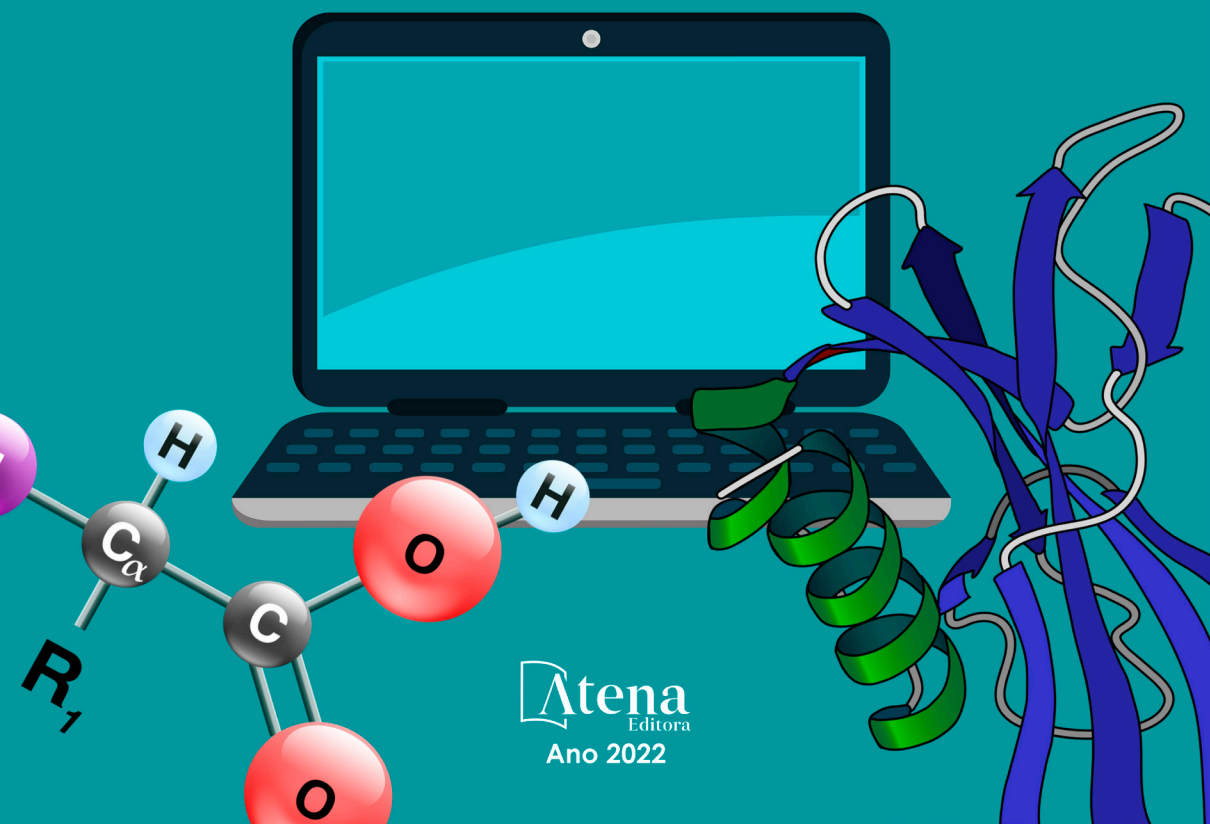


MÉTODOS DE MODELAGEM DA ESTRUTURA TRIDIMENSIONAL DE PROTEÍNAS

Editado por Kauê Santana e Cláudio Nahum

UM GUIA TEÓRICO E PRÁTICO



Atena
Editora
Ano 2022

MÉTODOS DE MODELAGEM DA ESTRUTURA TRIDIMENSIONAL DE PROTEÍNAS

Editado por Kauê Santana e Cláudio Nahum

UM GUIA TEÓRICO E PRÁTICO



Atena
Editora
Ano 2022

Editora chefe

Profª Drª Antonella Carvalho de Oliveira

Editora executiva

Natalia Oliveira

Assistente editorial

Flávia Roberta Barão

Bibliotecária

Janaina Ramos

Projeto gráfico

Bruno Oliveira

Camila Alves de Cremo

Luiza Alves Batista

Natália Sandrini de Azevedo

Imagens da capa

iStock

Edição de arte

Luiza Alves Batista

2022 by Atena Editora

Copyright © Atena Editora

Copyright do texto © 2022 Os autores

Copyright da edição © 2022 Atena Editora

Direitos para esta edição cedidos à Atena Editora pelos autores.

Open access publication by Atena Editora



Todo o conteúdo deste livro está licenciado sob uma Licença de Atribuição Creative Commons. Atribuição-Não-Comercial-NãoDerivativos 4.0 Internacional (CC BY-NC-ND 4.0).

O conteúdo dos artigos e seus dados em sua forma, correção e confiabilidade são de responsabilidade exclusiva dos autores, inclusive não representam necessariamente a posição oficial da Atena Editora. Permitido o *download* da obra e o compartilhamento desde que sejam atribuídos créditos aos autores, mas sem a possibilidade de alterá-la de nenhuma forma ou utilizá-la para fins comerciais.

Todos os manuscritos foram previamente submetidos à avaliação cega pelos pares, membros do Conselho Editorial desta Editora, tendo sido aprovados para a publicação com base em critérios de neutralidade e imparcialidade acadêmica.

A Atena Editora é comprometida em garantir a integridade editorial em todas as etapas do processo de publicação, evitando plágio, dados ou resultados fraudulentos e impedindo que interesses financeiros comprometam os padrões éticos da publicação. Situações suspeitas de má conduta científica serão investigadas sob o mais alto padrão de rigor acadêmico e ético.

Conselho Editorial**Ciências Biológicas e da Saúde**

Profª Drª Aline Silva da Fonte Santa Rosa de Oliveira – Hospital Federal de Bonsucesso

Profª Drª Ana Beatriz Duarte Vieira – Universidade de Brasília

Profª Drª Ana Paula Peron – Universidade Tecnológica Federal do Paraná

Prof. Dr. André Ribeiro da Silva – Universidade de Brasília

Profª Drª Anelise Levay Murari – Universidade Federal de Pelotas

Prof. Dr. Benedito Rodrigues da Silva Neto – Universidade Federal de Goiás



Prof. Dr. Cirênio de Almeida Barbosa – Universidade Federal de Ouro Preto
Prof^o Dr^a Daniela Reis Joaquim de Freitas – Universidade Federal do Piauí
Prof^o Dr^a Débora Luana Ribeiro Pessoa – Universidade Federal do Maranhão
Prof. Dr. Douglas Siqueira de Almeida Chaves – Universidade Federal Rural do Rio de Janeiro
Prof. Dr. Edson da Silva – Universidade Federal dos Vales do Jequitinhonha e Mucuri
Prof^o Dr^a Elizabeth Cordeiro Fernandes – Faculdade Integrada Medicina
Prof^o Dr^a Eleuza Rodrigues Machado – Faculdade Anhanguera de Brasília
Prof^o Dr^a Elane Schwinden Prudêncio – Universidade Federal de Santa Catarina
Prof^o Dr^a Eysler Gonçalves Maia Brasil – Universidade da Integração Internacional da Lusofonia Afro-Brasileira
Prof. Dr. Ferlando Lima Santos – Universidade Federal do Recôncavo da Bahia
Prof^o Dr^a Fernanda Miguel de Andrade – Universidade Federal de Pernambuco
Prof. Dr. Fernando Mendes – Instituto Politécnico de Coimbra – Escola Superior de Saúde de Coimbra
Prof^o Dr^a Gabriela Vieira do Amaral – Universidade de Vassouras
Prof. Dr. Gianfábio Pimentel Franco – Universidade Federal de Santa Maria
Prof. Dr. Helio Franklin Rodrigues de Almeida – Universidade Federal de Rondônia
Prof^o Dr^a Iara Lúcia Tescarollo – Universidade São Francisco
Prof. Dr. Igor Luiz Vieira de Lima Santos – Universidade Federal de Campina Grande
Prof. Dr. Jefferson Thiago Souza – Universidade Estadual do Ceará
Prof. Dr. Jesus Rodrigues Lemos – Universidade Federal do Piauí
Prof. Dr. Jônatas de França Barros – Universidade Federal do Rio Grande do Norte
Prof. Dr. José Aderval Aragão – Universidade Federal de Sergipe
Prof. Dr. José Max Barbosa de Oliveira Junior – Universidade Federal do Oeste do Pará
Prof^o Dr^a Juliana Santana de Curcio – Universidade Federal de Goiás
Prof^o Dr^a Lívia do Carmo Silva – Universidade Federal de Goiás
Prof. Dr. Luís Paulo Souza e Souza – Universidade Federal do Amazonas
Prof^o Dr^a Magnólia de Araújo Campos – Universidade Federal de Campina Grande
Prof. Dr. Marcus Fernando da Silva Praxedes – Universidade Federal do Recôncavo da Bahia
Prof^o Dr^a Maria Tatiane Gonçalves Sá – Universidade do Estado do Pará
Prof. Dr. Maurilio Antonio Varavallo – Universidade Federal do Tocantins
Prof^o Dr^a Mylena Andréa Oliveira Torres – Universidade Ceuma
Prof^o Dr^a Natiéli Piovesan – Instituto Federaci do Rio Grande do Norte
Prof. Dr. Paulo Inada – Universidade Estadual de Maringá
Prof. Dr. Rafael Henrique Silva – Hospital Universitário da Universidade Federal da Grande Dourados
Prof^o Dr^a Regiane Luz Carvalho – Centro Universitário das Faculdades Associadas de Ensino
Prof^o Dr^a Renata Mendes de Freitas – Universidade Federal de Juiz de Fora
Prof^o Dr^a Sheyla Mara Silva de Oliveira – Universidade do Estado do Pará
Prof^o Dr^a Suely Lopes de Azevedo – Universidade Federal Fluminense
Prof^o Dr^a Vanessa da Fontoura Custódio Monteiro – Universidade do Vale do Sapucaí
Prof^o Dr^a Vanessa Lima Gonçalves – Universidade Estadual de Ponta Grossa
Prof^o Dr^a Vanessa Bordin Viera – Universidade Federal de Campina Grande
Prof^o Dr^a Welma Emídio da Silva – Universidade Federal Rural de Pernambuco



Métodos de modelagem da estrutura tridimensional de proteínas: um guia teórico e prático

Diagramação: Camila Alves de Cremo
Correção: Maiara Ferreira
Indexação: Amanda Kelly da Costa Veiga
Revisão: RevisAtena
Organizadores: Kauê Santana
Cláudio Nahum

Dados Internacionais de Catalogação na Publicação (CIP)

M593 Métodos de modelagem da estrutura tridimensional de proteínas: um guia teórico e prático / Organizadores Kauê Santana, Cláudio Nahum. – Ponta Grossa - PR: Atena, 2022.

Formato: PDF

Requisitos de sistema: Adobe Acrobat Reader

Modo de acesso: World Wide Web

Inclui bibliografia

ISBN 978-65-258-0582-5

DOI: <https://doi.org/10.22533/at.ed.825222810>

1. Proteínas. I. Santana, Kauê (Organizador). II. Nahum, Cláudio (Organizador). III. Título.

CDD 613.282

Elaborado por Bibliotecária Janaina Ramos – CRB-8/9166

Atena Editora
Ponta Grossa – Paraná – Brasil
Telefone: +55 (42) 3323-5493
www.arenaeditora.com.br
contato@arenaeditora.com.br



DECLARAÇÃO DOS AUTORES

Os autores desta obra: 1. Atestam não possuir qualquer interesse comercial que constitua um conflito de interesses em relação ao artigo científico publicado; 2. Declaram que participaram ativamente da construção dos respectivos manuscritos, preferencialmente na: a) Concepção do estudo, e/ou aquisição de dados, e/ou análise e interpretação de dados; b) Elaboração do artigo ou revisão com vistas a tornar o material intelectualmente relevante; c) Aprovação final do manuscrito para submissão.; 3. Certificam que os artigos científicos publicados estão completamente isentos de dados e/ou resultados fraudulentos; 4. Confirmam a citação e a referência correta de todos os dados e de interpretações de dados de outras pesquisas; 5. Reconhecem terem informado todas as fontes de financiamento recebidas para a consecução da pesquisa; 6. Autorizam a edição da obra, que incluem os registros de ficha catalográfica, ISBN, DOI e demais indexadores, projeto visual e criação de capa, diagramação de miolo, assim como lançamento e divulgação da mesma conforme critérios da Atena Editora.



DECLARAÇÃO DA EDITORA

A Atena Editora declara, para os devidos fins de direito, que: 1. A presente publicação constitui apenas transferência temporária dos direitos autorais, direito sobre a publicação, inclusive não constitui responsabilidade solidária na criação dos manuscritos publicados, nos termos previstos na Lei sobre direitos autorais (Lei 9610/98), no art. 184 do Código Penal e no art. 927 do Código Civil; 2. Autoriza e incentiva os autores a assinarem contratos com repositórios institucionais, com fins exclusivos de divulgação da obra, desde que com o devido reconhecimento de autoria e edição e sem qualquer finalidade comercial; 3. Todos os e-book são *open access*, *desta forma* não os comercializa em seu site, sites parceiros, plataformas de *e-commerce*, ou qualquer outro meio virtual ou físico, portanto, está isenta de repasses de direitos autorais aos autores; 4. Todos os membros do conselho editorial são doutores e vinculados a instituições de ensino superior públicas, conforme recomendação da CAPES para obtenção do Qualis livro; 5. Não cede, comercializa ou autoriza a utilização dos nomes e e-mails dos autores, bem como nenhum outro dado dos mesmos, para qualquer finalidade que não o escopo da divulgação desta obra.



Dedico este livro à minha esposa, Lidiane Diniz, pois sem a sua dedicação e amor não seria possível concluir este trabalho.

Prof. Dr. Kauê Santana

APRESENTAÇÃO

A predição da estrutura de proteínas é, sem dúvida, um assunto-chave para quem atua nas áreas de modelagem molecular, bioinformática e biologia estrutural, dada a versatilidade de funções desempenhada por estes biopolímeros em processos biológicos, assim como suas diferentes aplicações farmacêuticas, industriais e biotecnológicas. Com este livro, pretendemos não somente atender às necessidades práticas relacionadas à modelagem e à análise das estruturas tridimensionais de proteínas, conteúdo ainda carente em língua portuguesa, mas também fornecer os principais conceitos necessários para a compreensão dos métodos, suas aplicações, assim como limitações. Focamos o conteúdo e a linguagem do texto para a graduação e devido ao grande número de expressões da língua inglesa presentes na literatura científica, incluímos os termos mais comumente utilizados dessa língua, sempre dando preferência à tradução deles para o português.

O livro traz uma linguagem acessível, incluindo referências atualizadas sobre os métodos, além de estar ricamente ilustrado, ainda traz, no final, um glossário de termos técnicos que enriquecerão a leitura e a compreensão do leitor.

Temos certeza que será uma importante fonte de consulta para alunos de graduação e pós-graduação, assim como para todos os interessados na área.

Prof. Kauê Santana
Prof. Cláudio Nahum

PREFÁCIO

A obra intitulada *Métodos de Modelagem da Estrutura Tridimensional de Proteínas: Um Guia Teórico-prático*, produzido por Kauê Santana e Cláudio Nahum, faz um apanhado geral das técnicas e abordagens computacionais mais avançadas usadas na predição e estudo das estruturas das proteínas. O livro está organizado de forma que ao leitor, inicialmente, sejam apresentados aos conceitos básicos inerentes aos métodos usados em diversos programas e algoritmos empregados na modelagem. Apesar de o embasamento teórico oferecido ser bastante amplo, a leitura do texto não se torna densa, isso faz com que o leitor, mesmo pertencendo ao público leigo, não tenha dificuldade em avançar no entendimento.

A segunda parte do livro traz um manual aplicado com um detalhamento muito rico e pontual de como proceder para a obtenção de uma determinada estrutura tridimensional. Notadamente, a parte de aplicação do livro mostra tudo o que é necessário para que aqueles que se aventuram nas tarefas de modelar e avaliar as estruturas tridimensionais, sintam-se bastante seguros quanto ao manuseio dos parâmetros e interpretação dos resultados obtidos através deste guia prático.

Na última parte do livro, os autores mostram diversas aplicações da modelagem computacional em distintos contextos da biologia. Vários exemplos são apresentados através de artigos científicos, o que se configura como uma forma de ampliar os horizontes do leitor quanto à vasta gama de possibilidades de aplicação dos métodos de modelagem. Além disso, os exemplos de aplicabilidade servem para instigar os leitores mais audazes a se aventurarem na laboriosa tarefa de predizer a estrutura de uma proteína.

Particularmente, o último capítulo do livro é bastante interessante, pois além de mostrar as perspectivas da predição tridimensional, ele discute, de forma geral, as atuais limitações. Os autores discorrem sem detalhar um método em particular e trazem à tona uma faceta da modelagem que é seu carácter de aproximação. Isso não poderia ser diferente, pois a modelagem computacional é uma inferência que, como tal, nos proporciona uma forma de idealizar uma entidade biológica desconhecida que é a real estrutura da proteína.

Em suma o livro *Métodos de Modelagem da Estrutura Tridimensional de Proteínas: Um Guia Teórico-prático* é um excelente referencial teórico-prático de fácil leitura que está sendo oferecido ao público acadêmico na língua portuguesa.

Prof. PhD. Élcio de Souza Leal
Docente do Instituto de Ciências Biológicas, Universidade Federal do Pará

SUMÁRIO

CAPÍTULO 1..... 1

A ESTRUTURA E O PROBLEMA DE DOBRAMENTO DE PROTEÍNAS

Kauê Santana da Costa

Anderson Henrique Lima e Lima

Jerônimo Lameira

 <https://doi.org/10.22533/at.ed.8252228101>

CAPÍTULO 2..... 7

MÉTODOS DE PREDIÇÃO POR HOMOLOGIA

Kauê Santana da Costa

João Marcos Pereira Galúcio

José Rogério de Araújo Silva

 <https://doi.org/10.22533/at.ed.8252228102>

CAPÍTULO 3..... 16

MODELAGEM DE PROTEÍNAS NO *MODELLER*

Kauê Santana da Costa


João Marcos Pereira Galúcio

 <https://doi.org/10.22533/at.ed.8252228103>

CAPÍTULO 4..... 29

MODELAGEM DE PROTEÍNAS NO *EASYMODELLER*

João Marcos Pereira Galúcio

 <https://doi.org/10.22533/at.ed.8252228104>


CAPÍTULO 5..... 33

MÉTODOS DE PREDIÇÃO *THREADING* E *AB INITIO*

Anderson Henrique Lima e Lima

Kauê Santana da Costa

Alberto Monteiro dos Santos

 <https://doi.org/10.22533/at.ed.8252228105>


CAPÍTULO 6..... 41

ALGORÍTMOS DE MODELAGEM *THREADING* E *AB INITIO*

Anderson Henrique Lima e Lima

Kauê Santana da Costa

Alberto Monteiro dos Santos

 <https://doi.org/10.22533/at.ed.8252228106>

CAPÍTULO 7..... 47

MODELAGEM DE PROTEÍNAS NO *I-TASSER*

Kauê Santana da Costa

Anderson Henrique Lima e Lima

 <https://doi.org/10.22533/at.ed.8252228107>


CAPÍTULO 8..... 53

APLICAÇÕES DA MODELAGEM POR HOMOLOGIA, *AB INITIO* E *THREADING*

Anderson Henrique Lima e Lima

Alberto Monteiro dos Santos

Kauê Santana da Costa

 <https://doi.org/10.22533/at.ed.8252228108>

CAPÍTULO 9..... 60

ATUAIS LIMITAÇÕES E PERSPECTIVAS DOS MÉTODOS *THREADING* E *AB INITIO*

Kauê Santana da Costa

Anderson Henrique Lima e Lima

Alberto Monteiro dos Santos

 <https://doi.org/10.22533/at.ed.8252228109>

GLOSSÁRIO DE TERMOS TÉCNICOS E SIGLAS 64

REFERÊNCIAS 70

SOBRE OS AUTORES 79

CAPÍTULO 1

A ESTRUTURA E O PROBLEMA DE DOBRAMENTO DE PROTEÍNAS

Kauê Santana da Costa

Anderson Henrique Lima e Lima

Jerônimo Lameira

A determinação da estrutura de proteínas mostra-se relevante em diferentes áreas científicas, entre elas, engenharia, química, biologia e medicina devido à participação destes biopolímeros em diferentes processos biológicos, assim como suas diversas aplicações industriais, farmacêuticas e biotecnológicas. Diferentes abordagens e métodos *in silico* têm sido desenvolvidos, ao longo dos anos, para a solução do problema de predição da estrutura destas macromoléculas, como alternativas aos métodos experimentais, por exemplo, espectroscopia de Ressonância Magnética Nuclear (RMN) e a cristalografia de difração de raios X. Além disso, os métodos que utilizam somente ferramentas computacionais se destacam, em alguns aspectos, em relação aos experimentais, como o fato de requererem menor investimento em infraestrutura e recursos humanos.

Usados na predição de estrutura de proteínas, os métodos *in silico* têm recebido considerável atenção da comunidade acadêmica, devido ao desenvolvimento de algoritmos e computadores com maior

capacidade de processamento de dados, e pelo aumento no número de sequências nucleotídicas e de aminoácidos depositados em bases de dados biológicos de livre acesso que favorecem a sua utilização em diferentes propósitos científicos (BENSON et al., 2017; MAGRANE; CONSORTIUM, 2011). Neste contexto, o termo *modelagem de proteínas* tem sido aplicado, atualmente, para referir-se à predição da estrutura tridimensional destas macromoléculas por algoritmos computacionais, partindo do conhecimento prévio da sua sequência de aminoácidos e o termo *desenho de proteínas* para designar a engenharia de proteínas, que consiste na construção de novas estruturas que exibam aplicações, principalmente, de cunho biotecnológico ou características diferentes daquelas encontradas na natureza.

PARADOXO DE LEVINTHAL E A HIPÓTESE DE ANFISEN

O processo físico que leva o dobramento de proteínas continua sendo um dos maiores desafios da Biologia Estrutural. O problema de dobramento de proteínas (PFP – do inglês *Protein Folding Problem*) levantado, inicialmente, por Cyrus Levinthal em 1968, colocou em evidência a difícil resolução da estrutura tridimensional dessas macromoléculas. O paradoxo de Levinthal, como é chamado, parte da observação de que não existe um tempo

suficientemente adequado, avaliando os processos biológicos para determinada proteína atingir sua conformação estável, considerando randomicamente todos os seus possíveis estados conformacionais (LEVINTHAL, 1968). No entanto, em 1973, os estudos pioneiros de Christian Anfisen sobre o processo de dobramento da ribonucleases A lançaram luz para a compreensão inicial da relação existente entre a sequência de aminoácidos de determinada proteína e a sua conformação nativa tridimensional (conformação funcional). A hipótese de Anfisen implica que a estrutura nativa estável dessas macromoléculas correspondem ao mínimo global de energia livre, sendo que esta depende de um conjunto de interações da proteína com o meio que a circunda, por exemplo, as interações iônicas, interações não ligadas, interações proteína-solvente, efeitos estéricos, torsionais, entre outros (ANFISEN, 1973).

Segundo a hipótese termodinâmica, a estrutura nativa das proteínas é determinada exclusivamente por um processo biofísico, isto é, pela interação da sequência com o meio de solvatação, sendo sua (1) conformação única: não há para determinada sequência duas configurações conformacionais com energias livres comparáveis; (2) estável: sua estrutura não está propensa a mudanças conformacionais bruscas devido às alterações ambientais pequenas, tais como, pH, pressão e temperatura (3) cineticamente acessível ao nível mínimo de energia, isto é, a diferença entre os níveis de energia e as formas enoveladas e não enoveladas é pequena do ponto de vista cinético, podendo as primeiras serem acessíveis em escalas de tempo comparáveis aos demais os processos biológicos que ocorrem em nível molecular.

A conformação funcional de uma proteína também depende um conjunto de fatores biológicos intrínsecos presentes em cada organismo, não podendo a sua determinação, ser reduzida, portanto, unicamente pela interação biofísica com meio abiótico. Entre estes fatores pode-se citar: modificações pós-traducionais, tais como glicosilação, metilação, clivagem proteolítica etc., a formação de complexos proteína-proteína, e a interação com moléculas de cofatores que levam a alterações alostéricas. Determinados eucariotos, proteínas de modificação denominadas chaperonas, também estão envolvidas nas alterações pós-traducionais e auxiliam no correto enovelamento da proteína na sua conformação final (BEN-ZVI; GOLOUBINOFF, 2001).

ESTRUTURA DAS PROTEÍNAS E AMINOÁCIDOS

A estrutura de uma proteína compreende diferentes níveis de organização que estão diretamente envolvidos em sua conformação nativa. Dentre os níveis organizacionais, distinguimos quatro estruturas: primária, secundária, terciária e quaternária.

A estrutura primária refere-se somente a sequência de resíduos de aminoácidos; a estrutura secundária corresponde ao dobramento da sequência pela formação de interações intermoleculares intracadeia, tais como, ligações de hidrogênio formadas por átomos da

cadeia principal (no inglês *backbone*) que levam a formação de estruturas denominadas de α -hélices (pronúncia: alfa-hélices) e β -folhas (beta-folhas). Conectando os elementos da estrutura secundária, há as regiões de alça (em inglês, *loop*), que são do ponto de vista conformacional muito variáveis. A estrutura terciária corresponde ao dobramento adicional destas estruturas na conformação nativa pela formação de novas interações formadas pelas cadeias laterais (no inglês *side chains*) dos resíduos de aminoácidos, tais como, interações hidrofóbicas, pontes dissulfeto, ligações de hidrogênio, etc. O quarto, e último nível, corresponde à estrutura quaternária que leva a formação de complexos, isto é, a formação de dímeros, trímeros, tetrâmeros, etc., e devido às interações intercadeia dos átomos das cadeias laterais e da cadeia principal de diferentes monômeros (Figura 1A).

Quimicamente, as proteínas são compostas de longas cadeias de aminoácidos que são unidos covalentemente entre si por ligações peptídicas. Cada aminoácido é composto por um o átomo de carbono quiral denominado carbono alfa (C_{α}), que forma ligações com um grupo amina, um grupo carboxila, uma cadeia lateral que varia de acordo com o aminoácido e um átomo de hidrogênio (Figura 1B).

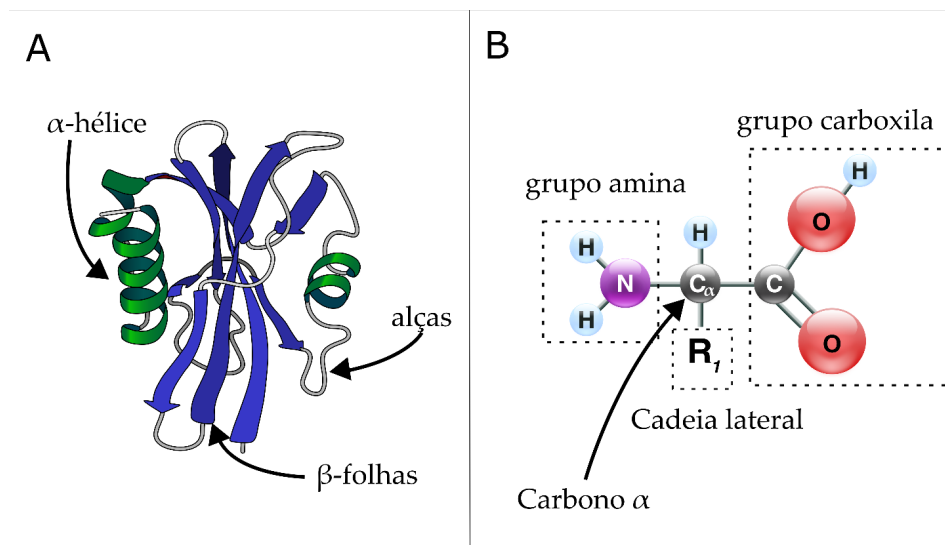


Figura 1. (A) Estrutura terciária de uma proteína exibindo as regiões de alfa-hélice, beta-folhas e alças. (B) Estrutura química do aminoácido mostrando os grupos químicos e localização do carbono central.

