2012

Algoritmos evolutivos e modelo HP para predição de estruturas de proteínas

Revista de Controle e Automação, Porto Alegre, RS, v. 23, n. 1, p. 25-37, 2012
http://www.producao.usp.br/handle/BDPI/39091

ALGORITMOS EVOLUTIVOS E MODELO HP PARA PREDIÇÃO DE ESTRUTURAS DE PROTEÍNAS

Paulo H. R. Gabriel[*]
phrg@icmc.usp.br

Vinícius V. de Melo[†]
vinicius.melo@unifesp.br

Alexandre C. B. Delbem[*]
acbd@icmc.usp.br

[*]Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo
São Carlos (SP), Brasil

[†]Departamento de Ciência e Tecnologia – Universidade Federal de São Paulo
São José dos Campos (SP), Brasil

ABSTRACT

Evolutionary algorithms and HP Model for protein structure prediction
Protein structures prediction (PSP) is a computationally complex problem. Simplified models of the protein molecule (such as the HP Model) and the use of evolutionary algorithms (EAs) are among the most investigated techniques for PSP. However, the evaluation of a structure represented by the HP model considers only the number of hydrophobic contacts, which doesn't enable the EA to distinguish between structures with the same number of contacts. This paper presents a new multi-objective formulation for PSP in HP Model. Two metrics are evaluated: the number of hydrophobic contacts and the distance between the hydrophobic amino acids. Both metrics are used by the Multi-objective EA in Tables. We showed that the algorithm is fast and robust.

KEYWORDS: Protein structure prediction, Evolutionary algorithms, Multi-objective optimization, HP Model.

RESUMO

Predição de estruturas de proteínas (PSP) é um problema computacionalmente complexo. Modelos simplificados da molécula proteica (como o Modelo HP) e o uso de Algoritmos Evolutivos (AEs) estão entre as principais técnicas investigadas para PSP. Entretanto, a avaliação de uma estrutura representada pelo Modelo HP considera apenas o número de contatos hidrofóbicos, não possibilitando distinguir entre estruturas com o mesmo número de contatos hidrofóbicos. Neste trabalho, é apresentada uma nova formulação multiobjetivo para PSP em Modelo HP. Duas métricas são avaliadas: o número de contatos hidrofóbicos e a distância entre os aminoácidos hidrofóbicos, as quais são tratados pelo AE Multiobjetivo em Tabelas (AEMT). O algoritmo mostrou-se rápido e robusto.

PALAVRAS-CHAVE: Predição de estrutura de proteínas, Algoritmos evolutivos, Otimização multiobjetivo, Modelo HP.

1 INTRODUÇÃO

A predição de estrutura de proteínas (PSP, do inglês *Protein Structure Prediction*) é atualmente uma das áreas mais pesquisadas na Biologia Molecular Computacional. Porém, nenhum método computacional simples e preciso para a determinação da estrutura tridimensional é conhecido. Por outro lado, a determinação, em laboratório, da sequência de aminoácidos que compõem uma proteína é relativamente simples e pouco dispendiosa. Por esse motivo, tem-se investigado diversos algoritmos computacionais para PSP partindo da informação da sequência de aminoácidos.

Artigo submetido em 05/05/2010 (Id.: 01143)
Revisado em 01/08/2010, 03/01/2011, 18/02/2011, 09/03/2011
Aceito sob recomendação do Editor Associado Prof. Ivan Nunes Da Silva

Os métodos computacionais para PSP são divididos basicamente em duas linhas (Ye, 2007): métodos baseados em conhecimento *a priori* sobre o domínio do problema e métodos de predição por primeiros princípios ou *Ab initio*. Os métodos baseados em conhecimento têm sido amplamente empregados, porém possuem uma série de limitações: (1) dependem de grandes bancos de dados de estruturas previamente determinadas, (2) demandam constante atualização desses bancos, (3) requerem algoritmos eficientes para alinhamento de múltiplas sequências, (4) necessitam de métodos computacionais eficientes que avaliem a similaridade entre estruturas.

Por outro lado, predições *Ab inito* não possuem tais restrições, dependendo unicamente da sequência de aminoácidos. Esses métodos baseiam-se no princípio de que a sequência dobra-se em um estado em que sua *energia livre* é mínima. A energia de uma proteína (sequência dobrada) é função das posições dos seus átomos e pode ser calculada utilizando modelos de campo de forças (Unger, 2004; Lopes, 2008). A energia calculada a partir de um campo de forças pode ser minimizada, permitindo que seja encontrada a forma tridimensional que corresponde ao estado de mínima energia. Portanto, a solução desse problema pode ser mapeada em um *problema de otimização* de uma função de energia, que descreve a interação entre os átomos componentes das proteínas e desses com o solvente (em geral, a água).

De modo similar a muitos problema de otimização do mundo real, a busca pela energia livre mínima de uma proteína apresenta complexidade computacional muito alta, não podendo ser solucionada em tempo viável por técnicas de força bruta (Berger and Leighton, 1998; Crescenzi et al., 1998). Uma das abordagens mais investigadas para esse tipo de problema é a utilização de técnicas que, em muitos casos, descobrem um conjunto de ótimos locais, a partir dos quais se pode encontrar o ótimo global ou uma solução aproximada para tal problema. Dentre essas técnicas, merecem destaque na literatura os Algoritmos Evolutivos (AEs) (Eiben and Smith, 2003).

Diversas pesquisas têm sido realizadas no sentido de aplicar AEs para minimizar a energia livre das proteínas. Devido à complexidade das estruturas moleculares, modelos simplificados têm sido largamente empregados para esse cálculo. Um dos modelos mais bem estudado é o Modelo Hidrofóbico–Hidrofílico, ou *Modelo HP* (Lau and Dill, 1989). Nesse modelo, as interações hidrofóbicas são consideradas como sendo a principal força no processo de dobramento e a estrutura é representada em uma malha bi- ou tridimensional.

Diversos trabalhos têm aplicado AEs ao Modelo HP (Lopes, 2008). Esses trabalhos visam minimizar a função de energia do Modelo HP sem, contudo, considerar outras características das conformações geradas. Isso resulta em uma avaliação relativamente grosseira das conformações. Consequentemente, esses AEs requerem grande número de avaliações da função objetivo a fim de obter soluções apropriadas sem ficarem presos em ótimos locais.

Este trabalho propõe uma formulação multiobjetivo para PSP com Modelo HP. Considera-se o número total de interações hidrofóbicas e o grau de compactação dos aminoácidos hidrofóbicos da proteína. A fim de lidar de maneira adequada e eficiente com esses dois objetivos, foi utilizado um AE para Otimização Multiobjetivo (AEOM) baseado em subpopulações, chamado Algoritmo Evolutivo Multiobjetivo em Tabelas (AEMT), proposto por Delbem (2002). Resultados obtidos em problemas combinatoriais multiobjetivos mostram a eficácia do AEMT (Santos et al., 2010). Recentemente, um novo método (Ishibuchi et al., 2009) similar ao AEMT demonstrou a superioridade desse tipo de abordagem em relação aos AEOM baseados em dominância de Pareto. O diferencial do AEMT em relação à proposta apresentada por Ishibuchi et al. (2009) está na sua simplicidade de implementação (possuindo poucas alterações em relação ao AE padrão) e, principalmente, ao fato de trabalhar com populações pequenas. Trabalhos envolvendo o AEMT demostraram empiricamente que o tratamento de múltiplos objetivos combinado à sua característica de populações pequenas possibilita uma convergência relativamente rápida (Santos et al., 2010).

O AE implementado foi comparado com outros trabalhos encontrados na literatura (Patton et al., 1995; Johnson and Katikireddy, 2006; Custódio, 2008). Primeiramente, validou-se o modelo, comparando-o a um algoritmo genético (AG) padrão (Patton et al., 1995), utilizado como referência em diversos trabalhos. Também foram realizadas comparações com o AG padrão desenvolvido na etapa inicial do trabalho de Custódio (2008). Posteriormente, comparou-se a abordagem proposta com duas abordagens não tradicionais e mais recentes. A primeira (Johnson and Katikireddy, 2006) consiste em um AG híbrido e a segunda é uma hiper-heurística (Custódio, 2008). Resultados expostos neste artigo motivam trabalhos futuros combinando o AEMT proposto para PSP com aspectos de AG híbrido e de hiper-heurísticas.

