Pós-Graduação em Ciência da Computação

# “Estudo de processos para modelagem, avaliação, alteração e seleção da estrutura de proteínas”

**Por**

***Erico Souza Teixeira***

**Dissertação de Mestrado**

Universidade Federal de Pernambuco
posgraduacao@cin.ufpe.br
www.cin.ufpe.br/~posgraduacao

RECIFE, Junho/2005

Pós-Graduação em Ciência da Computação

**Dissertação de Mestrado**

# Estudo de processos para modelagem, avaliação, alteração e seleção da estrutura de proteínas.

**Por**
Erico Souza Teixeira

**Orientadora**
Katia Silva Guimarães

Universidade Federal de Pernambuco
posgraduacao@cin.ufpe.br
www.cin.ufpe.br/~posgraduacao

RECIFE, Junho/2005

UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

ERICO SOUZA TEIXEIRA

# "Estudo de processos para modelagem, avaliação, alteração e seleção de estruturas de proteínas"

*ESTE TRABALHO FOI APRESENTADO À PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO DO CENTRO DE INFORMÁTICA DA UNIVERSIDADE FEDERAL DE PERNAMBUCO COMO REQUISITO PARCIAL PARA OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIA DA COMPUTAÇÃO.*

ORIENTADOR(A): Katia Silva Guimarães

RECIFE, Junho/2005

Teixeira, Erico Souza
Estudo de processos para modelagem, avaliação, alteração e seleção de estruturas de proteínas / Erico Souza Teixeira . – Recife : O Autor, 2005.
126 folhas. : il., fig., tab.

Dissertação (mestrado) – Universidade Federal de Pernambuco. CIn. Ciência da Computação, 2005.

Inclui bibliografia, glossário e apêndices.

1.Ciência da computação – Bioinformática. 2. Modelagem de proteínas – Otimização de estruturas – Metodologia Ab Initio. 3. Modelagem – Anális estrutural – Aplicação na bioquímica . I. Título.

| | | |
|---|---|---|
| 004.4 | CDU (2.ed.) | UFPE |
| 005.5 | CDD (22.ed.) | BC2005-659 |

Este trabalho é dedicado à memória de Erico Teixeira da Silva

# Agradecimentos

Gostaria de agradecer a algumas pessoas que foram importantes para o desenvolvimento desse trabalho, pois sem cada um deles não encontraria a *energia* necessária para atingir a *melhor conformação* do projeto.

Inicialmente quero agradecer àqueles que me motivaram a seguir a carreira acadêmica. Em primeiro lugar a minha mãe Mª do Socorro, que confiou na minha decisão de seguir uma área tão pouco reconhecida como é a pesquisa. Às minhas irmãs, Erika, Elayne, Laura e Januana, que sempre apoiaram a minha luta pelos meus sonhos. Aos meus amigos, Luís Antônio, Mathias Rufino, Paulo Barros, Sílvio Cortez e Gustavo Bastos, e minha namorada, Mayra Melo, incentivadores diários do meu curso de mestrado.

Quero agradecer também àqueles que de alguma forma ajudaram a construção do projeto. Mais uma vez a Mathias, físico excepcional, com quem passei muito tempo resolvendo equações trigonométricas e discutindo a movimentação dos átomos. A Laura e Mayra, designs que construíram muitas das imagens presentes nesse trabalho, que não representam 1% da capacidade de artista gráfico dessas garotas. Aos revisores de minha dissertação, Paulo, Taciana Pontual, Fred Pontes, Jeane Cecília, Mathias e Mayra, que tiverem paciência de ler todo o material, comentá-lo e ainda dizer que eu não estava atrapalhando as suas vidas. A Gustavo, bioinformata de grande potencial e companheiro deste os tempos da graduação, sempre disposto a discutir os problemas de implementação do meu projeto. A Obionor Oliveira, Everton Rennê e Jordep, mestres na arte do suporte de computadores, que me ajudaram na instalação dos vários programas utilizados no projeto. Ao professor Ricardo Longo, do Departamento de Química da UFPE, que teve bastante paciência em desvendar alguns dos mistérios da química.

**Guerra é Paz**
**Liberdade é Escravidão**
**Ignorância é Força**
Os três lemas do Partido – 1984 – (George Orwell)

# Resumo

O projeto consiste no estudo de diversos sistemas aplicados no processo de construção de modelos de estruturas terciárias de proteínas, considerando etapas de modelagem e avaliação da estrutura, assim como mecanismos para a modificação visando a melhora das conformações desenvolvidas.

Ao longo desse estudo foi também desenvolvido um conjunto de interfaces que permite a comunicação entre os diversos tipos de programas utilizados nas etapas citadas acima. Além disso, um método para a alteração e melhora de pequenas regiões de um modelo da proteína foi implementado e adicionado no fluxo do processo de modelagem.

Por fim, o algoritmo de construção de novas conformações foi avaliado em estruturas originadas da modelagem por homologia, criadas pela ferramenta MODELLER, sempre apresentado uma melhora nas regiões modificadas, enquanto elas apresentassem uma distância superior a 1,7 Å da estrutura nativa.

# Índice

**Lista de Figuras** ..... **8**
**Lista de Tabelas** ..... **10**
**Capítulo 1. Introdução** ..... **11**
**Capítulo 2. Estado da arte** ..... **16**
Seção 2.1. Bioinformática ..... 16
Seção 2.2. Conceitos preliminares ..... 17
2.2.1. Ligação Covalente ..... 17
2.2.2. Ligações de hidrogênio ..... 19
2.2.3. Interação de Van der Waals ..... 19
2.2.4. Interação eletrostática ..... 21
2.2.5. Ângulos Diedros ..... 21
2.2.6. Minimização Local e Global ..... 22
2.2.7. Raiz quadrada da média das distâncias (RMSD) ..... 23
2.2.8. Alinhamento de seqüências ..... 25
Seção 2.3. Aminoácidos ..... 25
Seção 2.4. Proteínas ..... 29
2.4.1. Formação da Proteína ..... 29
2.4.2. Do DNA à proteína ..... 31
2.4.3. Estruturas das Proteínas ..... 32
2.4.4. Ângulos diedros das proteínas ..... 37
Seção 2.5. Banco de Proteínas ..... 39
Seção 2.6. Predição de Estruturas de Proteínas ..... 40
Seção 2.7. Modelagem por homologia ..... 43
2.7.1. MODELLER ..... 47
Seção 2.8. Avaliação de Estruturas de Proteínas ..... 51
2.8.1. ANoLEA ..... 55
2.8.2. ProSa ..... 56
2.8.3. PROVE ..... 57
2.8.4. PROCHECK ..... 59
Seção 2.9. Reconstrução de uma cadeia polipeptídica ..... 60
Seção 2.10. Gerando conformações de polipeptídios ..... 61
Seção 2.11. Construção das cadeias laterais ..... 62
**Capítulo 3. Metodologia** ..... **65**
Seção 3.1. Modelagem de proteínas ..... 66
Seção 3.2. Avaliação das estruturas ..... 67
Seção 3.3. Construção dos novos modelos ..... 68
3.3.1. Obtenção do conjunto de pares de ângulos diedros ψ/Φ ..... 69
3.3.2. Representação da estrutura da proteína ..... 70
3.3.3. Técnica de construção das conformações ..... 73
3.3.4. Inserção de um átomo pelo ângulo diedro ..... 75
3.3.5. Inserção do átomo de oxigênio pela planaridade ..... 81
3.3.6 Inserção do átomo de Cβ pela quiralidade ..... 83
3.3.7. Validando as novas conformações ..... 86
3.3.8. Algoritmo para construção do conjunto das conformações ..... 88
3.3.9. Seleção das melhores conformações ..... 89
3.3.10. Seleção do modelo ..... 90
**Capítulo 4. Testes e discussão** ..... **91**
Seção 4.1. Precisão do Conjunto de conformações ..... 91
Seção 4.2. Modelagem de estruturas ..... 94