Ao todo existem vinte tipos de aminoácidos naturais denominados de padrões ou primários, são eles: glutamato e aspartato (polares ácidos); arginina, lisina e histidina (polares básicos); asparagina, glutamina, serina, treonina e tirosina (polares sem carga); e cisteína, fenilalanina, glicina, isoleucina, leucina, lisina, metionina, prolina, triptofano, alanina e valina (apolares). A tabela 1 exhibe a abreviação e o símbolo utilizado para a representação dos vinte tipos de aminoácidos.

Nome	Abreviação	Símbolo
Aminoácidos apolares		
Alanina	Ala	A
Glicina	Gly	G
Fenilalanina	Phe	F
Leucina	Leu	L
Valina	Val	V
Isoleucina	Ile	I
Prolina	Pro	P
Metionina	Met	M
Triptofano	Trp	W
Serina	Ser	S
Aminoácidos polares e sem carga		
Cisteína	Cys	C
Asparagina	Asn	N
Treonina	Thr	T
Tirosina	Tyr	Y
Asparagina	Asn	N
Glutamina	Gln	Q
Aminoácidos negativamente carregados (ácidos)		
Aspartato	Asp	D
Glutamato	Glu	E
Aminoácidos positivamente carregados (básicos)		
Arginina	Arg	R
Lisina	Lys	K
Histidina	His	H

Tabela 1. Abreviações e símbolos de representação dos 20 tipos diferentes de aminoácidos

INTRODUÇÃO AOS MÉTODOS DE PREDIÇÃO DA ESTRUTURA

Os métodos de predição da estrutura de proteínas podem ser subdivididos em duas categorias: os métodos baseados em molde (TBM – do inglês *Template-based Modeling*) que utilizam uma ou mais estruturas como referência previamente resolvidas por métodos experimentais para a modelagem. Neles estão incluídos a modelagem por homologia, também chamada de modelagem comparativa (FISER, 2010; FISER; ŠALI, 2003; SANTOS FILHO; ALENCASTRO; BICCA DE ALENCASTRO, 2003) e os métodos de predição por *threading* – ou método por reconhecimento de dobramento (CHENG; BALDI, 2006; JONES; MILLER; THORNTON, 1995; JONES; TAYLOR; THORNTON, 1992). Por último, há os métodos de predição *ab initio*, conhecidos também por *de novo* (BONNEAU et al., 2002) primeiros princípios (FLOUDAS et al., 2006) ou métodos livres de molde (TFM, do inglês *Template-free Modeling*) (KINCH et al., 2011).

A escolha do método adequado para a predição estrutural dependerá principalmente de dois critérios: a existência de uma estrutura molde e a aplicação na qual o modelo se destina. De modo geral, os métodos baseados em molde são preferidos para aplicações que exigem detalhes ultraestruturais mais precisos, tais como planejamento *in silico* de inibidores e análise do mecanismo catalítico de enzimas. Já os métodos *ab initio* têm sido aplicados com relativo sucesso no desenho de proteínas com novas topologias e atividades catalíticas (Figura 2).

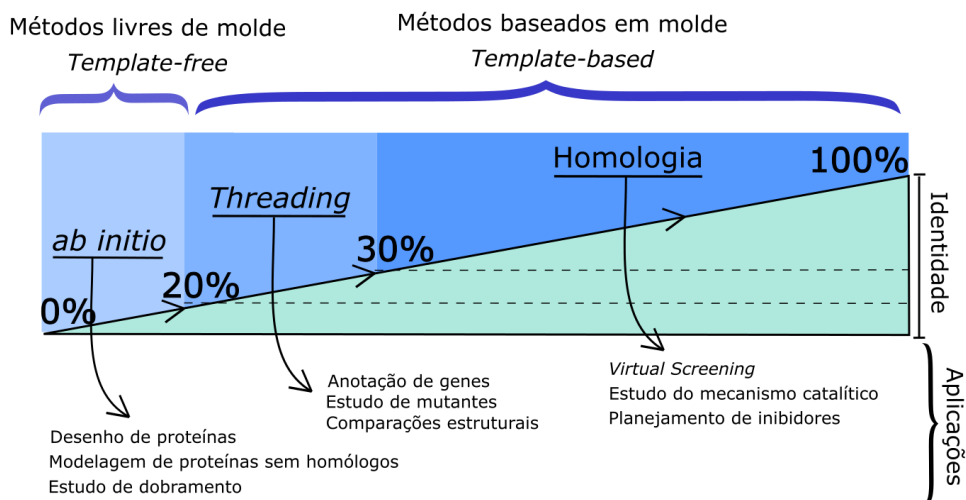


Figura 2. Relação entre métodos de modelagem da estrutura de proteínas e o valor de identidade da sequência. No diagrama, são exibidas algumas aplicações comumente direcionadas aos modelos obtidos por cada método, embora mais de um método possa ser usado para a mesma aplicação.

Embora, os métodos baseados em molde recebam maior atenção nos eventos de competição formal que ocorrem bienalmente, denominados Avaliação Crítica das Técnicas de Predisão da Estrutura de Proteínas (CASP, do inglês, *Critical Assessment of Techniques for Protein Structure Prediction*) (HUANG et al., 2014) – em que diferentes métodos e algoritmos são testados considerando a sua capacidade de prever a estrutura tridimensional destas macromoléculas na sua conformação nativa – os métodos *ab initio* destacam-se, atualmente, como as estratégias mais ambiciosas para a predição estrutural, assim como para o desenho de proteínas com novas topologias e atividades catalíticas (BONNEAU et al., 2001, 2002; SIMONS et al., 1999; ZHANG, 2007; 2009).

CAPÍTULO 2

MÉTODOS DE PREDIÇÃO POR HOMOLOGIA

Kauê Santana da Costa

João Marcos Pereira Galúcio

José Rogério de Araújo Silva

A predição por homologia, também conhecida por modelagem comparativa, visa determinar a estrutura da proteína de interesse, utilizando como molde (referência), estruturas de proteínas homólogas previamente elucidadas por métodos experimentais, tais como espectroscopia de Ressonância Magnética Nuclear (RMN) e difração de raio X. Atualmente, é considerada como um dos métodos mais fidedignos de predição da estrutura nativa de proteínas, porque utiliza, como referência, uma estrutura similar, sendo por isso, o método computacional preferido objetiva obter estruturas com maior precisão estrutural, bem como estruturas experimentais estão ausentes em estudos de planejamento *in silico* de fármacos e no estudo do mecanismo catalítico de enzimas (CAVASOTTO; PHATAK, 2009).

O termo homologia é utilizado na Biologia Evolutiva para se referir a duas estruturas que compartilham semelhanças por descenderem de um ancestral comum. Desta forma, a predição por homologia se baseia no pressuposto de que a estrutura terciária é mais

conservada que a estrutura primária (sequência) devido à conservação da função molecular ao longo da evolução. Nesta linha de raciocínio, considerando que duas proteínas pertencentes a mesma família apresentam estrutura mais preservada que a sequência, podemos assumir que caso esta similaridade seja previamente detectada no nível de sequência, então a estrutura tridimensional terá similaridade maior. De modo semelhante, caso a similaridade entre as duas sequências seja pequena, então, a correspondência entre a estrutura de ambas também será baixa.

CONCEITUANDO IDENTIDADE E SIMILARIDADE

Na comparação entre a sequências de aminoácidos de duas proteínas, dois parâmetros são fundamentais: similaridade e identidade. Ambos são obtidos através do alinhamento das sequências. A similaridade corresponde ao percentual de aminoácidos que compartilham entre si as mesmas características físico-químicas, por exemplo, leucina, valina e isoleucina são ambos aminoácidos hidrofóbicos apolares e, portanto, nas matrizes de alinhamento aplicado para sequências de proteínas apresentarão pontuações similares quando ocuparem as mesmas regiões de correspondência, conhecidas como *matches*. Já a identidade corresponde ao percentual de aminoácidos idênticos entre as duas sequências

analisadas (SÁNCHEZ; ŠALI, 1997).

Ambos os conceitos podem ser melhor expressos por meio de equações. A similaridade (S) é expressa pela Equação 1, conforme abaixo:

$$S = \frac{Ls \times 2}{La + Lb} \times 100 \quad (1)$$

Onde:

S = similaridade

Ls = número de aminoácidos similares da sequência.

La e Lb = comprimento da sequência A e B, respectivamente.

A identidade (I) é representada por equação similar, porém o denominador Li corresponde ao número de aminoácidos idênticos da sequência (Equação 2).

$$I = \frac{Li \times 2}{La + Lb} \times 100 \quad (2)$$

Onde:

I = identidade

Li = número de aminoácidos idênticos da sequência.

La e Lb = comprimento da sequência A e B, respectivamente.

Diferentes programas e alinhamento calculam os valores de identidade para sequências de aminoácidos. Entre estes pode-se citar o Muscle (EDGAR, 2004) e o T-Coffee (DI TOMMASO et al., 2011).

ETAPAS ENVOLVIDAS DA MODELAGEM POR HOMOLOGIA

A modelagem por homologia envolve cinco passos consecutivos e complementares que incluem: (1) pesquisa de estruturas homólogas em base de dados; (2) alinhamento da sequência alvo com as sequências de estrutura conhecida (moldes); (3) construção do modelo tridimensional; (4) a otimização estrutural, e por último; (5) a validação do modelo. Uma visão geral do processo de modelagem por homologia é mostrada na figura 3.

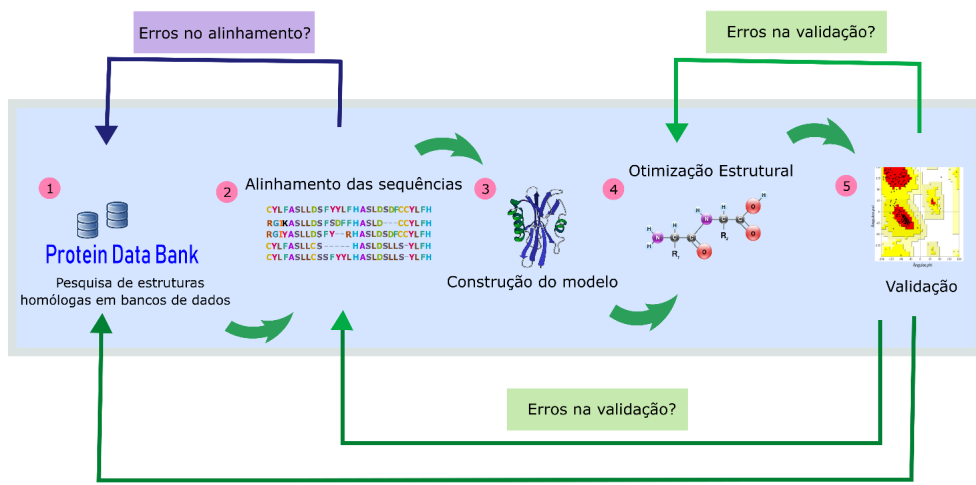


Figura 3. Visão geral do processo de modelagem por homologia. Erros em uma das etapas podem acarretar em inconsistências nos resultados da etapa subsequente, sendo, deste modo, necessário o retorno à uma etapa anterior para a correção.

BUSCA DE ESTRUTURAS HOMÓLOGAS

No primeiro passo, a seqüência alvo (de interesse) é comparada por meio de alinhamento com outras existentes em base de dados de estruturas experimentais. Atualmente, o RCSB *Protein Data Bank* (<https://www.rcsb.org/>) é o maior e mais abrangente base de dados público com informações estruturais de proteínas, ácidos nucleicos e seus ligantes (BERMAN et al., 2005). Para realizar a pesquisa de estruturas no PDB, pode-se utilizar o *Protein BLAST* (BLASTp sigla do inglês *Basic Local Alignment Search Tool*), um algoritmo de alinhamento simples e local (ALTSCHUL et al., 1997; BENSON et al., 2017) que realiza a busca sistemática em base de dados de seqüências. O BLASTp é disponibilizado para uso *on-line* pelo *National Center for Biotechnology Information* (NCBI) e fornece valores de cobertura, identidade e código de acessos de seqüências que melhor se alinham. A identidade corresponde ao porcentual de resíduos similares entre duas seqüências, sendo elas um dos parâmetros importantes para a seleção de estruturas moldes que serão utilizados na modelagem. Além de identidade satisfatória, devemos avaliar diferentes critérios em conjunto e, de modo geral, as estruturas devem:

1. Pertencer aos organismos filogeneticamente próximos.
2. Pertencer às proteínas que apresentam a mesma função molecular ou mesma família.
3. Se obtidas por meio da técnica de difração de raio X, devem apresentar alta qualidade (resolução baixa, preferencialmente $\leq 2,0 \text{ \AA}$). Valores maiores poderão ser permitidos na ausência de outras estruturas homólogas, porém devido à alta resolução, o modelo poderá apresentar qualidade inferior e, portanto, mais etapas

de otimização estrutural serão necessárias.

4. Estar o mais completas possível, isto é, com poucos ou nenhum fragmento ausente.
5. Apresentar cobertura satisfatória com a sequência alvo ($\geq 90\%$).
6. Apresentar identidade satisfatória no alinhamento ($\geq 30\%$), isto é, quanto maior a semelhança na sequência, maior será a semelhança na estrutura. Não havendo identidade mínima (30%) não é recomendável aplicar a modelagem por homologia.

ALINHAMENTO DAS SEQUÊNCIA ALVO E MOLDE

É realizado, na segunda etapa da modelagem por homologia, o alinhamento entre as sequências. Esta etapa é essencial no processo de modelagem, haja vista que permite a extração do molde informações estruturais que serão utilizadas na criação do modelo tridimensional. Estruturas podem ser modeladas, partindo de um alinhamento simples, isto é, alinhamento em que são consideradas somente duas sequências (sequência alvo e a sequência molde) ou de alinhamento múltiplo, em que mais de duas sequências são comparadas (a sequência alvo, dois ou mais moldes). O alinhamento múltiplo é indicado quando determinadas regiões da proteína-alvo carecem de uma estrutura de referência no molde, sendo este obtido, portanto, de outras proteínas homólogas. Diferentes programas podem ser utilizados no alinhamento das sequências, tais como o ClustalW (CHENNA, 2003), Muscle (EDGAR, 2004) e T-Coffee (DI TOMMASO et al., 2011). O Modeller, por exemplo, utiliza a matriz de similaridade BLOSUM62.

Uma alta identidade entre a sequência da estrutura molde com a sequência de interesse (alvo) é, preferencialmente, desejada, pois maior será a qualidade e, portanto, a acurácia do modelo gerado. Desta forma, modelos de alta identidade (de 95% a 100%), após validados, podem ser satisfatoriamente usados no estudo de mecanismos catalíticos de enzimas, na triagem virtual de fármacos e no planejamento *in silico* de inibidores (WEBB; SALI, 2016).

CONSTRUÇÃO DO MODELO

Na terceira etapa, realizamos a predição da estrutura propriamente, isto é, a construção do modelo tridimensional. Atualmente, vários programas e servidores aplicam abordagem por homologia, entre eles, podemos citar o servidor Swiss-model (BIASINI et al., 2014) e o pacote computacional Modeller (FISER; ŠALI, 2003) que aplicam estratégias diferentes na modelagem comparativa.

O Modeller é um programa desenvolvido em Fortran 90 e Python pelo grupo do pesquisador Andrej Sali e realiza a predição da estrutura de proteínas por homologia, implementando uma técnica inspirada por ressonância magnética nuclear conhecida como

satisfação das restrições espaciais, pela qual um conjunto de ligações angulares (entre três átomos) e diedrais (entre quatro átomos) são retiradas de uma estrutura molde (usada como referência) obtida experimentalmente e, então, aplicadas na estrutura alvo que será modelada (FISER; ŠALI, 2003). Além destas restrições, o campo de força CHARMM-22 fornece informações sobre os comprimentos e ângulos de ligação dos átomos da estrutura modelada, incorporando uma funcionalidade limitada para previsão *ab initio* das regiões de alças de proteínas, que são frequentemente muito variáveis e, portanto, difíceis de prever por modelagem por homologia (FISER; ŠALI, 2003). Como parâmetro de seleção das estruturas otimizadas, o Modeller aplica o conceito de Energia Descontínua de Proteína Optimizada (Dope, do inglês *Discrete Optimized Protein Energy*), um potencial estatístico obtido da minimização de funções de energia da estrutura que permite selecionar entre os modelos criados o de menor energia. No Modeller, a energia Dope é aplicada por meio de scripts em linguagem *Python*. Além da modelagem, o programa também realiza funções adicionais úteis na modelagem comparativa, tais como, o alinhamento simples (entre duas sequências) ou múltiplo (mais de duas sequências) e a otimização de regiões de alça da estrutura modelada (FISER; ŠALI, 2003).

O Swiss-model é um servidor que realiza a predição e a validação da estrutura de proteínas, além de apresentar uma interface intuitiva e fácil de usar. O servidor aplica a modelagem por homologia utilizando o conceito de corpos rígidos (BIASINI et al., 2014; KIEFER et al., 2009) que, por sua vez, consistem na modelagem da estrutura por partes, sendo que as regiões estruturalmente conservadas da proteína alvo são definidas através de predição das estruturas secundárias. Essas regiões são alinhadas com o molde, levando em conta a média das posições dos carbonos α das regiões estruturalmente conservadas da sequência.

As regiões que não satisfazem as exigências estruturais, geralmente, são porções de alças que, por apresentarem estruturas flexíveis, adquirem diferentes conformações. Informações para a modelagem da cadeia principal, dessas regiões variáveis, são obtidas em bases de dados de estruturas que apresentam conjuntos de alças classificados pelo número de resíduos de aminoácidos e pelo tipo de estruturas secundárias que estas conectam. Após a modelagem das regiões de alças, um modelo inicial do esqueleto peptídico é gerado e, em seguida, procede-se a inserção das cadeias laterais dos resíduos através de busca em bibliotecas de rotâmeros (BIASINI et al., 2014).

OTIMIZAÇÃO ESTRUTURAL

A quarta etapa consiste na otimização estrutural, isto é, o refinamento da estrutura tridimensional. Invariavelmente, após a construção do modelo, há erros esteroquímicos que necessitam de correção, pois influenciam na estabilidade da estrutura e na interpretação de dados estruturais que podem ser obtidos. A otimização compreende a alteração

guiada da estrutura de modo a melhorar sua qualidade estereoquímica e perfil de energia. Além de métodos de minimização de energia, são aplicados métodos de amostragem conformacional que empregam simulações de dinâmica molecular clássica com campos de força específicos (FAN, 2004; RAVAL et al., 2012). Diferentes algoritmos de otimização foram desenvolvidos com esta finalidade, porém uma discussão detalhada está fora do escopo deste trabalho.

VALIDAÇÃO DO MODELO

Após a etapa de otimização estrutural, a quinta e última etapa corresponde da modelagem por homologia corresponde à validação e permite avaliar o nível de qualidade e confiabilidade da estrutura obtida na modelagem. Modelos tridimensionais de alta qualidade, depois de validados, podem revelar dados biológicos relevantes para avaliar a atividade e a função da proteína, bem como realizar estudos sobre sua estrutura. Ao longo dos anos, diferentes ferramentas foram desenvolvidas e auxiliam na validação do modelo fornecendo parâmetros sobre a qualidade da estrutura.

Erros encontrados nas etapas de modelagem por homologia podem, por exemplo, serem devidos:

1. Aos alinhamentos errôneos das sequências, como a existência de muitos *mismatches*, baixa cobertura e identidade.
2. À seleção de estruturas-molde de baixa qualidade ou sem correspondência estrutural (ex.: estruturas-molde resolvidas por cristalografia de raios-X a alta resolução criará modelos de baixa qualidade);
3. À modelagem de regiões sem a estrutura-molde que levam invariavelmente à formação de regiões de alça.

A validação utiliza três principais critérios: (1) correlação das estruturas secundárias e terciárias da proteína de modo a inferir a conservação da sua função; (2) estereoquímica e (3) a energia (WEBB; SALI, 2016).

Na avaliação da correlação da estrutura secundária e terciária da proteína são aplicados métodos de alinhamento estrutural e métricas de comparação, como o desvio da raiz quadrada média das posições atômicas (RMSD do inglês *Root Mean Square Deviation*) calculado através da distância média entre os átomos de duas estruturas sobrepostas. A maioria dos programas utiliza o carbono α (C α), como referência para o cálculo do RMSD, porém outros átomos da cadeia principal podem também ser utilizados como nitrogênio e oxigênio. A equação utilizada para o cálculo do RMSD é mostrada a seguir (Equação 3):

$$RMSD = \sqrt{\frac{1}{N} \sum_{i=1}^N (\|x_i^A - x_i^B\|^2)} \quad (3)$$

Onde:

N = número de átomos comparados.

x_i^A e x_i^B = coordenadas espaciais dos átomos representativos das estruturas A e B, respectivamente.

Para cada valor de RMSD-C α (medido em angstroms, Å), é possível estabelecer um comparativo de quão semelhantes são as duas estruturas em questão – estrutura-molde (do inglês *template*) e a estrutura-alvo modelada (do inglês, *target*) (Tabela 1). Ao se comparar a estrutura molde com a estrutura alvo, os valores de RMSD-C α desejáveis devem ser preferencialmente $\leq 3,0$ Å. Valores maiores que 3,0 Å trazem dúvidas com relação à escolha da estrutura-molde e a qualidade do modelo final. Já estruturas com RMSD $\leq 0,4$ Å mostram que ambas são estruturalmente indistinguíveis, podendo indicar que são conformações da mesma proteína. Através do alinhamento estrutural, é possível analisar a disposição e a conservação de resíduos em sítios catalíticos, motivos proteicos, inferir a geometria de coordenação dos resíduos com íons metálicos, assim como prever a função da proteína, comparando-a com outras estruturas previamente elucidadas por métodos experimentais (BAXEVANIS; OUELLETTE, 2003).

RMSD (Å)	Comparação Estrutural
>12,0	Completamente não relacionadas
7,0	Não relacionadas
5,0	Podem estar estruturalmente relacionadas
3,0	Boa relação estrutural
2,0	Proximamente relacionadas
1,5	Muito proximamente relacionadas
0,8	Muito similares
$\leq 0,4$	Estruturalmente indistinguíveis

Tabela 1. Valores de RMSD-C α e a respectiva interpretação com relação à semelhança estrutural das proteínas.

Fonte: Adaptado de Baxevanis e Ouellette (2003).

Na validação, a avaliação da qualidade estereoquímica do modelo pode ser realizada por diferentes parâmetros estruturais. O gráfico de Ramachandran é, sem dúvida, a ferramenta de avaliação mais popular para este fim, pois permite a representação da proteína por meio de seus ângulos de torção (LOVELL et al., 2003). O gráfico de Ramachandran descreve as proteínas pelos seus ângulos de torção, conhecidos por ϕ

(pronúncia: *phi*) e ψ (*psi*) (Figura 4). O ângulo ϕ é definido com o ângulo diedro presente na ligação N-C α , enquanto que o ângulo ψ é definido como o ângulo diedro presente na ligação C α -C. Cada resíduo, portanto, pode ser definido por um ângulo ϕ e um ψ , quando submetidos num gráfico de dispersão com os ângulos ϕ no eixo X e os ângulos ψ no eixo Y.

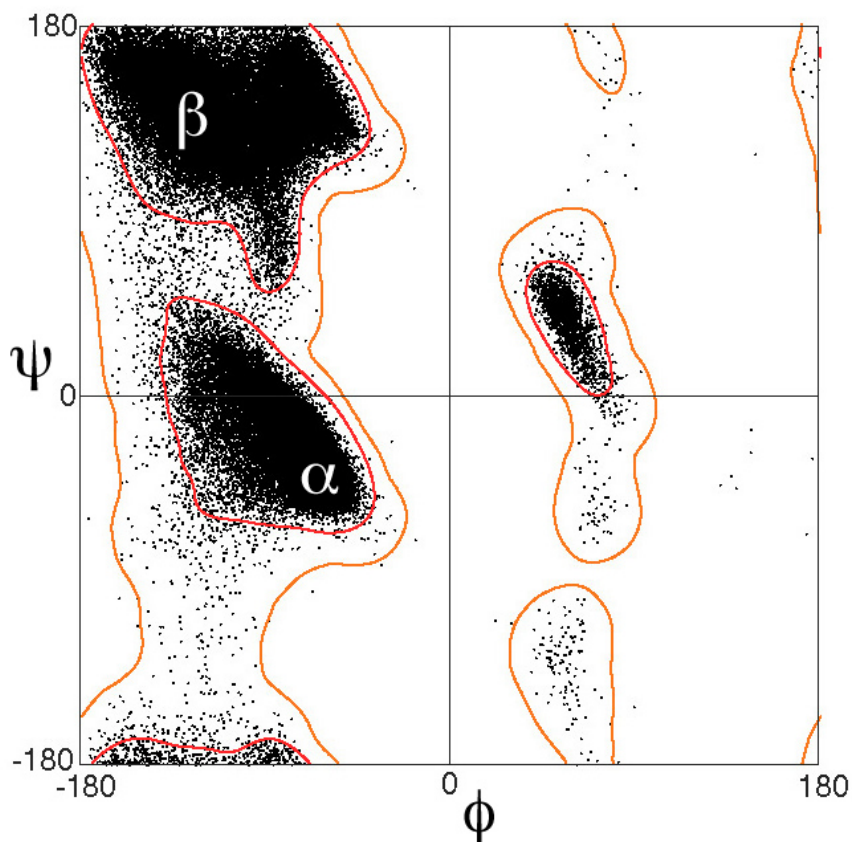


Figura 4. Gráfico de Ramachandran que mostra as regiões aceitáveis para os ângulos de torção ϕ e ψ localizados nas estruturas secundárias (α -hélices e β -folhas). Cada ponto do gráfico representa um resíduo de aminoácido da proteína analisada.

A representação de Ramachandran permite verificar os valores permitidos dos ângulos de torção, indicando quais resíduos se encontram em regiões mais favoráveis ou desfavoráveis dos ângulos, permitindo, assim, prever conflitos estéricos entre as cadeias (LOVELL et al., 2003). As regiões permitidas para cada aminoácido dependem da extensão da cadeia lateral, sendo assim, para a glicina que apresenta menor cadeia lateral são aplicadas menores restrições, visto que ela é formada por um único átomo de hidrogênio, assim como, para a alanina em que a cadeia lateral é formada por um grupo metila. Em

cadeias laterais que envolvem vários átomos, a mobilidade dela é restringida pelo tamanho. As regiões em cinza exibem os ângulos de torção permitidos para as β -folhas (indicados pela letra grega β) e α -hélices (regiões indicadas por α), as regiões demarcadas por linhas tracejadas exibem regiões de repulsão e choque entre diferentes átomos e representam, portanto, resíduos com valores errôneos para ângulos de torção. Entre os programas que geram o gráfico de Ramachandran, podem ser citados o Rampage, o Procheck (LASKOWSKI et al., 1993) e o servidor MolProbity (LOVELL et al., 2003).

Na avaliação da energia são aplicados métodos que aplicam valores normalizados obtidos da minimização de energia. Ao avaliar a energia, objetiva-se verificar a estabilidade da estrutura, isto é, sua propensão de formar interações intra e intermoleculares favoráveis que as tornem estáveis, assim como as conformações com enovelamento adequado. Entre os programas aplicados na avaliação da energia de modelos se destacam o Qmean (BENKERT; SCHWEDE; TOSATTO, 2009), ANOLEA (MELO; FEYTMANS, 1998) e o servidor ProSA (WIEDERSTEIN; SIPPL, 2007).