Este artigo está organizado da seguinte maneira: na Seção 2 é descrito o problema de PSP. Os principais conceitos relativos a AEs são expostos na Seção 3. Na Seção 4 são revisados os principais trabalhos envolvendo AEs e Modelo HP e na Seção 5 é apresentada a metodologia proposta neste trabalho. Resultados experimentais são descritos na Seção 6 e a Seção 7 apresenta-se as conclusões deste artigo.

2 PREDIÇÃO DE ESTRUTURA TERCIÁRIA DE PROTEÍNAS

Uma proteína é uma macromolécula composta por uma sequência de um alfabeto de 20 aminoácidos. Essa sequência forma um encadeamento chamado *cadeia principal*, composta de átomos de carbono e nitrogênio interligados. A sequência $\mathrm{N}-\mathrm{C}_\alpha-\mathrm{C}$ é chamada *unidade peptídica* e a ligação entre dois aminoácidos recebe o nome de *ligação peptídica*. Os planos de ligação entre esses átomos podem dobrar, como consequência das possíveis rotações de seus ângulos, formando uma estrutura tridimensional específica.

Fatores como interações entre resíduos e forças intramoleculares tornam o processo de PSP computacionalmente complexo, exigindo a criação de modelos computacionais elaborados (Dill et al., 2007). Além disso, os cálculos requeridos para PSP utilizando esses modelos estão além dos limites dos computadores atuais ou de um futuro próximo (Berger and Leighton, 1998). Por esses motivos, diversas pesquisas estão focadas no desenvolvimento de modelos simplificados de proteínas: abstrações matemáticas que ocultam muitos aspectos e enfatizam o efeito de outros (Hart and Newman, 2006). Por meio de análises e simulações computacionais de tais modelos, avaliam-se resultados experimentais comparando-os a estruturas de moléculas reais, a fim de determinar se os aspectos enfatizados têm, de fato, o efeito esperado. Nesse contexto, receberam destaque na literatura os *modelos em redes*, descritos a seguir.

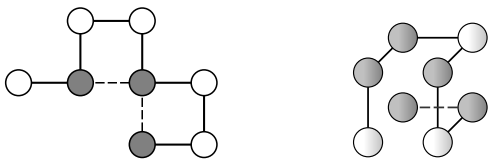
2.1 Modelos em Redes

Os modelos em redes[1], são modelos de proteínas com as seguintes simplificações (Clote and Backofen, 2000): (1) os monômeros (aminoácidos) são representados por pontos posicionados em vértices de uma malha; (2) as posições dos monômeros são restritas a posições em uma malhas (ou *lattice*) regulares; (3) o comprimento das ligações é único; (4) a distância mínima entre cada par de vértices é igual a 1; (5) a função de energia é simplificada.

O modelo em rede mais estudado é o *Modelo Hidrofóbico–Hidrofílico*, ou *Modelo HP*, desenvolvido por Lau and Dill (1989). Nesse modelo, o alfabeto dos 20 aminoácidos é reduzido a um alfabeto de apenas duas letras, H e P, onde H representa os aminoácidos *hidrofóbicos* e P representa os aminoácidos *polares* ou *hidrofílicos*. As topologias mais simples são a quadrática (Figura 1a), para duas dimensões, e a cúbica (Figura 1b), para três dimensões. Outras representações podem ser vistas em Hart and Newman (2006).

Nos modelos em redes, os aminoácidos podem ser classificados em *conectados* ou *vizinhos*. Dois aminoácidos nas posições i e j em uma sequência são conectados se $j = i + 1$ ou $j = i - 1$. O número de aminoácidos conectados é fixo e independente da conformação da sequência. Por outro lado, dois aminoácidos nas posições i e j são vizinhos topológicos se não são conectados e a distância Euclidiana entre i e j é igual a 1.

[1] Do inglês *lattice models.*



(a) Malha quadrática. (b) Malha cúbica.

Figura 1: Exemplo de conformação em malha quadrática (a) e malha cúbica (b) em Modelo HP. Esferas cinzas representam os aminoácidos H e as brancas, os P.

Assim, no Modelo HP, a energia livre E de uma conformação válida de comprimento n é dada pelo número negativo de vizinhos do tipo H–H. Com isso, a conformação de energia mínima é a mesma que maximiza o número de contatos H–H (Hart and Istrail, 1997). Formalmente, o estado nativo de uma molécula em malha é dada por:

$$E(n) = -\sum_{1 \leq i+1 < j \leq n} B_{i,j}\delta(r_i, r_j), \qquad (1)$$

sendo que

$$B_{i,j} = \begin{cases} 1 & \text{se os monômeros } i \text{ e } j \text{ são do tipo H;} \\ 0 & \text{caso contrário;} \end{cases}$$

$$\delta(r_i, r_j) = \begin{cases} 1 & \text{se os resíduos } i \text{ e } j \text{ não são conectados;} \\ 0 & \text{caso contrário.} \end{cases}$$

Por exemplo, para ambas conformações mostradas na Figura 1, o valor da energia livre é $E(8) = -2$.

Conceitualmente, o Modelo HP é relativamente simples se comparado com outras abordagens, além de possibilitar algumas extensões (Clote and Backofen, 2000; Hart and Newman, 2006). Apesar disso, computar o estado nativo que maximiza o número de vizinhos H–H é um problema $\mathcal{NP}$-Completo (Crescenzi et al., 1998; Berger and Leighton, 1998). Por essa razão, diversos métodos baseados em heurísticas de otimização têm sido aplicados, incluindo AEs (Unger, 2004), Algoritmos Meméticos (Bazzoli and Tettamanzi, 2004), algoritmos de Monte Carlo (Liang and Wong, 2001), Otimização por Colônias de Formigas (Shmygelska and Hoos, 2005) e por Enxame de Partículas (Kanj et al., 2009).

3 ALGORITMOS EVOLUTIVOS

AEs são meta-heurísticas populacionais de otimização que utilizam mecanismos inspirados na Teoria da Evolução, como seleção natural e sobrevivência de indivíduos mais adaptados, e na Genética, como mutação e recombinação (Eiben and Smith, 2003). A mimetização em computador do processo de evolução natural é um eficiente e sistemático método de busca por valores ótimos no espaço de possíveis soluções de um problema de otimização.

A vantagem dos AEs está na possibilidade de modelar um problema pela simples descrição de uma potencial solução do mesmo. Isso possibilita que AEs possam ser adaptados para uma grande diversidade de problemas complexos. Dentre os AEs atuais, os Algoritmos Genéticos (AGs) são os mais conhecidos (Goldberg, 1989; Holland, 1962). Essa técnica parte do seguinte princípio (Eiben and Smith, 2003): dada uma *população* de *indivíduos* (i.e., um conjunto de soluções), pressões do ambiente desencadeiam um processo de *seleção natural* (ou seja, um processo que privilegia em geral as melhores soluções até então encontradas), o que causa um incremento na taxa de adequação das soluções. Dada uma função a ser otimizada, gera-se aleatoriamente um conjunto de soluções que são elementos pertencentes ao domínio de uma função objetivo que, por sua vez, deve ser utilizada para avaliar a qualidade das soluções geradas. O valor resultante da avaliação da adequação é chamado *aptidão* (ou *fitness*).

Com base na aptidão, algumas das melhores soluções são selecionadas para constituírem uma nova *geração* pela aplicação de *operadores genéticos* (*recombinação* e *mutação*). A recombinação é um operador aplicado a duas ou mais soluções candidatas (chamadas *pais*) e resulta em uma ou mais novas soluções (*descendentes* ou *filhos*). Esse operador tenta combinar algumas partes do conteúdo genético de alguns indivíduos para gerar outros. Espera-se que as melhores partes sejam aproveitadas e, juntas, produzam melhores indivíduos. Já a mutação é aplicada em uma candidata a fim de gerar outra visando trazer novidades por meio de perturbações aleatórias. Ao final desse processo, as novas candidatas competem com as candidatas da geração anterior para, com base na aptidão, assumir um lugar na nova geração. Esse processo é repetido até que uma candidata apresente uma solução que seja suficientemente qualificada ou até que um número máximo de gerações seja obtido.

