut., Aprox., Espec.” (um dos termos onde houve o ajuste específico) foi menor que os outros testes. Da Figura A-6b, os testes com *fitness* específicos tiveram valores da energia do ângulo de ligação maiores que os outros testes, uma maneira que o AG encontrou de balancear os termos.

(a)

(b)

(c)

(d)

(e)

Figura A-6. Termos de energia da energia equivalente para os testes com a população inicial próxima da estrutura com estado nativo. (a) Energia do ângulo de ligação, (b) energia do ângulo diedral, (c) energia de van der Waals, (d) energia das ligações de hidrogênio, (e) legenda desta figura.

A Figura A-7 apresenta os resultados dos RMSDs dos testes e nota-se que, assim como nos testes da Seção A-2, os testes com *fitness* específico são sensíveis aos tipos de mutação empregados. Verifica-se que o RMSD para o teste “1 mut., Aprox., Espec.” fica estabilizado em 5,0 Å, que é um pouco maior que os 2,0 a 4,0 Å dos outros experimentos, porém ainda pode ser baixo o suficiente para lembrar as estruturas principais da conformação nativa [Unger, 2004].

Figura A-7. RMSDs para os testes com a população inicial próxima da estrutura com estado nativo.

Os resultados dos três experimentos com função de *fitness* específica para a 2GB1 restantes são apresentados na Figura A-8, comparando com os respectivos resultados dos

experimentos com a função de *fitness* padrão. Constatase que o menor RMSD com o *fitness* específico foi menor que o menor RMSD obtido pelo respectivo experimento com o *fitness* padrão. Obteve-se um valor de 6,321 Å para o MVG Espec., quando comparado aos 7,574 Å do MVG, aproximando do resultado desejado (5,0 Å) [Unger, 2004].

Figura A-8. RMSD ao longo das gerações da execução onde se obteve o menor RMSD para a proteína 2GB1.

4. Aumento de gerações

Neste teste, dois experimentos (um estático, Est-VdW e um dinâmico, MVG) foram executados por 150.000 gerações. Ao aumentar o número de gerações deseja-se verificar se as conclusões, para a 2GB1, dos experimentos dinâmicos e estáticos (MVG e algumas variantes terem melhor desempenho que os estáticos) podem ser generalizados para um número maior de gerações. Além disso, este teste serve como controle para verificar se o número de gerações utilizado nos experimentos (57500) é suficiente, ou seja, se os resultados não melhoram ao rodar três vezes mais gerações.

Figura A-9. Energia equivalente média em relação a 10 execuções dos melhores indivíduos ao longo das 150.000 gerações da proteína 2GB1.

Figura A-10. Média de 10 execuções do RMSD (em Å) dos melhores indivíduos ao longo das 150.000 gerações da proteína 2GB1.

Figura A-11. RMSD ao longo das 150.000 gerações da execução onde se obteve o menor RMSD para a proteína 2GB1.

A Figura A-9 apresenta a energia equivalente média para os dois experimentos, enquanto que as Figuras Figura A-10 e Figura A-11 mostram respectivamente os valores de RMSD médios e mínimos dos experimentos. Observa-se que os picos e vales do experimento MVG diminuem ao longo das gerações, enquanto que o experimento Est-VdW mantém os valores estabilizados, tanto para a energia equivalente quanto para o RMSD. No teste estatístico de Wilcoxon (similar ao realizado na Seção 4.2 para os outros experimentos) obteve-se um p-valor de 0,131, com os valores de mediana do MVG e Est-VdW respectivamente em 9,994 e 10,299. Comparando com a Tabela 7, o intervalo de confiança diminuiu, entretanto ainda não chegou aos 95%. Dessa forma, conclui-se que o MVG manteve o melhor desempenho que o Est-VdW para 150.000 gerações e que os resultados se estabilizaram.

