

Aplicação de Técnicas de Inteligência Artificial e Mineração de Dados no *Design* de Proteínas

Rafael K. Andrades (IC-CNPq)
Márcio Dorn (Doutorando)
Daniel S. Farenzena (Doutorando)
Luis C. Lamb (Orientador)

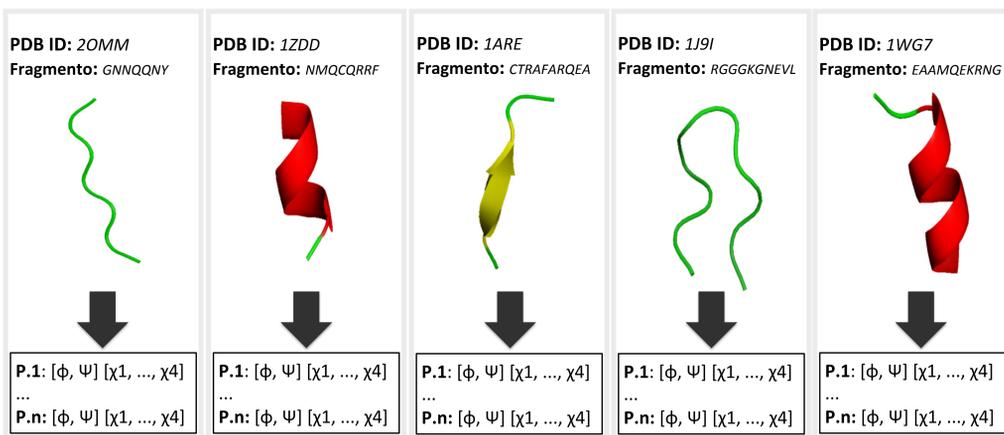
Instituto de Informática
Universidade Federal do Rio Grande do Sul
Porto Alegre, Brasil.

Agradecimentos:

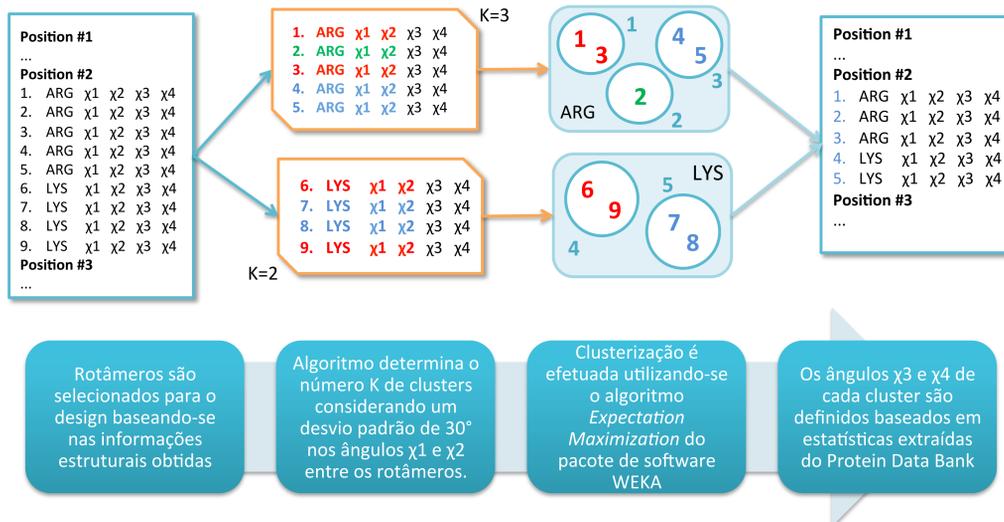


RESUMO: Um dos principais problemas pesquisados na Bioinformática Estrutural está associado ao *Design* de Proteínas [1]. Uma proteína é um polímero formado por vários resíduos de aminoácidos ligados por meio de uma ligação peptídica. A cadeia polipeptídica de uma proteína, em condições fisiológicas ideais, assume uma estrutura tridimensional (3D) única [2]. Esta estrutura associa funções biológicas à proteína tais como catalise em reações químicas, transporte e armazenamento e funções regulatórias [3]. As atuais técnicas laboratoriais para o estudo da estrutura 3D de proteínas, como Cristalografia e Ressonância Nuclear Magnética, possuem limitações. Estas limitações justificam a aplicação de técnicas computacionais para o estudo destas moléculas. O problema do *Design* de Proteínas, cujo objetivo é deduzir a seqüência linear de aminoácidos correspondentes a uma determinada estrutura 3D, é um problema de alta complexidade computacional classificado como NP-Difícil na Teoria da Computação [4]. Descobrir a seqüência de aminoácidos que gera uma determinada estrutura protéica permite investigar métodos de sintetização para as funções biológicas associadas a esta proteína. A indústria farmacêutica tem grande interesse no desenvolvimento de novos métodos computacionais para a síntese de novas proteínas e, por conseguinte, de novos fármacos. Neste trabalho, técnicas de Mineração de Dados e Inteligência Artificial são combinadas para reduzir o espaço-solução do problema de design de proteínas. O principal objetivo desta pesquisa é desenvolver novos algoritmos computacionais para a determinação de seqüências de aminoácidos de proteínas cujas estruturas 3D são conhecidas. O método proposto é composto por cinco etapas: (1) Extração de informações 3D da proteína-alvo; (2) Seleção e clusterização das informações estruturais obtidas de modelos do *Protein Data Bank* (PDB); (3) Codificação do espaço-solução em uma Forma Normal Conjuntiva (FNC) para representação ótima do espaço-solução [5]; (4) Restrição do espaço conformacional da proteína por meio do método *Dead-End-Elimination*; (5) Geração das seqüências-solução da FNC para determinação da seqüência de aminoácidos da proteína-alvo. Para avaliação do método proposto foram realizados testes com diferentes proteínas-alvo: PDB ID: 20MM, 1ZDD, 1ARE, 1J9I e 1WG7. Os resultados obtidos revelam uma taxa de sucesso de 60% a 85% na determinação das seqüências de aminoácidos. Os resultados alcançados pelo método também apresentam desempenho computacional compatível ou superior a outros métodos tipicamente utilizados e descritos na literatura.