É importante observar que vários componentes de um processo evolutivo são estocásticos: a *seleção* favorece indivíduos mais bem adaptados (ou seja, com melhor *fitness*), mas existe também a possibilidade de serem selecionados outros indivíduos. A probabilidade de ocorrência desse operador também é aleatória seguindo uma distribuição uniforme.

Uma classe de problemas que tem despertado interesse em pesquisa com AEs é a dos problemas de *otimização multiobjetivo*. Nesses, a qualidade da solução é definida com base na sua adequação a diversos objetivos possivelmente conflitantes (Deb, 2001). Na prática, muitos métodos simplesmente atribuem pesos às diferentes funções objetivo, produzindo uma função mono-objetivo, de forma a viabilizar a aplicação de algoritmos de otimização mono-objetivo.

Esses métodos, no entanto, são dependentes da escolha adequada dos pesos. Por essa razão, métodos que tentam encontrar soluções que apresentam um compromisso com os vários objetivos sem a utilização de pesos passaram a ser explorados. Essas técnicas consideram não somente uma solução para o problema, mas sim um conjunto de soluções ótimas, denominado *fronteira de Pareto* (Deb, 2001).

A principal diferença entre os AEs tradicionais e os AEOMs é o operador de seleção, pois a comparação entre duas soluções é realizada de acordo com o conceito de dominância. No entanto, Ishibuchi et al. (2009) demonstraram que AEOMs que trabalham com dominância podem ter seu desempenho prejudicado em problemas combinatoriais com mais de dois objetivos.

Nesses problemas, todos os indivíduos podem deixar de dominar algum dos objetivos, o que torna a seleção ineficiente. Como consequência, a diversidade populacional diminui. Uma alternativa eficiente consiste em tratar um problema multiobjetivo como um conjunto de problemas mono-objetivo independentes mais um problema mono-objetivo utilizando uma função ponderação de todos os objetivos do problema original.

Trabalhos anteriores a Ishibuchi et al. (2009) já haviam chegado, empiricamente, a conclusões semelhantes, utilizando o chamado Algoritmo Evolutivo Multiobjetivo em Tabelas (AEMT) (Delbem, 2002; Santos et al., 2010).

3.1 AE Multiobjetivo em Tabela

O AEMT trabalha com várias subpopulações em paralelo organizadas em tabelas. Constrói-se uma subpopulação para cada função objetivo, mais uma subpopulação para a função ponderação. Um novo indivíduo entra na subpopulação $SubPop_i$ se ele for melhor que o pior indivíduo de $SubPop_i$ para o objetivo associado a $SubPop_i$.

O indivíduo selecionado para reprodução pode ser proveniente de qualquer subpopulação. Essa seleção aumenta a diversidade entre os indivíduos que se reproduzem de modo que as características de um indivíduo de uma subpopulação possam migrar para as demais (Santos et al., 2010). Como

consequência, aumenta-se a possibilidade do algoritmo evitar ótimos locais.

Um parâmetro importante para o desempenho do AEMT é o tamanho das subpopulações. Experimentalmente, mostrou-se um problema de larga-escala para o qual subpopulações com tamanho entre 1 e 5 indivíduos possibilitaram a convergência de maneira relativamente rápida para soluções que se aproximam da fronteira de Pareto (Santos et al., 2010).

4 AEs APLICADOS A PSP

Nesta seção, são descritos aspectos de implementação adotados em outros trabalhos relativos a AEs para PSP que são importantes para a abordagem aqui proposta. Nos AEs, os indivíduos (ou seja, as estruturas geradas), podem ser representadas por *coordenadas Cartesianas* (Unger and Moult, 1993) ou por *coordenadas internas* (Patton et al., 1995; Cotta, 2003).

Nas representações por coordenadas internas, a localização de um aminoácido é especificada em relação à posição espacial do aminoácido anterior. Assim, uma conformação é descrita por uma sequência de *movimentos* a partir de um ponto definido aleatoriamente na malha. Essa representação é classificada em duas categorias (Ye, 2007): *representação absoluta* e *representação relativa*.

Na representação absoluta, supondo uma sequência s de n resíduos, uma possível conformação para essa sequência é gerada por $n-1$ deslocamentos. Nos modelos de duas dimensões, as escolhas para os movimentos ficam restritas aos seguintes valores: *acima* (*up*, U), *abaixo* (*down*, D), *direita* (*right*, R) e *esquerda* (*left*, L); enquanto que em três dimensões existem as posições *frente* (*front*, F) e *atrás* (*back*, B).

A representação absoluta, no entanto, permite o surgimento de soluções infactíveis por meio de *movimentos de retorno*, ou seja, por permitir que um determinado movimento seja exatamente o oposto do movimento anterior (por exemplo, U seguido de D, L seguido de R, etc.). De modo a diminuir o número de soluções infactíveis, foi proposta a *representação relativa* (Patton et al., 1995) que limita o total de possíveis movimentos para o resíduo seguinte. O sistema de referência não é fixo e cada movimento é representado em relação à direção do movimento anterior. Por exemplo, no caso de malhas quadráticas, são permitidas apenas três direções: *em frente* (*forward*, F), *à direita* (*right-turn*, R) e *à esquerda* (*left-turn*, L), sendo que o primeiro movimento é sempre F que posiciona o segundo resíduo na posição $(1, 0)$ da rede. No caso de malhas cúbicas, são possíveis ainda os movimentos *acima* (U) e *abaixo* (D).

É importante observar que ambos mecanismos não eliminam totalmente o aparecimento de soluções infactíveis, uma vez que essas podem ser geradas colidindo-se resíduos não adjacentes em outros pontos da malha. Nesse caso, uma das técnicas mais utilizadas consiste em aplicar uma *função de penalidade* (Patton et al., 1995; Cotta, 2003; Bazzoli and Tettamanzi, 2004) às conformações com colisões, atribuindo um alto valor para a energia das mesmas. Desse modo, os métodos de busca tendem a desfavorecer soluções infactíveis, eliminando-as durante o processo de busca.

Um dos primeiros trabalhos envolvendo AEs e Modelo HP foi o AG híbrido proposto para malhas quadráticas e estendido posteriormente para malhas cúbicas (Unger and Moult, 1993). Esse AG incorpora um método de busca de Monte Carlo, utilizando representação absoluta e codificação binária para descrever o cromossomo.

Posteriormente, Patton et al. (1995) propuseram um AG não híbrido, com representação relativa. O algoritmo mantém as conformações infactíveis na população, introduzindo uma função de penalidade. Os resultados obtidos foram comparados aos de Unger and Moult (1993), obtendo número menor de avaliações da função objetivo e, em alguns casos, soluções de menor energia.

Outro trabalho foi desenvolvido por Custódio (2008), que utiliza representação absoluta e explora o efeito de diferentes operadores de recombinação e mutação baseados em conhecimento específico do domínio de PSP. Esse tipo de algoritmo tem sido chamado de hiper-heurística (Burke et al., 2003), de modo a destacar que são AEs que possuem conhecimento sobre o domínio do problema. Custódio (2008) também buscou manter a diversidade populacional como forma de realizar uma maior investigação do espaço de busca. Além disso, desenvolveu um mecanismo de *reparo* que corrige as soluções infactíveis. Assim, se um movimento gera uma conformação inválida, novos movimentos são gerados aleatoriamente até que uma conformação válida seja encontrada. Caso não existam conformações válidas possíveis para uma determinada configuração, o valor da aptidão do indivíduo é mantido constante e igual a 0. Os resultados obtidos demonstraram que a proposta de operadores específicos proporcionou redução na quantidade de avaliações da função objetivo.