Referências Bibliográficas

- [Altschul *et al.*, 1990] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J. (1990) “Basic local alignment search tool.”, *Journal of Molecular Biology*, v. 215, p. 403-410
- [Anfinsen, 1973] Anfinsen, C. B. (1973) “Principles that govern the folding of protein chains”, *Science*, v. 181, p. 223-230
- [Arnold *et al.*, 2005] Arnold, K., Bordoli, L., Kopp, J., Schwede, T. (2005) “The SWISS-MODEL workspace: a web-based environment for protein structure homology modeling”, *Bioinformatics*, v. 22, p. 195-201
- [Baber *et al.*, 2009] Baber, J. C., Thompson, D. C., Cross, J. B., Humblet, C. (2009) “GARD: A Generally Applicable Replacement for RMSD”, *Journal of Chemical Information and Modeling*, v. 49, p. 1889-1900
- [Baxevanis & Oullette, 2001] Baxevanis, A., Oullette, B. (2001) “Bioinformatics – A practical guide to the analysis of genes and proteins”, Lawrence Erlbaum Associates Publishers
- [Benitez *et al.*, 2012] Benitez, C. M. V., Parpinelli, R. S., Lopes, H. S. (2012) “Parallelism, hybridism and coevolution in a multi-level ABC-GA approach for the protein structure prediction problem”, *Concurrency and Computation: Practice and Experience*, v. 24, p. 635-646
- [Berman *et al.*, 2000] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., Bourne, P. E. (2000) “The Protein Data Bank”, *Nucleic Acids Research*, v. 28, p. 235-242. www.pdb.org
- [Bordner, 2012] Bordner, A. J. (2012) “Force fields for homology modeling”, *Methods in molecular biology*, v. 857, p. 83-106
- [Branke, 2002] Branke, J. (2002) “Evolutionary Optimization in Dynamic Environments”, Boston, MA: Kluwer Academic Publishers
- [Buchner *et al.*, 2010] Buchner, G. S., Murphy, R. D., Buchete, N-V, Kubelka, J. (2010) “Dynamics of protein folding: Probing the kinetic network of folding-unfolding transitions with experiment and theory”, *BBA – Proteins & Proteomics*, v. 1814, p. 1001-1020
- [Cai-Yun *et al.*, 2009] Cai-Yun, W., Hao-Dong, Z., Le-Cai, C. (2009) “A new prediction protein structure method based on genetic algorithm and coarse-grained protein model”, *BMEI '09*, p. 1-5
- [Capriles, 2011] Capriles, P. V. Z. (2011) “Desenvolvimento e Implementação de um Modelo *Coarse-Grained* para Predição de Estruturas de Proteínas”, 100 f. Dissertação (Doutorado em Ciências em Modelagem Computacional) Laboratório Nacional de Computação Científica – LNCC.
- [Chandru *et al.*, 2003] Chandru, V., Dattasharma, A. D., Kumar, A. (2003) “The algorithmics of folding proteins on lattices”, *Discrete Applied Mathematics*, v. 127, p. 145-161
- [Clarke *et al.*, 1994] Clarke, N. D., Kissinger, C. R., Desjarlais, J., Gilliland, G. L., Pabo, C. O. (1994) “Structural studies of the engrailed homeodomain”, *Protein Science*, v. 3, p. 1779-1787
- [Clementi *et al.*, 2000] Clementi, C., Nymeyer, H., Onuchic, J. N. (2000) “Topological and energetic factors: What determines the structural details of the transition state ensemble and En-route