4.2.1. T0129 ..... 95
4.2.2. T0130 ..... 97
**Capítulo 5. Conclusão e Trabalhos futuros ..... 104**
**Bibliografia ..... 108**
**Glossário ..... 114**
**Apêndice FASTA ..... 118**
**Apêndice GenBank ..... 119**
**Apêndice MODELLER ..... 121**
**Apêndice PDB ..... 125**

# Lista de Figuras


| | | |
|---|---|---|
| 1.I | - DNA | 12 |
| 2.2.1.I | - Ligação Covalente | 18 |
| 2.2.1.II | - Energia | 18 |
| 2.2.1.III | - Eletronegatividade | 19 |
| 2.2.2.I | - Ligação de Hidrogênio | 20 |
| 2.2.3.I | - Interação de Van der Waals | 20 |
| 2.2.4.I | - Interação Eletrostática | 21 |
| 2.2.5.I | - Ângulo Diedro | 22 |
| 2.2.5.II | - Orientação do Diedro | 22 |
| 2.2.6.I | - Minimização | 23 |
| 2.2.7.I | - Superposição | 24 |
| 2.2.7.II | - RMSD | 24 |
| 2.3.I | - Aminoácido | 26 |
| 2.3.II | - 20 Aminoácidos | 27 |
| 2.3.III | - Nomenclatura | 27 |
| 2.3.IV | - Moléculas Especulares | 28 |
| 2.3.V | - Glicina | 28 |
| 2.3.VI | - Alanina L-D | 28 |
| 2.4.1.I | - Ligação Peptídica | 29 |
| 2.4.1.II | - Ressonância | 30 |
| 2.4.1.III | - Ligação Planar | 30 |
| 2.4.1.IV | - Anti-Syn | 30 |
| 2.4.2.I | - Tanscrição DNA | 31 |
| 2.4.2.II | - Tradução RNAm | 32 |
| 2.4.3.I | - Estrutura Covalente | 33 |
| 2.4.3.II | - Estrutura Secundária | 35 |
| 2.4.3.III | - Alça-β | 36 |
| 2.4.3.IV | - Estrutura Terciária | 36 |
| 2.4.3.V | - Dobras | 36 |
| 2.4.3.VI | - Estrutura Quaternária | 37 |
| 2.4.4.I | - Φ | 38 |
| 2.4.4.II | - ψ | 38 |
| 2.4.4.III | - ω | 38 |
| 2.7.I | - Modelagem por Homologia | 48 |
| 2.8.I | - AMBER | 53 |
| 2.8.1.I | - 40 átomos | 56 |
| 2.8.1.II | - Região não-local | 57 |
| 2.8.3.I | - Voronoi | 58 |
| 2.8.4.I | - $\chi_1$ | 60 |
| 2.11.I | - Energia SCWRL | 63 |
| 2.11.II | - Resolvendo Componente Biconexo | 64 |
| 3.I | - Ciclo | 65 |
| 3.2.I | - Regiões Individuais | 68 |
| 3.2.II | - Todas as Regiões | 68 |
| 3.3.2.I | - Grade2D | 72 |
| 3.3.3.I | - Ramachandran | 74 |
| 3.3.4.I | - Diedro $A_1$-$A_2$-$A_3$-$A_4$ | 76 |
| 3.3.4.II | - Posicionando $A_4$ | 77 |

| | | |
|---|---|---|
| 3.3.4.III | - Translação | 77 |
| 3.3.4.IV | - $\theta_2/\varphi_2$ | 78 |
| 3.3.4.V | - Rotação $\varphi_{2_}\pi$ | 78 |
| 3.3.4.VI | - Rotação Z | 78 |
| 3.3.4.VII | - Rotação $\theta_2$ | 79 |
| 3.3.4.VIII | - Rotação X | 79 |
| 3.3.4.IX | - $\varphi_1$ | 80 |
| 3.3.4.X | - Posição Final $A_4$ | 80 |
| 3.3.5.I | - Planaridade | 81 |
| 3.3.5.II | - Deslocamento da Planaridade | 82 |
| 3.3.5.III | - Planaridade do O | 82 |
| 3.3.6.I | - Deslocamento da quiralidade | 83 |
| 3.3.6.II | - Translação da quiralidade | 84 |
| 3.3.6.III | - Posicionando Cβ | 84 |
| 3.3.6.IV | - Orientação φ | 85 |
| 3.3.6.V | - CORN | 85 |
| 3.3.6.VI | - Determinando φ | 86 |
| 3.3.8.I | - Fluxo | 89 |
| 4.2.1.I | - Avaliação T0129-01 | 96 |
| 4.2.2.I | - Avaliação T0130-01 | 98 |
| 4.2.2.II | - Região 3D-01 | 99 |
| 4.2.2.III | - Avaliação T0130-02 | 100 |
| 4.2.2.IV | - Região 3D-02 | 100 |
| 4.2.2.V | - Região 3D-03 | 102 |
| 4.2.2.VI | - Região 3D-04 | 102 |

# Lista de Tabelas


| | | |
|---|---|---|
| 2.4.2.I | - Códon-Nucleotídeo | 32 |
| 2.4.4.I | - Ângulos Diedros das Cadeias Laterais | 39 |
| 2.7.1.I | - Modelagem | 48 |
| 2.8.3.I | - Prove | 59 |
| 3.3.1.I | - Conjuntos ψ/Φ | 70 |
| 3.3.2.I | - Raio de Van der Waals | 71 |
| 3.3.2.II | - Alfabeto | 72 |
| 3.3.3.I | - Engh & Huber | 73 |
| 4.1.I | - Laços Fiser | 92 |
| 4.1.II | - Média Dos Melhores RMSD | 93 |
| 4.2.2.I | - Seleção T0130-01 | 98 |
| 4.2.2.II | - Seleção T0130-02 | 100 |
| 4.2.2.III | - Análise T0130 | 101 |
| 4.2.2.vI | - Análise de tempo | 103 |

# Capítulo 1. Introdução

"Eu não vim aqui dizer como isso vai acabar.
Eu vim dizer como isso vai começar."
Neo – Matrix (1998)

O homem, ao longo de suas pesquisas e busca do conhecimento, procurou entender o funcionamento de um dos sistemas mais complexos já apreciados: a vida. Muitos cientistas e filósofos criaram as mais variadas teorias, não só para compreender os elaborados sistemas de atividade celular realizados pelos organismos, mas também para determinar o elemento responsável pela origem, criação e transmissão das características de cada um dos sistemas.

Por volta de 1940 já se sabia que as proteínas (macromoléculas presentes nas células) são as responsáveis pela determinação das funcionalidades e pelo controle do processo da vida. Mas foi apenas em 1947, com os trabalhos realizados por Erwin Chargaff, e em 1952, pelos projetos de Alfred Hershey e Martha Chase, que ficou evidente que o DNA (ácido desoxirribonucléico) é o material de hereditariedade dos organismos. Em 1953, com os projetos desenvolvidos por James D. Watson e Francis Crick, sua estrutura foi determinada **[WC53] (Figura 1.I: DNA)**.

Com o objetivo de entender como uma estrutura simples de DNA tem a propriedade de transmitir diversas informações para as mais distintas espécies de seres vivos, iniciaram-se projetos de determinação das seqüências dessas moléculas presentes nos diversos organismos. Os métodos aplicados foram desenvolvidos em 1977, por Maxam e Gilbert **[MG77]** e por Sanger et al **[SNC77]**, sendo a primeira base de dados contendo esses elementos construída pelo Laboratório Nacional de Los Alamos, no Novo México, por membros do GenBank e pelo Laboratório Europeu de Biologia Molecular (EMBL) **[EMBL]** em 1979. Era o princípio do relacionamento entre as técnicas da informática e os estudos da biologia. À medida que novas seqüências eram submetidas e armazenadas nos bancos de dados, essa interação crescia, pois a quantidade de dados gerada era enorme. Além disso, aumentava o interesse pela análise das seqüências através, principalmente, do desenvolvimento de programas computacionais.

Atualmente, o processo de seqüenciamento se tornou uma rotina nos laboratórios de biologia molecular, um mecanismo consolidado que não sofre muitas modificações.