Métodos híbridos utilizando AGs também foram propostos, como a utilização de *backtracking* para correção de colisões (Cotta, 2003; Johnson and Katikireddy, 2006) e busca local (Krasnogor, 2002; Bazzoli and Tettamanzi, 2004). Este trabalho explora uma modelagem evolutiva multiobjetivo para PSP em Modelo HP, descrita na Seção 5.

5 O AEMT PARA PSP COM MODELO HP

Nesta seção é descrita a abordagem proposta em relação à codificação das soluções, funções objetivo, e operadores genéticos. Além disso, é descrito o tratamento de conformações infactíveis, utilizando a Matriz de Conformações M_C.

No projeto do algoritmo proposto, foram utilizadas ideias descritas em trabalhos relacionados, apresentados anteriormente, de modo a desenvolver um AE eficiente e robusto. As decisões tomadas durante o desenvolvimento do algoritmo são apresentadas nas subseções seguintes.

5.1 Codificação das Soluções

O algoritmo proposto utiliza representação absoluta para malhas cúbicas. Apesar de não evitar movimentos de retorno, essa codificação tem sido utilizada em diversos trabalhos recentes (Custódio, 2008). Além disso, pode-se utilizar um procedimento simples para detectar colisões e decidir entre aceitá-las, porém penalizando a conformação gerada, ou corrigi-las, utilizando um mecanismo de correção adequado.

Foi adotada a codificação inteira para representar os indivíduos, com base nos movimentos ($U = 0$, $L = 1$, $F = 2$, $B = 3$, $R = 4$ e $D = 5$). Por meio dessa representação, os operadores genéticos tendem a modificar as conformações de maneira mais acentuada (Bazzoli and Tettamanzi, 2004; Custódio, 2008). Assim, uma sequência de n monômeros é codificada por meio de cromossomos de comprimento $n - 1$ (sendo que o primeiro monômero está na posição $(0, 0, 0)$ da malha). Uma vez que a sequência de monômeros é lida, pode-se descobrir, com base no cromossomo, se o resíduo é do tipo H ou P: basta observar que o elemento na posição i do cromossomo corresponde ao elemento $i-1$ da sequência.

5.2 Funções Objetivo

Quando um AE é utilizado, é necessário escolher uma função adequada para o cálculo de aptidão dos indivíduos. A primeira função busca maximizar o número de contatos H–H. Soluções infactíveis são permitidas na população, porém são penalizadas de modo que um menor *fitness* seja associado a essas soluções (Patton et al., 1995; Krasnogor, 2002; Cotta, 2003). De acordo com esse método, o *fitness* f de uma dada conformação c é determinado pela Equação 2 (Bazzoli and Tettamanzi, 2004).

$$f(c) = n_H(c) - \rho \cdot n_C(c), \tag{2}$$

tal que $n_H(c)$ é o número de contatos hidrofóbicos e $n_C(c)$ é o número de pontos da malha em que ocorrem colisões (i.e., posições ocupadas por dois ou mais monômeros). Por simplicidade, é atribuído o valor 1 ao termo ρ e a energia livre, nesse caso, é dada por $-f(c)$.

A fim de reduzir o número de cálculos necessários para estabelecer o valor da aptidão de um indivíduo, calcula-se o valor de $n_C(c)$ e, então, caso esse valor seja igual a 0 (ou seja, caso não existam colisões), calcula-se o valor de $n_H(c)$. Caso haja colisões (ou seja, caso $n_C(c)$ seja maior que 0), é atribuído 0 ao termo $n_H(c)$ (Gabriel and Delbem, 2009). Dessa maneira, evita-se calcular a aptidão para todos os indivíduos durante parte do processo evolutivo, e esses passam a ser classificados em dois grupos: os de *fitness* negativo (infactíveis) e os de *fitness* positivo (factíveis).

É importante observar que a função $f(c)$ atribui um valor inteiro à aptidão da conformação c. Apesar de amplamente adotado na literatura, esse método tem como limitação o fato de não identificar duas soluções distintas, porém com o mesmo número de contatos H–H. Como consequência, uma conformação c_1 mais compacta que c_2, mas ambas com o mesmo número de contatos, terá a mesma probabilidade de ser selecionada que c_2, apesar de ser potencialmente melhor.

Isso motivou o acréscimo de novos objetivos ao Modelo HP tradicional, de modo a avaliar o grau de compactação das conformações geradas. Assim, utiliza-se neste trabalho um objetivo adicional baseado na distância Euclidiana entre os monômeros hidrofóbicos da malha. Diferentemente do número de iterações hidrofóbicas, esse valor (um número real positivo) deve ser minimizado. Duas medidas de distância são investigadas nesse caso: (1) a maior distância $d_{max}(c)$ entre os resíduos presentes na malha e (2) a média das distâncias $d_{avg}(c)$ entre os resíduos.

Uma função de ponderação $\delta(c)$ é adotada para ponderar os diferentes objetivos considerados, conforme mostrado na Equação 3.

$$\delta(c) = \frac{f(c)}{d(c)}, \tag{3}$$

tal que $f(c)$ é dado pela Equação 2 e $d(c)$ é um valor escolhido entre $d_{max}(c)$ e $d_{avg}(c)$. Essa escolha é processada com base na taxa de variação desses valores entre duas gerações subsequentes. Assim, seja $d^t_{max}(c)$ o valor de $d_{max}(c)$ na t-ésima geração, a taxa de variação de $d_{max}(c)$ é dada por $\frac{d^t_{max}(c)}{d^{t-1}_{max}(c)}$. Analogamente, a taxa de variação de $d_{avg}(c)$ é dada por $\frac{d^t_{avg}(c)}{d^{t-1}_{avg}(c)}$. Uma vez processados esses cálculos, escolhe-se como $d(c)$ o valor de distância que apresentar maior variação de uma geração para a outra. Desse modo, evita-se, por exemplo, casos em que $d_{max}(c)$ não varia de uma geração para a outra, apesar da conformação c ser mais compacta uma vez que houve diminuição de outras distâncias consideradas no cálculo de $d_{avg}(c)$.

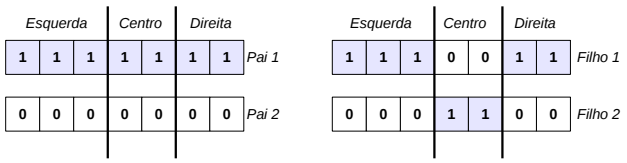


Figura 2: Exemplo de recombinação envolvendo dois pontos de corte.

5.3 Operadores Genéticos

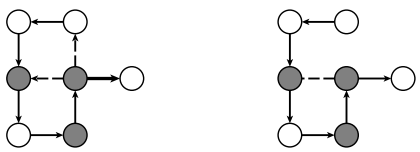
Krasnogor (2002) e Flores and Smith (2003) demonstram que um operador de recombinação de múltiplos pontos provê uma melhor troca de informações entre os indivíduos gerados. Um método proposto para escolher os pontos onde ocorrerá a troca de informação genética consiste em dividir o cromossomo em relação ao comprimento $l(s)$ da sequência s. Essa divisão é dada por $n = int(\frac{l(s)}{10})$, tal que $int(\bullet)$ é uma função que devolve o valor inteiro mais próximo ao resultado da divisão (Custódio, 2008). Assim, por exemplo, as sequências de 27 monômeros serão divididas em três pontos e as de 64 em seis pontos. Uma vez definidos os n pontos de corte, a recombinação é realizada conforme ilustrado na Figura 2, para $n = 2$.

Já o operador de mutação é aplicado aos novos indivíduos gerados pela recombinação. Neste trabalho, os indivíduos têm dois genes consecutivos modificados, a partir de um determinado ponto escolhido aleatoriamente. Isso impede, por exemplo, que a mutação afete um único gene, produzindo filhos com poucas informações novas, conforme demonstrado por Krasnogor (2002). Além disso, a taxa de mutação deve ser relativamente maior que o adotado em outros problemas resolvidos por AEs. Após selecionar-se aleatoriamente o ponto onde ocorrerá a mutação, aplica-se um operador chamado *random resetting* (Eiben and Smith, 2003), que substitui o gene selecionado por um valor escolhido aleatoriamente dentre os possíveis valores que os genes podem assumir.