CAPÍTULO 3

MODELAGEM DE PROTEÍNAS NO *MODELLER*

Kauê Santana da Costa

João Marcos Pereira Galúcio

O Modeller é um programa desenvolvido em Fortran 90 e *Python* pelo grupo do pesquisador Andrej Šali da Universidade da Califórnia, nos Estados Unidos, e realiza a predição da estrutura de proteínas por homologia. Neste capítulo, mostraremos o passo a passo necessário para a modelagem por homologia.

Programas e bases de dados usados

Modeller

Editor de Texto

UCSF Chimera

RSCB *Protein Data Bank*

As informações deste capítulo podem ser adaptadas para os sistemas Windows e Linux, e é recomendável executar cada uma das etapas em um diretório diferente que contenha os arquivos de input para cada script executado. Para fins de organização, separamos as etapas em diretórios nomeados com os seguintes nomes: *encontrando*, *comparando*, *alinhando*, *construindo* e *otimizando*.

1 | PROCURANDO MOLDES

ESTRUTURAS

Para buscar estruturas homologas as que desejamos modelar, utilizaremos o alinhamento realizado pelo BLASTp. Acesse ao servidor pelo endereço: <https://blast.ncbi.nlm.nih.gov/>.

Escolha a opção *Protein* BLAST (BLASTp). Na página que aparece (Figura 5), (1) adicione a sequência de aminoácidos na caixa do quadro *Enter Query Sequence*. São aceitos os códigos de acesso no GenBank e formato FASTA da sequência. Em seguida; (2) selecione o base de dados em que será realizada a pesquisa: no quadro *Choose Search Set* na opção *Database* selecione *Protein Data Bank* e (3) selecione o nome científico da espécie de interesse na opção *Organism*. Mantenha as outras opções inalteradas e clique em BLAST ao final da página.

Your search is limited to records that include: Homo sapiens (taxid:9606)

[Edit and Resubmit](#) | [Save Search Strategies](#) | [Formatting options](#) | [Download](#) | [YouTube](#) [How to read this page](#) | [Blast report description](#) | **NEW** [Click here to see the new BLAST results page](#)

Job title: Protein Sequence

RID: [EK6ZDRAN014](#) (Expires on 05-27 03:47 am)

Query ID: [lc|Query_349781](#) | Database Name: [pdb](#)
 Description: [None](#) | Description: [PDB protein database](#)
 Molecule type: [amino acid](#) | Program: [BLASTP 2.9.0+](#) | [Citation](#)
 Query Length: [303](#)

Other reports: [Search Summary](#) | [Taxonomy reports](#) | [Distance tree of results](#) | [Multiple alignment](#) | [MSA viewer](#)

Analyze your query with [SmartBLAST](#)

Graphic Summary

Show Conserved Domains

A Putative conserved domains have been detected, click on the image below for detailed results.

B Distribution of the top 100 Blast Hits on 100 subject sequences. Mouse over to see the title, click to show alignments.

C Sequences producing significant alignments:

Select: [All](#) [None](#) Selected: 0

Alignments: [Download](#) | [GenPept](#) | [Graphics](#) | [Distance tree of results](#) | [Multiple alignment](#)

Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
<input type="checkbox"/> Chain A, SERINE/THREONINE-PROTEIN KINASE PAK 4 (Homo sapiens)	622	622	100%	0.0	99.67%	2C0D_A
<input type="checkbox"/> Chain A, Serine/threonine-protein kinase PAK 4 (Homo sapiens)	618	618	99%	0.0	99.67%	2QDN_A
<input type="checkbox"/> Chain A, Serine/threonine-protein Kinase Pak 4 (Homo sapiens)	617	617	99%	0.0	99.67%	4E1E_A
<input type="checkbox"/> Chain A, Serine/threonine-protein kinase PAK 4 (Homo sapiens)	617	617	99%	0.0	99.67%	5UPL_A
<input type="checkbox"/> Chain A, Serine/threonine-protein kinase PAK 4 (Homo sapiens)	616	616	99%	0.0	99.67%	5UVD_A
<input type="checkbox"/> Chain A, Serine/threonine-protein Kinase Pak 4 (Homo sapiens)	616	616	99%	0.0	99.67%	4E1F_A
<input type="checkbox"/> Chain A, Serine/threonine-protein Kinase Pak 4 (Homo sapiens)	615	615	99%	0.0	99.67%	4SRU_A
<input type="checkbox"/> Chain A, Protein Fam212a serine/threonine-protein Kinase Pak 4 (Homo sapiens)	615	615	99%	0.0	99.67%	4SRB_A
<input type="checkbox"/> Chain A, Serine/threonine-protein Kinase Pak 4 (Homo sapiens)	611	611	99%	0.0	99.00%	4J0J_A
<input type="checkbox"/> Chain A, P21-ACTIVATED KINASE 4 (Homo sapiens)	600	600	96%	0.0	99.66%	2EVA_A
<input type="checkbox"/> Chain A, Serine/threonine-protein kinase PAK 4 (Homo sapiens)	600	600	96%	0.0	99.66%	5UVA_A
<input type="checkbox"/> Chain A, SERINE/THREONINE-PROTEIN KINASE PAK 4 (Homo sapiens)	599	599	96%	0.0	99.32%	2K4Z_A
<input type="checkbox"/> Chain A, Serine/threonine-protein Kinase Pak 4 (Homo sapiens)	599	599	96%	0.0	99.66%	4D0V_A
<input type="checkbox"/> Chain A, Serine/threonine-protein kinase PAK 4 (Homo sapiens)	593	593	95%	0.0	99.65%	4DVE_A
<input type="checkbox"/> Chain A, Serine/threonine-protein kinase PAK 7 (Homo sapiens)	519	519	96%	0.0	84.93%	4E87_A
<input type="checkbox"/> Chain A, SERINE/THREONINE-PROTEIN KINASE PAK 4 (Homo sapiens)	461	461	96%	3e-165	74.74%	2C36_A
<input type="checkbox"/> Chain A, Serine/threonine-protein kinase PAK 4 (Homo sapiens)	459	459	95%	3e-165	75.69%	4K87_A
<input type="checkbox"/> Chain A, Serine/threonine-protein Kinase Pak 4 (Homo sapiens)	326	326	94%	2e-112	53.50%	4JLI_A
<input type="checkbox"/> Chain A, Serine/threonine-protein Kinase Pak 1 (Homo sapiens)	326	326	94%	2e-112	53.50%	4JLO_A

Atina e Windows

Figura 6. Parâmetros de alinhamento obtidos pelo BLASTp para a seleção da sequência da melhor estrutura homóloga.

Nesta seção, as etapas de criação da biblioteca, a comparação e a seleção da estrutura-molde no Modeller são opcionais, pois os critérios analisados podem ser suficientes para a escolha da estrutura molde. No entanto, aplicaremos os scripts visando a comparação das sequências.

1.1 Criando Biblioteca de Sequências em Formato PIR

Coloque todas as sequências em um único arquivo no formato PIR e, em seguida, remova todos os espaços entre os aminoácidos. No Linux, você poderá utilizar os editores Vim ou Gedit. No Windows, é aconselhável usar o NotePad++.

O formato das sequências deve seguir o mesmo do base original disponibilizado pelo Modeller, sendo que cada linha da sequência polipeptídica deve conter exatos 75 resíduos. O formato de sequências em PIR que são aceitos pelo programa Modeller seguem o seguinte padrão:

Cabeçalho

```
>P1;1cf9A
structureX:1cf9:1:A:753:A:MOL_ID1;MOLECULELYSOZYME;CHAINA;SYNONYML?
```

- Em “structureX:**1cf9**”, o número em negrito corresponde ao código de identificação da sequência no PDB.
- Em “1:A:**753**”, o número em negrito corresponde à quantidade de aminoácidos presente na sequência.

Sequência de resíduos

Cada linha deve conter 75 resíduos e ao final devem terminar com um asterisco (*).

```
MSEAAHVLIITGAAGQIGYILSHWIASGELYGDRQVYLHLLDIPPAMNRLTALTMELEDCAFPHLAGFVATDPK.
AFKIDIDCAFLVASMPLKPGQVRADLISSNSVIFKNTGEYLSKWAKE SVKVLVIGNEPDNTNCEIAMLHAKNLKPE:
FSSL SMLDQNRAYYE VASKLGVDVKDVHDIIVMGNHGESMVADLTQATFTREGKTKQKVVVDVLDHDYVFDTFKK:
GHRAWDILEHRGFTSAASPTKAAIQHMKAWLFGTAPGEVLSMGIPVPEGNEFYGIKPGVVF SFCNVNDRKGGKIHV
EGFRKVNWLREKLDFTKDLFHEKEIALNHLAQGG*
```

1.2 Selecionando as Estruturas Moldes Pesquisadas

Pasta 1: Encontrando

Arquivos de entrada:

TvLDH.ali (sequência da estrutura de interesse, *target*)

pdb_95PIR (biblioteca de sequências)

build_profile.py (script)

Arquivos de saída:

build_profile.ali

build_profile.log

Modificando o script “build_profile.py”

O script *build_profile.py* deve ser editado nos campos mostrados (Figura 7) para realizar a leitura da sequência alvo em formato .ALI (em azul) e da biblioteca em formato PIR (em vermelho). Note que dois arquivos nos retângulos vermelhos aparecem no formato .BIN, no entanto, será necessário modificar somente o nome do arquivo.

```

#-- Prepare the input files

#-- Read in the sequence database
sdb = sequence_db(env)
sdb.read(seq_database_file='pdb_95.pir', seq_database_format='PIR',
         chains_list='ALL', mirmax_db_seq_len=(30, 4000), clean_sequences=True)

#-- Write the sequence database in binary form
sdb.write(seq_database_file='pdb_95.bin', seq_database_format='BINARY',
         chains_list='ALL')

#-- Now, read in the binary database
sdb.read(seq_database_file='pdb_95.bin', seq_database_format='BINARY',
         chains_list='ALL')

#-- Read in the target sequence/alignment
aln = alignment(env)
aln.append(file='TvLDH.ali', alignment_format='PIR', align_codes='ALL')

#-- Convert the input sequence/alignment into
# profile format
prf = aln.to_profile()

#-- Scan sequence database to pick up homologous sequences
prf.build(sdb, matrix_offset=-450, rr_file='${LIB}/blosum62.sim.mat',

```

Figura 7. Campos de edição do script *build_profile.py*.

Executando o Modeller (versão 9.21) no Terminal

Execute os comandos abaixo no terminal do sistema. Entre no diretório faça:

```
$ cd diretorio_arquivo/execucao
```

No *prompt* do Windows para listar os arquivos de dentro do diretório, faça:

```
$ dir
```

No terminal do Linux, digite:

```
$ ls
```

Execute o Modeller:

```
$ mod9.21 build_profile.py
```

Observações:

Para obter informações de erro de execução, verifique o arquivo de *.log* gerado em cada etapa.

Este tutorial foi criado para a versão 9.21, no entanto, caso sua versão seja superior ou inferior, como 9.14, utilize o comando *mod9.14*, etc.

Na pasta em que o arquivo *.py* foi executado, abra o arquivo *build_profile.log* no editor de texto. Em *HITS FOUND ITERATION*, encontramos informações relacionadas ao alinhamento. São desejáveis as sequências que apresentam maior identidade e menor *E-value* (Figura 8).

1	2	3	4	5	6	7	8	9	10	11			
1cf9A	1	1	37700	753	502	38.27	0.0	2	467	13	488	75	560
1CF9B	1	2	37700	753	502	38.27	0.0	3	467	13	488	75	560
1E93A	1	3	46100	484	502	42.77	0.0	4	467	14	493	2	478
1E93A	1	4	46100	484	502	42.77	0.0	5	467	14	493	2	478
1gg9A	1	5	37600	753	502	38.27	0.0	6	467	13	488	75	560
1GG9B	1	6	37600	753	502	38.27	0.0	7	467	13	488	75	560
1a4eA	1	158	66350	488	502	53.81	0.0	8	474	14	491	4	488
4b1cA	1	1284	53650	499	502	49.15	0.0	9	465	19	492	25	496
1e93A	1	1663	45450	475	502	43.10	0.0	10	461	19	493	4	474
1p80A	1	1736	37650	727	502	38.27	0.0	11	467	13	488	49	534
1d9fA	1	2225	52500	497	502	48.09	0.0	12	466	19	492	24	495
1gweA	1	3988	41150	498	502	40.80	0.0	13	461	19	492	9	481
1q11A	1	7470	45450	491	502	46.80	0.0	14	428	19	455	9	446
1s18A	1	7917	47900	474	502	44.68	0.0	15	463	19	492	4	473
1sy7A	1	8105	35550	698	502	43.13	0.0	16	349	39	399	27	390
1ye9A	1	9533	21700	223	502	44.09	0.0	17	215	13	230	1	220
1ye9E	1	9534	16100	256	502	34.52	0.0	18	238	245	488	1	252
1m7sA	1	9766	36100	484	502	40.00	0.0	19	413	15	440	1	420

Figura 8. Colunas do arquivo de saída do Modeller (*build_profile.log*).

As colunas mais importantes correspondem à 1ª, 3ª, 5ª e 7ª e 8ª que correspondem, respectivamente, ao código de identificação das sequências no PDB, ordem de disposição das sequências no base PIR, ao número de resíduos da sequência, valor de identidade em relação à sequência alvo e aos valores de *E-value* (valores ideais iguais ou próximos de 0.0).

1.3 Comparando as Estruturas Homólogas

Esta etapa consiste em comparar os modelos (em formato .PDB) selecionados no passo anterior de acordo com os valores de identidade. Acesse o *Protein Data Bank* (<http://www.rcsb.org>) e faça o *download* de quatro modelos selecionados para serem comparados de acordo com os parâmetros analisados no arquivo .log e adicione à pasta juntamente com os scripts.

Modificando o script "compare.py"

No script *compare.py*, o nome do arquivo (azul) e a cadeia (vermelho) de cada arquivo .PDB devem ser especificados. Certifique-se que escreveu corretamente o nome dos arquivos. É importante atentar que o programa, utiliza sintaxe rígida e, portanto, irá diferenciar letras maiúsculas e minúsculas.

```
env = environ()
aln = alignment(env)
for (pdb, chain) in (('1d9f', 'A'), ('1q9w', 'A'), ('2xq1', 'A'),
                    ('4BLC', 'A'), ('1A4E', 'A')):
    m = model(env, file=pdb, model_segment=('FIRST:'+chain, 'LAST:'+chain))
    aln.append_model(m, atom_files=pdb, align_codes=pdb+chain)
```

Pasta 2: Comparando
Arquivos de entrada:
 melhores moldes em formato PDB
compare.py (Script)
Arquivo de saída:
compare.log
family.mat

Arquivo de saída compare.log

No arquivo *compare.log* em *Weighted pair-group average clustering based on a distance matrix*, verificamos a identidade da sequência alvo em relação às sequências das estruturas que serão usadas como moldes (verde), assim como a resolução (azul) da estrutura obtida por raios-X (Figura 9). Os modelos ideais devem apresentar alta identidade e baixa resolução.

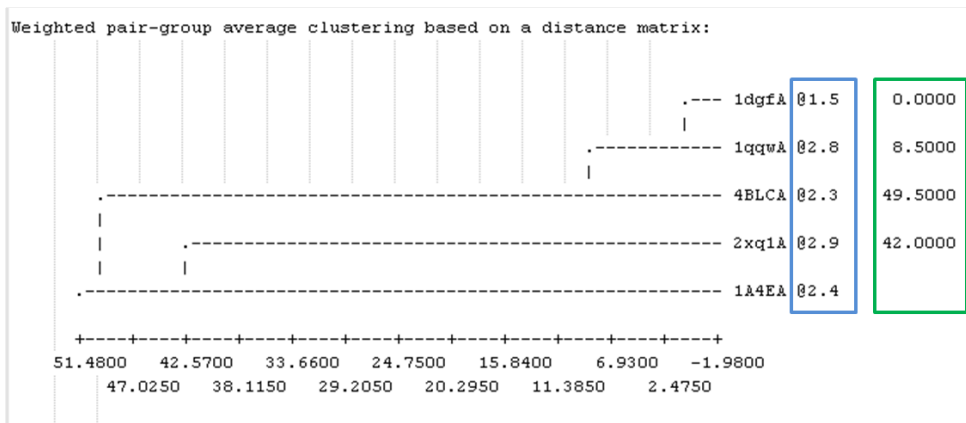


Figura 9. Arquivo de saída *compare.log* do Modeller.

2 | ALINHAMENTO DAS SEQUÊNCIAS

Pasta 3: Alinhando
Arquivos de entrada:
TvLDH.ali (sequência alvo)
1A4E.PDB (sequência molde)
align2d.py (script)
Arquivos de saída:
TvLDH-1bdmA.ali
TvLDH-1bdmA.pap

Modificando o script align2d.py

1. Indique o nome do arquivo em formato PDB (retângulo vermelho) e a cadeia (círculos) em que será realizado o alinhamento.

```
mdl = model(env, file='lbdm', model_segment=('FIRST', 'LAST'))
```

2. Modifique nome do arquivo da sequência alvo TvLDH.ali (azul), assim como o código especificado dentro do mesmo (verde).

```
aln.append(file='TvLDH.ali', align_codes='TvLDH')
```

3. Modifique o nome do arquivo .PDB que será usado como molde (verde) e o nome que será colocado no arquivo gerado (azul).

```
aln.append_model(mdl, align_codes='lbdmA', atom_files='lbdm.pdb')
```

Os arquivos de saída do programa podem ou não serem alterados (fica a critério do usuário).

```
aln.write(file='TvLDH-lbdmA.ali', alignment_format='PIR')
aln.write(file='TvLDH-lbdmA.pap', alignment_format='PAP')
```

Formato do arquivo ALI da sequência alvo

```
>P1;Q8X220
```

```
sequence:Q8X220:::0.00: 0.00
```

```
MALLTAETFRLLQFNKRRLLRRPYPRKALLCYQLTPQNGSTPTRGYFENKKKCHAEICFINEIKSMGL
DETQCYQVTCYLTWSPCSSCAWELVDFIKAHDHLNLRIFASRLYYHWCKPQQDGLRLLCGSQVPVE
VMGFPEFADCWENFVDHEKPLSFNPNYKMLEELDKNRSRAIKRRLDRIKIPGVRAQGRYMDILCDAEV*
```

Um alinhamento desejável deve exibir poucos *gaps* (lacunas), e muitos *matches* (correspondência) entre os aminoácidos comparados das duas sequências. A Figura 10 exibe um alinhamento ideal entre a sequência molde (referência) e alvo (de interesse).



Figura 10. Alinhamento desejável entre a sequência molde e alvo (de interesse). Poucos ou nenhum *gap* e muitos *matches*. Aminoácidos similares, como a isoleucina (I) e a valina (V), assim como os idênticos estão destacados com a mesma cor no alinhamento. Neste exemplo, usou-se o programa MUSCLE.

3 | CONSTRUINDO O MODELO

Pasta 4: Construindo

Arquivos de entrada:

TvLDH-1a4e.ali (alinhamento)

TvLDH.ali (sequência alvo)

Melhor modelo escolhido em formato .PDB

Model-single.py (script)

Arquivos de saída:

São gerados vários arquivos em formato .PDB

Modificando o script *model-single.py*

Modifique o nome do arquivo de alinhamento (.ALI)

```
a = automodel(env, alnfile='TvLDH-1bdmA.ali',
```

Indique o nome da sequência “alvo” (vermelho) e o nome da sequência “molde” (azul). O nome de ambas está no arquivo de alinhamento gerado no passo anterior (.ALI).

```
knowns='1bdmA', sequence='TvLDH',
```

Indique quantos modelos serão gerados na execução do programa (círculo vermelho).

```
a.starting_model = 1  
a.ending_model = 5
```

Observação: o nome deve ser o mesmo encontrado no alinhamento.

4 | OTIMIZANDO O MODELO

A otimização das alças no Modeller é realizada pelo script *loop_refine.py*. Para saber qual região deve ser otimizada submeta seu melhor modelo no servidor *Swiss-Model* (<http://swissmodel.expasy.org/>) e veja no gráfico do QMEAN as regiões de alta energia.

Através do Chimera podemos identificar as regiões de alça da sequência para discriminá-las no *script* de otimização *loop_refine.py*. É importante ressaltar que o sítio catalítico de enzimas, de modo geral, é conservado e, deste modo, deve-se atentar ao realizar a alteração das conformações para não alterar a posição dos resíduos nestas regiões.

Pasta 5: Otimização das Alças

Arquivos de entrada:

loop_refine.py (script)

Modelo selecionado em formato .PDB

Arquivos de saída:

Arquivos no formato .PDB (modelos)

loop_refine.log

Editando o *script loop_refine.py*

1. Adicione o intervalo dos resíduos que serão otimizados (vermelho):

```
def select_loop_atoms(self):
    # 10 residue insertion
    return selection(self.residue_range('273', '283'))
```

2. Adicione o nome do melhor modelo gerado (ou pelo script “*multi-model.py*” ou pelo “*AutoModelMulty.py*”) (azul) e o nome da sequência alvo que será gerado (verde):

```
m = MyLoop(env,
           inimodel='TvLDH-mult.pdb', # initial model of the target
           sequence='TvLDH')        # code of the target
```

Avaliando os Modelos

Os modelos gerados serão avaliados de acordo com a pontuação Dope (do inglês *Discrete Optimized Protein Energy*), um potencial estatístico usado para avaliar modelos gerados por modelagem comparativa no Modeller. Quanto menor a pontuação em valores negativos, melhor o modelo.

A análise da energia de cada um dos modelos da proteína alvo é mostrado pelo script *model_energies.py*. Através deste script, selecionaremos o modelo com menor energia através dos valores obtidos para o Dope.

Editando o arquivo *model_energies.py*

Adicione o intervalo dos modelos gerados (em azul) durante o passo 4 (construindo modelos). O símbolo de porcentagem “%” significa que serão lidos todos os arquivos de *TvLDH.BL%04d0001.PDB*.

```
log.verbose() # request verbose output
env = environ()
env.libs.topology.read(file='${LIB}/top_heav.lib') # read topology
env.libs.parameters.read(file='${LIB}/par.lib') # read parameters

for i in range(1, 11):
    # read model file
    code = 'TvLDH.BL%04d0001.pdb' % i
    mdl = complete_pdb(env, code)
    s = selection(mdl)
    s.assess_dope(output='ENERGY_PROFILE_NO_REPORT', file='TvLDH.profile',
                 normalize_profile=True, smoothing_window=15)
```

Deve-se escolher um dos modelos gerados que exibe a menor energia Dope (azul).

Abra o arquivo *loop_refine.log*.

```
<< end of ENERGY.  
DOPE score : -50066.785156  
openf__224_> Open TvLDH.BL00080001.pdb
```

ATENÇÃO: Não pertence ao modelo 08, e sim ao 07

Analizando as Regiões de Alça pelo Chimera

Abra o arquivo .PDB da estrutura usando o programa UCSF Chimera (Figura 11). O modelo será exibido em diagrama de Ribbon (1) e para colori-lo, vá na barra de menu e faça:

Action > *Color*

Pra visualizar a sequência de aminoácidos que formam β -folhas e α -hélices, faça:

Tools > *Sequence* > *Sequence*.

As estruturas secundárias são mostradas nos quadros coloridos e alças não são destacadas (2). Para verificar a numeração dos resíduos da sequência, na caixa que abrir, faça:

Numberings > *Overall alignment*

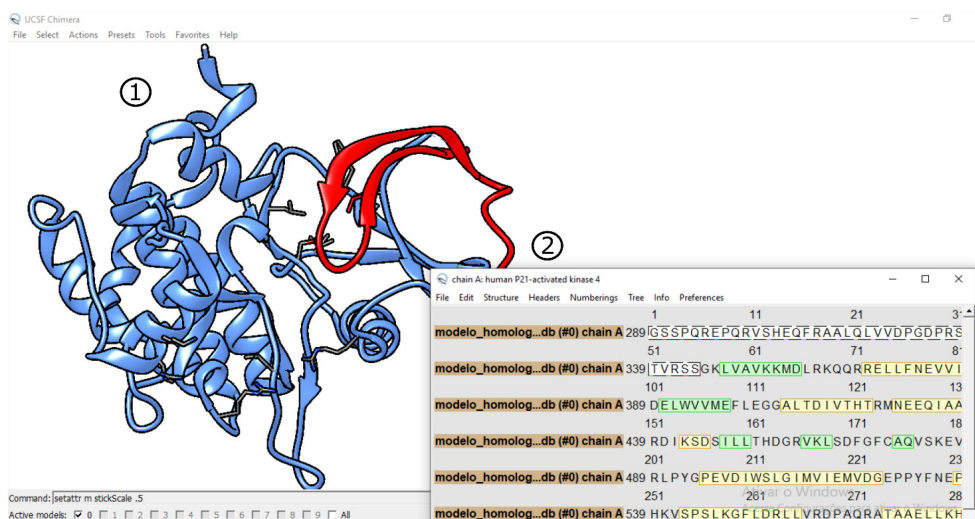


Figura 11. Visualização do modelo tridimensional pelo UCSF Chimera. As regiões em verde exibem as β -folhas em amarelo as α -hélices. Passando o *mouse* sobre os resíduos, vemos a posição de cada um no lado inferior direito da tela.

5 | VALIDANDO O MODELO

Usaremos o servidor suíço *Swiss-model* mantido pela Universidade de Bazel para a validação do modelos gerado nas etapas anteriores. Acesse o servidor pelo endereço abaixo:

<https://swissmodel.expasy.org/assess>

Na página, faça o upload do arquivo .PDB gerado nas etapas anteriores e nomeie a atividade no campo *Structure File*, (1) clicando em *Upload Coordinate File* (Figura 12). Em seguida, (2) clique em *Start Assessment* para executar a análise da estrutura.

Structure Assessment

Start a new Structure Assessment Project

Structure File: modelo_homologia.pdb ✓

Project Title (Optional):

Email (Optional):

Ativar o Windows
Acesse Configurações para ativar o Windows.

Swiss Institute of Bioinformatics | Schwede Group | Contact Us | Terms of use | [Back to the Top](#)

Figura 12. Página inicial do Swiss-model exibe as regiões para *upload* das estruturas que serão analisadas.

Após a modelagem, o gráfico de Ramachandran da estrutura tridimensional é exibida do lado esquerdo e a estrutura em diagrama de Ribbon do lado direito. Clicando nos pontos referentes a cada resíduo, o painel contendo a estrutura é atualizado para exibição de detalhes estruturais.

CAPÍTULO 4

MODELAGEM DE PROTEÍNAS NO *EASYMODELLER*

João Marcos Pereira Galúcio

- Python versão 2.7, disponível em python.org (versões 3.X não são suportadas).

O *EasyModeller* é uma versão gráfica simples e intuitiva do Modeller desenvolvido por Kuntal Kumar Bhusan na Universidade de Hyderabad (Índia) e está disponível gratuitamente para uso acadêmico.

REQUISITOS

- Modeller versão 9.16 ou superior.
- EasyModeller

TUTORIAL

1. Abra o EasyModeller na pasta de instalação do programa. É possível verificar que há quatro abas que representam o processo de modelagem por homologia no painel principal: *Load Inputs*, *Align Templates*, *Align Query* e *Build Model* (Figura 15).