Em contrapartida, as análises estão em pleno desenvolvimento. Um exemplo disso é a busca das regiões do DNA que controlam a produção das proteínas.

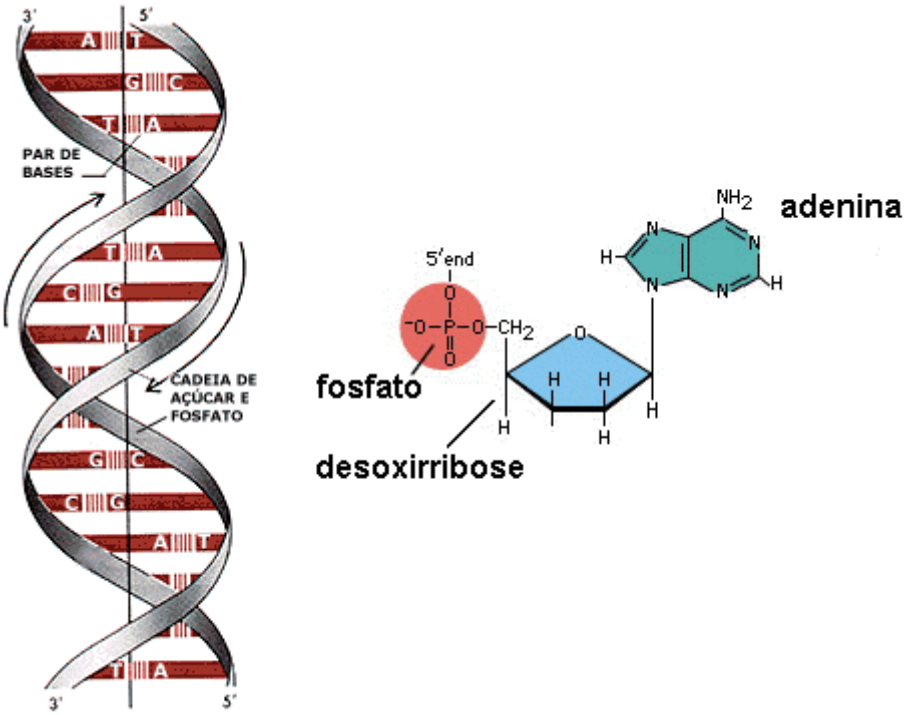



**Figura 1.I: DNA** – Do lado esquerdo, a dupla-hélice da cadeia de DNA, mostrando a ligação entre os pares de base Adenina (A) e Timina (T), Guanina (G) e Citosina (C). Do lado direito, uma representação de um nucleotídeo, a molécula fundamental do DNA, distinguindo três regiões: a base nitrogenada (neste caso a Adenina), a desoxirribose (açúcar) e o fosfato.

Da mesma forma que ocorreu com o DNA, também se buscou determinar e armazenar as seqüências de proteínas. O primeiro método de sequenciamento das proteínas surgiu ainda em 1952, construído por Sanger e Tuppy **[ST51]**, sendo a primeira base de dados um livro que reunia as seqüências, criado por Margaret Dayhoff, em 1972 **[Day72]**.

Obter a seqüência de uma proteína era apenas uma fase inicial, assim como foi para o DNA. Projetos para análise foram e ainda são desenvolvidos, a fim de compreender o mecanismo de atividade e a finalidade de cada uma delas. Mas para que esses estudos possam ser executados da melhor maneira possível, é preciso criar mecanismos para descobrir as estruturas das proteínas, pois existe uma forte relação entre suas conformações tridimensionais e suas funcionalidades **[Mou01]**.

Há vários métodos disponíveis para a obtenção das conformações 3D, também chamadas de estruturas terciárias. Dentre elas, as técnicas laboratoriais, como a Ressonância Magnética Nuclear (RMN) e a Cristalografia por Raios-X, destacam-se por

constituírem as metodologias mais precisas. Uma maior precisão garante um melhor conhecimento das funções biológicas das proteínas e de suas interações com outros substratos, e, por conseguinte, uma melhor determinação do mecanismo de ação das mesmas. No entanto, tais metodologias são lentas e caras para se resolver o problema, podendo alguns projetos durarem até um ano.

Outros mecanismos mais rápidos foram desenvolvidos, como as abordagens por homologia, *threading* e *ab initio*. A modelagem por homologia determina o modelo da estrutura de uma proteína a partir de estruturas já conhecidas, baseada na observação de que proteínas com seqüências similares apresentam estruturas também similares. No procedimento *ab initio* busca-se o novo modelo de uma proteína por minimização de energia, sob o principio de que a estrutura real de uma proteína é a que apresenta a menor energia. A metodologia de *threading* é uma combinação de homologia e *ab initio*. Todavia, essas metodologias não são bastante precisas quando comparadas aos métodos experimentais.

Mas para qualquer que seja a metodologia aplicada, o modelo criado da estrutura da proteína apresenta trechos defeituosos, ou seja, que não estão de acordo com a conformação real. Assim, mecanismos de análise de estruturas foram desenvolvidos com o objetivo de identificar as regiões problemáticas. As análises podem estar relacionadas com propriedades estereoquímicas (ver **Glossário**), como distância e ângulos entre os átomos, análise estatística de banco de dados ou energias baseadas em campos de força.

Uma vez determinados os trechos defeituosos, pode-se modificar suas conformações a fim de tornar o modelo mais preciso, ou seja, com menos regiões indicadas na análise de estruturas. A alteração da conformação dessas regiões pode ser feita através da resolução de polinômios de alta ordem, busca em banco de dados ou minimização de energia.

Ao atingir a precisão aceitável para o modelo estrutural da proteína, dar-se continuidade a análise da proteína, com a determinação das funcionalidades bioquímicas e dos mecanismos de interação com outras moléculas.

Tendo em vista dar aos profissionais que lidam com a modelagem melhores condições para se obter conformações mais próximas das reais, propõe-se um modo de eliminar os pequenos defeitos encontrados em estruturas de proteínas. A ferramenta desenvolvida permite a construção da conformação de uma proteína por homologia, avaliação da estrutura, seleção de regiões defeituosas, criação de novas conformações

para estas regiões, seleção das conformações das regiões mais próximas do real, construção de um novo modelo com as conformações selecionadas e escolha do melhor modelo. Para atingir esse objetivo, foram implementados:

- Um sistema que facilita a utilização dos principais programas aplicados na modelagem de proteínas, através da criação de interfaces de comunicação entre eles, já que existe bastante dificuldade em conciliar os diversos programas com suas mais variadas saídas e parâmetros de entrada;
- Um mecanismo para gerar um consenso entre as regiões defeituosas indicadas pelas ferramentas de avaliação de estruturas, pois nem sempre há uma concordância entre os trechos determinados por cada um dos softwares;
- Uma técnica para a construção de conjuntos de conformações para seqüência de aminoácidos (conjunto *decoy*), que pode ser aplicada para a análise, a determinação e a predição de estruturas de proteínas;
- Um mecanismo de seleção das conformações mais próximas de uma estrutura real em um conjunto *decoy*, dando opções para a alteração do modelo da proteína em estudo;
- Um sistema para a seleção do melhor modelo, dentre os vários construídos, após a alteração das estruturas das regiões defeituosas, substituídas pelas estruturas selecionadas no conjunto *decoy*.

Outros projetos que se dedicam à construção e melhora das estruturas das proteínas o fazem atacando apenas blocos do problema: modelagem, ou validação ou modificação ou seleção de estruturas. Isto dificulta bastante o desenvolvimento de novos modelos, pois requer do usuário um trabalho adicional para conciliar os diferentes tipos de arquivo com o qual cada software lida.

Além disso, projetos dedicados à construção de novas conformações e seleção das melhores estruturas produzem testes para demonstrar a validade da nova ferramenta apenas sob modelos originados dos processos experimentais, deixando de lado a metodologia mais usada atualmente para modelagem: a homologia. Nesse aspecto, um dos testes realizados no projeto aqui descrito considera as estruturas originadas da homologia.