- intermediates for protein folding? An investigation for small globular proteins”, *Journal of Molecular Biology*, v. 298, p. 937-953
- [Cobb & Grefenstette, 1993] Cobb, H. G., Grefenstette, J. J. (1993) “Genetic algorithms for tracking changing environments”, In *Proc. of the Fifth Intl. Conf. on Genetic Algorithms*, p. 529-530
- [Cooper *et al.*, 2003] Cooper, L. R., Corne, D. W., Crabbe, J. C. (2003) “Use of novel Hill-Climbing Genetic Algorithm in protein folding simulations”, *Computational Biology and Chemistry*, v. 27, p. 575-580
- [Cruz *et al.*, 2011] Cruz, C., González, J. R., Pelta, D. A. (2011) “Optimization in dynamic environments: a survey on problems, methods and measures”, *Soft. Comput.*, v. 15, p. 1427-1448
- [DePristo *et al.*, 2004] DePristo, M. A., Bakker, P. I. W., Blundell, T. L. (2004) “Heterogeneity and inaccuracy in protein structures solved by X-Ray crystallography”, *Structure*, v. 12, p. 831-838
- [Eswar *et al.*, 2006] Eswar, N., Marti-Renom, A., Webb, B., Madhusudhan, M. S., Eramian, D., Shen, M., Pieper, U., Sali, A. (2006) “Comparative protein structure modeling with MODELLER”, *Current Protocols in Bioinformatics*, c. 5, u. 5-6
- [Gopalakrishnan *et al.*, 2007] Gopalakrishnan, K., Sheik, S. S., Ranjani, C. V., Udayakumar, A., Sekar, K. (2007) “Conformational angles database (CADB 3.0)”, *Protein and Peptide Letters*, v. 14, p. 665-668
- [Grant *et al.*, 2006] Grant, B. J., Rodrigues, A. P., ElSawy, K. M., McCammon, J. A., Caves, L. S. (2006) “Bio3D: an R package for the comparative analysis of protein structures.”, *Bioinformatics*, v. 22, p. 2695-2696
- [Gronenborn *et al.*, 1991] Gronenborn, A. M., Fipula, D. R., Essig, N. Z., Achari, A., Whitlow, M., Wingfield, P. T., Clore, G. M. (1991) “A novel, highly stable fold of the immunoglobulin binding domain of streptococcal protein G.”, *Science*, v. 253, p. 657-661
- [Guntsch *et al.*, 2001] Guntsch, M., Middendorf, M., Schmeck, H. (2001) “An ant colony optimization approach to dynamic TSP”, In *Proceedings of the Genetic and Evolutionary Computation Conference*, p. 860-867
- [Hao & Scheragat, 1999] Hao, M. G. & Scheragat, H. A. (1999) “Designing potential energy functions for protein folding”, *Current Opinion in Structural Biology*, v. 9, p. 184-188
- [Herraez, 2006] Herraez, A. (2006) “Biomolecules in the computer: Jmol to the rescue”, v. 34, p. 255-261
- [Hoque *et al.*, 2011] Hoque, M. T., Chetty, M., Lewis, A., Sattar, A. (2011) “Twin Removal in Genetic Algorithms for Protein Structure Prediction Using Low-Resolution Model”, *Trans. Computational Biology and Bioinformatics*, v. 8, p. 234
- [Hornak *et al.*, 2006] Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A., Simmerling, C. (2006) “Comparison of multiple Amber Force Fields and development of improved protein backbone parameters”, *Proteins: Structure, Function and Bioinformatics*, v. 65, p. 712-725
- [Hubbard & Park, 1995] Hubbard, T. J., Park, J. (1995) “Fold recognition and ab initio structure predictions using hidden markov models and β -strand pair potentials”, *Proteins: Structure, Function and Bioinformatics*, v. 23, p. 398-402

- [Ihaka & Gentleman, 1996] Ihaka, R., Gentleman, R. (1996) “R: A language for data analysis and graphics”, *Journal of Computational and Graphical Statistics*, v. 5, p. 299-314
- [Ishivatari, 2009] Ishivatari, L. H. U. (2009) “Comparação de diferentes contribuições em campos de força minimalistas para o problema de determinação de estruturas de proteínas através de algoritmos genéticos”, Monografia de Conclusão de Curso em Informática Biomédica, Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, Universidade de São Paulo
- [Jana & Sil] Jana, N. D., Sil, J. (2012) “Protein Structure Prediction in 2D HP Lattice Model Using Differential Evolutionary Algorithm”, *Advances in Intelligent and Soft Computing*, v. 132, p. 281-290
- [Jiang *et al.*, 2003] Jiang, T., Cui, Q., Shi, G., Ma, S. (2003) “Protein folding simulations of the hydrophobic-hydrophilic model by combing tabu search with genetic algorithms”, *Journal of Chemical Physics*, v. 119, p. 4592-4596
- [Joosten *et al.*, 2010] Joosten, R. P., Te Beek, T. A. H., Krieger, E., Hekkelman, M. L., Hooft, R. W. W., Schneider, R., Sander, C., Vriend, G. (2010) “A series of PDB related databases for everyday needs”, *Nucleic Acids Research* 2010, p. 1-9
- [Kabsch & Sander, 1983] Kabsch, W., Sander, C. (1983) “Dictionary of protein secondary structure: patter recognition of hydrogen-bonded and geometrical features”, *Biopolymers*, v. 22, p. 2577-2637
- [Karplus *et al.*, 1999] Karplus, K., Barrett, C., Cline, M., Diekhans, M., Grate, L., Hughey, R. (1999) “Predicting protein structure using only sequence information”, *Proteins: Structure, Function, and Bioinformatics*, v. 37, p. 121-125
- [Kashtan *et al.*, 2007] Kashtan, N., Noor, E., Alon, U. (2007) “Varying environments can speed up evolution”, *Proc. of the National Academy of Sciences (INAS)*, v. 104, p. 13711-13716
- [Lam *et al.*, 2011] Lam, S. Y., Yeung, R. C., Yu, T. H., Sze, K. H., Wong, K. B. (2011) “A rigidifying salt-bridge favors the activity of thermophilic enzyme at high temperatures at expense of low-temperature activity”, *PLoS Biology*, v. 9, e1001027
- [Lammert *et al.*, 2009] Lammert, H., Schug, A., Onuchic, J. N. (2009) “Robustness and generalization of structure based models for protein folding and function”, *Proteins: Structure, Function and Bioinformatics*, v. 77, p. 881-891
- [Leach, 2001] Leach, A. R. (2001) “Molecular Modelling: Principles and Applications”, 2^a ed., Prentice Hall
- [Leaver-Fay *et al.*, 2011] Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, O. F., Thompson, J., Jacak, R., Kaufman, K., Renfrew, P. D., Smith, C. A., Sheffler, W., Davis, I. W., Cooper, S., Treuille, A., Mandell, D. J., Richter F., Ban, Y. E., Fleishman, S. J., Corn, J. E., Kim, D. E., Lyskov, S., Berrondo, M., Mentzer, S., Popovic, Z., Havranek, J. J., Karanicolas, J., Das, R., Meiler, J., Kortemme, T., Gray, J. J., Kuhlman, B., Baker, D., Bradley, P. (2011) “ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules”, *Methods Enzymol*, v. 487, p. 545-574
- [Lehninger *et al.*, 2004] Lehninger, A. L., Nelson, D. L., Cox, M. M. (2004) “Principles of Biochemistry”, 4th ed., WH Freeman
- [Liang & Wong, 2001] Liang, F., Wong, W. H. (2001) “Evolutionary Monte Carlo for protein folding simulations”, *Journal of Chemical Physics*, v. 115, p. 3374-3380