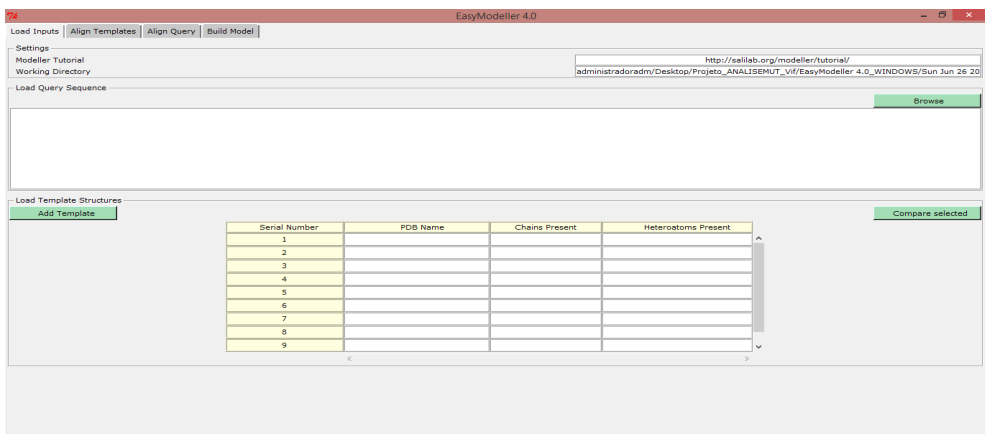


Figura 15. Painel inicial do EasyModeller

Na primeira aba, *Load Inputs*, insira a sequência que deseja modelar em formato FASTA na caixa *Load Query Sequence*, clicando em *Browse* (Figura 16). Em *Load Template Structures*, insira o arquivo(s) .PDB(s) da(s) estrutura(s) usadas como molde. Nesse exemplo, apenas uma estrutura será usada como referência. Note que o programa reconhece heteroátomos e diferentes cadeias da proteína. Selecione a caixa em *Serial Number*.

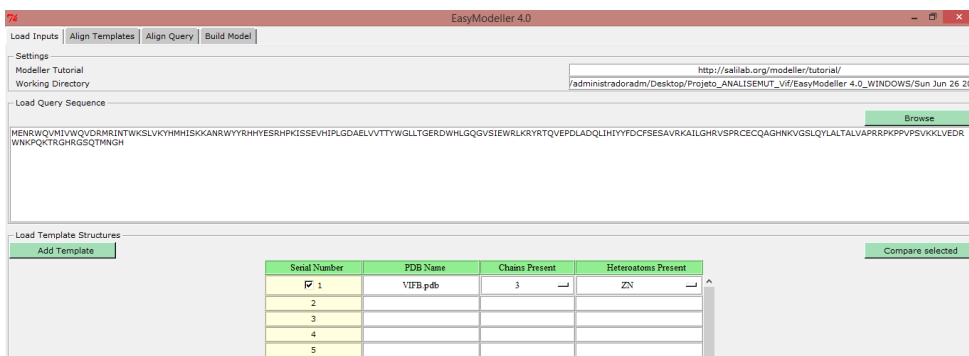


Figura 16. Área de entrada da sequência de resíduos de aminoácidos.

2. Em *Align Templates*, são alinhadas as estruturas molde selecionadas. No exemplo, não há alinhamento, pois utilizamos somente um molde (Figura 17).

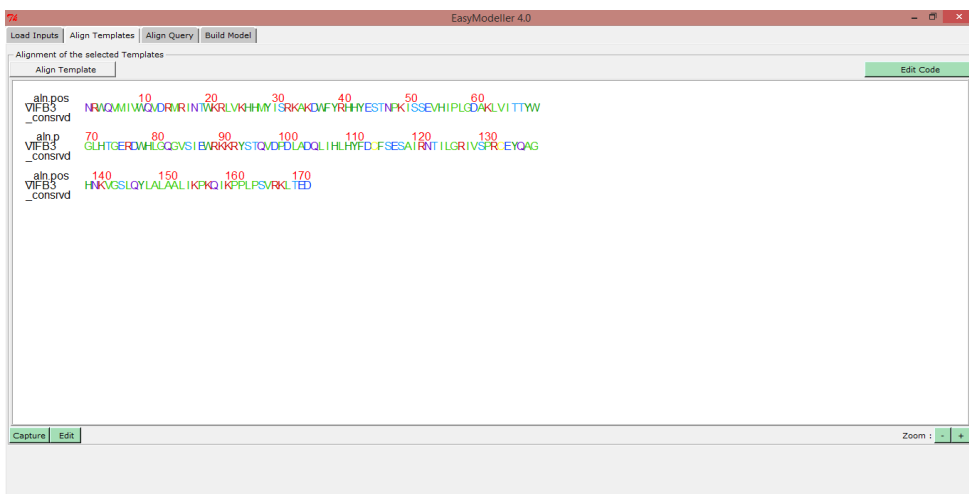


Figura 17. Alinhamento da sequência alvo com o molde.

Na aba *Align Query*, alinha-se as sequências alvo e da estrutura molde. Clique no

botão *Align Query with Templates*. Note que a estrutura é indicada em cores: em vermelho, regiões conservadas. Abaixo das sequências são exibidos, em laranja, as regiões de α -hélice e na última linha, também em laranja, as β -folhas (Figura 18). As alças são indicadas em branco.

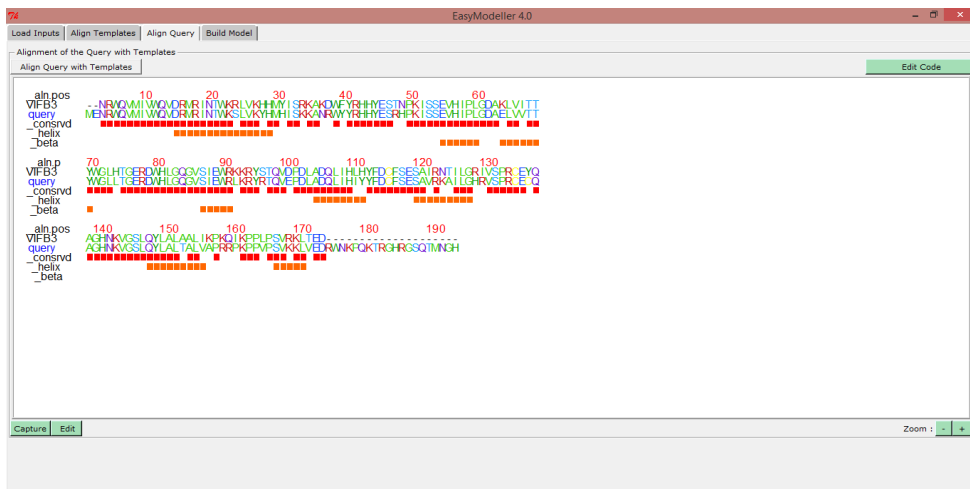


Figura 18. Painel mostrando regiões conservadas da sequência e a correspondência das estruturas secundárias

4. Na aba *Build Model*, gera-se o modelo da estrutura. Clique em *Generate Model* (Figura 19).

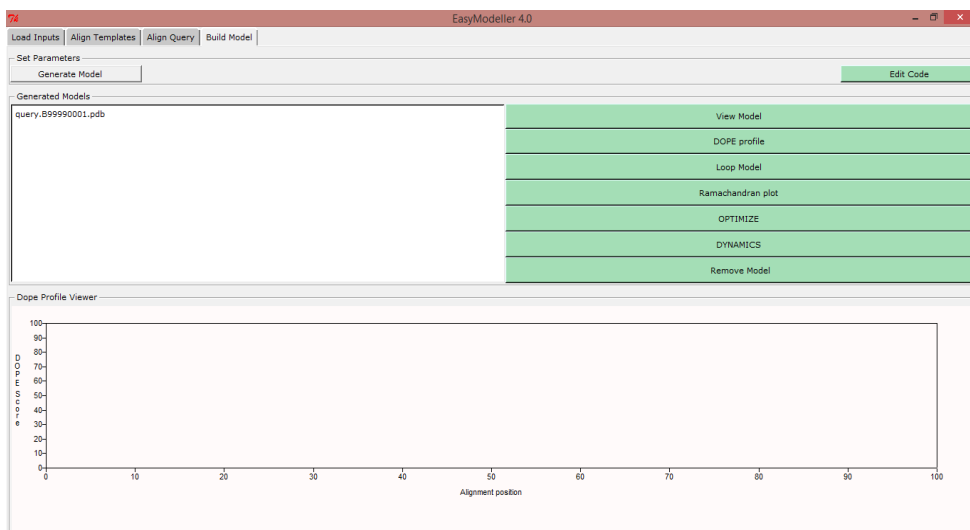


Figura 19. Área de modelagem da estrutura do EasyModeller

Após obtido o modelo tridimensional, podem ser exploradas opções adicionais para análise da estrutura (Figura 12). Na opção *View Model*, apresenta-se a visualização do modelo obtido. Há também opções para gerar o gráfico de Ramachandran, realizar a otimização da estrutura, avaliar a pontuação de energia Dope etc. A estrutura pode ser aberta diretamente pelo programa Pymol.

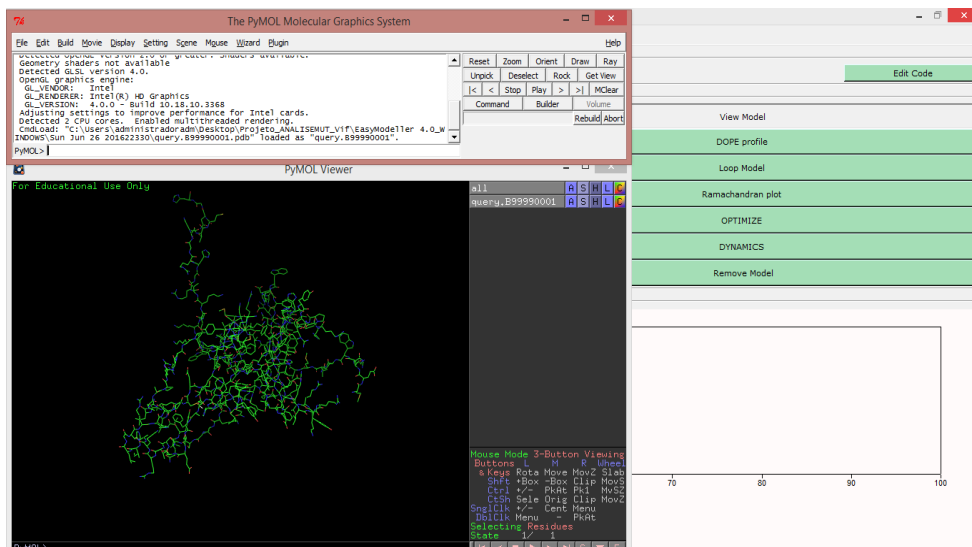


Figura 20. Análise do modelo tridimensional gerado no Pymol.

O arquivo em formato .PDB da estrutura modelada é salvo na pasta do EasyModeller, em uma subpasta gerada automaticamente e nomeada com a data e dia em que a modelagem foi realizada.

Anderson Henrique Lima e Lima

Kauê Santana da Costa

Alberto Monteiro dos Santos

MÉTODOS DE PREDIÇÃO POR *THREADING*

O termo *threading* foi inicialmente cunhado em 1992 por David Jones e colaboradores, para a predição da estrutura de proteínas através do reconhecimento de dobramentos (do inglês *protein folding*) semelhantes com ou sem relação evolutiva (JONES; MILLER; THORNTON, 1995; JONES; TAYLOR; THORNTON, 1992). O *threading* mostra-se útil para a predição de estruturas de proteínas que apresentam baixa identidade (valores $\leq 30\%$) com outras sequências disponíveis em bases de dados para a modelagem comparativa, no entanto, seleciona estruturas que compartilham semelhanças entre si (FLOUDAS et al., 2006).

Esta abordagem baseia-se na conjectura que o número de possíveis dobramentos é finito para todas as proteínas existentes e, desta forma, proteínas não relacionadas também podem assumir estruturas similares devido à evolução convergente (WANG, 1998). Um fator determinante para o desenvolvimento deste método foi a descoberta, através da análise espacial da organização dos aminoácidos nas

proteínas, de preferências na interação entre estes, isto é, resíduos hidrofóbicos, por exemplo, podem ser mais suscetíveis à interação com resíduos de mesma natureza hidrofóbica. Esta observação permitiu o desenvolvimento de potenciais de interação explícita, na qual, analisando a estrutura primária, é possível determinar, através de matrizes de pontuação, diferentes características envolvidas em modelar a estrutura da proteína, tais como preferências de interação de pares de resíduos, regiões favoráveis à formação de estruturas secundárias, regiões acessíveis à solvatação, entre outras.²⁰ Atualmente, diferentes servidores empregam a modelagem de proteínas por *threading* e apresentam uma interface amigável e de fácil utilização, estes incluem os servidores Phyre2 (KELLEY et al., 2015), I-TASSER (ROY; KUCUKURAL; ZHANG, 2010; ZHANG, 2009) e Raptor-X (KÄLLBERG et al., 2012).

Os métodos *threading* podem ser subdivididos em dois grupos: *threading* baseado em perfil (em inglês *profile-based threading*) e o *threading* baseado em potenciais de pares (em inglês *pair potentials-based threading*) (KORETKE et al., 1999). Na primeira metodologia, os algoritmos criam a estrutura utilizando, como referência, a topologia de estruturas conhecidas, baseando-se em matrizes de pontuação que definem o “perfil” tridimensional da sequência (ROST; SCHNEIDER; SANDER, 1997). Estas

matrizes podem ser derivadas de diferentes informações, entre as quais, matrizes de distância evolutiva (como a de Dayhoff para proteínas), do alinhamento de seqüências e da predição da estrutura secundária (JONES et al., 1999; ZHOU; ZHOU, 2005). O processo de modelagem utilizado por *threading* baseado em perfil é mostrado esquematicamente na Figura 21. O segundo baseia-se em potenciais de pares de resíduos que informam a probabilidade de determinados resíduos encontrarem-se a certa distância, baseando-se em energias de contato de pares de resíduos e energias hidrofóbicas, sendo alguns derivados de análise estatística de estruturas cristalográficas de proteínas (BRYANT; LAWRENCE, 1993; TAYLOR, 1997).

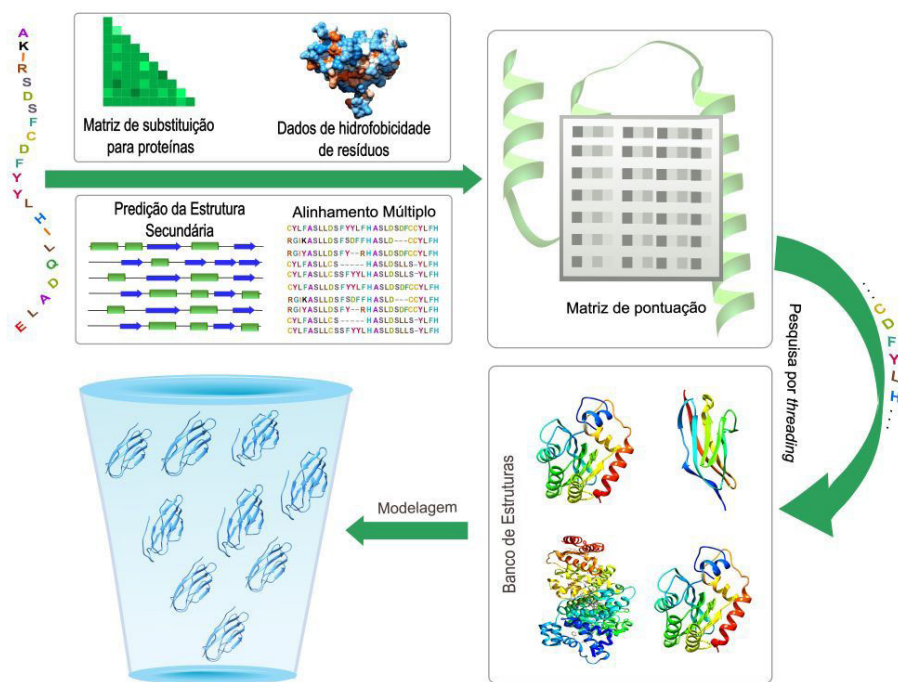


Figura 21. Processo esquemático da modelagem de proteínas pela abordagem *threading* baseado em perfil (em inglês, *profile-based threading*). Etapas de agrupamento estrutural, avaliação e minimização de energia e validação podem estar presentes em determinados algoritmos de modelagem por *threading*

Algoritmos que aplicam a predição por *threading* têm se mostrado bons resultados nos eventos CASP e incluem: THREADER (JONES et al., 1999), Phyre2 (KELLEY et al., 2015) e I-TASSER (ROY; KUCUKURAL; ZHANG, 2010; ZHANG, 2009). A literatura, com modelagem de proteínas por *threading*, é extremamente vasta e utiliza diferentes algoritmos, programas e pipelines (IHM et al., 2009; NOCUA et al., 2014; TSANG et al., 2011). Abordaremos a estratégia de predição do I-TASSER, que recentemente apresentou

resultados extremamente satisfatórios nas últimas competições do CASP (ZHANG, 2007, 2009), em especial, o CASP7 ocorrido em 2006, na cidade de Pacific Grove, na Califórnia, Estados Unidos (TRAPANE; LATTMAN, 2007) e o CASP8 na ilha de Sardenia na Itália (KRYSHTAFOVYCH et al., 2009).

MÉTODOS DE PREDIÇÃO *AB INITIO*

Os métodos de modelagem *ab initio* são usados para predição da estrutura tridimensional, utilizando como ponto de partida, apenas a estrutura primária da proteína alvo. Neste método, são utilizados uma análise do espaço conformacional da estrutura unida a uma função de pontuação de energia a fim de encontrar as estruturas com o mínimo de energia potencial global. Normalmente, os algoritmos que empregam esta abordagem geram um grande número de modelos teóricos com diferentes conformações estruturais e um dos desafios é selecionar o modelo final (GOPAL; KLENIN; WENZEL, 2009).

Esta abordagem se destaca por não depender de estruturas homólogas previamente conhecidas que possam ser utilizadas como referência na modelagem, e emerge do fato de que muitas proteínas ainda não apresentam identidade suficiente ($\geq 30\%$) com seqüências de estruturas depositadas em bases de dados para a utilização de métodos de modelagem comparativa. Os métodos *ab initio* podem ser subdivididos em duas categorias: *ab initio* com informações de base de dados também chamados de baseados em conhecimento (em inglês *knowledge-based*), e *ab initio* sem informações de base de dados (FLOUDAS et al., 2006).

Os métodos *ab initio* baseados em conhecimento não comparam à proteína alvo com a proteína molde, mas sim curtos fragmentos entre ambas, e uma vez que os fragmentos compatíveis são identificados. Estes algoritmos iniciam a montagem dos mesmos para a formação da estrutura-alvo com o auxílio de funções de pontuação de energia e algoritmos de pesquisa conformacional. Os métodos *ab initio* sem informação de base de dados consideram que a conformação de uma proteína depende unicamente da sua estrutura primária. Estes métodos utilizam potenciais físicos para a resolução da estrutura e selecionam aquelas que melhor representam o estado nativo, por meio de algoritmos que empregam pesquisa conformacional baseada nos potenciais. Com relação aos pacotes de programas que podem realizar predição *ab initio* sem informações de base de dados, podemos citar os aplicados à simulação de dinâmica molecular, tais como CHARMM (BROOKS et al., 1983) (*Chemistry at Harvard Molecular Mechanics*), AMBER (*Assisted Model Building with Energy Refinement*), (CASE et al., 2005; SALOMON-FERRER; CASE; WALKER, 2013) e GROMACS (*GRoningen MACHine for Chemical Simulation*) (VAN DER SPOEL et al., 2005). É importante ressaltar que embora estes pacotes tenham sido desenvolvidos para simulação de dinâmica molecular, atualmente, eles podem ser utilizados na elucidação estrutural e para análise do enovelamento de proteínas.

Os métodos *ab initio* consideram que uma proteína pode adquirir diferentes conformações, porém a estrutura de menor energia global corresponde ao seu estado nativo (KUHLMAN; BAKER, 2000). Deste modo, os algoritmos de modelagem, normalmente, geram uma grande quantidade de modelos teóricos com o propósito de se obter diferentes conformações e empregam um tratamento estatístico para selecionar as estruturas que melhor representam a estrutura nativa em relação às demais que são geradas pelo processo de modelagem. Desta forma, é comum programas aplicarem um agrupamento (em inglês *clustering*) das estruturas similares, além de uma análise baseada em centroide, isto é, a seleção das conformações que melhor representam as demais do agrupamento (em inglês *cluster*).

Os algoritmos de agrupamento utilizam um critério de similaridade estrutural, como a Raiz do Desvio Quadrático Médio (RMSD do inglês *Root Mean Square Deviation*) do carbono alfa (C α), para a formação dos agrupamentos. Assim, é possível reduzir a quantidade de modelos teóricos gerados e analisar somente as estruturas mais significativas (ROHL et al., 2004; SHORTLE; SIMONS; BAKER, 1998; ZHANG; KOLINSKI; SKOLNICK, 2003; ZHANG; SKOLNICK, 2004). No geral, programas que realizam predição *ab initio* disponibilizam sua ferramenta de agrupamento, contudo, pacotes externos já foram desenvolvidos para esta finalidade como o MaxSub (SIEW et al., 2000).

A figura 22 representa esquematicamente o processo de agrupamento utilizado por programas *ab initio*. Além destes, são realizadas otimizações nas estruturas proteicas, a fim de melhorar a qualidade estereoquímica e o perfil de energia dos modelos obtidos. Devido suas peculiaridades, normalmente, os métodos *ab initio* requerem demasiado tempo de processamento computacional quando comparados aos outros métodos de predição *in silico*.

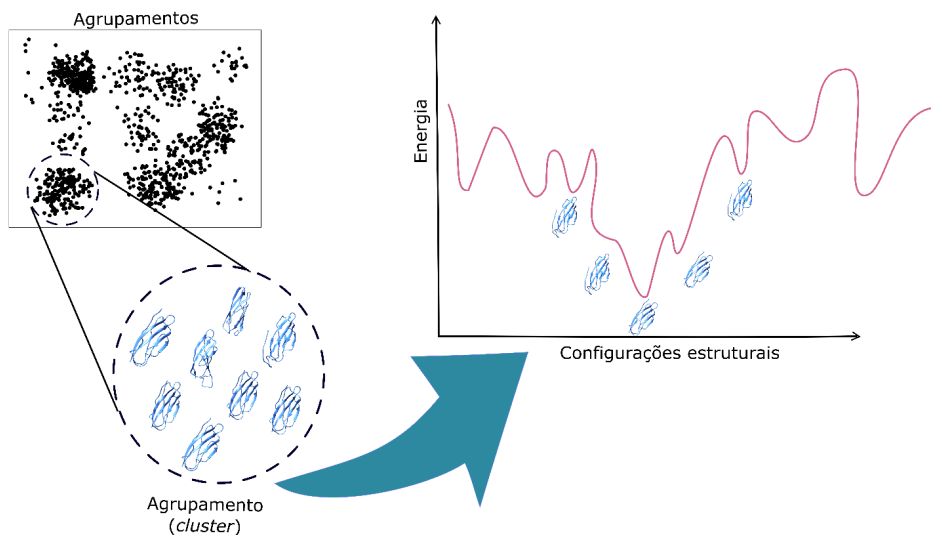


Figura 22. Visão geral do processo de agrupamento utilizado pelos métodos *ab initio*. Modelos de baixa energia global são estruturalmente semelhantes à forma nativa (KUHLMAN; BAKER, 2000).

O sucesso da predição *ab initio* depende de três componentes essenciais para seu funcionamento: (1) uma função de pontuação de energia capaz de discriminar corretamente a forma nativa das demais (denominadas *decoy*); (2) uma representação geométrica da proteína e (3) uma técnica de pesquisa conformacional capaz de discriminar, rapidamente, as formas de baixa energia que representam um estado termodinamicamente estável das outras formas, isto, realizar uma varredura na superfície de energia obtida pela geração das estruturas propostas. Atualmente, há diferentes programas que realizam a modelagem *ab initio* e aplicam diferentes estratégias de pesquisa conformacional e campos de força (KIHARA et al., 2001; KLEPEIS; FLOUDAS, 2003; SIMONS et al., 1999; XU; ZHANG, 2012; ZHANG; KOLINSKI; SKOLNICK, 2003). A tabela 1 sumariza os principais programas. Abordaremos, nas seções seguintes, separadamente cada um dos componentes da predição *ab initio*.

Algoritmo	Campo de Força	Seleção do modelo	Pesquisa Conformacional	Referências
ROSETTA	Físico e baseado em conhecimento	Agrupamento e menor energia potencial	Monte Carlo	(SIMONS et al., 1997)
UNRES	Físico	Agrupamento/ menor energia livre	<i>Conformational space annealing</i>	(LIWO; KHALILI; SCHERAGA, 2005)
I-TASSER	Baseado em conhecimento	Agrupamento e menor energia livre	Monte Carlo	(ZHANG, 2008, 2009)
ASTRO-FOLD	Físico	Menor energia	<i>Conformational space annealing</i> e dinâmica molecular	(KLEPEIS; FLOUDAS, 2003)
QUARK	Baseado em conhecimento	Agrupamento e menor energia livre	Monte Carlo	(XU; ZHANG, 2012)
TOUCHSTONE	Baseado em conhecimento	Agrupamento e menor energia livre	Monte Carlo	(KIHARA et al., 2001; ZHANG; KOLINSKI; SKOLNICK, 2003)
AMBER e CHARMM	Físico	Menor energia livre	Dinâmica molecular	(BROOKS et al., 1983; CASE et al., 2005)

Tabela 1. Resumo dos principais algoritmos utilizados para predição *ab initio* da estrutura de proteínas discriminados pelo tipo de campo de força, método de seleção do modelo, e o algoritmo de pesquisa conformacional. É importante ressaltar que pacotes de simulação de DM também podem ser usados na predição *ab initio* de proteínas.

REPRESENTAÇÃO GEOMÉTRICA

De modo geral, uma representação geométrica, que leve em consideração todos os átomos da cadeia polipeptídica, demandaria enorme tempo computacional devido ao elevado número de partículas e ângulos de torção necessários para a representação das estruturas. Dessa maneira, uma representação geométrica de proteínas devidamente adotada pode mostrar-se extremamente eficiente para os cálculos de energia potenciais do sistema analisado, sem comprometer o tempo de cálculo computacional.

Alguns algoritmos de predição *ab initio* diminuem os graus de liberdade através de modelos semirreduzidos, considerando somente os átomos pesados da cadeia

polipeptídica (ex.: N, C α , C, O, C β) para representar a cadeia principal e o centro de massa das cadeias laterais (XU; ZHANG, 2012). O algoritmo Rosetta utiliza os átomos pesados da cadeia principal e representa a cadeia lateral pelo C β mantendo os ângulos e distâncias de ligação constantes (SIMONS et al., 1997). O algoritmo Touchstone representa a estrutura da cadeia principal pelos átomos de carbono alfa (C α) com ligações laterais de átomos de carbono beta (C β) e centros de massa da cadeia lateral (KIHARA et al., 2001).

MÉTODOS DE PESQUISA CONFORMACIONAL

Há três principais métodos utilizados para a pesquisa conformacional da estrutura de proteínas: dinâmica molecular (DM), o método estocástico Metrópolis Monte Carlo (MC) (METROPOLIS et al., 1953) e os algoritmos genéticos. Os métodos de pesquisa conformacional analisam diferentes conformações possíveis que uma proteína pode adquirir de modo a encontrar aquela de menor energia e são cruciais, quando se deseja atingir modelos de ótima resolução atômica e conformação semelhante à forma nativa. É, por este motivo, novos métodos de pesquisa têm sido conduzidos ao longo dos anos, e tem se mostrado um campo ativo de estudo (DING et al., 2008; LIWO et al., 1998; THACHUK; SHMYGELSKA; HOOS, 2007; ZHANG, 1999).