Este projeto é o primeiro passo para a criação de um sistema que facilite a modelagem e a avaliação de qualquer estrutura de proteína. O objetivo principal

entender todos os processos presentes no processo de modelagem e construir o primeiro esqueleto de um mecanismo que promova o trabalho dos bioquímicos e biólogos.

A ferramenta foi desenvolvida em três blocos independentes e intercomunicáveis: modelagem, avaliação e alteração das estruturas, o que permite modificações pontuais. Futuras versões deverão incluir outros programas utilizados na modelagem de estruturas, facilidades gráficas e inserção de maior quantidade de informações e propriedades bioquímicas.

O documento está dividido em quatro partes: o **Capítulo 2** descreve as ferramentas que foram aplicados no projeto e os principais conceitos estudados para o entendimento do processo de determinação da estrutura de proteína; **o Capitulo 3** apresenta como foram desenvolvidas as três etapas do projeto: modelagem, avaliação e alteração das estruturas, relatando os algoritmos aplicados, assim como as interfaces de comunicação; **Capítulo 4** traz os resultados e suas análises para os dois diferentes tipos testes realizados; por fim, o **Capítulo 5** compreende as conclusões deste projeto.

# Capítulo 2. Estado da arte

"I want him know what I know.
I want him to know I want him to know."
A Noiva – Kill Bill: Vol. 1 (2003)

Neste capítulo há uma introdução que descreve a comunicação existente entre os campos da Informática, Química e Biologia, definindo a área multidisciplinar denominada por Bioinformática (Seção 2.1. Bioinformática). Em seguida, há a definição de alguns termos necessários para o entendimento dos procedimentos descrito neste documento (Seção 2.2. Conceitos preliminares). As seções 2.3 e 2.4 definem os principais elementos focados neste projeto: os aminoácidos e as proteínas, respectivamente. As seções restantes mostram o estado da arte para as várias etapas envolvidas no processo de modelagem: armazenamento de seqüência, predição, avaliação, modificação e seleção de estruturas.

## Seção 2.1. Bioinformática

A pequena quantidade de dados originados nos primeiros estudos feitos pela Química e Biologia para o entendimento das estruturas e funcionalidades do DNA e proteínas não tornava necessário o desenvolvimento de complexos mecanismos de armazenamento, pois as metodologias de obtenção das seqüências ainda se encontravam num estágio embrionário. Entretanto, após o desenvolvimento de técnicas de seqüenciamento mais rápidas e precisas, e com o avanço da Biologia Molecular e da Bioquímica, a quantidade de informação apresentou um crescimento bastante significativo, não sendo mais possível ser organizada e avaliada apenas através do trabalho manual.

Havia agora a necessidade de empregar estruturas capazes de armazenar grandes quantidades de dados, além de disponibilizar meios para um fácil acesso, atualização, visualização e análise dos mesmos. Isto promoveu o crescimento do relacionamento e da troca de informações entre as disciplinas de Química, Biologia e Ciência da Computação.

Esse trabalho envolvendo conceitos relativos à Informática, à Biologia e à Química, além de outras áreas como a Matemática e a Estatística, deu origem à Bioinformática. Inicialmente, havia apenas a preocupação com a construção de bancos de dados contendo as informações originadas dos projetos de seqüenciamento, cujos

problemas eram relacionados à organização interna dos dados (*design* do projeto) e a interfaces de comunicação direcionadas aos pesquisadores para manuseio dos dados (visualização, submissão e atualização). A partir do momento em que o interesse pela análise dos dados aumentou, outros objetivos foram traçados na Bioinformática, com a aplicação de ferramentas e algoritmos especializados, explorando técnicas de mineração de dados e inteligência artificial.

De uma forma geral, a Bioinformática é principalmente empregada nas indústrias farmacêutica e agropecuária, proporcionando a construção de ferramentas computacionais, que tornem eficiente o acesso, uso e gerenciamento das informações, além da criação de novos algoritmos e da aplicação de técnicas estatísticas e matemáticas para avaliar a grande massa de dados.

## Seção 2.2. Conceitos preliminares

Durante o desenvolvimento deste projeto, surgiram vários termos técnicos e definições da Biologia, Química e Informática, sendo muitos deles mencionados ao longo deste texto. Por se tratar de um documento direcionado a profissionais e estudantes de áreas muito distintas, essa seção foi dedicada ao esclarecimento desses termos, a fim de facilitar a plena compreensão da dissertação.

### 2.2.1. Ligação Covalente

A ligação covalente ocorre quando dois átomos se unem e atingem um menor estado energético pelo compartilhamento de pares de elétrons de modo que cada par seja formado por um elétron de cada um dos átomos. O número de pares de elétrons compartilhados é chamado covalência.

Num átomo, os elétrons se encontram distribuídos em diversos níveis (camadas eletrônicas) e para que atinja estabilidade é preciso, na maioria dos casos, seguir a regra do octeto. Nesta regra, o átomo adquire energia mínima quando possui 8 elétrons na camada eletrônica mais externa (de valência), ou 2 elétrons, quando possui uma única camada . Por exemplo: o Carbono (C) possui quatro elétrons na última camada **(Figura 2.2.1.I: Ligação Covalente)**, precisando de mais quatro para uma melhor estabilidade. O Oxigênio (O) apresenta seis elétrons na última camada, necessitando apenas de mais dois. Desta forma, o C se liga a dois átomos de O por duas ligações covalentes.

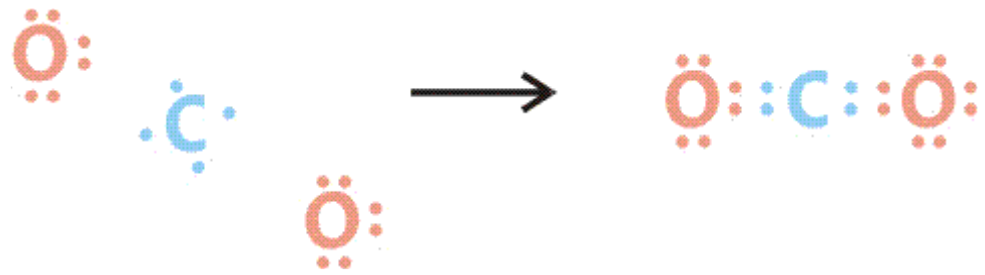


**Figura 2.2.1.I: Ligação Covalente** – Formação do dióxido de carbono ($CO_2$) por ligação covalente.

A energia mínima ou a situação de estabilidade entre dois átomos ocorre quando há um equilíbrio entre as forças de repulsão e atração elétricas existentes entre os núcleos e os elétrons dos próprios átomos. A função que rege o nível de energia entre dois átomos varia de acordo com a distância entre seus núcleos **(Figura 2.2.1.II: Energia)**. Quando a distância entre os átomos é maior do que na posição de menor energia, a força de atração é predominante (parte contínua da curva); ao diminuir a distância em relação à posição de equilíbrio, a força de repulsão apresenta uma maior influência (parte pontilhada-tracejada da curva).