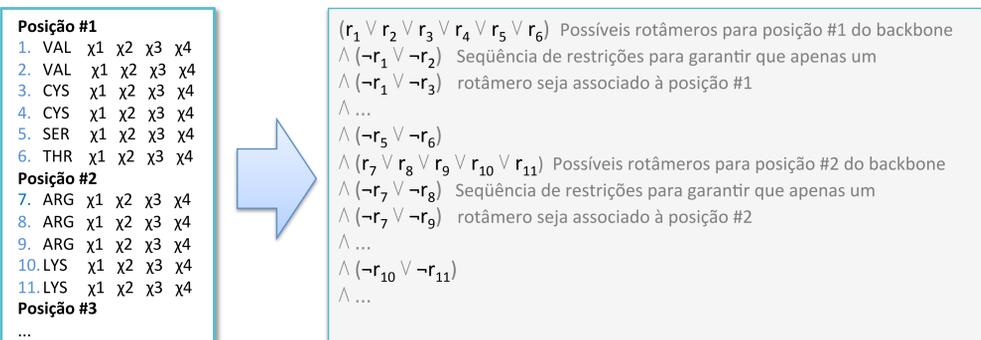
Passo I: Extração de Informações Estruturais



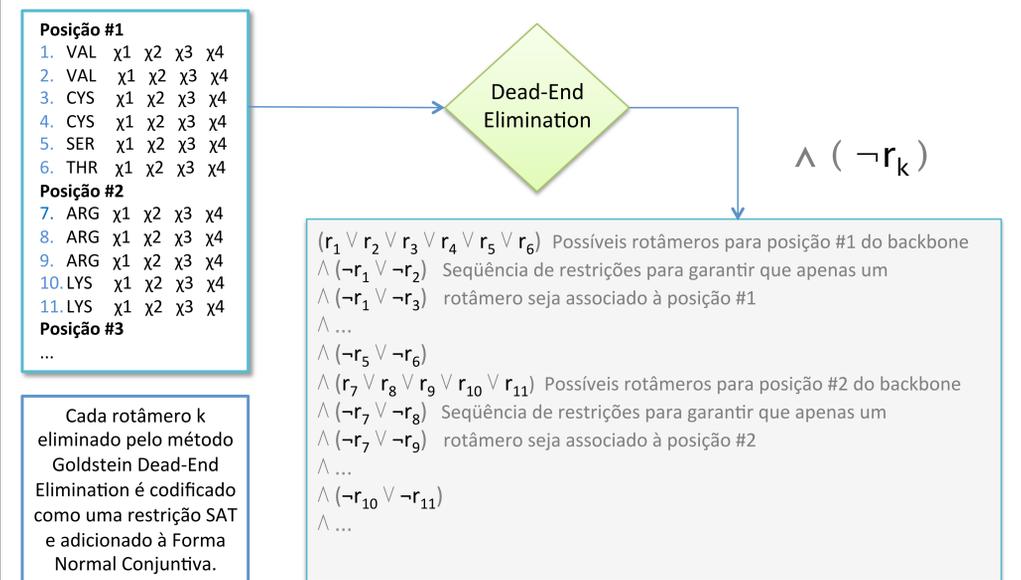
Passo II: Seleção e Clusterização das Informações Estruturais



Passo III: Codificação do Espaço-solução em Forma Normal Conjuntiva (FNC)



Passo IV: Restrição do Espaço Conformacional da Proteína pelo método Dead-End Elimination



Passo V: Geração das Seqüências-Solução da FNC



Resultados

20MM	1ZDD	1ARE
Seqüência original: GNNQQNY	Seqüência original: NMQCRRF	Seqüência original: CTRAFARQEA
Seqüência gerada: GNNQQND	Seqüência gerada: NEQCRRN	Seqüência gerada: CTRGNGRQEG
Precisão: 85,7% - 100%	Precisão: 75% - 75%	Precisão: 60% - 60%

LEGENDA	1J9I	1WG7
DESIGN CORRETO		
*DESIGN COMPATÍVEL		
DESIGN INCORRETO		
	Seqüência original: RGGGKGNVEVL	Seqüência original: EAAMQEKRRNG
	Seqüência gerada: RGGGRGNETN	Seqüência gerada: EGGEQERRNG
	Precisão: 70% - 70%	Precisão: 60% - 70%

* Aminoácido com polaridade compatível (hidrofóbico / hidrofílico).

Referências

- [1] Lippov, S., Tidor, B., 1997. Progress in computational protein design. *Curr. Opin. Biotechnol.* 18, 1.
- [2] Lehninger, A., Nelson, D., Cox, M., 2005. *Principles of Biochemistry*, 4th Edition. W.H. Freeman, New York.
- [3] Lesk, A. M., 2002. *Introduction to Bioinformatics*, 1st Edition. Oxford University Press Inc., New York.
- [4] Floudas, C., Fung, H., McAllister, S., Moennigmann, M., Rajgaria, R., 2006. Advances in protein structure prediction and de novo protein design: A review. *Chem. Eng. Sci.* 61 (3), 966.
- [5] Selman, B., Levesque, H., Mitchell, D., 1992. A new method for solving hard satisfiability problems. *Proceedings of the National Conference on Artificial Intelligence*, San Jose, USA, 440.