Diferentes algoritmos Metrópolis MC foram desenvolvidos (ZHANG, 1999) e existem boas revisões sobre o assunto (HANSMANN; OKAMOTO, 1999). Provavelmente, a versão mais popular de algoritmo MC é o *Simulated Annealing* (SA) (KIRKPATRICK; GELATT; VECCHI, 1983) aplicado com sucesso em programas de predição *ab initio*, em especial, o algoritmo Rosetta (KAUFMANN et al., 2010; ROY; KUCUKURAL; ZHANG, 2010).

O SA é um algoritmo de otimização global que aplica o Metrópolis MC para gerar as conformações da proteína, seguindo a distribuição canônica de energia de Boltzman para determinada temperatura. Este algoritmo inicia a geração de perturbações aleatórias nas conformações, através do aumento da temperatura seguida das diminuições progressivas, e mantém as melhores soluções conformacionais baseadas no mínimo de energia da estrutura. Em resumo, objetiva encontrar as combinações de rotâmeros que ocupam o mínimo de energia global, de acordo com uma função de pontuação de energia (OKAMOTO, 2009). Os algoritmos genéticos são representados principalmente pela técnica de *Conformational space annealing* (CSA) implementados no ASTRO-FOLD (KLEPEIS; FLOUDAS, 2003) e UNRES (LIWO et al., 1998).

CAMPOS DE FORÇA E FUNÇÃO DE PONTUAÇÃO DE ENERGIA

Os campos de força são um conjunto de parâmetros físico-químicos e de funções matemáticas utilizadas para descrever a energia potencial de um sistema, estes contêm informações utilizadas pela função de pontuação de energia (do inglês *energy score function*) que irão determinar se a estrutura do modelo final obtida se mostra fidedigna

para representar a conformação nativa da proteína. Uma função de pontuação de energia (também referida por função de mérito) deve, portanto, auxiliar o método de pesquisa conformacional a discriminar corretamente as estruturas semelhantes à forma nativa daquelas não enoveladas ou mal enoveladas geradas no processo de modelagem, sendo desta forma, úteis para avaliar a qualidade dos modelos gerados. Baseando-se na utilização ou não de informações estatisticamente extraídas de estruturas resolvidas experimentalmente, os campos de força são divididos em duas categorias: baseados em conhecimento e campos de força físicos. Atualmente, os algoritmos *ab initio* aplicam um (KLEPEIS; FLOUDAS, 2003) ou ambos na predição da estrutura (RAMAN et al., 2009; ROHL et al., 2004).

Métodos de predição, que empregam campos de força físicos aliado aos métodos de pesquisa conformacional, podem ser considerados estritamente *ab initio*, por não utilizarem nenhuma informação proveniente de estruturas proteicas experimentalmente resolvidas. A utilização destes campos de força de potenciais físicos surge da hipótese termodinâmica de Anfinsen, pois tentam resolvê-la baseando-se na obtenção da sua menor energia livre global (PILLARDY et al., 2001). Com relação aos campos de força físico, podemos citar os já bem conhecidos OPLS (JORGENSEN; MAXWELL; TIRADO-RIVES, 1996), AMBER, (CASE et al., 2005) e CHARMM (BROOKS et al., 1983). Quando associados aos modelos de solvatação e à simulação de dinâmica molecular, estes campos de força podem se tornar eficientes não somente na elucidação da estrutura tridimensional, mas também no estudo do enovelamento de proteínas (VOELZ et al., 2010).

Com relação aos campos de força baseados em conhecimento, eles contêm termos de energia empíricos derivados de estruturas de proteínas resolvidas experimentalmente (SIPPL, 1990). Em alguns casos, estes campos abrangem potenciais de contato de pares de resíduos e propensões da sequência em desenvolver determinada estrutura secundária (HAO; SCHERAGAT, 1999; SKOLNICK, 2006). O algoritmo Rosetta, por exemplo, (SIMONS et al., 1999) aplica um campo de força baseado em conhecimento unido à montagem de fragmentos de proteínas de estrutura conhecida e, no final, depois de obtidos os modelos, estes são submetidos ao refinamento atômico através de um potencial físico (RAMAN et al., 2009).

ALGORÍTMOS DE MODELAGEM *THREADING* E *AB INITIO*

Anderson Henrique Lima e Lima

Kauê Santana da Costa

Alberto Monteiro dos Santos

Neste capítulo, abordaremos em mais detalhes as estratégias utilizadas por dois principais algoritmos que mostram resultados satisfatórios nos eventos do CASP: o I-TASSER (ZHANG, 2007;5 2009) que utiliza um misto de modelagem *ab initio* e *threading* e o Rosetta (BONNEAU et al., 2001; SIMONS et al., 1999) que emprega a abordagem *ab initio* baseada em conhecimento.

ALGORITMO ROSETTA

O algoritmo Rosetta foi desenvolvido pelo grupo do pesquisador David Baker, da Universidade de Washington (Estados Unidos), e emprega a modelagem *ab initio* baseada em conhecimento que, quando comparada às demais ferramentas usadas para a modelagem de novo, tem mostrado avanços significativos para a predição da estrutura tridimensional de proteínas durante os eventos de competição do CASP (RAMAN et al., 2009; TAI et al., 2014). Devido ao seu sucesso na predição *ab initio* de proteínas, o método empregado tem sido estendido aos outros problemas abordados na modelagem, entre eles: *docking* (em inglês

docking) de proteínas e o refinamento de regiões de alças (ROHL et al., 2004). O programa está disponível na versão instalável e *on-line* (<http://rosetta.bakerlab.org/>) no servidor Robetta (do inglês *Full-chain Protein Structure Prediction Server*) também desenvolvido e mantido pelo mesmo grupo.

A estratégia para predição de proteínas utilizada pelo Rosetta baseia-se no pressuposto de que a distribuição das conformações assumidas por um fragmento pequeno da sequência alvo pode ser aproximada pela distribuição de estruturas adotadas em outras proteínas sem relação evolutiva (BONNEAU et al., 2001; KAUFMANN et al., 2010; SIMONS et al., 1999), sendo assim, os modelos construídos são compostos de fragmentos de baixa energia de estruturas conhecidas. O Rosetta utiliza duas bibliotecas de fragmentos, porém, não homólogas: uma contém sequências de três e a outra de nove resíduos de aminoácidos. A pesquisa de fragmentos para a modelagem ocorre em base de dados de estruturas obtidas por difração de raios-X com resolução de 2,5 Å, ou melhor, e com identidade menor que 50%. Para a pesquisa de fragmentos ideais, também é utilizada a predição da estrutura secundária da proteína-alvo, sendo esta realizada em três servidores: PSIPRED, Sam-T99 e JUFO. Depois de predita a estrutura secundária, o nível de similaridade dos fragmentos em relação à estrutura alvo é calculado para cada posição

(KAUFMANN et al., 2010; ROHL et al., 2004). Atualmente, a forma mais prática para a obtenção das bibliotecas de fragmentos de três e nove resíduos é através do servidor Robetta (KIM; CHIVIAN; BAKER, 2004).

Os fragmentos selecionados são então utilizados para a construção do modelo que inicia, arbitrariamente, com a proteína na forma completamente distendida. Inicialmente, uma janela de fragmentos de nove resíduos da proteína-alvo é selecionada e, em seguida, um fragmento do base de dados contendo o mesmo tamanho é escolhido dentre as 25 melhores sequências classificadas para esta posição. Depois de selecionado, ocorre a substituição dos ângulos de torção ϕ , ψ e ω do fragmento da proteína-alvo pelos dos ângulos de torção do fragmento de estrutura conhecida e a energia resultante é avaliada a cada nova inserção. O processo de montagem dos fragmentos é realizado por meio de um processo de otimização baseado no algoritmo *Monte Carlo-Simulated annealing*, e utiliza um campo de força baseado em conhecimento que descreve todos os átomos de maneira explícita (em inglês, *all-atom*). Este leva em conta potenciais de Lennard Jones para forças de van der Waals, termos eletrostáticos, potenciais de ligações de hidrogênio, aproximações de interações de solvatação e termos de energia livre interna para a conformação de aminoácidos. Depois de finalizado o processo de montagem, este é reiniciado com janelas de fragmentos da proteína alvo contendo três resíduos (RAMAN et al., 2009; ROHL et al., 2004).

O Rosetta utiliza uma representação do espaço de torção em que o esqueleto proteico é especificado, exclusivamente, por modificações aleatórias dos ângulos de torção ϕ , ψ e ω . Desta forma, os comprimentos e ângulos de ligação de resíduos individuais são mantidos fixos e o espaço cartesiano é gerado pelas coordenadas atômicas dos átomos pesados do esqueleto proteico (ROHL et al., 2004). Esta estratégia, evidentemente, reduz o tempo computacional por diminuir os graus de liberdade da estrutura proteica e, conseqüentemente, o espaço de pesquisa conformacional. O funcionamento do algoritmo *ab initio* empregado pelo Rosetta é mostrado resumidamente na figura 23.

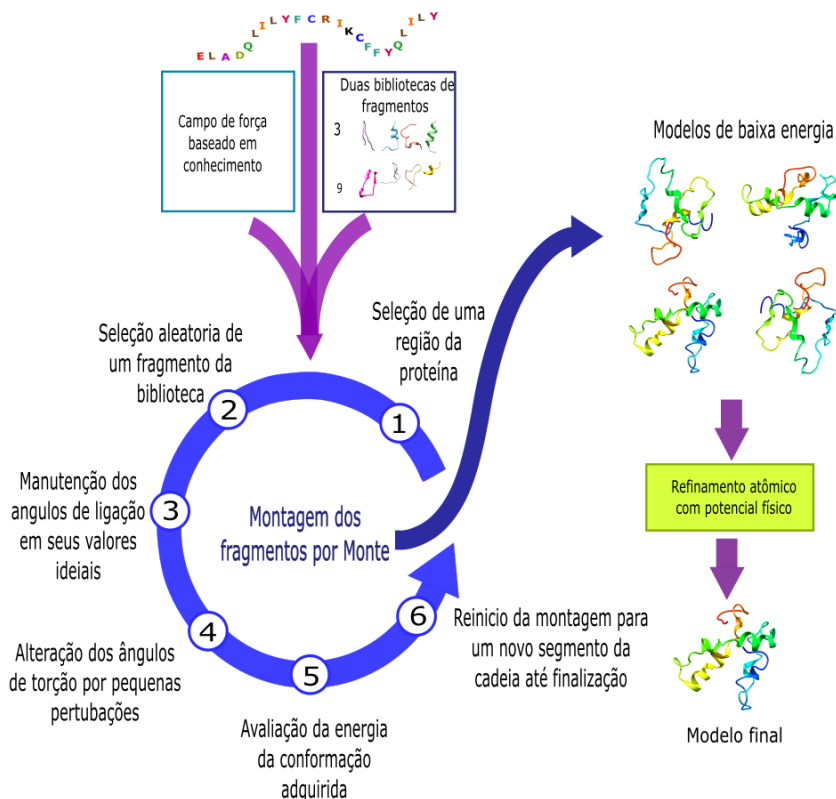


Figura 23. Resumo do funcionamento do algoritmo *ab initio* do Rosetta

A seleção dos modelos no Rosetta é baseada no menor perfil de energia da estrutura resultante e, para isto, se utiliza da função de pontuação de energia (ROHL et al., 2004; SIMONS et al., 1999). A maximização da função de pontuação do Rosetta segue a separação Bayesiana da energia total em componentes que descrevem a vizinhança da estrutura analisada, durante a simulação que emprega Monte Carlo. Baseando-se no fato de que diferentes estruturas apresentam conformações mínimas de energia próximas à estrutura nativa, o Rosetta também agrupa os modelos de acordo com a similaridade estrutural (com $C\alpha$ RMSD $< 5,0 \text{ \AA}$) e seleciona aqueles que aparecem no centro de cada um, isto é, seleciona aquelas estruturas que melhor representam as demais com a menor configuração de energia potencial (centroides), sendo esta, calculada através de todos os termos que compõem a sua função de pontuação de energia (KAUFMANN et al., 2010). Para informações mais detalhadas do protocolo de agrupamento utilizado pelo Rosetta, o leitor pode consultar Shortle et al. (2018) (SHORTLE; SIMONS; BAKER, 1998).

Para obter características de baixa resolução das estruturas nativas e melhor discriminar os modelos nativos dos demais, as conformações de baixa resolução

selecionadas são submetidas a um refinamento da estrutura que aplica um campo de força com funções de potenciais físicos (RAMAN et al., 2009).

O Rosetta utiliza um protocolo denominado *AbinitioRelax* que consiste em duas etapas principais na modelagem: etapa de montagem *ab initio*, dos fragmentos e a etapa de otimização (relaxamento) estrutural (BRADLEY; MISURA; BAKER, 2005). Na primeira etapa, é realizada uma pesquisa de granulação-grossa (do inglês *coarse-grained*) baseada em fragmentos no qual o espaço conformacional da proteína de interesse é pesquisado usando uma função de pontuação de energia “centroide” que permite determinar características, que se adequem à conformações nativas de proteínas (etapa *ab initio*). O segundo passo consiste no refinamento de todos os átomos usando o campo de força explícito detalhado do Rosetta (etapa *relax*). A segunda etapa, como envolve alterações nos ângulos de torção da proteína, demanda maior tempo computacional e também utiliza o protocolo de Minimização Monte Carlo (MCM) que em cada movimento da estrutura emprega: perturbações aleatórias dos ângulos de torção da cadeia principal da molécula (ϕ e ψ) e a otimização combinatória das conformações e rotações da cadeia principal e das cadeias laterais (RAMAN et al., 2009).

ALGORITMO I-TASSER

O algoritmo I-TASSER (do inglês *Iterative Threading ASSEmbly Refinement*) foi desenvolvido pelo grupo de Yang Zhang e consiste num algoritmo de modelagem avançada que emprega, em linhas gerais, um misto de predição por *threading* seguida de refinamento da estrutura proteica e predição *ab initio*. Comparado a outros métodos, tem mostrado resultados convincentes em consecutivos eventos do CASP (ZHANG, 2007, 2009). O programa está atualmente disponível para utilização acadêmica no servidor *on-line* mantido pela Universidade de Michigan (EUA) e na forma instalável compatível com sistemas Linux.

O funcionamento do algoritmo I-TASSER pode ser subdividido em três principais etapas.

Primeiramente, as sequências alvo são pesquisadas através de uma biblioteca de estruturas em formato PDB usando como valor de *cut-off* 70% em relação à identidade da sequência. Em primeiro lugar, o servidor faz uma pesquisa em base de dados não redundante pelo PSI-BLAST com o objetivo de encontrar aquelas que apresentam semelhanças estruturais. Baseando-se nos múltiplos alinhamentos com o seu homólogo, um perfil da sequência é então criado e o alinhamento também é utilizado para prever a estrutura secundária pelo servidor Psipred (MCGUFFIN; BRYSON; JONES, 2000). Obtidos o perfil da proteína e a estrutura secundária, as sequências analisadas são pesquisadas por *threading* em base de estruturas em formato PDB, usando o servidor local Lomets que combina sete programas de pesquisa *threading* que incluem entre outros, o Fugue (SHI; BLUNDELL; MIZUGUCHI, 2001), Prospect (GUO et al., 2004) e o Muster (WU; ZHANG,

2008). Com base nos perfis obtidos por estes servidores, as estruturas de referência são pontuadas e a qualidade do alinhamento é avaliada pelo Z-score que mostra a significância estatística da pontuação de energia (ROY; KUCUKURAL; ZHANG, 2010).

Na segunda etapa, as regiões pesquisadas por *threading* com alta pontuação que se alinham com a sequência alvo são, em seguida, removidas para posterior montagem deles na estrutura da proteína alvo utilizando simulações de Monte Carlo. As regiões da sequência alvo, que não alinham com as sequências pesquisadas por *threading*, são construídas por modelagem *ab initio*. O algoritmo I-TASSER utiliza para a representação da proteína, um modelo reduzido, em que cada resíduo na cadeia principal é especificado pelo carbono α e as cadeias laterais, pelo seu centro de massa. As estruturas modeladas (modelos) de menor energia potencial são agrupadas pelo algoritmo SPIKER e aquelas que melhor representam as demais do agrupamento (denominadas centroides) são selecionadas para posterior otimização.

Na terceira etapa, ocorre um segundo ciclo de modelagem, utilizando como ponto de partida, o centroide, obtido na etapa anterior e, nesta etapa, pretende-se obter estruturas de menor energia. Restrições espaciais que correspondem aos ângulos diedrais do primeiro centroide são extraídas e novamente estruturas experimentais em PDB são pesquisadas pelo programa de alinhamento estrutural TM-align (ZHANG, 2005).

O propósito desta segunda rodada de modelagem é remover possíveis conflitos estéricos da primeira, melhorando assim, a topologia da estrutura. A pontuação TM (em inglês, *TM-score*) é utilizada como critério quantitativo para mensurar a similaridade estrutural de duas proteínas. Valores no intervalo entre $0,5 < TM\text{-score} < 1,0$ exibem estruturas com alta similaridade e, portanto, de predição confiável, enquanto que valores de *TM-score* $< 0,5$ exibem proteínas de topologias diferentes (XU; ZHANG, 2010). O objetivo de propor o *TM-score* é resolver o problema de erros locais nos cálculos de RMSD da estrutura. Como o RMSD é uma distância média de todos os pares de resíduos entre duas estruturas, um erro local, por exemplo, causado pela desorientação numa região de alça (variável do ponto de vista conformacional) aumentará, demasiadamente, o valor do RMSD, embora o alinhamento estrutural global entre ambas as proteínas esteja condizente.

No cálculo do *TM-score*, no entanto, as pequenas variações nas distâncias obtidas do alinhamento entre ambas as estruturas serão melhor consideradas que as variações maiores, o que torna a pontuação por esta abordagem insensível ao erro de modelagem local. Adicional ao *TM-score*, o I-TASSER também utiliza uma pontuação de confiabilidade denominada *C-score* que estima a qualidade dos modelos gerados pelo programa. O *C-score* é calculado com base na significância dos alinhamentos dos moldes da estrutura obtidos pela pesquisa de *threading* (estruturas de referência) e nos parâmetros de convergência das simulações de construção da estrutura nativa. Os valores de *C-score* variam num intervalo de -5 a 2, onde uma pontuação maior indica um modelo de maior confiabilidade e os valores menores indicam um modelo de baixa confiabilidade. De modo

geral, a pontuação do *TM-score* está intimamente correlacionada à pontuação do *C-score*, porque indica que quanto maior a similaridade da estrutura alvo com a estrutura (ou estruturas) utilizada(s) como referência(s) (molde), maior será a confiabilidade de predição do modelo final.

Kauê Santana da Costa

Anderson Henrique Lima e Lima

ACESSANDO A SEQUÊNCIA PARA A MODELAGEM

Acesse o base de sequência de proteínas UniProt através do endereço: <https://www.uniprot.org/>

Pesquise no base do UniProt, uma sequência de aminoácidos de seu interesse. No campo “Sequence”, copie a sequência dos resíduos de aminoácidos e salve no seu computador em formato Fasta (Figura 24).

Neste capítulo, mostraremos a predição da estrutura pela abordagem do servidor I-TASSER (do inglês *Iterative Threading ASSEmbly Refinement*) e analisaremos os resultados de predição. O servidor, assim como a versão instalável do programa, foi desenvolvido pelo grupo de Yang Zhang, da Universidade de Michigan (EUA), e ambos empregam um algoritmo de modelagem avançada que utiliza um misto de predição por *threading* com predição *ab initio* de regiões sem cobertura.

The screenshot displays the UniProt Sequence page for entry A0A3Q8I9F4-1. The page is titled "Sequence" and shows the sequence status as "Complete". The sequence is displayed in FASTA format, with the following amino acid sequence: MSAENGNAI VRVSSNKRKF GYDVTKRL HEGYPEVVIS ALGTAIADAV SVVELLNQG VVEVKRICTS RAQFDDVRST TDKIEVVVV KSPDFDALYD QQQKREIRAK ARAEADEADD E. The sequence is shown in a table format with column markers at 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, and 120. On the right side, the sequence length is 121, the mass (Da) is 13,360, the last modified date is April 10, 2019 - v1, and the checksum is 703641A6EE9182BA. A "BLAST" button is visible. On the left side, there is a sidebar with feature view options, including "Function", "Names & Taxonomy", "Subcellular location", "Pathology & Biotech", "PTM / Processing", and "Expression".

Figura 24. Acesso às sequências de aminoácidos no base UniProt.

PESQUISANDO POR ESTRUTURAS HOMÓLOGAS

Utilizando o *BLASTp* realize uma busca de estruturas homólogas no *Protein Data Bank* que apresentem similaridade $\geq 30\%$ que possam ser utilizadas como molde para a modelagem da estrutura. Após realizar a pesquisa, caso o servidor retorne que não foram encontradas sequências similares na base do PDB (Figura 25) teremos que utilizar uma nova estratégia de modelagem.

The screenshot shows the BLASTp search results interface. The top navigation bar includes the NIH logo and 'U.S. National Library of Medicine National Center for Biotechnology Information'. The search title is 'BLASTp » blastp suite » results for RID-19DBUEYP016'. The search parameters are listed on the left:

Job Title	Protein Sequence
RID	19DBUEYP016 Search expires on 01-08 23:44 pm Download All
Program	Citation
Database	pdb See details
Query ID	Ic Query_48512
Description	None
Molecule type	amino acid
Query Length	121
Other reports	

The 'Filter Results' section on the right includes an 'Organism' filter (set to 'only top 20 will appear'), a search box for organism names, and an 'Add organism' button. Below this are filters for 'Percent Identity', 'E value', and 'Query Coverage', each with input fields and 'to' labels. 'Filter' and 'Reset' buttons are at the bottom of the filter section.

A yellow warning banner at the bottom of the page reads: **⚠** No significant similarity found. For reasons why, [click here](#).

Figura 25. Resultado de pesquisa de similaridade realizado no Protein Data Bank usando a ferramenta BLASTp do GenBank.


A abordagem *threading* é indicada para a predição da estrutura quando a similaridade da sequência alvo se mostra menor que 30% e a *ab initio* quando a identidade/similaridade entre ambas é inexistente (0%). No caso acima, como não foram encontradas similaridades suficientes ($< 30\%$), realizaremos a predição usando o servidor I-TASSER.

REALIZANDO A PREDIÇÃO PELO I-TASSER

Acesse o servidor I-TASSER, utilizando o endereço: <https://zhanglab.ccmb.med.umich.edu/I-TASSER/>

O servidor apresenta interface intuitiva e amigável, exibindo na primeira página os campos para preenchimento com as informações necessárias para a modelagem (figura 26, campos indicados pelas setas). Na primeira caixa, deve-se adicionar a sequência de aminoácidos da proteína em formato Fasta, seguido do e-mail de contato no qual os resultados serão enviados (obrigatório), a senha de registro criada no momento do cadastro do usuário (obrigatório) e um código identificador para reconhecimento da submissão (opcional).

Home Research Services Publications People Teaching Jobs



(The server completed predictions for [514429 proteins](#) submitted to the server. [\(The template library was updated on 2014-01-22\)](#))

I-TASSER (Iterative Threading ASSEMBly Refinement) is a hierarchical approach to protein structure and function prediction. It generates full-length atomic models constructed by iterative template-based fragment assembly simulations. Function insights of the 'Zhang-Server' was ranked as the No. 1 server for protein structure prediction in recent community-wide [CASP7](#), [CASP8](#), and [CASP9](#). The server is in active development with the goal to provide the most accurate structural and functional predictions. [\(The template library was updated on 2014-01-22\)](#) and our developers will study and answer the questions accordingly. ([> More about the server ...](#))

[\[Queue\]](#) [\[Forum\]](#) [\[Download\]](#) [\[Search\]](#) [\[Registration\]](#) [\[Statistics\]](#) [\[Help\]](#)

I-TASSER On-line Server ([View an example of I-TASSER output](#)):

Copy and paste your sequence below ([10, 1500] residues in **FASTA format**). [Click here for a sample input.](#)

```
MSAENGNGNAIVRVSSNKRKFGYVDYTKHRLHEGYPEVVVISALGTAIADAVSVVVELLKNQG
VVEVKKICTSRAQFDDVVRSTTTDKIEVVVVKSPDFDAIYDQQQKDREIAKARAEADEADD
E
```

Or upload the sequence from your local computer:

No file selected.

Email: (mandatory, where results will be sent to)

Password: (mandatory, please [click here](#) if you do not have a password)

Figura 26. Campos de preenchimento obrigatório para a modelagem utilizando o servidor I-TASSER.

O servidor também dispõe da opção de incluir uma estrutura cristalográfica como referência e definir restrições para a formação de interações intramoleculares da estrutura para realizar a modelagem (Figura 26). Na caixa *Specify template without alignment*, é possível fazer o upload do arquivo PDB da estrutura que será usada como referência e na caixa *Assign contact/distance restraints* é possível adicionar as restrições referentes às interações entre os resíduos. Segue abaixo um exemplo

```
DIST 12 HG21 50 HB1 8.1
DIST 14 HA 57 1HE 6.2
DIST 21 HB2 43 HD11 4.0
DIST 124 CA 84 CA 17.4
DIST 36 UNK 120 CA 17.4
```

O arquivo também pode especificar o contato entre os resíduos especificando a numeração de cada um:

CONTACT 33 6
 CONTACT 60 29
 CONTACT 37 345
 CONTACT 109 42

▼ **Option I: Assign additional restraints & templates to guide I-TASSER modeling.**

(Read more explanation on how to add restraints)

- o Assign contact/distance restraints No file selected. Atom-atom restraints [Explanation](#)
- o Specify template without alignment Type a PDB ID [Explanation](#)
- o Specify template without alignment No file selected. Upload a structure [Explanation](#)
- o Specify template with alignment No file selected. Template & alignment [Explanation](#)

Figura 26. Opções de modelagem usando o servidor I-TASSER

ANÁLISE DOS RESULTADOS DE PREDIÇÃO

Os resultados de modelagem do I-TASSER são exibidos em diferentes caixas na página do servidor (Figura 27). Na primeira caixa é exibida a sequência de aminoácidos em formato Fasta da proteína submetida (*Submitted Sequence in FASTA format*), na segunda a predição da estrutura secundária da sequência (*Predicted Secondary Structure*) e na terceira as regiões dos resíduos acessíveis ao solvente (*Predicted Solvent Accessibility*).

Nesta terceira caixa, cada resíduo é indicado abaixo com uma numeração que varia de 0 a 9 que indica o grau de exposição ao solvente, sendo 0 os resíduos menos expostos e 9 os resíduos mais expostos.

Submitted Sequence in FASTA format

```

>Protein
HQKVLVNSIKDRNTIQDNSTLEVTLSKYSTSPYLLEAAHLSFENPVEDEAKICVQELQ
CFKIITKNSPLPSSII SRREKNDVFLLEAGRAEKIVISRSTPTFNKQTKRVSNWSPNS
LQVFTGKIPKATPELGSSENSASPPRFKTEKMEKVLPTFDKCESNLTVNTS

```

Predicted Secondary Structure

	20	40	60	80	100	120	140
Sequence	HQKVLVNSIKDRNTIQDNSTLEVTLSKYSTSPYLLEAAHLSFENPVEDEAKICVQELQCFKIITKNSPLPSSII SRREKNDVFLLEAGRAEKIVISRSTPTFNKQTKRVSNWSPNSLQVFTGKIPKATPELGSSENSASPPRFKTE						
Prediction	CCCCCCCCCCCCCCCCCCSSSSSSCCCHHHHHHHHCCCCCHHHHHHHHCCCHHHCCCCCCCCCCCCCCCCCHHHHHHHHHHCCCCCCCCCCCCCCCCCCCHHHCCCCCCCCCCCCCCCCCCCCCC						
Conf. Score	9520024532234335676625677743786267876652058877421246676765236234338989876433223587078898875404267314788623445534553330334303530236752215433366776310001						
	H:Helix; S:Strand; C:Coil						

Predicted Solvent Accessibility

	20	40	60	80	100	120	140
Sequence	HQKVLVNSIKDRNTIQDNSTLEVTLSKYSTSPYLLEAAHLSFENPVEDEAKICVQELQCFKIITKNSPLPSSII SRREKNDVFLLEAGRAEKIVISRSTPTFNKQTKRVSNWSPNSLQVFTGKIPKATPELGSSENSASPPRFKTE						
Prediction	844336334764443666433130222444333212433443347334656241214425234133644523552444566323235216454331333444414543562414224323420344035334441434743444441444						
	Values range from 0 (buried residue) to 9 (highly exposed residue)						

Acesse Configurações para ativar o Windows.