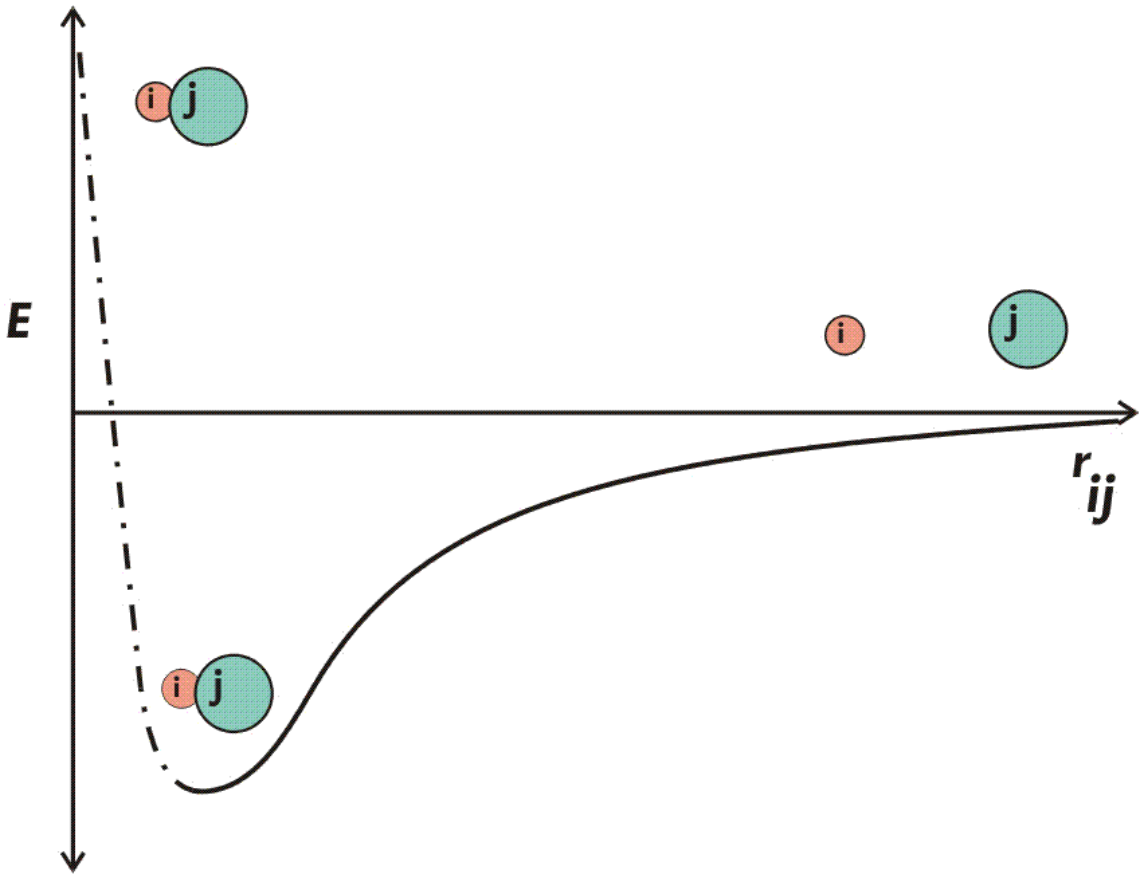



**Figura 2.2.1.II: Energia** – Representação da função de energia ( **E** ) entre dois átomos **i** e **j** em função da distância que separa seus núcleos ($\mathbf{r}_{ij}$).

Numa ligação covalente formada por átomos diferentes, é comum um deles atrair mais fortemente o par eletrônico do que o outro. Esta capacidade que um átomo tem de atrair para si o par eletrônico que compartilha é denominada de eletronegatividade. Na **Figura 2.2.1.III: Eletronegatividade**, as duas moléculas à esquerda apresentam átomos do mesmo tipo; conseqüentemente, a atração que os átomos exercem sobre o par de elétrons compartilhado é a mesma. Para a molécula da direita, o Cloro (Cl) apresenta uma maior eletronegatividade e atrai para mais próximo de si o par de elétrons compartilhado, formando uma molécula polar.

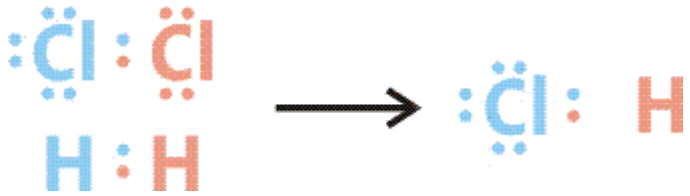



**Figura 2.2.1.III: Eletronegatividade** – Ação da eletronegatividade em moléculas diatômicas.

## 2.2.2. Ligações de hidrogênio

As ligações de hidrogênio ocorrem quando um átomo de Hidrogênio (H) está ligado covalentemente a átomos eletronegativos, como o Flúor (F), o Oxigênio (O) e o Nitrogênio (N), e apresenta uma forte atração com o par de eletros desses mesmos tipos de átomos presentes na molécula vizinha (intermoleculares) ou na própria molécula (intramoleculares). A **Figura 2.2.2.I: Ligação de Hidrogênio A** mostra moléculas de água formando um dipolo elétrico, com o lado negativo ($\delta^-$) representado pelo elemento mais eletronegativo, caracterizando, assim, uma molécula polar. Esse dipolo criado é essencial na formação das ligações de hidrogênio, onde o lado positivo (Hidrogênio) é atraído pelo lado negativo (neste exemplo, o Oxigênio). Em alguns casos, a ligação de hidrogênio pode ocorrer dentro da própria molécula **(Figura 2.2.2.I: Ligação de Hidrogênio B)**.

## 2.2.3. Interação de Van der Waals

As interações de Van der Waals, ou forças de dispersão de London, são outro tipo de forças intermoleculares responsáveis por manter moléculas eletricamente neutras ou apolares (como $H_2$, $F_2$, $Cl_2$) próximas. Essas forças resultam do seguinte: num determinado instante, uma molécula pode apresentar mais elétrons num lado que em outro, estando, momentaneamente polarizada. Por indução elétrica, esta molécula

provoca a polarização da molécula vizinha, resultando numa fraca atração entre ambas **(Figura 2.2.3.I: Interação de Van der Waals)**.

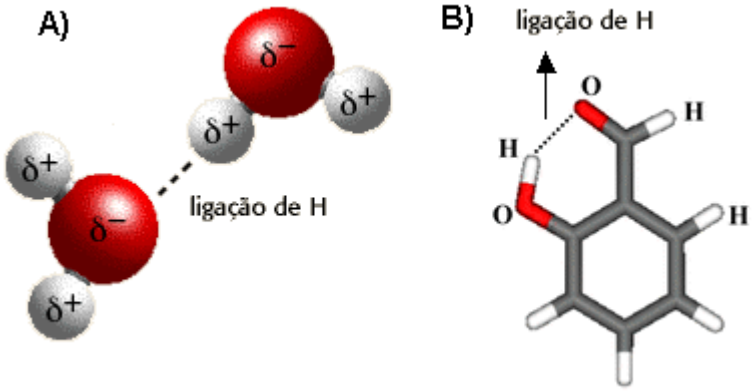



**Figura 2.2.2.I: Ligação de Hidrogênio** – a) Ligação de hidrogênio em moléculas de água ($H_2O$). b) Representação de uma ligação de hidrogênio intramolecular.

Essa força entre os átomos é normalmente modelada de acordo com a função de Lennard-Jones 12-6:

$$v(r) = 4\varepsilon\left(\left(\frac{\sigma}{r}\right)^{12} - 2\left(\frac{\sigma}{r}\right)^{6}\right) \qquad (\text{Eq. 2.2.3.I})$$

onde **ε** e **σ** são parâmetros que dependem do par de átomos interagentes, e o termo ***r*** representa a distância entre os átomos. A **Figura 2.2.1.II: Energia** corresponde à representação gráfica dessa função, sendo o termo elevado a potência 12 associado à repulsão e o termo a potência 6 à atração.

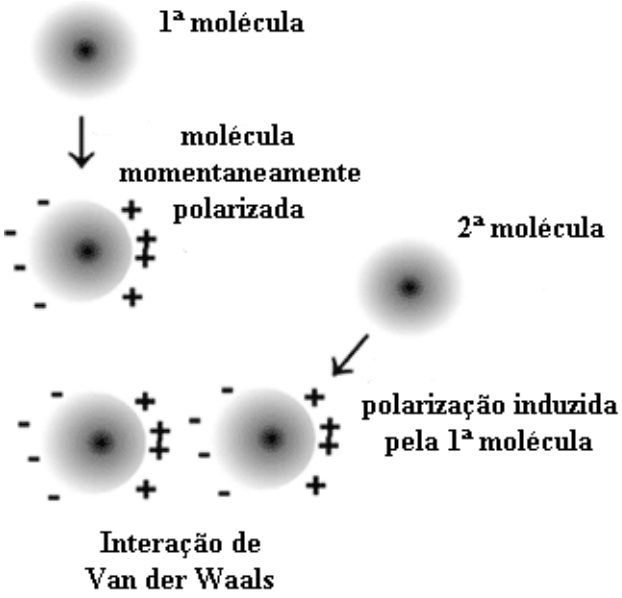



**Figura 2.2.3.I: Interação de Van der Waals** - Mecanismo do surgimento da interação de Van der Waals.

## 2.2.4. Interação eletrostática

Interação eletrostática é a força que age num átomo de carga elétrica $\mathbf{q_1}$ como resultado da presença de um outro átomo de carga $\mathbf{q_2}$, podendo ser de repulsão quando as cargas apresentam o mesmo sinal (negativa-negativa ou positiva-positiva) ou de atração, quando os átomos apresentam cargas opostas (positiva-negativa) **(Figura 2.2.4.I: Energia Interação Eletrostática)**.