5.4 Tratamento de Soluções Infactíveis

Usualmente, utiliza-se uma função de penalidade que reduz a probabilidade de indivíduos infactíveis serem selecionados para as próximas gerações. Experimentalmente, no entanto, pôde-se observar que a aptidão média da primeira geração é muito baixa, o que significa que todos os indivíduos apresentam colisões, ou seja, são infactíveis (Gabriel and Delbem, 2009). Essa observação motivou o desenvolvimento de um método para corrigir colisões e iniciar o processo de busca em regiões mais promissoras do espaço.

O mecanismo de correção de soluções infactíveis aqui proposto utiliza uma matriz (chamada Matriz de Conformações, M_C) para descodificar os cromossomos, de modo a evitar soluções infactíveis e, ao mesmo tempo, avaliar o número de contatos H–H. Nesse método, cada gene indexa uma posição de M_C, representando uma malha 3D. Caso a posição indexada já esteja ocupada, ou seja, caso já exista um resíduo na coordenada (x, y, z) de M_C, ocorre uma colisão. Para evitar a colisão, um novo valor é gerado e descodificado em uma nova posição espacial. A Figura 3 ilustra parte de uma conformação em duas dimensões.



(a) Detecção. (b) Correção.

Figura 3: Ilustração do processo de busca por soluções factíveis. Antes de definir a posição de j, testa-se as colisões: (3a) mostra as potenciais colisões (setas tracejadas) e a nova posição escolhida (seta grossa); (3b) mostra a conformação resultante e o contato H–H formado (linha tracejada).

Pode-se observar pela Figura 3 que, em caso de colisão, é verificado também o tipo do resíduo presente na posição, comparando-o com o resíduo anterior. Caso ambos os resíduos sejam do tipo H, significa que existe um contato H–H, que é contabilizado no cálculo de n_H (Equação 2). Desse modo, em uma única varredura do cromossomo, corrige-se as soluções infactíveis e calcula-se o valor da aptidão. Para garantir correções em todo o cromossomo em apenas uma varredura, deve-se aplicar um procedimento eficiente. Assim, utiliza-se um procedimento de correção que é limitado por um número constante de passos por resíduo, conforme descrito a seguir.

Considera-se um vetor de permutações que consiste em um vetor com os possíveis movimentos, codificados em valores inteiros. Para cada correção, gera-se uma permutação aleatória nesse vetor, de modo a evitar que a mesma sequência de movimentos seja aplicada em todo o cromossomo. Assim, quando se identifica uma colisão, cria-se uma nova permutação do vetor de deslocamentos e utiliza-se o valor contido na primeira posição do vetor como uma alternativa de movimento para eliminar a colisão. Caso esse movimento seja inválido, utiliza-se o movimento da próxima posição dentro do mesmo vetor, e assim por diante. Se todos os possíveis valores do vetor forem testados sem que uma solução factível seja encontrada, o indivíduo é eliminado. Com isso, a construção da estrutura é cancelada e uma nova estrutura começa a ser gerada.

A reinicialização da matriz 3D para avaliação de um novo indivíduo requer um tempo computacional cúbico em relação no tamanho do cromossomo. Para reduzir esse tempo, a cada

indivíduo gerado é atribuído um valor chamado *data de nascimento*, que é um contador de indivíduos. Assim, o primeiro indivíduo gerado tem data de nascimento igual a 1, o segundo igual a 2 e assim por diante. Na varredura de M_C para o indivíduo i, salva-se em cada posição de M_C um monômero com o valor i. Se a posição de um monômero em M_C já contiver i, então ocorreu uma colisão. Para um novo indivíduo j ($j \neq i$), pode-se utilizar a mesma M_C sem reinicializá-la, pois agora a colisão ocorrerá ao se tentar salvar j em uma posição de M_C que já contenha j.

Com isso, o custo da análise de cada cromossomo é linear em relação ao tamanho do mesmo. Além disso, utiliza-se um espaço de memória cúbico, em relação ao tamanho do indivíduo, para cada execução do AG. O tempo para alocação dessa memória é cúbico, porém é amortizado pelo total de indivíduos analisados sem necessidade de novas alocações ao longo de todo o processo de busca (Gabriel and Delbem, 2009).

6 EXPERIMENTOS E RESULTADOS

Experimentos computacionais foram conduzidos de modo a validar a abordagem proposta. Nesta seção, inicialmente, são apresentados os resultados que motivaram o desenvolvimento de um algoritmo multiobjetivo. Em seguida, apresenta-se os resultados utilizando o AEMT, o qual é avaliado para sequências de 27 e 64 monômeros utilizadas como *benchmark* na literatura. Faz-se uma comparação com trabalhos que propuseram AGs mono-objetivo (Patton et al., 1995; Custódio, 2008). Essa avaliação inicial busca comparar o impacto que o acréscimo de novos objetivos traz ao modelo. Nessa etapa também é avaliado o efeito da matriz de conformações (M_C) sobre o AE proposto. Em seguida, compara-se o AEMT com resultados mais recentes da literatura, que empregam estratégias híbridas (Johnson and Katikireddy, 2006) e de hiper-heurísticas (Custódio, 2008).

6.1 Testes com AEs Mono-objetivo

A adoção de uma estratégia multiobjetivo foi motivada por um conjunto de experimentos preliminares, que considerou duas implementações de um AG mono-objetivo. Na primeira, denominada $\mathcal{I}1$, utilizou-se apenas o número de contatos hidrofóbicos como forma de avaliar os indivíduos (Equação 2). A segunda, $\mathcal{I}2$, considerou a ponderação entre o número de contatos H–H e a distância $d(c)$ entre os monômeros da malha (Equação 3).

Foram utilizadas sequências de 27 resíduos (Unger and Moult, 1993), taxa de mutação de 3% e de recombinação de 80%. Para cada caso, o algoritmo foi executado 50 vezes com diferentes sementes para o gerador de números aleatórios. A população era de 120 indivíduos. Esses parâmetros foram escolhidos empiricamente, após diferentes avaliações.

A Tabela 1 compara os resultados obtidos por ambas implementações em relação ao número de contatos H–H, ao número de avaliações da função objetivo (quando o algoritmo atinge valores ótimos) e à taxa de acertos (o percentual de execuções nas quais o AG obteve a energia mínima). Em relação a $\mathcal{I}2$, é apresentado, ainda, o menor valor de $d(c)$ e o valor da energia mínima correspondente a esse[2].

Tabela 1: Comparação entre $\mathcal{I}1$ e $\mathcal{I}2$.

	$\mathcal{I}1$			$\mathcal{I}2$			
Seq.	Aval.	Min.	Ac.	Aval.	Min.	d	Ac.
273d.1	205.096	-9	6	318.961	-9	1,5 (-8)	76
273d.2	–	-9	0	38.401	-10	1,8 (-10)	92
273d.3	34.858	-8	30	47.041	-8	2,4 (-8)	100
273d.4	–	-14	0	1.000.723	-15	3,2 (-14)	76
273d.5	36.458	-8	16	29.281	-8	0,8 (-8)	100
273d.6	160.378	-11	6	734.641	-11	2,2 (-10)	16
273d.7	905.604	-13	2	948.721	-13	2,9 (-12)	46
273d.8	13.176	-4	92	17.041	-4	0,7 (-4)	100
273d.9	12.314	-7	16	73.681	-7	0,9 (-7)	100
273d.10	10.164	-11	14	152.401	-11	1,4 (-11)	94
Média	172.256		18	336.089		1,77	80
Mediana	35.658		10	113.041		1,65	93

É possível observar que a função de ponderação $\delta(c)$ em $\mathcal{I}2$ foi suficiente para aumentar a taxa de acertos do algoritmo (chegando a 100% em alguns casos). No entanto, $\mathcal{I}2$ não obteve maior eficiência em relação ao número de avaliações da função objetivo em muitos casos; além disso, observa-se que o menor valor de $d(c)$ não implica, necessariamente, um maior valor de $f(c)$. Isso indica que ambas métricas (f e d) são, em certo nível, independentes e que $d(c)$ tem grande influência sobre o aumento da robustez no modelo, porém ao custo de aumentar o número de avaliações.