Figura 27. Resultados de predição do I-TASSER mostrando a predição da estrutura secundária e as regiões dos resíduos acessíveis ao solvente.

Na quarta caixa é exibido um gráfico do *B-factor* (*Predicted normalized B*) distribuído ao longo da sequência de aminoácidos juntamente com a predição da estrutura secundária

(Figura 28). Os valores de *B-factor* são um importante indicativo da mobilidade e flexibilidade de regiões da estrutura (resíduos e átomos da proteína). No I-TASSER, os valores do gráfico são deduzidos das estruturas das proteínas obtidas do base PDB utilizadas como molde, assim como, dos perfis de sequência derivados dos bases de dados de sequência.

Predicted normalized B-factor

(B-factor is a value to indicate the extent of the inherent thermal mobility of residues/atoms in proteins. In I-TASSER, this value is deduced from threading template proteins from the PDB in combination with the sequence profiles derived from sequence databases. The reported B-factor profile in the figure below corresponds to the normalized B-factor of the target protein, defined by $B=(B-u)/s$, where B is the raw B-factor value, u and s are respectively the mean and standard deviation of the raw B-factors along the sequence. [Click here to read more about predicted normalized B-factor](#))

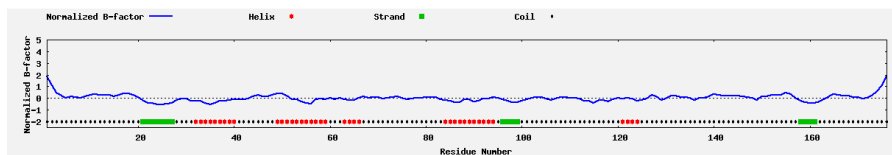


Figura 28. Gráfico do *B-factor* plotado em função da sequência de resíduos de aminoácidos da proteína.

A quinta caixa (Figura 29) exibe o resultado da pesquisa das estruturas experimentais obtidas no PDB utilizadas, como molde para modelagem (*Top 10 threading templates used by I-TASSER*). Na Figura 29, a seta no topo indica a predição da estrutura secundária ao longo da sequência de resíduos, na qual “H” indica as regiões de alfa-hélice, “S” as regiões de beta-folha, e “C” as regiões de alça da estrutura; logo abaixo é exibido o alinhamento *threading* das sequências pesquisadas como molde.

As sete colunas que exibem os resultados indicam da esquerda para a direita: (1) *rank* das melhores estruturas (coluna *Rank*), onde as melhores são exibidas primeiro e são seguidas das demais; (2) os códigos de acesso do PDB das estruturas usadas como moldes, seguido da cadeia (coluna *PDB Hits*); (3) o valor percentual de identidade da sequência dos moldes usados para a região alinhada com a sequência indagada (*query*) (coluna *Iden1*); (4) o valor percentual de identidade da sequência de todas as cadeias das estruturas moldes (*templates*) com a sequência de indagada (*query*); (5) cobertura do alinhamento, que representa o número de resíduos alinhados dividido pelo comprimento da sequência proteína indagada (*query*) (coluna *Cov*); (6) valor de Z-score referente ao alinhamento, sendo que o ideal são valores maiores que 1, pois indicam alinhamento satisfatório; (7) link para *download* das estruturas pesquisadas pelo alinhamento do *threading* utilizado pelo I-TASSER.

Rank	PDB Hit	Ident1	Ident2	Cov	Norm. Z-score	Download	Align.
1	6cmk3	0.13	0.15	0.82	1.11	Download	
2	2pft	0.15	0.40	0.92	1.21	Download	
3	8kzA	0.04	0.10	0.94	1.27	Download	
4	5mc9	0.04	0.41	0.97	1.08	Download	
5	1tt4A	0.17	0.23	0.69	1.21	Download	
6	2ma5	0.10	0.17	0.88	1.12	Download	
7	3ica5	0.10	0.44	0.99	1.08	Download	
8	2pfa	0.12	0.22	0.97	1.10	Download	
9	3jcsN	0.09	0.20	0.96	1.10	Download	
10	6fthD	0.09	0.33	0.98	1.04	Download	

Sec.Str	Seq
CCCCCCCCCCCCCCCCCCCCSSSSSSSSCCCCNNNNNNNNNNCCCCCCCCNNNNNNNNNNCCCCCCCCCCCCCCCCCCCCNNNNNNNNNNCCCCCCCCCCCCCCCCCCCC	HQRVLLVIVIKRRTNIQDNRESTLEVTSLSKYSTPFLLEAAASHLSPEHFVEDEAKICVQLQCPKIIITKNSPLPSSIIIRREKKNDFVLEEAGRAEIVISRSSTSPFHNGIKRVVSWSSFN
-----SSLHYAYRGRYLICVHGR-----ELIKFRG--NTCVLALMGHEQLHLLQLPFRINRGEGRPAIHIACNDYVQCLSLIIGVGLVMDVINDTFLHY-----CLEYGSIE	AQLVAIFGQRI-----TIDY---FEELIKVQVTFVMDLRFKAEFL-----SELIITIDAEVTFEWEENR-----TFQDELIVIVVIALELFTFGAGHSIAEAEDEFF
-----PAPARAAASAAVATADAAADATASASAVAAATADADASATIHASASAVAAATADADASATIHASASAVAAATADADASAAVAVASAAALEITVVAAYAAATANT	-----PAPARAAASAAVATADAAADATASASAVAAATADADASATIHASASAVAAATADADASATIHASASAVAAATADADASAAVAVASAAALEITVVAAYAAATANT
-----LRLTIGEMVYGVAVFVSNLFPVAYAQKFLNATVENITFESFVWQVYVYVYIAC--LQPDIDILFHGDTQIGERGI--LSGQGRVRSVAVARLYQGVNVPFLDQFSL-----E	-----LRLTIGEMVYGVAVFVSNLFPVAYAQKFLNATVENITFESFVWQVYVYVYIAC--LQPDIDILFHGDTQIGERGI--LSGQGRVRSVAVARLYQGVNVPFLDQFSL-----E
-----RSIVDHYNIKIF-----DSDDLIPVLDQVY-----IGHLKHIEHERM-----LFRMVFDFP-----FLKIRHDSILK	-----RSIVDHYNIKIF-----DSDDLIPVLDQVY-----IGHLKHIEHERM-----LFRMVFDFP-----FLKIRHDSILK
-----MDAIAKKNMQLKIDENALDRAEQADKKAADRDRSQDELVSLQKIKGTEFLDRYSALKDAQKLELAEKAT--DAEAVASIMRRIQVVEELDRAQERLA	-----MDAIAKKNMQLKIDENALDRAEQADKKAADRDRSQDELVSLQKIKGTEFLDRYSALKDAQKLELAEKAT--DAEAVASIMRRIQVVEELDRAQERLA
-----HFDSSNKIVLNLMIQGRINVIIEEPESEDLTHLAQKFPILNLIKIDSYVNVKSYIV--NDFISLNGAMTVSRVDRLIKLCERLDLIFONGINKDQLISCFAGAIGE	-----HFDSSNKIVLNLMIQGRINVIIEEPESEDLTHLAQKFPILNLIKIDSYVNVKSYIV--NDFISLNGAMTVSRVDRLIKLCERLDLIFONGINKDQLISCFAGAIGE
-----LGGAKVVTTSRFSK--QTDVYQSIYAKYG--AKSSTLVVFPNQSQKQVLEALIEFLYLDALFPFAAEGEILDSFARHMLINKNSARGIETFRMNHGDSSE--ANQL	-----LGGAKVVTTSRFSK--QTDVYQSIYAKYG--AKSSTLVVFPNQSQKQVLEALIEFLYLDALFPFAAEGEILDSFARHMLINKNSARGIETFRMNHGDSSE--ANQL
-----IGGAVANLYWDDRQSELEHQGVAVKARIALEKIIEGVNINTRKAEVKKLVLDLRGSRG--VRMDLASSAKVWVKTSENNNAVSDGSAVS---HDEVAEMRG	-----IGGAVANLYWDDRQSELEHQGVAVKARIALEKIIEGVNINTRKAEVKKLVLDLRGSRG--VRMDLASSAKVWVKTSENNNAVSDGSAVS---HDEVAEMRG
-----GVQQLTKLTKRPRGETVGTAVAFSSDFRFLFLOGEGGF--LKCSLAPAFQFSPHGS--FRRNLSLQAGTDGRVHLSMLQAPPLTSLQLSLFLFAVRKSI-----VRLVFAA--GR	-----GVQQLTKLTKRPRGETVGTAVAFSSDFRFLFLOGEGGF--LKCSLAPAFQFSPHGS--FRRNLSLQAGTDGRVHLSMLQAPPLTSLQLSLFLFAVRKSI-----VRLVFAA--GR

Figura 29. Alinhamento das seqüências utilizadas como molde com a seqüência indagada (*query*); e resultados associados, tais como valores de identidade, cobertura e Z-score.

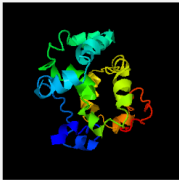
A sexta caixa (*Top 5 final models predicted by I-TASSER*) exibe o resultado da modelagem do servidor, isto é, as estruturas preditas para a seqüência submetida (Figura 30). Nela são exibidas as cinco melhores estruturas com base nos valores de *TM-score* (ver capítulo 6) e *C-score* utilizado para estimar a qualidade dos modelos gerados pelo I-TASSER, sendo calculado com base nos alinhamentos do modelo e nos parâmetros de convergência das simulações de montagem da estrutura. O C-score é normalmente encontrado no intervalo que varia entre os valores de -5 a 2, onde quanto maior a pontuação maior será a confiança em relação ao modelo gerado. Esta pontuação tem utilidade semelhante aos valores de similaridade estrutural calculados pelo *TM-score* e RMSD-Ca (ver Capítulo 2 e Capítulo 5) para prever a proximidade da estrutura modelada com o molde gerado. Preferencialmente, deve-se escolher o primeiro modelo, devido aos valores satisfatórios de *C-score* e *TM-score* obtidos na modelagem.

Top 5 final models predicted by I-TASSER

(For each target, I-TASSER simulates a large ensemble of structural conformations, called decoys. To select the final model, I-TASSER uses the SPICKER program to cluster all the decoys based on the pair-wise structure similarity, and reports up to five models which corresponds to the five largest structure clusters. The confidence of each model is quantitatively measured by C-score that is calculated based on the significance of detecting template alignments and the convergence parameters of the structure assembly simulation. C-score is typically in the range of (-5, 2), where a C-score of a higher value signifies a model with a higher confidence and vice-versa. TM-score and RMSD are estimated based on C-score and protein length following the correlation observed between these quantities. Since the top 5 models are ranked by the cluster size, it is possible that the lower-rank models have a higher C-score in some cases. Although the first model has a better quality in most cases, it is also possible that the lower-rank models have a better quality than the higher-rank models in some benchmark tests. If the I-TASSER simulation coverage, it is possible to have less than 5 clusters generated; this is usually an indication that the models have a good quality because of the converged simulation.)

[View About C-score](#)
[Learn structure quality: profile of the top five models](#)

(By right-click on the images, you can export image files or change the configurations, e.g. modifying the background color or stopping the spin of your model.)



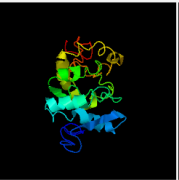
[Reset to initial orientation](#) Spin On/Off

[Download Model 1](#)

C-score = 4.39 (Read more about C-score)

Estimated TM-score = 1.25-0.89

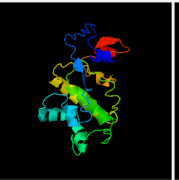
Estimated RMSD = 15.7-3.3 Å



[Reset to initial orientation](#) Spin On/Off

[Download Model 2](#)

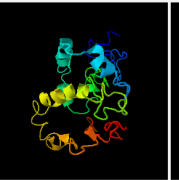
C-score = -4.47



[Reset to initial orientation](#) Spin On/Off

[Download Model 3](#)

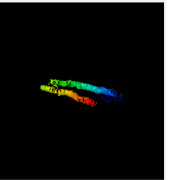
C-score = -4.64



[Reset to initial orientation](#) Spin On/Off

[Download Model 4](#)

C-score = -4.87



[Reset to initial orientation](#) Spin On/Off

[Download Model 5](#)

C-score = -3.60

[Acesse Configurações para ativar o Windows.](#)

Figura 30. Estruturas modeladas pelo I-TASSER. Por padrão o servidor libera as cinco melhores estruturas, seguida do valor de *C-score* e *TM-score*.

APLICAÇÕES DA MODELAGEM POR HOMOLOGIA, *AB INITIO* E *THREADING*

Anderson Henrique Lima e Lima

Alberto Monteiro dos Santos

Kauê Santana da Costa

Nas seções a seguir, serão discutidas algumas aplicações inteligentes dos métodos que utilizam a abordagem *ab initio* e *threading* em diferentes estudos.

As aplicações do conhecimento da estrutura de proteínas são inúmeros e incluem: (1) estudo conformacional de enzimas (COSTA et al., 2019; SITTEL; JAIN; STOCK, 2014); (2) análise do mecanismo catalítico enzimático (LAMEIRA et al., 2019; MORAES et al., 2012); (3) predição da função de proteínas; (4) estudo de mutantes (NEVES CRUZ et al., 2019); (5) planejamento *in silico* de fármacos e outros compostos bioativos (DA COSTA et al., 2019; LI et al., 2017) (6) estudo do modo de ligação de inibidores (ZHOU et al., 2013).

Por utilizarem uma estrutura de referência, os métodos de modelagem comparativa se mostram, de modo geral, mais fiéis à representação estrutural dos modelos em relação àqueles gerados por métodos livres de molde. Deste modo, as estruturas obtidas por estes métodos têm sido aplicadas com sucesso para explicar diferentes questões biológicas, de modo similar aos modelos obtidos por métodos experimentais. Em outros casos, os métodos de modelagem comparativa são aplicados, como ferramentas auxiliares, para corrigir e completar modelos gerados por métodos experimentais.

APLICAÇÕES DOS MÉTODOS *AB INITIO* E *THREADING*

Os métodos *ab initio* se destacam pelas suas aplicações no estudo de enovelamento de proteínas e no desenho de proteínas com novas topologias e atividades catalíticas. O desenho, também chamado de *problema inverso* de *modelagem de proteínas*, desponta como um campo promissor devidos as suas aplicações biotecnológicas e tem mostrado relativo sucesso quando aplicado em conjunto com técnicas experimentais que validam os achados (KHOURY et al., 2014). A figura 24 exhibe uma visão geral da aplicação dos métodos *ab initio* na modelagem e no desenho da estrutura de proteínas.

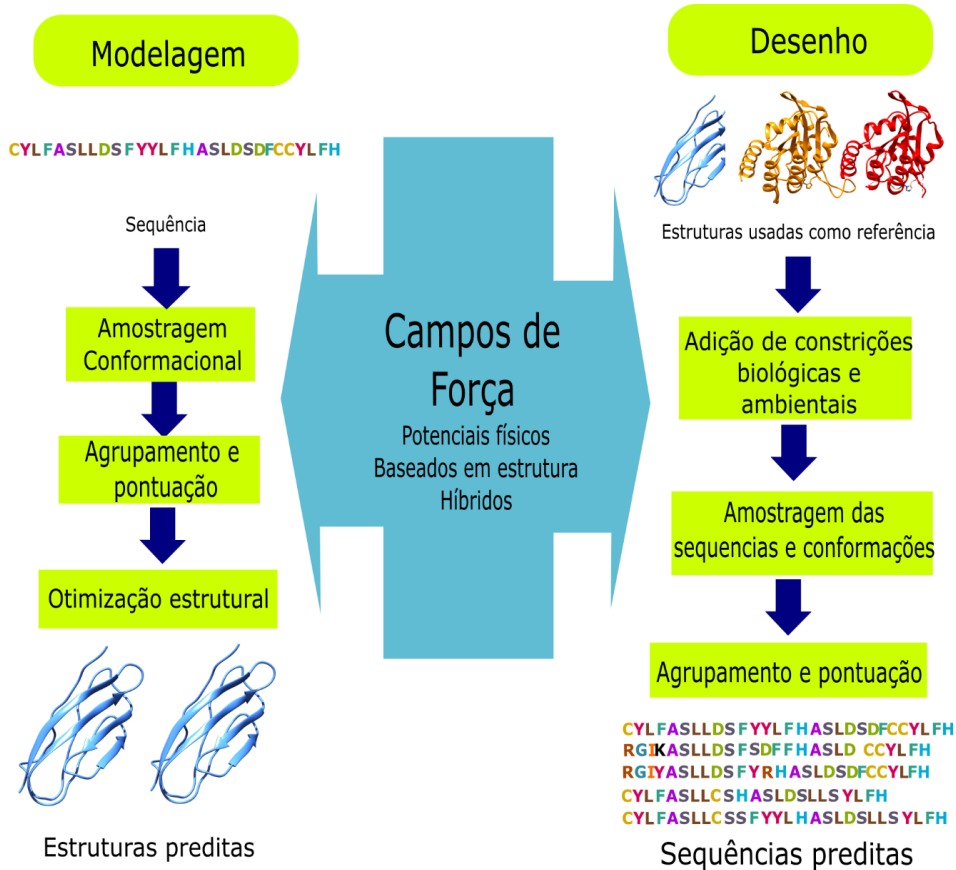


Figura 24. Visão geral do funcionamento e aplicação dos métodos *ab initio* quando usados no desenho e na modelagem de estruturas de proteínas.

DESENHO DE PROTEÍNAS COM NOVAS TOPOLOGIAS

Os métodos *ab initio* são atualmente a única solução para o desenho de novas proteínas (proteínas engenheiradas). Os métodos, portanto, representam um importante passo para o desenvolvimento da biologia sintética através do desenho computacional de proteínas com novas topologias e funções (CORREIA et al., 2014; JIANG et al., 2008; KUHLMAN et al., 2003; RICHTER et al., 2011).

Correia et al. (2014) utilizaram a modelagem *ab initio* para o desenho de uma nova topologia de proteínas, contendo epítomos específicos para a composição de vacinas com o propósito de neutralizar o Vírus Respiratório Sincicial (RSV). Através do Rosetta, os autores utilizaram um protocolo denominado *Fold From Loops* (FFL) que consiste, resumidamente, nos seguintes passos: (1) seleção do motivo proteico funcional e topologia alvo que será unida ao motivo; (2) modelagem *ab initio* para a construção de diferentes esqueletos proteicos que se mostrem compatíveis com a topologia alvo; (3) refinamento das estruturas

obtidas com o objetivo de selecionar aquelas de menor energia para dada conformação; (4) a filtragem dirigida por escolha humana para a correção de falhas remanescentes e seleção dos melhores arcabouços que preservem o epítipo viral.

Neste estudo, os autores produziram pequenos arcabouços proteicos, contendo um novo tipo de dobramento, mas que preservavam um epítipo viral do RSV. Estes oligopeptídeos foram unidos em partículas multivalentes utilizadas na composição de vacinas, e estas, utilizadas na imunização de macacos. As proteínas desenhadas conseguiram, de maneira acurada, mimetizar os epítopos nativos do vírus, sendo capazes desta forma, de induzir uma resposta imune humoral satisfatória nos animais.

PREDIÇÃO EM LARGA ESCALA DE PROTEÍNAS E ANOTAÇÃO DE GENOMAS

Os métodos *ab initio* podem auxiliar a anotação funcional de matrizes abertas de leitura (ORFs, do inglês *Open Read Frame*) de genes, assim como na determinação da função e mecanismo de ação de proteínas e permitir a análise dos diferentes tipos de dobramento de proteínas em estudos de genômica estrutural (JONES, 2000; JOSHI; JYOTHI, 2003; KIHARA et al., 2002; SIMONS; STRAUSS; BAKER, 2001). Kihara et al. (2002), por exemplo, acessaram, em larga escala, a estrutura de diferentes proteínas expressas pelo genoma *Mycoplasma genitalium*, utilizando para isto, a plataforma de predição Touchstone (KIHARA et al., 2001). Esta plataforma aplica a predição *ab initio* e utiliza restrições obtidas por *threading* como o objetivo de reduzir o espaço de pesquisa para predição da estrutura, além disso aplica como método de pesquisa conformacional e refinamento, o Metrópolis MC.

Neste estudo, foi possível prever a ultraestrutura de 85 pequenas proteínas com tamanho igual ou menor a 150 resíduos de comprimento, sendo que destas 34 mostraram-se possíveis de serem preditas pelo método *threading*, das 51 proteínas restantes, 29 convergiram para cinco agrupamentos, destas, se baseado no conjunto treino, foi possível obter 24 proteínas (~84,8% das 29 proteínas) com dobramento correto e correspondente à forma nativa (KIHARA et al., 2002). Outros estudos utilizam abordagens mistas que unem a modelagem por homologia, *threading* e *ab initio*. Singh et al. (2014), por exemplo, empregaram estes métodos com o objetivo de realizar a anotação baseada em estrutura de genes de *Helicobacter pylori*, para isto os autores utilizaram as plataformas I-TASSER (servidor que realiza um misto de *threading* e *ab initio*), (ZHANG, 2008).

O base de dados ModBase (com modelos obtidos por modelagem comparativa) e Phyre2 (servidor para modelagem por *threading*) (KELLEY et al., 2015), além de servidores de alinhamento e servidores de predição estrutural, tais como: o Psipred (MCGUFFIN; BRYSON; JONES, 2000) para a predição da estrutura secundária, o pGenTHREADER (LOBLEY; SADOWSKI; JONES, 2009) para reconhecimento de dobramento e identificação

de homólogos distantes, e o Fugue (SHI; BLUNDELL; MIZUGUCHI, 2001) para alinhamento também baseado em *threading*. Os autores conseguiram anotar 464 proteínas de um total de 557 analisadas e conhecidas previamente como não caracterizadas (SINGH; GUTTULA; GURUPRASAD, 2014).

PREDIÇÃO DE MUDANÇAS CONFORMACIONAIS

Recentemente, os métodos *ab initio* foram aplicados para prever conformações induzidas por regulação alostérica de determinadas moléculas efectoras. Embora, atualmente, a análise, em detalhes de efeitos alostéricos, possa ser realizada por meio de simulação de dinâmica molecular, esta ferramenta é limitada para alterações conformacionais na escala de nanosegundos. Partindo disso, Kidd et al. (KIDD; BAKER; THOMAS, 2009) utilizaram a metodologia de predição do Rosetta para prever alterações conformacionais em três proteínas alostéricas bem conhecidas (CheY, Integrina α L domínio I, e Ras) com o objetivo de determinar os estados não ligados (sem o ligante) a partir de estruturas ligadas (complexadas ao ligante) (KIDD; BAKER; THOMAS, 2009).

Neste trabalho, os autores utilizaram, como critério de seleção das proteínas do conjunto de teste, aquelas que apresentavam estruturas ligadas e não ligadas no *RCSB Protein Data Bank*, com resolução $\leq 2,5$ Å, sequências com < 200 aminoácidos, e um rearranjo estrutural C α -C α entre as duas formas maior que 3,5 Å. O protocolo de predição consistiu em três principais passos: (1) a geração da diversidade de estrutural através do algoritmo de montagem de fragmentos, o Metrópolis Monte Carlo e utilizando como estrutura de partida, a forma nativa ligada (holo), (2) o refinamento das cadeias laterais de todos os resíduos (3) minimização de todos os átomos explícitos da proteína.

Para seleção, utilizaram um algoritmo de agrupamento e definiram como limiar para as estruturas similares do mesmo agrupamento um RMSD de não mais que 1,0 Å e selecionaram aquelas de menor energia que se encontravam nos agrupamentos com o maior número de estruturas. Além disso, calcularam a energia de interação de pares de resíduos. Os autores especificaram em quais regiões da proteína eram mais susceptíveis de alterações induzidas por efetores alostéricos e obtiveram maior sucesso na predição de alterações conformacionais quando analisaram estruturas secundárias e alças e conseguiram ainda prever com acurácia de 0,3–3,4 Å (C α -RMSD) estruturas de baixa energia próximas à conformação cristalográfica das formas não ligadas.

DESENHO DE ENZIMAS CONTENDO NOVOS SÍTIOS CATALÍTICOS

Um campo extremamente promissor e recente na Biologia Estrutural é o desenho computacional de proteínas com sítios ativos capazes de executar novas atividades catalíticas (RICHTER et al., 2011; ZANGHELLINI et al., 2006). O desenho computacional normalmente é referido como um problema inverso de predição, pois inicia pela análise das

estruturas de proteínas conhecidas em que se deseja extrair características interessantes, para então se modelar sequências que possam adquirir a estrutura desejável para estas características. Como diferentes sequências podem adquirir a mesma estrutura, podemos considerar que existe uma redundância no espaço de pesquisa das sequências (KHOURY et al., 2014).

Os métodos *ab initio* são os preferidos nas estratégias de desenho computacional, a exemplo disto, Richter et al. (2011), utilizando a estratégia de modelagem do Rosetta, desenvolveram um protocolo para o desenho de enzimas que catalisam reações químicas novas (RICHTER et al., 2011). O protocolo consiste resumidamente nos seguintes passos: (1) escolha do mecanismo catalítico (utilizando métodos de mecânica quântica) e do arcabouço mínimo do sítio ativo envolvido na reação; (2) a seleção de sítios de um conjunto de esqueletos proteicos existentes que possam representar o sítio ativo idealizado para a catálise da reação química através do módulo RosettaMatch (ZANGHELLINI et al., 2006); (3) desenho do sítio ativo através dos arcabouços proteicos de modo a permitir a complementaridade espacial com o substrato (e seus estados de transição) e de estabilizar as cadeias laterais dos resíduos catalíticos nas suas conformações de ligação; e o último passo consiste na (4) validação e pontuação das sequências obtidas com relação à suas características intrínsecas, tais como número de ligações de hidrogênio e contatos não locais entre resíduos (Figura 25) (RICHTER et al., 2011).

A evolução dirigida pode ser utilizada como técnica complementar ao desenho computacional de proteínas com o objetivo de melhorar a atividade catalítica de enzimas engenheiradas (ALTHOFF et al., 2012; GIGER et al., 2013; KHERSONSKY et al., 2011). A evolução dirigida refere-se às alterações aleatórias na sequência de DNA de determinado gene, induzidas por técnicas como Error-prone PCR e DNA *shuffling*, seguida de clonagem e seleção dos clones recombinantes que exibam o fenótipo desejado. Este processo é repetido de maneira cíclica com o objetivo de incrementar características estruturais de proteínas selecionadas na fase inicial (JÄCKEL; KAST; HILVERT, 2008). No entanto, por ser um processo randômico que não analisa diretamente a estrutura da proteína, técnicas computacionais podem favorecer a análise de determinantes envolvidos no mecanismo de ação molecular e, portanto, auxiliar no desenho racional de proteínas com atividade melhorada.

Similarmente Röthlisberger et al. (2008) desenharam enzimas capazes de catalisar a eliminação Kemp, um modelo de reação que consiste na transferência de próton de um carbono que normalmente é restringida, pois requer enormes barreiras de energia para se processar naturalmente (RÖTHLISBERGER et al., 2008).