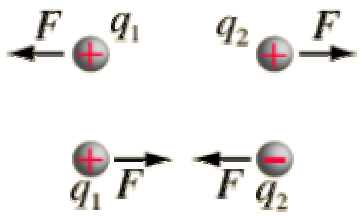



**Figura 2.2.4.I: Interação Eletrostática** – Representação das forças que agem nos átomos quando eles apresentam cargas de mesmo sinal (positivo-positivo) e quando se encontram com cargas de sinais opostos (positivo-negativo).

Para determinar a força que age nos átomos, aplica-se a Lei de Coulomb:

$$F = \frac{q_1 q_2}{4\pi\epsilon_0 d^2} \quad (\text{Eq. 2.2.4.I})$$

onde ***d*** é a distância entre as cargas e $\boldsymbol{\epsilon_0}$ é a constante elétrica. Quando as cargas estão imersas num meio dielétrico de constante $\boldsymbol{\epsilon}$, a equação de força torna-se,

$$F = \frac{q_1 q_2}{4\pi\epsilon_0 \epsilon d^2} \quad (\text{Eq. 2.2.4.II})$$

## 2.2.5. Ângulos Diedros

Ângulo diedro ou ângulo diédrico, ou simplesmente diedro, é o menor ângulo formado a partir da intersecção entre dois planos **(Figura 2.2.5.I: Ângulo Diedro)**.

Um diedro formado a partir da intersecção dos planos determinados pelos pontos 1-2-3-4 **(Figura 2.2.5.II: Orientação do Diedro A)** será positivo caso a rotação do eixo 3-4 seja horária em relação a 1-2, quando se visualiza a ligação 2-3 para baixo. Além disso, o ângulo é associado ao valor de 0º quando os pontos se apresentam numa conformação *syn* (em forma de U), ou seja, 1 e 4 estão no plano descrito por 1-2-3 do mesmo lado em relação à ligação 2-3 **(Figura 2.2.5.II: Orientação do Diedro B)**; por fim, o ângulo apresenta o valor de 180º na conformação *anti* **(Figura 2.2.5.III:**

**Orientação do Diedro C**), quando 1 e 4 se encontram em lados opostos. Dessa forma, o valor assumido por um diedro está compreendido no intervalo (-180º, +180º].

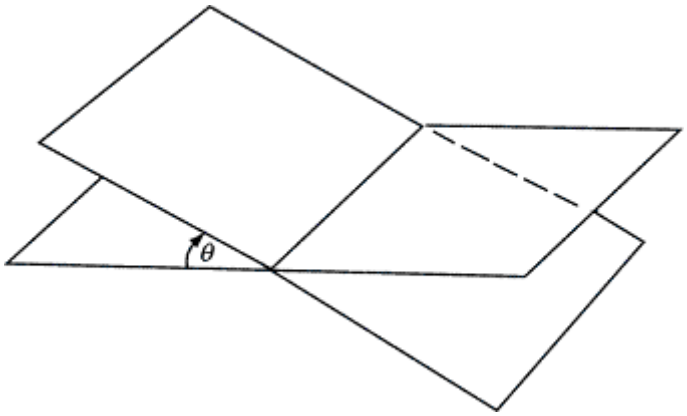



**Figura 2.2.5.I: Ângulo Diedro** - A partir da intersecção entre dois planos, o menor ângulo é determinado como o diedro, neste exemplo o ângulo é θ.

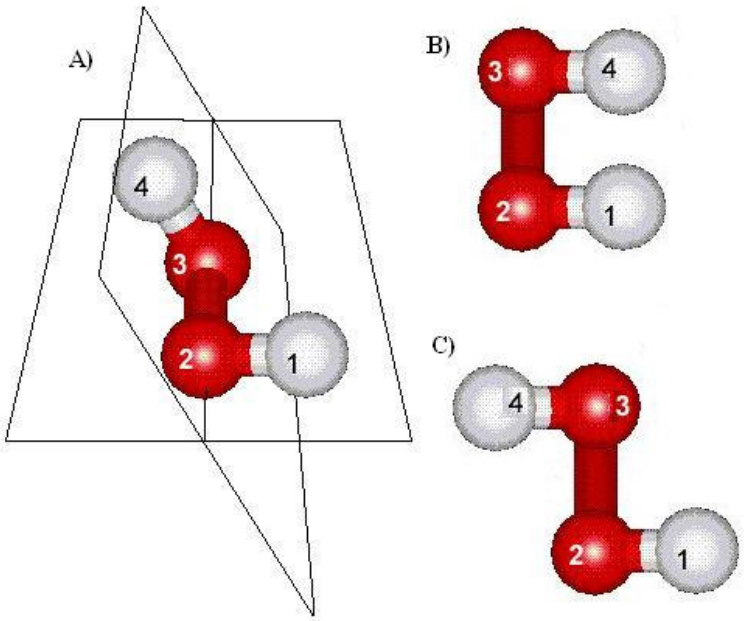



**Figura 2.2.5.II: Orientação do Diedro** – A) O ângulo diedro, formado a partir da intersecção dos planos determinados pelos pontos 1-2-3-4, será positivo, neste exemplo, caso a rotação do eixo 3-4 seja orientada para a direita; B) Conformação *syn*, em forma de U, associada ao valor de 0º do diedro; C) Conformação *anti*, referente a 180º.

## 2.2.6. Minimização Local e Global

Muitas vezes busca-se obter a estrutura de uma molécula a partir do processo de minimização da função de energia da molécula. Essa minimização pode ocorre de duas formas: localmente e globalmente.

A minimização local, ou otimização local, é o algoritmo que busca a melhor solução para um problema dentro de um conjunto de soluções próximas entre si, solução

esta que depende do ponto de início do processo **(Figura 2.2.6.I: Minimização)**. Por outro lado, a minimização global busca a melhor solução entre todas as possíveis. Infelizmente, nem sempre é possível obtê-la, sendo impraticável em algumas situações.

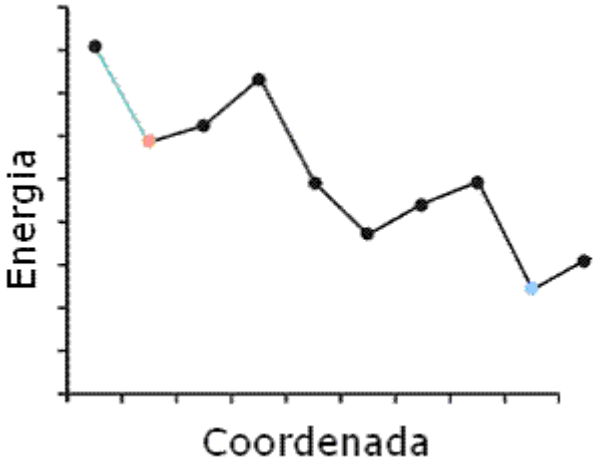



**Figura 2.2.6.I: Minimização** – Dado o gráfico acima, uma busca pelo menor valor local realizada num processo de minimização, que tomasse como valor inicial qualquer ponto pertencente ao segmento em verde, inevitavelmente retornaria o valor correspondente ao ponto em laranja. Caso o mesmo procedimento fosse realizado com um algoritmo de otimização Global, era bem provável que o valor de retorno correspondesse ao ponto azul.