De modo a explorar melhor as informações fornecidas por $f(c)$ e $d(c)$, foi proposta uma abordagem multiobjetivo que otimize separadamente ambos os valores, descrita a seguir.

6.2 Testes com o AEMT Proposto

Para otimizar os diferentes objetivos, foi utilizado o AEMT (Seção 3.1), para o qual foram criadas quatro subpopulações:

1. $SubPop_1$: armazena os indivíduos avaliados em relação a $f(c)$. Esse objetivo deve ser maximizado;
2. $SubPop_2$: armazena os indivíduos avaliados em relação a $d_{max}(c)$. Esse objetivo deve ser minimizado;

[2]Nessa e nas demais tabelas apresentadas neste artigo, as médias e medianas do número de avaliações consideram apenas os valores nos casos em que o ótimo foi obtido. Nos cálculos das taxas de acertos, os casos em que não se obteve o ótimo (taxa de acerto igual a zero) também são computados.

Tabela 2: Comparação dos resultados obtidos para sequências de 27 monômeros com o trabalho de $\mathcal{P}95$ e de $\mathcal{C}08$.

	$\mathcal{P}95$			$\mathcal{C}08_1$			AEMT			AEMT+M_C		
Seq.	Aval.	Min.	Ac.[3]	Aval.	Min.	Ac.	Aval.	Min.	Ac.	Aval.	Min.	Ac.
273d.1	27.786	−9	-	160.200	−9	48	8.956	−9	80	318.961	−9	76
273d.2	81.900	−10	-	49.800	−10	96	12.891	−10	80	38.401	−10	92
273d.3	16.757	−8	-	57.000	−8	98	2.294	−8	100	47.041	−8	100
273d.4	85.447	−15	-	207.400	−15	70	19.837	−15	62	1.000.723	−15	76
273d.5	8.524	−8	-	16.600	−8	100	6.660	−8	76	29.281	−8	100
273d.6	44.053	−11	-	403.000	−12	34	38.317	−11[4]	66	734.641	−11	16
273d.7	85.424	−13	-	504.600	−13	52	18.019	−13	72	948.721	−13	46
273d.8	3.603	−4	-	4.400	−4	100	1.848	−4	100	17.041	−4	100
273d.9	10.610	−7	-	67.000	−7	94	8.077	−7	100	73.681	−7	100
273d.10	16.282	−11	-	292.400	−11	62	11.006	−11	86	152.401	−11	94
Média	38.038,6		-	176.240		75,40	12.790,5		82,2	336.125,2		70
Mediana	22.271,5		-	113.600		82,00	9.981		80	113.041		84

[1] Valores não reportados por Patton et al. (1995).

[2] Observa-se que o valor -11 é o mínimo registrado pelos autores. Em 33% das repetições com a sequência `273d.6` obteve-se o valor -12.

3. $SubPop_3$: armazena os indivíduos avaliados em relação a $d_{avg}(c)$. Esse objetivo deve ser minimizado;
4. $SubPop_4$: armazena os indivíduos avaliados em relação a $\delta(c)$. Esse objetivo deve ser maximizado.

Por convenção (Santos et al., 2010), são utilizados tamanhos pequenos para as subpopulações. Nos experimentos foram utilizadas sequências de 27 e 64 monômeros (Unger and Moult, 1993). Duas categorias de experimentos foram conduzidas:

1. AEMT $\times$ AGs convencionais (Seção 6.2.1): comparação com trabalho de Patton et al. (1995) ($\mathcal{P}95$) e com a primeira etapa do trabalho de Custódio (2008) ($\mathcal{C}08_1$), no qual um AG padrão foi proposto utilizando-se os operadores genéticos descritos na Seção 5.3. Busca-se, assim, validar o modelo multiobjetivo, destacando sua contribuição frente a abordagens mono-objetivo;
2. AEMT $\times$ AGs com orientação adicional (Seção 6.2.2): comparação com o trabalho de Johnson and Katikireddy (2006) ($\mathcal{J}06$) e com a segunda etapa do trabalho de Custódio (2008) ($\mathcal{C}08_2$). Nesse caso, busca-se comparar o AEMT com duas abordagens mais recentes: um híbrido de AG com *backtracking* e uma hiperheurística (Burke et al., 2003), situando a proposta em relação aos últimos avanços. Porém, observa-se que o trabalho de Custódio (2008) não se trata de PSP *Ab initio*, uma vez que utiliza conhecimento adicional sobre o domínio do problema.

6.2.1 AEMT *versus* AGs convencionais

Resultados obtidos pelo AE multiobjetivo foram comparados com o AG desenvolvido por Patton et al. (1995) ($\mathcal{P}95$), cujos resultados são melhores que os do trabalho pioneiro de Unger and Moult (1993), e com o AG de Custódio (2008), que utiliza os mesmos operadores genéticos adotados neste artigo. Patton et al. (1995) utilizam recombinação de um ponto com probabilidade de 95% e mutação de um gene, com probabilidade de 0,1%. A taxa de recombinação em Custódio (2008) também é de 95% e a de mutação é de 5%.

No caso das sequências de 27 monômeros, duas estratégias de geração da população inicial foram investigadas: (1) gerar os indivíduos de forma aleatória sem utilizar métodos de correção; (2) utilizar M_C para corrigir os indivíduos infactíveis da população inicial. O método de correção não foi aplicado para os indivíduos de outras gerações, por provocar mudanças bruscas nas conformações geradas e por amenizar o efeito do operador de mutação, responsável pela variabilidade genética das populações. Utilizando o AEMT sem M_C, avaliou-se apenas 8 indivíduos, adotando-se os seguintes tamanhos de subpopulação: $SubPop_1 = 1$, $SubPop_2 = 2$, $SubPop_3 = 2$ e $SubPop_4 = 3$. No caso da utilização de M_C (AEMT+M_C), foram necessárias populações maiores, como será descrito posteriormente. Os resultados obtidos são comparados na Tabela 2.

É possível concluir que o AEMT foi capaz de encontrar o mesmo número de contatos hidrofóbicos que o AG padrão para todas as sequências menores. É importante destacar que Patton et al. (1995) e Custódio (2008) utilizam populações de 500 indivíduos enquanto que o AEMT utilizou apenas 8 indivíduos. Isso demonstra o efeito positivo da abordagem multiobjetivo proposta em relação à manutenção da diversidade populacional. A taxa de acerto manteve-se superior a 60% em todos os casos nos quais M_C não foi aplicado.

No entanto, deve-se observar que os resultados envolvendo M_C não foram muito satisfatórios. Isso porque o mecanismo de correção de soluções infactíveis gera um conjunto de indivíduos com aptidão baixa, ou seja, que possuem poucas (ou nenhuma) interações H–H. A combinação AEMT+M_C cria uma tendência de iniciar a busca em regiões de aptidão melhor que no caso de totalmente aleatório. Porém, essas regiões provavelmente correspondem a ótimos locais, o que

contribui para prender a busca em regiões subótimas. Tal comportamento foi sugerido por (Flores and Smith, 2003), mas pode ser amenizado pela utilização de populações iniciais grandes, contendo cerca de 10.000 indivíduos, por exemplo. No entanto, a escolha de grande número de indivíduos resulta em um grande número de avaliações da função objetivo.

Ainda em relação a comparações com os AGs convencionais, foram utilizadas sequências maiores, de 64 monômeros, com os seguintes tamanhos de subpopulação: $SubPop_1 = 1$, $SubPop_2 = 1$, $SubPop_3 = 1$ e $SubPop_4 = 2$. A taxa de mutação foi de 40% e a de recombinação de 50%. A Tabela 3 apresenta os resultados obtidos. Nesse caso, a matriz M_C não foi empregada.