Para escolha do mecanismo catalítico, os autores utilizaram cálculos de mecânica quântica do estado de transição para idealizar um possível sítio ativo para a catálise da reação, em seguida, utilizando o algoritmo RosettaMatch, pesquisaram por possíveis esqueletos proteicos que poderiam, em teoria, representar este sítio ativo. Depois de obtidas

as estruturas, todas as proteínas foram analisadas com relação a parâmetros cinéticos, em seguida, o gene da enzima desenhada *in silico* com melhor atividade catalítica (denominada KE07) foi submetido à metagênese e ao DNA *shuffling*, e as formas variantes, em seguida, clonadas em plasmídeos de expressão e transformadas em *E. coli* (RÖTHLISBERGER et al., 2008). As colônias recombinantes foram triadas quanto sua capacidade de hidrolisar o 5-nitrobenzoisoxazol, seguida pela formação dos seus produtos.

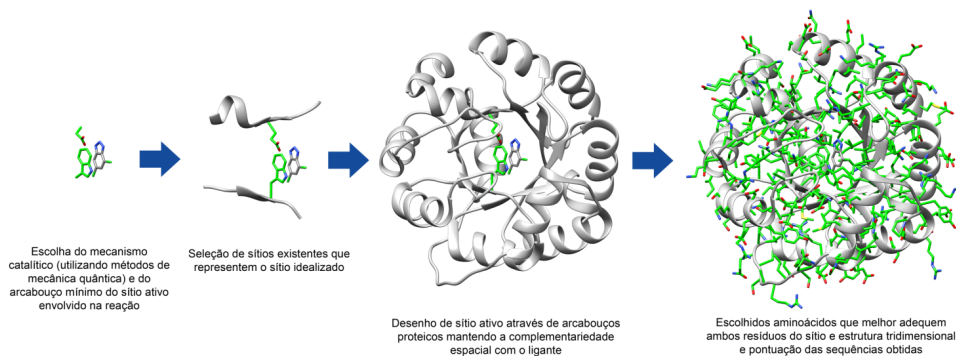


Figura 25. Esquema representativo do desenho computacional de proteínas com novos sítios catalíticos desenvolvido por Richter et al. (2011).

PREDIÇÃO DE ESTRUTURA DE PROTEÍNAS OU SEGMENTOS SEM HOMÓLOGOS

Em alguns casos, nos quais a sequência da proteína-alvo (ou regiões desta) não apresenta qualquer identidade com a sequência de estruturas previamente resolvidas por métodos experimentais e disponíveis em base de dados, os métodos *ab initio* podem ser a solução. Atualmente, esta é uma das mais importantes aplicações dos métodos *ab initio* (BAKER et al., 2006; BAR-ON et al., 2011; KEMEGE et al., 2011).

Em um estudo, realizamos a modelagem *ab initio* para prever a estrutura tridimensional do Fator de Infecção Viral (Vif do inglês, *virion infectivity factor*), uma proteína acessória do Vírus da Imunodeficiência Humana (DA COSTA et al., 2014). Para isto, utilizamos o protocolo *AbinitioRelax* disponível no programa Rosetta (BRADLEY; MISURA; BAKER, 2005) e definimos a relação espacial existente entre as cadeias laterais dos resíduos de cisteína e histidina que compõem um motivo dedo-dezinco essencial para o mecanismo de ação molecular desta proteína. Ao todo, foram gerados 50 mil modelos que, destes 500 foram selecionados e, em seguida, submetidos a uma análise alternativa de centroide e à inspeção visual com relação à formação do motivo e a disposição dos resíduos presentes nos domínios de ligação da proteína.

No referido trabalho, conseguimos prever a localização de domínios de ligação e,

além disso, realizamos docagem molecular com outras proteínas, simulação de dinâmica molecular para análise de mudanças conformacionais e cálculos de energia livre de ligação da Vif complexada com as proteínas celulares EloB e EloC. O modelo teórico foi comparado com um fragmento do BC-box do Vif obtido por cristalografia e obtivemos um RMSD satisfatório de 0,904 Å para esta região (STANLEY et al., 2008).

Com relação aos métodos que utilizam a abordagem *threading* para predição da estrutura, estes são recomendados para proteínas ou regiões da cadeia que não apresentam identidade mínima de 30% com sequências utilizadas como molde que permitam a aplicação de métodos de modelagem por homologia. Com relação a estes, recentemente, publicamos um estudo em que analisamos as alterações conformacionais da 3-Hydroxy-3-Methylglutaryl Coenzyme Reductase, uma enzima envolvida na biossíntese do colesterol com o propósito de identificar a relevância da região do domínio Flap para a primeira etapa de reação desempenhada pela enzima. Neste estudo, esta região ausente na cristalografia foi modelada por *threading* utilizando o programa I-TASSER (COSTA et al., 2019) we constructed a model of human HMGR (hHMGR. Em outro estudo, a abordagem *threading* foi utilizada para predizer a estrutura completa da Alba3 de *Leishmania infatum*, uma enzima que apresenta identidade menor que 30% com seus homólogos próximos, mas exibe um domínio Alba envolvido na ligação ao DNA/RNA comum com outras enzimas (DA COSTA et al., 2017).

ATUAIS LIMITAÇÕES E PERSPECTIVAS DOS MÉTODOS *THREADING* E *AB INITIO*

Kauê Santana da Costa

Anderson Henrique Lima e Lima

Alberto Monteiro dos Santos

Apesar de terem evoluído consideravelmente nos últimos anos mostrando resultados satisfatórios nas estruturas obtidas, os métodos de modelagem da estrutura tridimensional de proteínas também apresentam suas limitações, em especial no que diz respeito aos métodos que utilizam a abordagem *threading* e *ab initio*. Este capítulo irá apresentar as principais limitações e discutir perspectivas de desenvolvimento, e algumas novas aplicações que estão sendo abordadas para estes métodos.

ATUAIS LIMITAÇÕES E DESAFIOS

Conciliar um baixo custo computacional com uma simulação que obtenha modelos mais realísticos com baixa resolução e baixa energia mostra-se um grande desafio para a área de predição de estruturas de proteínas. Com relação aos algoritmos *ab initio*, apesar de eles terem evoluído ao longo destes anos – progresso notado durante os eventos do CASP – estes apresentam algumas limitações. Uma delas refere-se ao demasiado tempo computacional em relação às abordagens baseadas em

molde, uma vez que a grande parte da sua aplicação, tem sido na obtenção de estruturas que não apresentam qualquer identidade ou não apresentam valores satisfatórios desta com outras sequências de proteínas (BAKER et al., 2006; BAR-ON et al., 2011; KEMEGE et al., 2011).

No entanto, os métodos *ab initio*, que não eram capazes de prever de maneira acurada a topologia de estruturas que continham mais de 100 resíduos (SIMONS; STRAUSS; BAKER, 2001), na última década, nos mostram, com certa precisão, a estrutura de proteínas com 100-120 resíduos de aminoácidos. É importante ressaltar ainda que já foi relatado na literatura, a modelagem bem sucedida de estruturas raras, em que se obtiveram modelos de resolução satisfatória (<2,0 Å de átomos de Ca) (BRADLEY; MISURA; BAKER, 2005). Com relação a isso, o algoritmo aplicado no programa de predição Quark, disponibilizado e mantido pela Universidade de Michigan (EUA) permite sequências com tamanhos maiores de até 200 resíduos de aminoácidos (XU; ZHANG, 2012).

Tomando como base os resultados apresentados nos últimos eventos do CASP, os modelos, algumas vezes, são distantes ou apresentam determinadas deficiências em relação às estruturas nativas, o que requer que inspeções visuais sejam realizadas com o propósito de estabelecer características

importantes no alvo que se deseja obter, pois pelo menos até o momento, as funções de pontuação ou métodos de comparação – como o GDT_TS (do inglês *Global Distance Test Total Score*), largamente utilizados para a seleção de estruturas – não são capazes de reproduzir as decisões humanas na escolha de modelos teóricos próximos à conformação nativa (BEN-DAVID et al., 2009; KRYSHTAFOVYCH; FIDELIS; MOULT, 2014; TAI et al., 2014).

O GDT_TS é um método métrico que determina a acurácia global de predição do modelo em relação à estrutura nativa experimentalmente obtida, pois mede a porcentagem de átomos de Ca espacialmente alinhados entre ambas as estruturas. É considerado mais robusto que o RMSD na comparação de estruturas. Valores de GDT_TS entre 90-100 representam sucesso na predição. Modelos com valores de 89-60 representam proteínas com similaridade estrutural satisfatória. Modelos com valores de 30-20 representam estruturas aleatórias e sem significância obtidas no processo de modelagem. Durante os eventos do CASP, temos percebido progressos na qualidade dos modelos, por exemplo, no CASP 5, somente um dos cinco alvos estudados se mostrou acima de 60, e os outros com pontuação acima de 40. Em contraste, durante o CASP 10, três estruturas alvo com menos de 120 resíduos mostraram GDT maior que 60 (KINCH et al., 2011).

Métodos de predição *ab initio* que utilizam campos de força físicos aliados à simulação de dinâmica molecular, estão longe de se tornar rotineiros na predição da estrutura de proteínas devido ao elevado custo computacional, embora sucessos de predição já tenham ocorrido com esta abordagem, utilizando super-computadores (NGUYEN et al., 2014). Considerando que a proteína de interesse apresente uma longa cadeia polipeptídica e que o dobramento desta na sua conformação nativa, ao longo da dinâmica molecular, pode requerer escalas superiores ao de nanosegundos, uma capacidade e tempo de processamento maiores serão necessários para a simulação. Recentemente, Voelz et al., (2010) simularam o dobramento da região N-terminal da proteína ribossomal L9 (NTL9) por abordagem *ab initio*, utilizando o sistema de computadores Folding@Home e uma versão modificada do pacote computacional GROMACS escrita para GPUs utilizando um tempo total de simulação 1.52 microssegundos (ms) (VOELZ et al., 2010).

Nesta simulação, utilizaram o campo de força AMBER ff96 e o modelo de solvatação GBSA e iniciaram os cálculos de DM em diferentes temperaturas e com fragmento da proteína em diferentes estados conformacionais (incluindo a forma distendida). No referido estudo, apesar de simularem um fragmento de somente 39 resíduos para a proteína que apresenta um tempo de dobramento de ~1,5 ms obtiveram um RMSD-Ca de 3,1 Å com relação à forma nativa obtida por difração de raios-X. Deste modo, percebemos que os métodos *ab initio* que utilizam simulação de dinâmica molecular aliado ao campos de força com potenciais físicos e modelos de solvatação, podem ser úteis na compreensão do enovelamento de proteínas quando comparados aos métodos de predição baseados em molde, como o *threading* e modelagem por homologia, visto que estes últimos, utilizam

somente estruturas de referência como molde, de onde se extraem um arcabouço estrutural para construção do modelo.

Com relação ao sucesso do desenho de enzimas, embora tenhamos notado avanços interessantes nesta área (KHERSONSKY et al., 2011; RÖTHLISBERGER et al., 2008), as enzimas engenheiradas ainda apresentam pouca eficiência na atividade catalítica quando comparadas àquelas do estado nativo. É por isso que o desenho de enzimas tem sido utilizado em conjunto com a evolução dirigida, como forma de corrigir possíveis erros na estrutura catalítica não previstos pelos métodos computacionais. Neste sentido, avanços no desenvolvimento dos campos de força, na seleção da conformação das alças, ou esqueletos proteicos adequados ao sítio catalítico, ainda serão necessários para aperfeiçoar os métodos *in silico* aplicados ao desenho destas estruturas (BAKER, 2010).

PERSPECTIVAS E DESENVOLVIMENTOS FUTUROS

Apesar das limitações apresentadas no desenho de proteínas, a biologia sintética é, atualmente, um dos mais empolgantes e recentes campos de desenvolvimento científico e encontra-se em crescente expansão. Com relação a isto, os métodos *ab initio* são, sem dúvida, os mais versáteis em aplicações biotecnológicas e a gama de produtos (derivados de proteínas) desenhados por estes métodos não param de crescer. Atualmente, um grande número de trabalhos com este propósito tem sido publicados na literatura, incluindo propósitos e abordagens diferentes, tais como: obtenção novas topologias proteicas para o desenho de vacinas (CORREIA et al., 2014), enzimas desenhadas para exercer novas catálises (ALTHOFF et al., 2012), bem como, nanomateriais com subunidades proteicas capazes de automontagem que são interessantes para fins biotecnológicos (KING et al., 2014).

Com o aumento da qualidade de predição dos métodos *ab initio*, é possível que grupos de pesquisa comecem a adotar estes métodos na sua rotina, podendo se tornar ferramentas não somente auxiliares de predição, mas também concorrentes aos métodos baseados em molde, como a modelagem por homologia que hoje é largamente utilizada no estudo de mutantes e no planejamento *in silico* de fármacos. Com relação ao avanço na qualidade de predição, por exemplo, o I-TASSER tem mostrado resultados convincentes de predição que o colocam em primeiro lugar nos eventos de competição do CASP (do inglês, *Critical Assessment of Techniques for Protein Structure Prediction*), tais como CASP7, CASP8, CASP9, CASP10, CASP11, CASP12, e CASP13 (KRYSHTAFOVYCH et al., 2019). Além disso, recentemente foi reportado que os modelos obtidos pelo algoritmo I-TASSER superam em qualidade as estruturas obtidas por cristalografia por difração de raios-X e, podem desta forma, auxiliar em métodos de triagem virtual baseados em estrutura (SBVS do inglês, *Structure-based Virtual Screening*) (DU et al., 2015; RODRIGUES et al., 2012).

Similarmente, devido ao grande avanço das metodologias de modelagem, é possível que futuramente os métodos *in silico* não baseados em molde, no qual se inclui exclusivamente a predição *ab initio* sejam, gradativamente, mais aplicados com propósitos que requerem detalhes estruturais mais precisos como o planejamento *in silico* de fármacos e no estudo de mecanismos catalítico-enzimáticos.

GLOSSÁRIO DE TERMOS TÉCNICOS E SIGLAS

A

Ab initio: Na Biologia Estrutural, refere-se a uma abordagem de modelagem tridimensional que não depende de estruturas de referência. Neste método, são utilizadas a análise do espaço conformacional unida a uma função de pontuação de energia, a fim de encontrar as estruturas com o mínimo de energia potencial global.

Aminoácido: São os blocos construtores das proteínas. Cada aminoácido é composto por um átomo de carbono quiral denominado de carbono alfa (Ca) que forma ligações com o grupo amina, o carboxila, uma cadeia lateral que variam de acordo com o aminoácido e o hidrogênio.

Alça: Regiões conformacionalmente variáveis da estrutura tridimensional de proteínas que conectam às estruturas secundárias.

Alinhamento: Método computacional de organização de duas ou mais sequências ou estruturas cujo propósito é evidenciar similaridades ou dissimilaridades entre elas.

Algoritmo: Conjunto ordenado e lógico de instruções para execução de determinada tarefa computacional.

B

BLAST: Do inglês *Basic Local Alignment Search Tool*, trata-se de um algoritmo de comparação de sequências utilizado para busca em bases de sequências a fim de gerar alinhamentos locais, a partir de uma sequência de interesse (no inglês, *query sequence*). É a ferramenta de pesquisa padrão do NCBI GenBank.

BLOSUM: Do inglês *Blocks Substitution Matrix*. É um conjunto de matrizes de substituição aplicada às proteínas, as quais as pontuações para cada posição são derivadas de frequências de substituições em blocos de alinhamentos locais em proteínas relacionadas. Cada matriz é adaptada para uma dada distância evolutiva. Exemplo: As matrizes BLOSUM62 derivam de alinhamentos que compartilham 62% de identidade.

C

Campo de força: Conjunto de parâmetros e funções matemáticas que são utilizados para descrever a energia potencial total de um sistema.

CASP: Do inglês, *Critical Assessment of Techniques for Protein Structure Prediction*, são eventos científicos que ocorrem bianualmente em que diferentes métodos e algoritmos são testados levando em consideração a capacidade de prever a estrutura tridimensional de proteínas na sua conformação nativa.

Carbono alfa (Ca): Carbono quiral presente em todos os aminoácidos. Forma quatro grupos de ligações simples: carboxila, amina, hidrogênio e o radical variável que varia de acordo com o aminoácido.

Compilador: Programa utilizado para transformar código fonte escrito em determinada linguagem de programação em um programa executável.

Conformação: Arranjo tridimensional de átomos e de ligações em uma molécula que descrevem sua geometria.

Conformação nativa: Corresponde à conformação final de uma proteína (ou ácido nucleico) na qual esta desempenha sua função biológica. É a conformação funcional.

CPU: Do inglês, *Central Process Unit* (Unidade Central de Processamento). É o componente principal do computador, também é conhecido como processador. A CPU é responsável por calcular e realizar tarefas determinadas pelo usuário.

C-terminal: Porção final da cadeia polipeptídica da proteína que contém um grupo carboxila.

D

Docagem: Do inglês *docking*. É o método computacional que tenta prever a melhor afinidade e complementariedade de duas moléculas quando ligadas a um complexo estável.

Domínio: Em Biologia Estrutural, se trata de regiões ou porções da proteína que exibe sequência conservada evolutivamente. Normalmente, são regiões funcionais da estrutura.

Diagrama de Ribbon: Representação esquemática tridimensional da estrutura de proteínas, em que as alfa-hélices são representadas por fitas enroladas e as beta-folhas, por setas.

DOPE: Do inglês *Discrete Optimized Protein Energy*. Trata-se de uma unidade de energia potencial utilizada pelo programa de modelagem Modeller.

E

Enovelamento: Processo físico no qual uma proteína por meio de interações com o meio ou ação de outras proteínas ou cofatores atingem sua conformação nativa.

Enzima: Proteínas que apresentam função catalítica, isto é, atuam como catalisadoras de reações químicas.

F

FASTA: Formato de representação de sequências biológicas largamente aplicado em bases de dados. É aplicado para a representação de sequências de nucleotídeos e aminoácidos.

Forças de Solvatação: Forças de atração e repulsão causado pelas moléculas dispersas-se em uma solução contra os íons ou outras moléculas que constituem o soluto. Nas proteínas, estas forças normalmente são exercidas pelas moléculas de água.

Função de pontuação de energia: Função pela qual se avalia o estado conformacional de uma proteína por meio de cálculos de energia potencial. As funções variam para cada algoritmo.

G

Gráfico de Ramachandran: Gráfico de dispersão que representa a estrutura de proteínas por meio dos ângulos de torção ϕ e ψ dos aminoácidos.

GUI: Do inglês *Graphical User Interface* (Interface Gráfica ao Usuário). É a interface gráfica que permite a interação do usuário com o sistema operacional.

Genoma: Conjunto haplóide completo de todos os genes e demais sequências presentes em um organismo.

Gap: Na tradução lacuna. Em Bioinformática, se refere às regiões de não correspondência entre duas sequências em alinhamentos e são representadas por traços.

GPU: Do inglês *Graphics Processing Unit*. São unidades de processamento especializado em processar gráficos.

H

Homologia: Relação evolutiva entre duas estruturas em que estas compartilham um ancestral em comum (mesma origem filogenética). Na Biologia Estrutural, duas proteínas são consideradas homólogas quando descendem de um mesmo ancestral.

I

Input: Expressão em inglês que na linguagem da área computacional se refere a todo dado utilizado como entrada para execução de um processo.

Identidade: Porcentagem de caracteres similares entre duas sequências quando alinhadas (excluindo-se os *gaps*).

Iteração: Em computação, se refere à uma série de passos em um algoritmo por meio do qual o processamento de dados é executado repetitivamente até o resultado exceder um limite particular.

L

Ligante: Molécula que se liga a um determinado bioreceptor, podendo agir como um co-fator, substrato ou inibidor.

Loop: ver alça.

M

Matches: Na tradução do inglês *correspondência*. Caracteres idênticos em duas sequências alinhadas que ocupam a mesma coluna.

Metilação: Modificação pós-traducional que correm em proteínas. Corresponde à adição de grupos metil à estrutura.

Mismatches: Na tradução do inglês *não correspondência*. Caracteres diferentes em duas

seqüências alinhadas que ocupam a mesma coluna.

Molde: Estrutura utilizada como referência em métodos de modelagem por homologia. Termo deriva do inglês *template*.

Modelagem por homologia: Método de predição da estrutura de proteínas que utiliza estruturas de referência que apresentam similaridade na seqüência devido à ancestralidade comum (homologia).

Motivos: Regiões curtas e conservadas estruturalmente de uma seqüência de proteínas. Geralmente são partes altamente conservadas dos domínios.

N

NGS: Do inglês *Next Generation Sequencing* (Sequenciamento de Nova Geração). Refere-se a todas as metodologias de sequenciamento nascidas após 2005 e incluem, entre outros, os métodos de sequenciamento por ligação, sequenciamento por detecção de prótons, e o pirosequenciamento.

Nucleotídeo: Correspondem aos blocos construtores dos ácidos nucleicos. São constituídos por um fosfato, açúcar e uma base nitrogenada.

N-terminal: Porção inicial da seqüência polipeptídica da proteína que contém o grupo amina. Normalmente, inicia com um resíduo de metionina.

O

Open source: Expressão em inglês que se refere aos programas nos quais a código fonte é aberto, isto é, as informações são livres.

Output: Expressão em inglês que se refere aos dados liberados da execução de um processo computacional.

P

Protein Data Bank: Base de dados que contém a estrutura tridimensional de macromoléculas biológicas elucidadas por métodos experimentais, tais como, cristalografia de raios-X, espectroscopia de Ressonância Magnética Nuclear (RMN) e crio-microscopia eletrônica.

PDB: Formato de arquivo utilizado para a representação da estrutura tridimensional de moléculas. Atualmente é o formato amplamente difundido e interpretado por diferentes programas de análise de estrutura. É representado pela extensão .PDB.

PCR: Do inglês *Polymerase Chain Reaction*, trata-se de uma técnica da biologia molecular que realiza a amplificação de segmentos específicos de DNA em uma mistura complexa, na qual estão presentes também curtos iniciadores (denominados primers) oligonucleotídeos para o segmento de interesse, além de reagentes para síntese de DNA.

R

Refinamento: Corresponde à otimização na estrutura tridimensional de uma macromolécula de modo a corrigir distorções ou erros esterequímicos.

Root Mean Square Deviation (RMSD): Cálculo métrico que exhibe a distância média entre duas proteínas estruturalmente alinhadas. Normalmente, utilizam-se como referência os átomos de carbono α .

Rotâmero: Conformação mais comum da cadeia lateral de resíduos de aminoácidos. São gerados por modificações nos ângulos de rotação.

Ribbon: Ver *Diagrama de Ribbon*.

S

Similaridade: Porcentual de resíduos de aminoácidos similares entre duas proteínas obtidos após o alinhamento de duas sequências. Resíduos similares são aqueles que compartilham características físico-químicas semelhantes.

Simulated Annealing (SA): Na predição estrutural de proteínas, trata-se de um algoritmo que aplica o Metropolis MC para a minimização da estrutura. Este gera perturbações aleatórias nos rotâmeros por meio de ciclos de aquecimento e resfriamento seguindo a distribuição canônica de energia de Boltzman para determinada temperatura com objetivo de se encontrar a menor energia global da estrutura.

Script: Conjunto de códigos escritos em uma linguagem de programação específica que são interpretados por meio de um programa.

SDF: Do inglês, *Structure Data File*. Formato de arquivo de estrutura desenvolvido pela empresa MDL. É representado pelas extensões .sd ou .sdf.

Simulação de Dinâmica Molecular (DM): Método computacional que utiliza a mecânica molecular para descrever o movimento atômico ao longo do tempo. Neste método, átomos e moléculas são descritos como partículas unidas por forças harmônicas ou elásticas, e o movimento atômico é resolvido pelas equações newtonianas.

T

Threading: Método de predição da estrutura de proteínas que reconhece o tipo de dobramento de proteínas homólogas distantes ou sem homologia, mas que conservam estruturas similares devido à convergência evolutiva. Assim como a modelagem por homologia, o *threading* é um método baseado em molde.

TM-score: métrica aplicada pelo programa TM-align utilizado pelo I-TASSER para medir a similaridade estrutural entre duas proteínas.

Tradução: Em Biologia Molecular, se refere ao processo no qual a molécula de RNA mensageiro é interpretada pelo ribossomo para a produção da cadeia polipeptídica da proteína.

V

Virtual Screening: Do inglês, triagem virtual, trata-se de uma abordagem computacional que realiza a triagem/filtragem de compostos com aplicações no desenvolvimento de novos fármacos. Há a abordagem baseada em estrutura e a baseada no ligante.

Z

Z-score: método estatístico que indica a separação de valores em relação à sua contraparte. É representado pela fórmula: $\text{valor} - \text{média} / \text{Desvio padrão}$. Na Biologia Estrutural é utilizado para acessar a significância energética de determinada estrutura com uma amostra de proteínas oriundas de base de dados ou preditas por métodos computacionais.

REFERÊNCIAS

ALTHOFF, E. A. *et al.* Robust design and optimization of retroaldol enzymes. **Protein Science**, v. 21, n. 5, p. 717-26, 2012.

ALTSCHUL, S. F. *et al.* **Gapped BLAST and PSI-BLAST: A new generation of protein database search programs** **Nucleic Acids Research**, 1997.

ANFINSEN, C. B. Principles that govern the folding of protein chains. **Science**, v. 181, n. 4096, p. 223-30, jul., 1973.

BAKER, D. An exciting but challenging road ahead for computational enzyme design. **Protein Science**, v. 19, n. 10, p. 1817-19, out., 2010.

BAKER, M. L. *et al.* Ab initio modeling of the herpesvirus VP26 core domain assessed by cryoEM density. **PLoS Computational Biology**, v. 2, n. 10, p. 1313-24, 2006.

BAR-ON, D. *et al.* Dynamic Conformational Changes in MUNC18 Prevent Syntaxin Binding. **PLoS Computational Biology**, v. 7, n. 3, p. e1001097, 3 mar., 2011.

BAXEVANIS, A. D.; OUELLETTE, B. F. F. **BIOINFORMATICS A Practical Guide to the Analysis of Genes and Proteins SECOND EDITION**. [s.l.: s.n.].

BEN-DAVID, M. *et al.* Assessment of CASP8 structure predictions for template free targets. **Proteins: Structure, Function and Bioinformatics**, v. 77, n. SUPPL. 9, p. 50-65, 2009.

BEN-ZVI, A. P.; GOLOUBINOFF, P. Review: Mechanisms of Disaggregation and Refolding of Stable Protein Aggregates by Molecular Chaperones. **Journal of Structural Biology**, v. 135, n. 2, p. 84-93, ago., 2001.

BENKERT, P.; SCHWEDE, T.; TOSATTO, S. C. QMEANclust: estimation of protein model quality by combining a composite scoring function with structural density information. **BMC structural biology**, v. 9, p. 35, jan., 2009.

BENSON, D. A. *et al.* GenBank. **Nucleic Acids Research**, v. 45, n. D1, p. D37-D42, jan. 2017.

BERMAN, H. M. *et al.* The Protein Data Bank. In: **Structural Bioinformatics**. [s.l.: s.n.]. v. 28p. 181-198.

BIASINI, M. *et al.* SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. **Nucleic Acids Research**, v. 42, n. W1, p. W252–W258, jul., 2014.

BONNEAU, R. *et al.* Rosetta in CASP4: Progress in ab initio protein structure prediction. **Proteins: Structure, Function and Genetics**, v. 45, n. SUPPL. 5, p. 119–126, 2001.

BONNEAU, R. *et al.* De Novo Prediction of Three-dimensional Structures for Major Protein Families. **Journal of Molecular Biology**, v. 322, n. 1, p. 65-78, set., 2002.