- [Lima, 2006], Lima, T. W. (2006) “Algoritmos evolutivos para predição de estruturas de proteínas”, 139 f. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) Instituto de Ciências Matemáticas e de Computação – ICMC, Universidade de São Paulo – USP, São Paulo
- [Liu *et al.*, 2011] Liu, T., Tang, G. W., Capriotti, E. (2011) “Comparative Modeling: The State of the Art and Protein Drug Target Prediction”, *Comb. Chem. & High Throughput Screening*, v. 14, pp. 532-547
- [Liu & Beveridge, 2002] Liu, Y., Beveridge, D. L. (2002) “Exploratory studies of ab initio protein structure prediction: Multiple copy simulated annealing, AMBER energy functions, and a generalized born/solvent accessibility solvation model”, *Proteins: Structure, Function and Bioinformatics*, v. 46, p. 128-146
- [Liu & Tao, 2010] Liu, Y., Tao, L. (2010) “An improved simulated annealing algorithm and its application to minimizing the energy of protein tertiary structure”, *World Automaton Congress*, p. 489
- [Lodish *et al.*, 2003] Lodish, H., Berk, A., Matsudaira, P., Kaiser, C. A., Krieger, M., Scott, M. P., Zipursky, L., Darnell, J. (2003) “Molecular Cell Biology”, 5^a ed., WH Freeman
- [Marcus & Ben Naim, 1985] Marcus, Y., Naim, A. A. (1985) “A study of the structure of water and its dependence of solutes based on the isotope effects on solvation thermodynamics in water”, *Journal of Chemical Physics*, v. 83, 4744-4759
- [Meyer *et al.*, 2006] Meyer, K. D., Nasuto, S. J., Bishop, M. (2006) “Stochastic diffusion search: partial function evaluation in swarm intelligence dynamic optimization”, *Studies in Computational Intelligence*, v. 31, p. 185-207
- [Mitchell, 1998] Mitchell, M. (1998) “An Introduction to Genetic Algorithms”, Cambridge: MIT Press
- [Montelione *et al.*, 2000] Montelione, G. T., Zheng, D., Huang, Y. J., Gunsalus, K. C., Szyperski, T. (2000) “Protein NMR spectroscopy in structural genomics.”, *Nature Structure Biology*, v. 7, p. 982-985
- [Monticelli *et al.*, 2008] Monticelli, L., Kandasamy, S. K., Periole, X., Larson, R. G., Tieleman, P. D., Marrink, S-J. (2008) “The MARTINI coarse-grained force field: extension to proteins”, *Journal of Chemical Theory and Computation*, v. 4, 819-834
- [Oliveira, 2011] Oliveira, L. L. (2011) “Uso de estratégias baseadas em conhecimento para algoritmos genéticos aplicados à predição de estruturas tridimensionais de proteínas”. Dissertação de Mestrado, Pós-Graduação em Bioinformática, Universidade de São Paulo (USP), Ribeirão Preto.
- [Pan, 2010] Pan, Y. (2010) “Keynote: Protein Structure Prediction and its Understanding Based on Machine Learning Methods”, *Computer and Information Technology*, p. 95
- [Park *et al.*, 2011] Park, Y-J., Chun, S-H., Kim, B-C. (2011) “Cost-sensitive case-based reasoning using a genetic algorithm: Application to medical diagnosis”, *Artificial Intelligence in Medicine*, v. 51, p. 133-145
- [Pelta *et al.*, 2009] Pelta D., Cruz, C., Verdegay, J. L. (2009) “Simple control rules in a cooperative system for dynamic optimization problems”, *Intl. Journal of Genetic Systems*, v. 38, p. 701-717
- [Ponder & Case, 2003] Ponder, J. W., Case, D. A. (2003) “Force fields for protein simulations”, *Adv. Protein Chem.*, v. 66, p. 27-85