Pode-se observar que, para as instâncias de 64 monômeros, a taxa de acertos caiu para 11% e o AEMT não chegou a encontrar o ótimo para a sequência `643d.6`. Destaca-se, no entanto, que a taxa de acerto não é apresentada por Patton et al. (1995). Em experimentos preliminares, que buscaram reproduzir o AG $\mathcal{P}95$, a taxa de acertos foi muito baixa, quase sempre inferior a 1%. Adicionalmente, observa-se que o algoritmo $\mathcal{C}08_1$ não encontrou o mínimo em três casos, além de apresentar taxa de acerto de 7,20%, em média. Isso se deve ao fato de o espaço de busca com 64 monômeros ser muito maior que o de 27 monômeros, tornando o problema mais difícil.

6.2.2 AEMT *versus* AGs com Orientação Adicional

Diversas técnicas têm sido empregadas de modo a melhorar o desempenho de AEs em relação a PSP apresentando, recentemente, resultados relevantes.

Johnson and Katikireddy (2006) ($\mathcal{J}06$) criaram um mecanismo de *backtracking* que possibilitou resultados superiores aos de Patton et al. (1995) para as sequências menores. Nos experimentos, utilizam recombinação de dois pontos com probabilidade de 95%, mutação de um gene, com probabilidade de 0,1% e população cujo tamanho varia entre 1000 e 1600 indivíduos. O AEMT foi executado utilizando a mesma configuração descrita na Seção 6.2.1, com 8 indivíduos. Na Tabela 4 é apresentada a comparação entre esses trabalhos.

Observa-se que o AEMT encontrou as mesmas soluções de Johnson and Katikireddy (2006) (inclusive em relação a `273d.6`), utilizando, na maioria dos casos, menos avaliações da função objetivo (apenas em quatro casos o número de avaliações foi superior). Porém, a média de avaliações de ambos os casos foi bastante similar. Deve-se observar que o AEMT não é híbrido com outra técnica. Assim, a hibridação do AEMT com um mecanismo de *backtracking* similar ao de Johnson and Katikireddy (2006) deve ser uma abordagem promissora para PSP com Modelo HP.

Finalmente, na Tabela 5 são comparados o AEMT e a segunda etapa do trabalho de Custódio (2008) ($\mathcal{C}08_2$), que utiliza populações de 500 indivíduos e os operadores genéticos descritos na Seção 5.3.

Merece destaque, novamente, o pequeno número de indivíduos avaliados em cada geração (apenas 5). Além disso, destaca-se que o foco do trabalho de Custódio (2008) foi o desenvolvimento de uma hiper-heurística (Burke et al., 2003), pois seus operadores genéticos utilizam conhecimento adicional do domínio do problema de PSP, realizando modificações consideradas mais promissoras segundo esse conhecimento. Isso justifica os melhores resultados. Deve-se observar que a utilização de conhecimento adicional sobre o problema é um aspecto positivo para a solução do problema de PSP. No entanto, isso não segue a proposta original de somente utilizar como informação a sequência de aminoácidos para se determinar a estrutura de proteína. Por outro lado, os resultados do AEMT e de Custódio (2008) em relação ao AG convencional sugerem que a combinação de ambos pode gerar uma abordagem promissora para PSP com Modelo HP que não seja *Ab initio*.

7 CONCLUSÕES

O principal objetivo deste trabalho foi o desenvolvimento de uma abordagem eficiente e robusta para PSP em Modelo HP. Pôde-se observar a limitação do Modelo HP em relação ao tratamento de soluções subótimas, uma vez que o modelo não considera que duas soluções com mesmo número de contatos hidrofóbicos possam ser distinguidas adequadamente. Experimentalmente, demonstrou-se que uma nova métrica, que calcula a compactação dos monômeros nas estruturas de proteínas, contribui para uma avaliação mais apropriada da estrutura. Com isso, a taxa de acertos do AE para PSP com Modelo HP aumenta consideravelmente.

O conhecimento adquirido com esses experimentos motivou a proposta e o desenvolvimento de um AEOM. Foi utilizado o AEMT que considera múltiplos objetivos e avalia, em geral, um número relativamente pequeno de indivíduos. As funções objetivo avaliadas pelo algoritmo consideram, simultaneamente, o número de contatos hidrofóbicos e a distância entre os monômeros das conformações geradas. Resultados obtidos pelo AEMT com sequências de *benchmark* pequenas apresentaram um número consideravelmente menor de avaliações da função de aptidão em relação ao AG padrão mono-objetivo.

Como forma de aumentar a robustez do AEMT, foi proposto um método de tratamento de soluções inválidas, de modo a

Tabela 3: Comparação dos resultados obtidos com o AEMT para sequências de 64 monômeros com $\mathcal{P}95$ e de $\mathcal{C}08_1$.

	$\mathcal{P}95$			$\mathcal{C}08_1$			AEMT		
Seq.	Aval.	Min.	Ac.[5]	Aval.	Min.	Acertos	Aval.	Min.	Ac.
643d.1	433.533	−27	-	-	−25	0	654.320	−27	4
643d.2	167.017	−30	-	1.753.000	−30	2	1.117.558	−30	22
643d.3	172.192	−38	-	583.400	−39	14	244.936	−38	12
643d.4	107.143	−34	-	4.007.600	−34	6	2.374.943	−34	10
643d.5	154.168	−36	-	-	−35	0	2.266.601	−36	2
643d.6	454.727	−31	-	513.800	−34	20	-	−30	0
643d.7	320.396	−25	-	1.988.200	−26	12	644.915	−25	26
643d.8	315.036	−34	-	-	−32	0	2.646.709	−34	4
643d.9	151.705	−33	-	604.000	−36	8	623.314	−33	12
643d.10	191.019	−26	-	1.724.800	−27	10	139.571	−26	18
Média	246.693,6		-	1.596.400		7,2	1.190.318,56		11
Mediana	181.605,5		-	1.724.800		7	654.320		11

[5] Valores não reportados por Patton et al. (1995).

buscar conformações adequadas com menor esforço computacional. Criou-se uma rotina que, auxiliada por uma matriz (M_C) e por um vetor de permutações, reconstrói eficientemente as estruturas infactíveis, gerando exclusivamente soluções factíveis. Resultados combinado essa estratégia com o AEMT, no entanto, não apresentaram melhora significante no desempenho do algoritmo aqui proposto em relação ao número de avaliações da função objetivo, devido à grande quantidade de indivíduos necessários na geração inicial para que o AE não ficasse preso em ótimos locais.

Com base nessas limitações e em outras características desejáveis, indicam-se os principais desenvolvimentos futuros a serem seguidos. Dentre esses, destacam-se:

1. Investigação de métodos de controle de diversidade populacional (Tragante do Ó, 2009), de modo que a busca por regiões promissoras não fique presa em ótimos locais quando se utiliza o AEMT+M_C;

2. Adoção de operadores de mutação diferenciados, de modo a gerar novas conformações seguindo heurísticas mais adequadas (Custódio, 2008), no caso em que se busque soluções não *Ab initio* para PSP;

3. Implementação de técnicas de busca local, cujo impacto positivo tem sido demonstrado em trabalhos relacionados (Krasnogor, 2002; Bazzoli and Tettamanzi, 2004);

4. Pesquisa do impacto de técnicas de aumento de desempenho de AEs desenvolvidas recentemente, como os chamados algoritmos de estimação de distribuição (Santana et al., 2008; Melo, 2009).

Além disso, os resultados da implementação do AEMT motivam a proposta desse algoritmo para modelos mais elaborados e que, portando, contribuem de forma mais significativa com o progresso na área de PSP com Modelo HP.

Tabela 4: Comparação do AEMT com $\mathcal{J}06$.