BOWIE, J. U.; LÜTHY, R.; EISENBERG, D. A method to identify protein sequences that fold into a known three-dimensional structure. **Science (New York, N.Y.)**, v. 253, n. 5016, p. 164-170, 1991.

- BRADLEY, P.; MISURA, K. M. S.; BAKER, D. Toward high-resolution de novo structure prediction for small proteins. **Science (New York, N.Y.)**, v. 309, n. 5742, p. 1868-71, 2005.
- BROOKS, B. R. *et al.* CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. **Journal of Computational Chemistry**, v. 4, n. 2, p. 187-217, 1983.
- BRYANT, S. H.; LAWRENCE, C. E. An empirical energy function for threading protein sequence through the folding motif. **Proteins**, v. 16, n. 1, p. 92-112, 1993.
- CASE, D. A. *et al.* The Amber biomolecular simulation programs. **Journal of Computational Chemistry**, v. 26, n. 16, p. 1668-88, dez., 2005.
- CAVASOTTO, C. N.; PHATAK, S. S. Homology modeling in drug discovery: current trends and applications. **Drug Discovery Today**, v. 14, n. 13-14, p. 676-83, jul., 2009.
- CHENG, J.; BALDI, P. A machine learning information retrieval approach to protein fold recognition. **Bioinformatics**, v. 22, n. 12, p. 1456-63, 2006.
- CHENNA, R. Multiple sequence alignment with the Clustal series of programs. **Nucleic Acids Research**, v. 31, n. 13, p. 3497-3500, jul., 2003.
- CORREIA, B. E. *et al.* Proof of principle for epitope-focused vaccine design. **Nature**, v. 507, n. 7491, p. 201-6, 2014.
- COSTA, C. H. S. *et al.* Computational study of conformational changes in human 3-hydroxy-3-methylglutaryl coenzyme reductase induced by substrate binding. **Journal of Biomolecular Structure and Dynamics**, v. 37, n. 16, p. 4374-83, 23 nov., 2019.
- DA COSTA, K. S. *et al.* Structural analysis of viral infectivity factor of HIV type 1 and its interaction with A3G, EloC and EloB. **PLoS ONE**, v. 9, n. 2, p. e89116, jan., 2014.
- DA COSTA, K. S. *et al.* Structural and evolutionary analyses of Leishmania Alba proteins. **Molecular and Biochemical Parasitology**, v. 217, p. 23-31, 2017.
- DA COSTA, K. S. *et al.* Exploring the Potentiality of Natural Products from Essential Oils as Inhibitors of Odorant-Binding Proteins: A Structure- and Ligand-Based Virtual Screening Approach To Find Novel Mosquito Repellents. **ACS Omega**, v. 4, n. 27, p. 22475-86, 31 dez., 2019.
- DI TOMMASO, P. *et al.* T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. **Nucleic acids research**, v. 39, n. Web Server issue, p. W13-7, 1 jul., 2011.
- DING, F. *et al.* Ab Initio Folding of Proteins with All-Atom Discrete Molecular Dynamics. **Structure**, v. 16, n. 7, p. 1010-18, 2008.
- DU, H. *et al.* Protein structure prediction provides comparable performance to crystallographic structures in docking-based virtual screening. **Methods**, v. 71, p. 77-84, 2015.
- EDGAR, R. C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. **Nucleic Acids Research**, v. 32, n. 5, p. 1792-7, 2004.

- FAN, H. Refinement of homology-based protein structures by molecular dynamics simulation techniques. **Protein Science**, v. 13, n. 1, p. 211-20, 1 jan., 2004.
- FISER, A. Template-based protein structure modeling. **Methods in molecular biology (Clifton, N.J.)**, v. 673, p. 73-94, 2010.
- FISER, A.; ŠALI, A. MODELLER: Generation and Refinement of Homology-Based Protein Structure Models. **Methods in Enzymology**, v. 374, p. 461-91, jan., 2003.
- FLOUDAS, C. A. *et al.* Advances in protein structure prediction and de novo protein design: A review. **Chemical Engineering Science**, v. 61, n. 3, p. 966-88, fev. 2006.
- GIGER, L. *et al.* Evolution of a designed retro-aldolase leads to complete active site remodeling. **Nature chemical biology**, v. 9, n. 8, p. 494-8, 2013.
- GOPAL, S. M.; KLENIN, K.; WENZEL, W. Template-free protein structure prediction and quality assessment with an all-atom free-energy model. **Proteins**, v. 77, n. 2, p. 330-41, 1 nov., 2009.
- GUO, J. -T. *et al.* PROSPECT-PSPP: an automatic computational pipeline for protein structure prediction. **Nucleic Acids Research**, v. 32, n. Web Server, p. W522–W525, 1 jul., 2004.
- HANSMANN, U. H. E.; OKAMOTO, Y. New Monte Carlo algorithms for protein folding. **Current Opinion in Structural Biology**, v. 9, n. 2, p. 177-83, abr., 1999.
- HAO, M.-H.; SCHERAGAT, H. A. Designing potential energy functions for protein folding. **Current Opinion in Structural Biology**, v. 9, n. 2, p. 184-8, abr., 1999.
- HUANG, Y. J. *et al.* Assessment of template-based protein structure predictions in CASP10. **Proteins: Structure, Function and Bioinformatics**, v. 82, n. SUPPL.2, p. 43-56, 2014.
- IHM, Y. *et al.* Structural Model of the Rev Regulatory Protein from Equine Infectious Anemia Virus. **PLoS ONE**, v. 4, n. 1, p. e4178, 12 jan., 2009.
- JÄCKEL, C.; KAST, P.; HILVERT, D. Protein design by directed evolution. **Annual review of biophysics**, v. 37, p. 153-73, 2008.
- JIANG, L. *et al.* De novo computational design of retro-aldol enzymes. **Science (New York, N.Y.)**, v. 319, n. 5868, p. 1387-91, 2008.
- JONES, D. T. *et al.* Successful recognition of protein folds using threading methods biased by sequence similarity and predicted secondary structure. **Proteins: Structure, Function and Genetics**, v. 37, n. SUPPL. 3, p. 104-11, 1999.
- JONES, D. T. Protein structure prediction in the postgenomic era. **Current opinion in structural biology**, v. 10, n. 3, p. 371-9, 2000.
- JONES, D. T.; MILLER, R. T.; THORNTON, J. M. Successful protein fold recognition by optimal sequence threading validated by rigorous blind testing. **Proteins: Structure, Function and Genetics**, v. 23, n. 3, p. 387-97, 1995.

JONES, D. T.; TAYLOR, W. R.; THORNTON, J. M. A new approach to protein fold recognition. **Nature**, v. 358, n. 6381, p. 86-9, 1992.

JORGENSEN, W. L.; MAXWELL, D. S.; TIRADO-RIVES, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. **Journal of the American Chemical Society**, v. 118, n. 45, p. 11225-36, jan., 1996.

JOSHI, R. R.; JYOTHI, S. Ab-initio prediction and reliability of protein structural genomics by PROPAINOR algorithm. **Computational Biology and Chemistry**, v. 27, n. 3, p. 241-52, 2003.

KÄLLBERG, M. *et al.* **Template-based protein structure modeling using the RaptorX web server** **Nature Protocols**, 2012.

KAUFMANN, K. W. *et al.* Practically useful: what the Rosetta protein modeling suite can do for you. **Biochemistry**, v. 49, n. 14, p. 2987-98, 13 abr., 2010.

KELLEY, L. A. *et al.* The Phyre2 web portal for protein modeling, prediction and analysis. **Nature protocols**, v. 10, n. 6, p. 845-58, 2015.

KEMEGE, K. E. *et al.* Ab initio structural modeling of and experimental validation for Chlamydia trachomatis protein CT296 reveal structural similarity to Fe(II) 2-Oxoglutarate-dependent enzymes. **Journal of Bacteriology**, v. 193, n. 23, p. 6517-28, 2011.

KHERSONSKY, O. *et al.* Optimization of the in-silico-designed Kemp eliminase KE70 by computational design and directed evolution. **Journal of Molecular Biology**, v. 407, n. 3, p. 391-412, 2011.

KHOURY, G. A. *et al.* Protein folding and de novo protein design for biotechnological applications. **Trends in Biotechnology**, v. 32, n. 2, p. 99-109, 2014.

KIDD, B. A.; BAKER, D.; THOMAS, W. E. Computation of Conformational Coupling in Allosteric Proteins. **PLoS Computational Biology**, v. 5, n. 8, p. e1000484, 28 ago., 2009.

KIEFER, F. *et al.* The SWISS-MODEL Repository and associated resources. **Nucleic Acids Research**, v. 37, n. SUPPL. 1, 2009.

KIHARA, D. *et al.* TOUCHSTONE: an ab initio protein structure prediction method that uses threading-based tertiary restraints. **Proceedings of the National Academy of Sciences of the United States of America**, v. 98, n. 18, p. 10125-30, 2001.

KIHARA, D. *et al.* Ab initio protein structure prediction on a genomic scale: application to the Mycoplasma genitalium genome. **Proceedings of the National Academy of Sciences of the United States of America**, v. 99, n. 9, p. 5993-8, 2002.

KIM, D. E.; CHIVIAN, D.; BAKER, D. Protein structure prediction and analysis using the Robetta server. **Nucleic Acids Research**, v. 32, n. Web Server, p. W526–W531, 1 jul. 2004.

KINCH, L. *et al.* CASP9 assessment of free modeling target predictions. **Proteins: Structure, Function and Bioinformatics**, v. 79, n. SUPPL. 10, p. 59-73, jan. 2011.

KING, N. P. *et al.* Accurate design of co-assembling multi-component protein nanomaterials. **Nature**, v. 510, n. 7503, p. 103-8, 2014.

KIRKPATRICK, S.; GELATT, C. D.; VECCHI, M. P. Optimization by simulated annealing. **Science (New York, N.Y.)**, v. 220, n. 4598, p. 671-80, 1983.

KLEPEIS, J. L.; FLOUDAS, C. A. ASTRO-FOLD: a combinatorial and global optimization framework for Ab initio prediction of three-dimensional structures of proteins from the amino acid sequence. **Biophysical journal**, v. 85, n. 4, p. 2119-46, 2003.

KORETKE, K. K. *et al.* Fold recognition using sequence and secondary structure information. **Proteins**, v. Suppl 3, p. 141-8, 1999.

KRYSHTAFOVYCH, A. *et al.* **Protein structure prediction center in CASP8** **Proteins: Structure, Function and Bioinformatics**, 2009.

KRYSHTAFOVYCH, A. *et al.* Critical assessment of methods of protein structure prediction (CASP)—Round XIII. **Proteins: Structure, Function, and Bioinformatics**, v. 87, n. 12, p. 1011-20, 23 dez., 2019.

KRYSHTAFOVYCH, A.; FIDELIS, K.; MOULT, J. CASP10 results compared to those of previous CASP experiments. **Proteins: Structure, Function and Bioinformatics**, v. 82, n. SUPPL.2, p. 164-74, fev., 2014.

KUHLMAN, B. *et al.* Design of a novel globular protein fold with atomic-level accuracy. **Science (New York, N.Y.)**, v. 302, n. 5649, p. 1364-68, 2003.

KUHLMAN, B.; BAKER, D. Native protein sequences are close to optimal for their structures. **Proceedings of the National Academy of Sciences of the United States of America**, v. 97, n. 19, p. 10383-8, 12 set., 2000.

LAMEIRA, J. *et al.* Predicting the affinity of halogenated reversible covalent inhibitors through relative binding free energy. **Physical Chemistry Chemical Physics**, v. 21, n. 44, p. 24723-30, 2019.

LASKOWSKI, R. A. *et al.* **PROCHECK: a program to check the stereochemical quality of protein structures**. Disponível em: <<http://scripts.iucr.org/cgi-bin/paper?S0021889892009944>>.

LEVINTHAL, C. Are there pathways for protein folding? **Journal de Chimie Physique et de Physico-Chimie Biologique**, v. 65, p. 44-5, 1968.

LI, K. *et al.* Design and synthesis of novel 2-substituted 11-keto-boswellic acid heterocyclic derivatives as anti-prostate cancer agents with Pin1 inhibition ability. **European Journal of Medicinal Chemistry**, v. 126, p. 910-19, jan., 2017.

LIWO, A. *et al.* United-residue force field for off-lattice protein-structure simulations: III. Origin of backbone hydrogen-bonding cooperativity in united-residue potentials. **Journal of Computational Chemistry**, fev., 1998.

LIWO, A.; KHALILI, M.; SCHERAGA, H. A. Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains. **Proceedings of the National Academy of Sciences**, v. 102, n. 7, p. 2362-67, 15 fev., 2005.

LOBLEY, A.; SADOWSKI, M. I.; JONES, D. T. pGenTHREADER and pDomTHREADER: New methods for improved protein fold recognition and superfamily discrimination. **Bioinformatics**, v. 25, n. 14, p. 1761-67, 15 jul., 2009.

LOVELL, S. C. *et al.* Structure validation by C α geometry: phi,psi and C β deviation. **Proteins**, v. 50, n. 3, p. 437-50, 15 fev., 2003.

MAGRANE, M.; CONSORTIUM, U. P. UniProt Knowledgebase: A hub of integrated protein data. **Database**, v. 2011, p. bar009–bar009, 29 mar., 2011.

MCGUFFIN, L. J.; BRYSON, K.; JONES, D. T. The PSIPRED protein structure prediction server. **Bioinformatics (Oxford, England)**, v. 16, n. 4, p. 404-5, abr., 2000.

MELO, F.; FEYTMANS, E. Assessing protein structures with a non-local atomic interaction energy. **Journal of Molecular Biology**, v. 277, n. 5, p. 1141-52, abr., 1998.

METROPOLIS, N. *et al.* Equation of State Calculations by Fast Computing Machines. **The Journal of Chemical Physics**, v. 21, n. 6, p. 1087-92, 1953.

MORAES, G. *et al.* Homology modeling, molecular dynamics and QM/MM study of the regulatory protein PhoP from *Corynebacterium pseudotuberculosis*. **Journal of molecular modeling**, v. 18, n. 3, p. 1219-27, mar., 2012.

NEVES CRUZ, J. *et al.* Measuring the structural impact of mutations on cytochrome P450 21A2, the major steroid 21-hydroxylase related to congenital adrenal hyperplasia. **Journal of Biomolecular Structure and Dynamics**, p. 1-10, 14 abr., 2019.

NGUYEN, H. *et al.* Folding Simulations for Proteins with Diverse Topologies Are Accessible in Days with a Physics-Based Force Field and Implicit Solvent. **Journal of the American Chemical Society**, v. 136, n. 40, p. 13959-62, 8 out., 2014.

NOCUA, P. A. *et al.* Leishmania braziliensis replication protein A subunit 1: molecular modelling, protein expression and analysis of its affinity for both DNA and RNA. **Parasites & Vectors**, v. 7, n. 1, p. 573, 2014.

OKAMOTO, Y. Monte-Carlo Simulated Annealing in Protein Folding. In: **Encyclopedia of Optimization**. Boston, MA: Springer US, 2009. p. 2323-37.

PILLARDY, J. *et al.* Development of Physics-Based Energy Functions that Predict Medium-Resolution Structures for Proteins of the α , β , and α/β Structural Classes. **The Journal of Physical Chemistry B**, v. 105, n. 30, p. 7299-311, ago., 2001.

RAMAN, S. *et al.* Structure prediction for CASP8 with all-atom refinement using Rosetta. **Proteins**, v. 77, p. 89-99, jan. 2009.

RAVAL, A. *et al.* Refinement of protein structure homology models via long, all-atom molecular dynamics simulations. **Proteins: Structure, Function, and Bioinformatics**, p. n/a-n/a, 2012.

RICHTER, F. *et al.* De novo enzyme design using Rosetta3. **PLoS ONE**, v. 6, n. 5, 2011.

RODRIGUES, R. P. *et al.* Virtual Screening Strategies in Drug Design. **Revista Virtual de Química**, v. 4, n. 6, 2012.

ROHL, C. A *et al.* Protein structure prediction using Rosetta. **Methods in enzymology**, v. 383, n. 2003, p. 66-93, jan. 2004.

ROST, B.; SCHNEIDER, R.; SANDER, C. Protein fold recognition by prediction-based threading. **Journal of molecular biology**, v. 270, n. 3, p. 471-80, 1997.

RÖTHLISBERGER, D. et al. Kemp elimination catalysts by computational enzyme design. **Nature**, v. 453, n. 7192, p. 190-5, 2008.

ROY, A.; KUCUKURAL, A.; ZHANG, Y. I-TASSER: a unified platform for automated protein structure and function prediction. **Nature Protocols**, v. 5, n. 4, p. 725-38, abr., 2010.

SALOMON-FERRER, R.; CASE, D. A.; WALKER, R. C. An overview of the Amber biomolecular simulation package. **Wiley Interdisciplinary Reviews: Computational Molecular Science**, v. 3, n. 2, p. 198-210, mar., 2013.

SÁNCHEZ, R.; ŠALI, A. Advances in comparative protein-structure modelling. **Current Opinion in Structural Biology**, v. 7, n. 2, p. 206-14, abr., 1997.

SANTOS FILHO, O. A.; ALENCASTRO, R. B. DE; BICCA DE ALENCASTRO, R. Modelagem de proteínas por homologia. **Química Nova**, v. 26, n. 2, p. 253-9, mar., 2003.

SHI, J.; BLUNDELL, T. L.; MIZUGUCHI, K. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. **Journal of molecular biology**, v. 310, n. 1, p. 243-57, 2001.

SHORTLE, D.; SIMONS, K. T.; BAKER, D. Clustering of low-energy conformations near the native structures of small proteins. **Proceedings of the National Academy of Sciences of the United States of America**, v. 95, n. 19, p. 11158-62, 1998.

SIEW, N. *et al.* MaxSub: an automated measure for the assessment of protein structure prediction quality. **Bioinformatics (Oxford, England)**, v. 16, n. 9, p. 776-85, 2000.

SIMONS, K. T. *et al.* Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. **Journal of Molecular Biology**, v. 268, n. 1, p. 209-25, 25 abr., 1997.

SIMONS, K. T. *et al.* Ab initio protein structure prediction of CASP III targets using ROSETTA. **Proteins**, v. Suppl 3, n. SUPPL. 3, p. 171-6, jan. 1999.

SIMONS, K. T.; STRAUSS, C.; BAKER, D. Prospects for ab initio protein structural genomics. **Journal of molecular biology**, v. 306, n. 5, p. 1191-99, 2001.

SINGH, S.; GUTTULA, P. K.; GURUPRASAD, L. Structure Based Annotation of Helicobacter pylori Strain 26695 Proteome. **PLoS ONE**, v. 9, n. 12, p. e115020, 2014.

SIPPL, M. J. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. **Journal of molecular biology**, v. 213, n. 4, p. 859-83, 1990.

SITTEL, F.; JAIN, A.; STOCK, G. Principal component analysis of molecular dynamics: On the use of Cartesian vs. internal coordinates. **The Journal of Chemical Physics**, v. 141, n. 1, p. 014111, 7 jul., 2014.

SKOLNICK, J. In quest of an empirical potential for protein structure prediction. **Current Opinion in Structural Biology**, v. 16, n. 2, p. 166-71, abr., 2006.

STANLEY, B. J. *et al.* Structural insight into the human immunodeficiency virus Vif SOCS box and its role in human E3 ubiquitin ligase assembly. **Journal of virology**, v. 82, n. 17, p. 8656-63, set., 2008.

TAI, C. H. *et al.* Assessment of template-free modeling in CASP10 and ROLL. **Proteins: Structure, Function and Bioinformatics**, v. 82, n. SUPPL.2, p. 57-83, 2014.

TAYLOR, W. R. Multiple sequence threading: an analysis of alignment quality and stability. **Journal of molecular biology**, v. 269, n. 5, p. 902-43, 1997.

THACHUK, C.; SHMYGELSKA, A.; HOOS, H. H. A replica exchange Monte Carlo algorithm for protein folding in the HP model. **BMC bioinformatics**, v. 8, p. 342, 2007.

TRAPANE, T. L.; LATTMAN, E. E. Seventh Meeting on the Critical Assessment of Techniques for Protein Structure Prediction. **Proteins: Structure, Function, and Bioinformatics**, v. 69, n. S8, p. 1-2, 2007.

TSANG, A. *et al.* Francisella tularensis 2-C-Methyl-D-Erythritol 4-Phosphate Cytidyltransferase: Kinetic Characterization and Phosphoregulation. **PLoS ONE**, v. 6, n. 6, p. e20884, 9 jun., 2011.

VAN DER SPOEL, D. *et al.* GROMACS: Fast, flexible, and free. **Journal of Computational Chemistry**, v. 26, n. 16, p. 1701-18, dez., 2005.

VOELZ, V. A. *et al.* Molecular Simulation of ab Initio Protein Folding for a Millisecond Folder NTL9(1-39). **Journal of the American Chemical Society**, v. 132, n. 5, p. 1526-28, 10 fev., 2010.

WANG, Z. X. A re-estimation for the total numbers of protein folds and superfamilies. **Protein engineering**, v. 11, n. 8, p. 621-6, 1998.

WEBB, B.; SALI, A. Comparative protein structure modeling using MODELLER. **Current Protocols in Bioinformatics**, v. 2016, n. 1, p. 5.6.1-5.6.37, set., 2016.

WIEDERSTEIN, M.; SIPPL, M. J. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. **Nucleic acids research**, v. 35, n. Web Server issue, p. W407-10, jul., 2007.

WU, S.; ZHANG, Y. MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information. **Proteins: Structure, Function, and Bioinformatics**, v. 72, n. 2, p. 547-56, 4 fev., 2008.

XU, D.; ZHANG, Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. **Proteins: Structure, Function and Bioinformatics**, v. 80, n. 7, p. 1715-35, 2012.

XU, J.; ZHANG, Y. How significant is a protein structure similarity with TM-score = 0.5? **Bioinformatics**, v. 26, n. 7, p. 889-95, 1 abr., 2010.

ZANGHELLINI, A. *et al.* New algorithms and an in silico benchmark for computational enzyme design. **Protein science : a publication of the Protein Society**, v. 15, n. 12, p. 2785-94, 2006.

ZHANG, H. A new Hybrid Monte Carlo algorithm for protein potential function test and structure refinement. **Proteins**, v. 34, n. 4, p. 464-71, 1999.

ZHANG, Y. TM-align: a protein structure alignment algorithm based on the TM-score. **Nucleic Acids Research**, v. 33, n. 7, p. 2302-9, 11 abr., 2005.

ZHANG, Y. Template-based modeling and free modeling by I-TASSER in CASP7. **Proteins: Structure, Function and Genetics**, v. 69, n. SUPPL. 8, p. 108-17, jan., 2007.

ZHANG, Y. I-TASSER server for protein 3D structure prediction. **BMC Bioinformatics**, v. 9, n. 1, p. 40, jan. 2008.

ZHANG, Y. I-TASSER: Fully automated protein structure prediction in CASP8. **Proteins: Structure, Function and Bioinformatics**, v. 77, n. SUPPL. 9, p. 100-13, jan., 2009.

ZHANG, Y.; KOLINSKI, A.; SKOLNICK, J. TOUCHSTONE II: a new approach to ab initio protein structure prediction. **Biophysical journal**, v. 85, n. 2, p. 1145-64, 2003.

ZHANG, Y.; SKOLNICK, J. SPICKER: A clustering approach to identify near-native protein folds. **Journal of Computational Chemistry**, v. 25, n. 6, p. 865-71, 2004.

ZHOU, H.; ZHOU, Y. Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. **Proteins: Structure, Function and Genetics**, v. 58, n. 2, p. 321-8, 2005.

ZHOU, M. *et al.* Exploring the binding mode of HIV-1 Vif inhibitors by blind docking, molecular dynamics and MM/GBSA. **RSC Advances**, v. 3, n. 44, p. 22532-43, 2013.

SOBRE OS AUTORES

KAUÊ SANTANA DA COSTA - Laboratório de Simulações e Desenvolvimento de Ferramentas Computacionais. Instituto de Biodiversidade da Universidade Federal do Oeste do Pará.

ALBERTO MONTEIRO DOS SANTOS - Laboratório de Bioinformática do Instituto de Biodiversidade, da Universidade Federal do Oeste do Pará.

ANDERSON HENRIQUE LIMA E LIMA - Laboratório de Planejamento e Desenvolvimento de Fármacos do Instituto de Ciências Exatas e Naturais, da Universidade Federal do Pará.

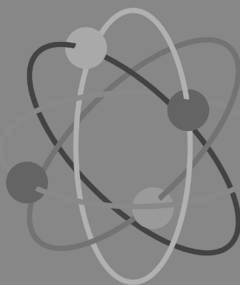
JOSÉ ROGÉRIO DE ARAÚJO SILVA - Laboratório de Planejamento e Desenvolvimento de Fármacos do Instituto de Ciências Exatas e Naturais da Universidade Federal do Pará.

JERÔNIMO LAMEIRA SILVA - Laboratório de Planejamento e Desenvolvimento de Fármacos do Instituto de Ciências Biológicas da Universidade Federal do Pará.

JOÃO MARCOS PEREIRA GALÚCIO - Laboratório de Simulações e Desenvolvimento de Ferramentas Computacionais do Instituto de Biodiversidade da Universidade Federal do Oeste do Pará.

A predição da estrutura de proteínas é, sem dúvida, um assunto chave para todos que atuam nas áreas de modelagem molecular, bioinformática e biologia estrutural, dada a versatilidade de funções desempenhada por estes biopolímeros em processos biológicos, assim como suas diferentes aplicações farmacêuticas, industriais e biotecnológicas. Com este livro, pretende-se não somente atender às necessidades práticas relacionadas à modelagem e análise das estruturas tridimensionais de proteínas, conteúdo ainda carente em língua portuguesa, mas também fornecer os principais conceitos necessários para a compreensão dos métodos, suas aplicações, assim como limitações. O livro traz um conteúdo e linguagem adaptada para alunos de graduação e, devido ao grande número de expressões originárias do inglês presentes na literatura científica, são incluídos os termos mais comumente utilizados na língua de origem, sempre dando preferência à tradução destes para o português.

O livro traz linguagem acessível, inclui referências atualizadas sobre os métodos, se encontra ricamente ilustrado e traz ao final um glossário de termos técnicos que enriquecerão a leitura e compreensão do leitor. Esta obra é uma importante fonte de consulta de alunos de graduação e pós-graduação de cursos relacionados às áreas de ciências Biológicas, Médicas, Exatas e Naturais, assim como, para todos interessados no assunto.



A predição da estrutura de proteínas é, sem dúvida, um assunto chave para todos que atuam nas áreas de modelagem molecular, bioinformática e biologia estrutural, dada a versatilidade de funções desempenhada por estes biopolímeros em processos biológicos, assim como suas diferentes aplicações farmacêuticas, industriais e biotecnológicas. Com este livro, pretende-se não somente atender às necessidades práticas relacionadas à modelagem e análise das estruturas tridimensionais de proteínas, conteúdo ainda carente em língua portuguesa, mas também fornecer os principais conceitos necessários para a compreensão dos métodos, suas aplicações, assim como limitações. O livro traz um conteúdo e linguagem adaptada para alunos de graduação e, devido ao grande número de expressões originárias do inglês presentes na literatura científica, são incluídos os termos mais comumente utilizados na língua de origem, sempre dando preferência à tradução destes para o português.

O livro traz linguagem acessível, inclui referências atualizadas sobre os métodos, se encontra ricamente ilustrado e traz ao final um glossário de termos técnicos que enriquecerão a leitura e compreensão do leitor. Esta obra é uma importante fonte de consulta de alunos de graduação e pós-graduação de cursos relacionados às áreas de ciências Biológicas, Médicas, Exatas e Naturais, assim como, para todos interessados no assunto.