## 2.2.7. Raiz quadrada da média das distâncias (RMSD)

Dadas duas estruturas espaciais, ou um trecho de cada uma das estruturas, com a mesma seqüência de átomos, a raiz quadrada da média das distâncias (RMSD) é calculada a partir da seguinte expressão:

$$d_{\mathrm{r.m.s.}} = \left( \frac{1}{N} \sum_{i=1}^{N} \left| \vec{r}_i(A) - \vec{r}_i(B) \right|^2 \right)^{1/2} \qquad \textbf{( Eq. 2.2.7.I )}$$

onde **N** é o número de átomos da seqüência, $\mathbf{r_i(A)}$ representa as coordenadas do átomo **i** na estrutura **A**. Pela definição fica evidente que o menor valor dessa distância é zero, enquanto que não existe um limite superior.

Existem duas formas de se calcular o RMSD: localmente ou globalmente. No primeiro caso, RMSD-Local, os átomos equivalentes das duas estruturas, ou dos dois trechos, são superpostos, através de rotação e ou translação das estruturas, como mostrado na **Figura 2.2.7.I: Superposição**. Nesse caso é preciso determinar quais os

átomos serão usados para guiar a superposição. Em seguida é efetuado o cálculo do RMSD. Para o RMSD-Global, o cálculo da distância é feito a partir das posições originais dos átomos das estruturas.

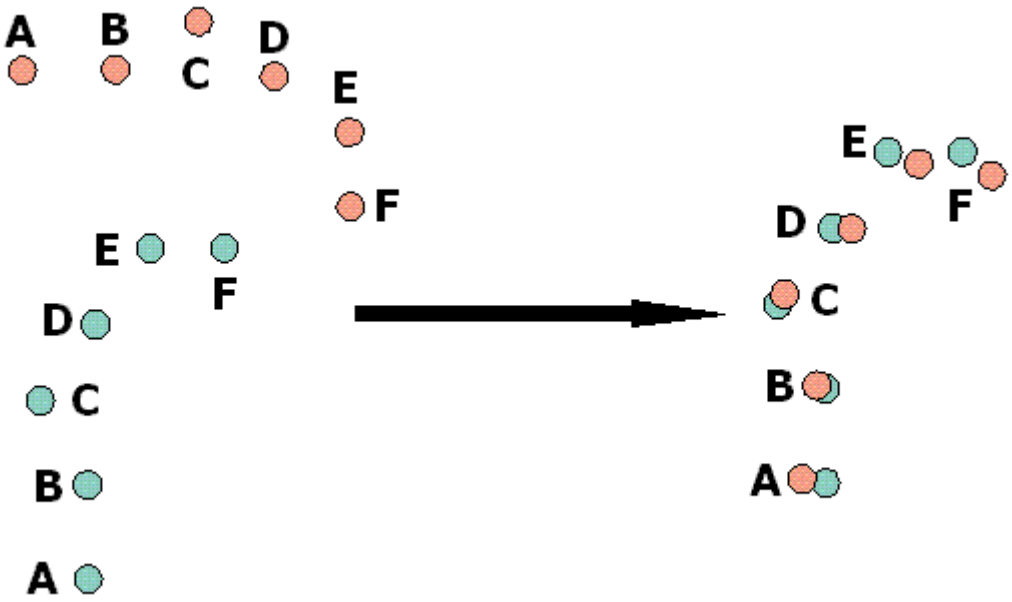



**Figura 2.2.7.I: Superposição –** Superposição dos átomos de duas estruturas

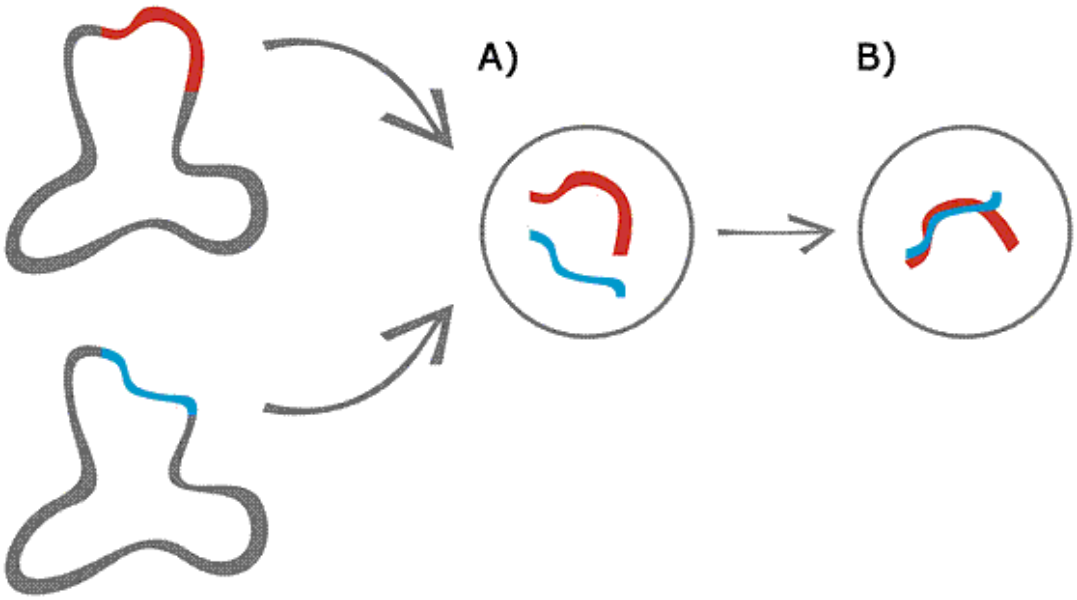



**Figura 2.2.7.I: RMSD –** No lado esquerdo há duas estruturas bastante semelhantes (a região em cinza das duas estruturas é idêntica.). A) Trechos de regiões correspondentes, mostrando que a direção é bastante diferente, gerando um alto valor para o RMSD-Global. B) Após o a superposição das regiões percebe-se uma semelhança entre elas, logo um valor baixo para o RMSD-Local.

O RMSD-Local é usado quando se deseja medir a similaridade interna entre trechos de estruturas, sem interesse nos outros átomos. Por outro lado, o RMSD-Global é principalmente utilizado na avaliação da orientação espacial de trechos de estruturas,

ou seja, considerando a disposição do trecho em relação aos demais átomos. Por exemplo, regiões correspondentes de duas conformações distintas mesmo apresentando um baixo valor no cálculo do RMSD-Local, podem apresentar direções bastante diferentes quando toda a estrutura é considerada, e conseqüentemente um alto valor de RMSD-Global, como acontece na **Figura 2.2.7.II: RMSD.**

### 2.2.8. Alinhamento de seqüências

Trata-se de uma comparação entre duas ou mais seqüências procurando caracteres individuais ou padrões de caracteres que estão em uma mesma ordem nas seqüências. Têm por objetivo determinar o grau de similaridade entre duas ou mais seqüências, ou semelhança entre regiões das seqüências.

De uma forma bastante simplificada, o método insere espaços entre os elementos unitários das seqüências com o objetivo de fazê-las coincidirem o máximo possível. A qualidade de um alinhamento é dada em função do número de espaços inseridos, os elementos pareados coincidentes e não coincidentes.

Existem dois tipos de alinhamento: global e local. O primeiro usa a seqüência por inteiro, aplicando tantos caracteres quantos forem possíveis. Mais usado na busca de alinhamentos ótimos, analisando seqüências com aproximadamente o mesmo tamanho e que apresentem um certo grau de similaridade.

O alinhamento local tem por objetivo comparar pequenos trechos entre as seqüências, formando sub-alinhamentos entre as mesmas. Focado especialmente quando se tem por interesse determinar regiões conservadas que indiquem, pelo menos, uma fraca relação entre as seqüências.

Além de um alinhamento entre seqüências, é possível também determinar a similaridade entre estruturas 3D. Neste caso, além da ordem em que os elementos se apresentam, é também preciso considerar a orientação dos trechos da estrutura.