	$\mathcal{J}06$			AEMT		
Seq.	Aval.	Min.	Ac.[6]	Aval.	Min.	Ac.
273d.1	15.854	−9	-	8.956	−9	80
273d.2	19.965	−10	-	12.891	−10	80
273d.3	7.991	−8	-	2.294	−8	100
273d.4	23.525	−15	-	19.837	−15	62
273d.5	3.561	−8	-	6.660	−8	76
273d.6	14.733	−11	-	38.317	−11	66
273d.7	23.112	−13	-	18.019	−13	72
273d.8	889	−4	-	1.848	−4	100
273d.9	5.418	−7	-	8.077	−7	100
273d.10	5.592	−11	-	11.006	−11	86
Média	12.064		-	12.790,5		82,2
Mediana	11.362		-	9.981		80

[6] Valores não reportados por Johnson and Katikireddy (2006).

Tabela 5: Comparação do AEMT com $\mathcal{C}08_2$.

	$\mathcal{C}08_2$			AEMT		
Seq.	Aval.	Min.	Ac.[7]	Aval.	Min.	Ac.
Seq.	Aval.	Min.	Ac.	Aval.	Min.	Ac.
643d.1	228.000	−31	98	654.320	−27	4
643d.2	115.000	−36	100	1.117.558	−30	22
643d.3	87.000	−44	100	244.936	−38	12
643d.4	159.000	−39	100	2.374.943	−34	10
643d.5	134.000	−40	100	2.266.601	−36	2
643d.6	177.000	−33	88	–	−30	0
643d.7	76.500	−28	100	644.915	−25	26
643d.8	178.500	−36	62	2.646.709	−34	4
643d.9	74.500	−38	100	623.314	−33	12
643d.10	82.500	−31	100	139.571	−26	18
Média	131.250		94,8	1.190.318,56		11
Mediana	124.750		100	654.320		11

[7] Valores reportados em relação aos novos mínimos obtidos.

AGRADECIMENTOS

Os autores agradecem à Profª Drª Telma Woerle de Lima Soares e a Daniel Rodrigo Ferraz Bonetti, por valiosas discussões. À FAPESP e à CAPES pelo apoio financeiro.

REFERÊNCIAS

Bazzoli, A. and Tettamanzi, A. (2004). A memetic algorithm for protein structure prediction in a 3D-lattice HP model, *EvoWorkshops 2004*, Vol. 3005 of *LNCS*, Springer, pp. 1–10.

Berger, B. and Leighton, T. (1998). Protein folding in the hydrophobic–hydrophilic (HP) model is NP-complete, *Journal of Computational Biology* **5**(1): 27–40.

Burke, E. K., Hart, E., Kendall, G., Newall, J., Ross, P. and Schulenburg, S. (2003). Hyper-heuristics: An emerging direction in modern search technology, *Handbook of Metaheuristics*, Kluwer Academic, pp. 457–474.

Clote, P. and Backofen, R. (2000). *Computational Molecular Biology: An Introduction*, John Wiley & Sons.

Cotta, C. (2003). Protein structure prediction using evolutionary algorithms hybridized with backtracking, *IWANN '03*, Springer, pp. 321–328.

Crescenzi, P., Goldman, D., Papadimitriou, C., Piccolboni, A. and Yannakakis, M. (1998). On the complexity of protein folding, *Journal of Computational Biology* **5**(3): 423–466.

Custódio, F. L. (2008). *Algoritmos genéticos para predição* Ab Initio *de Estruturas de Proteínas*, PhD thesis, LNCC, Petrópolis, RJ.

Deb, K. (2001). *Multi-objective Optimization Using Evolutionary Algorithms*, John Wiley & Sons.

Delbem, A. C. B. (2002). *Restabelecimento de Energia em Sistemas de Distribuição por Algoritmo Evolucionário Associado a Cadeias de Grafos*, PhD thesis, EESC/USP, São Carlos, SP.

Dill, K. A., Ozkan, S. B., Weikl, T. R., Chodera, J. D. and Voelz, V. A. (2007). The protein folding problem: when will it be solved?, *Curr. Opin. Struct. Biol.* **17**: 342–346.

Eiben, A. . E. and Smith, J. E. (2003). *Introduction to Evolutionary Computing*, Springer.

Flores, S. D. and Smith, J. (2003). Study of fitness landscapes for the HP model of protein structure prediction, *CEC 2003*, Vol. 4, IEEE, pp. 2338–2345.

Gabriel, P. H. R. and Delbem, A. C. B. (2009). Representations for evolutionary algorithms applied to protein structure prediction problem using HP model, *BSB 2009*, Vol. 5676 of *LNCS*, Springer, pp. 97–108.

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley.

Hart, W. E. and Istrail, S. (1997). Robust proofs of NP-hardness for protein folding: General lattices and energy potentials, *Journal of Computational Biology* **4**(1): 1–22.

Hart, W. E. and Newman, A. (2006). Protein structure prediction with lattice models, *Handbook of Molecular Biology*, CRC Press, chapter 30, pp. 30.1–30.24.

Holland, J. H. (1962). Outline for a logical theory of adaptive systems, *J. ACM* **9**(3): 297–314.

Ishibuchi, H., Sakane, Y., Tsukamoto, N. and Nojima, Y. (2009). Adaptation of scalarizing functions in MOEA/D: An adaptive scalarizing function-based multiobjective evolutionary algorithm, *EMO 2009*, Vol. 5467 of *LNCS*, Springer, pp. 438–452.

Johnson, C. M. and Katikireddy, A. (2006). A genetic algorithm with backtracking for protein structure prediction, *GECCO 2006*, Vol. 2, ACM, EUA, pp. 299–300.

Kanj, F., Mansour, N., Khachfe, H. and Abu-Khzam, F. (2009). Protein structure prediction in the 3D HP model, *AICCSA-2009*, EUA, pp. 732–736.

Krasnogor, N. (2002). *Studies on the Theory and Design Space of Memetic Algorithms*, PhD thesis, CMES/UWE, Bristol, Inglaterra.

Lau, K. F. and Dill, K. A. (1989). A lattice statistical mechanics model of the conformational and sequence spaces of proteins, *Macromolecules* **22**: 3986–3997.

Liang, F. and Wong, W. H. (2001). Evolutionary monte carlo for protein folding simulations, *Chemical Physics* **115**(7): 444–451.

Lopes, H. S. (2008). Evolutionary algorithms for the protein folding problem: A review and current trends, *Comp. Intel. in Biomed. & Bioinform.*, Vol. 151 of *SCI*, Springer, pp. 297–315.

Melo, V. V. (2009). *Técnicas de aumento de eficiência para metaheurísticas aplicadas a otimização global cuontínua e discreta*, PhD thesis, ICMC/USP, São Carlos, SP.

Patton, A. L., Punch III, W. F. and Goodman, E. D. (1995). A standard GA approach to native protein conformation prediction, *Proceedings of ICGA*, pp. 574–581.

Santana, R., Larrañaga, P. and Lozano, J. A. (2008). Protein folding in simplified models with estimation of distribution algorithms, *IEEE Trans. Evol. Comp.* **12**(4): 418–438.

Santos, A. C., Delbem, A. C. B., London, Jr., J. B. A. and Bretas, N. G. (2010). Node-depth encoding and multiobjective evolutionary algorithm applied to large-scale distribution system reconfiguration, *IEEE Trans. Power Syst.* **25**(3): 1254–1265.

Shmygelska, A. and Hoos, H. H. (2005). An ant colony optimization algorithm for the 2D and 3D hydrophobic polar protein folding problem, *BMC Bioinf.* **6**(30): 1–22.

Tragante do Ó, V. (2009). *Técnicas de controle de diversidade de populações em algoritmos genéticos para determinação de estruturas de proteínas*, Master's thesis, FFCLRP/USP, Ribeirão Preto, SP.

Unger, R. (2004). The genetic algorithm approach to protein structure prediction, *Struct. Bond.* **110**: 153–175.

Unger, R. and Moult, J. (1993). A genetic algorithm for 3D protein folding simulations, *Proceedings of ICGA*, pp. 581–588.

Ye, J. (2007). *A review of artificial intelligence techniques applied to protein structure prediction*, Master's thesis, School of Computing Science, SFU, Vancouver, CA.