- [Rafiei *et al.*, 2011] Rafiei, F. M., Manzari, S. M., Bostanian, S. (2011) “Financial health prediction models using artificial neural networks, genetic algorithm and multivariate discriminant analysis: Iranian evidence”, *Expert Systems with Applications*, v. 38, p. 10210-10217
- [Recktenwald *et al.*, 1992] Recktenwald, A., Schomburg, D., Schmid, R. D. (1992) “Protein engineering and design: Method and the industrial relevance”, *Journal of Biotechnology*, v.28, p. 1-23
- [Schrodinger, 2010] Schrodinger, L. L. C. (2010) “The PyMOL Molecular Graphics System”, version 1.5.0.1
- [Schulze-Kremer, 1992] Schulze-Kremer, S. (1992) “Genetic Algorithms for Protein Tertiary Structure Prediction”, *Parallel Problem Solving from Nature II*, (R. Männer, B. Manderick, eds.), p. 391-400
- [Söding *et al.*, 2005] Söding, J., Biegert, A., Lupas, A. N. (2005) “The HHpred interactive server for protein homology detection and structure prediction”, *Nucleic Acids Research*, v. 33, p. 244-248
- [Tozzini, 2005] Tozzini, V. (2005) “Coarse-grained models for proteins”, *Current Opinion in Structural Biology*, v. 15, 144-150
- [Trojanowski & Wierzchon, 2009] Trojanowski, K., Wierzchon, S. T. (2009) “Immune-based algorithms for dynamic optimization”, *Inf. Sci.*, v. 179, p. 1495-1515
- [Unger & Moult, 1993] Unger, R., Moult, J. (1993) “Genetic Algorithms for protein folding simulations”, *Journal of Molecular Biology*, v. 231, p. 75-81
- [Unger, 2004] Unger, R. (2004) “The building block approach to protein structure prediction”, *The New Avenues in Bioinformatics*, v.8, p. 177-188
- [Voet & Voet, 2003] Voet, D., Voet, J. G. (2003) “Biochemistry”, New York: J Wiley
- [Werhli & Lemke, 2003], Werhli, A. V., Lemke, N. (2003) “Um novo operador evolutivo para a determinação da estrutura tridimensional de proteínas”, *Anais do IV Encontro Nacional de Inteligência Artificial (ENIA)*, Campinas, v. 8, p. 287-295
- [Wilcoxon, 1945] Wilcoxon, F. (1945) “Individual comparisons by ranking methods”, *Biometrics Bulletin*, v. 1, p. 80-83
- [White & Heijne, 2004] White, S. H., von Heijne, G. (2004) “The machinery of membrane protein assembly”, *Current Opinion in Structural Biology*, v. 14, 397-404
- [Xu & Zhang, 2012] Xu, D., Zhang, Y. (2012) “Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field”, *Proteins*, v. 80, p. 1715-1735
- [Yap & Cosic, 2001] Yap, A., Cosic, Y. (2001) “Application of genetic algorithm for predicting tertiary structures o peptide chains”, In the *Proceedings of the First Joint Conference on Engineering in Medicine Biology and Annual Fall Meeting of the Biomedical Engineering Society (BMES / EMBS)*, Atlanta, USA, p. 12-14
- [Yap *et al.*, 2008] Yap, E-H., Fawzi, N. L., Head-Gordon, T. (2008) “A coarse-grained α -carbon protein model with anisotropic hydrogen-bonding”, *Proteins: Structure, Function and Bioinformatics*, v. 70, p. 626-638
- [Zhang, 2008] Zhang, Y. (2008) “I-TASSER server for protein 3D structure prediction”, *BMC Bioinformatics*, v. 9, p. 40

[Zhao & Wang, 2010] Zhao, J-Q., Wang, L. (2010) “Center Based Genetic Algorithm and its application to the stiffness equivalence of the aircraft wing”, Expert Systems with Applications, v. 38, p. 6254-6261