## Seção 2.3. Aminoácidos

Os aminoácidos são as unidades fundamentais das proteínas e são formados por átomos de Carbono (C), Oxigênio (O), Hidrogênio (H) e Nitrogênio (N), alguns com a presença também de átomos de Enxofre (S). Essas unidades podem ser divididas em 4 partes: a cadeia lateral (R), o grupo carboxila (COOH), que lhe atribui a característica de ácido orgânico, o grupo amina ($NH_2$), constituindo o caráter básico, e o Carbono alfa

(Cα), que se conecta com as 3 partes citadas e com mais um átomo de hidrogênio. Na **Figura 2.3.I: Aminoácido** é apresentada a representação da estrutura bidimensional de um aminoácido.

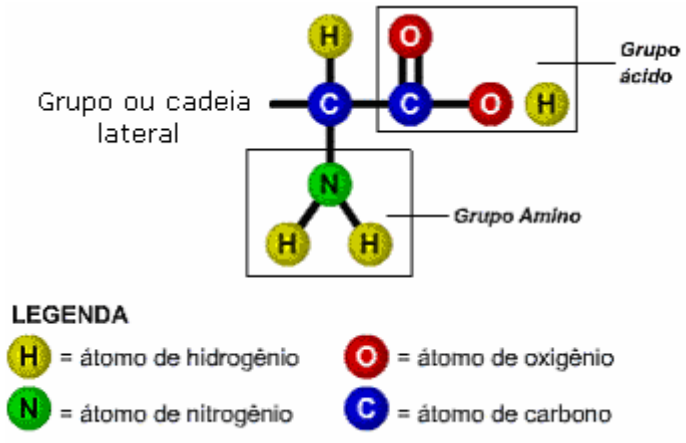



**Figura 2.3.I: Aminoácido** – Representação da estrutura bidimensional de um aminoácido.

O grupo lateral determina os diferentes tipos de aminoácidos encontrados na natureza, podendo ser constituído por um único átomo de H, como ocorre com a Glicina, ou por uma combinação de átomos de C, N, O, H e S, sendo este último presente apenas na Metionina e Cisteína. As estruturas dos 20 aminoácidos mais comuns estão presentes na **Figura 2.3.II: 20 Aminoácidos**. Para os átomos de C, N, O e S presentes na cadeia lateral, chamados de átomos pesados, há uma nomenclatura especial: o primeiro átomo do grupo lateral, na maioria dos casos um C, recebe a extensão β (beta), tornando-se o Cβ. Os demais recebem como sufixo outras letras gregas seguindo a ordem alfabética: γ (gama), δ (delta), ε (epsilon), ζ (zeta), η (eta). Na **Figura 2.3.III: Nomenclatura** há dois exemplos, Lisina e Tirosina, de como se procede a nomenclatura dos átomos da cadeia lateral.

Uma das principais propriedades apresentada pelos aminoácidos é determinada a partir de sua estrutura tridimensional ou estereoquímica. Dados dois aminoácidos, sendo um a imagem especular do outro, eles desviam um feixe de luz polarizada incidente neles em direções opostas. Tais aminoácidos quirais são então classificados como enantiômeros entre si.

No caso da molécula de aminoácido, ele e sua imagem especular não se superpõem, sendo também denominados de estereoisômeros. Fato semelhante ocorre

entre a mão direita e esquerda de um ser humano. Elas representam a imagem especular uma da outra, mas não se sobrepõem.

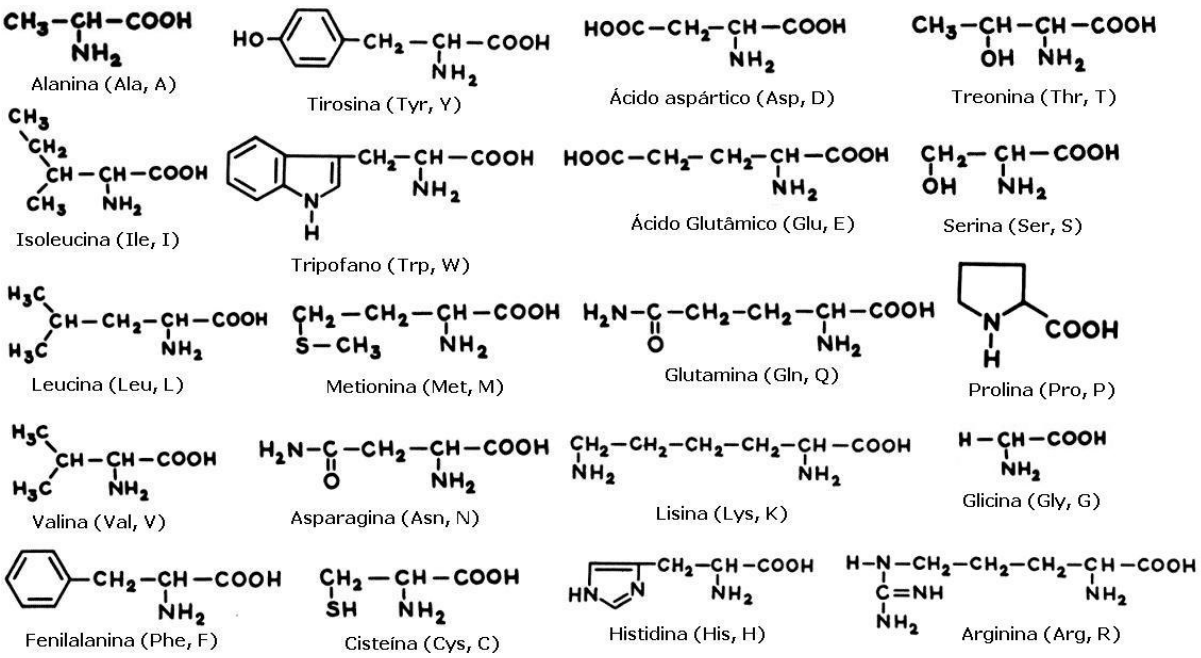



**Figura 2.3.II: 20 Aminoácidos** - Os 20 diferentes aminoácidos normalmente encontrados nas proteínas, com suas estruturas bidimensionais, nomes e abreviações.

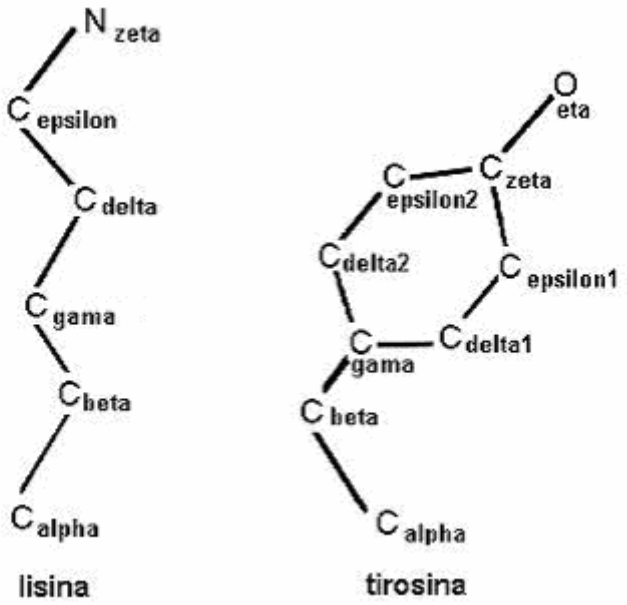



**Figura 2.3.III: Nomenclatura** - Exemplos da convenção da nomenclatura dos átomos dos grupos laterais dos aminoácidos Lisina e a Tirosina.

A propriedade da quiralidade das moléculas de aminoácidos só ocorre quando nelas encontra-se um centro de quiralidade, ou seja, um átomo ligado a quatro grupos diferentes **(Figura 2.3.IV: Moléculas Especulares)**. O centro de quiralidade dos

aminoácidos apresenta-se no Cα, com exceção da Glicina, que apresenta dois átomos de H conectados ao Cα, sendo então aquiral **(Figura 2.3.V: Glicina)**.

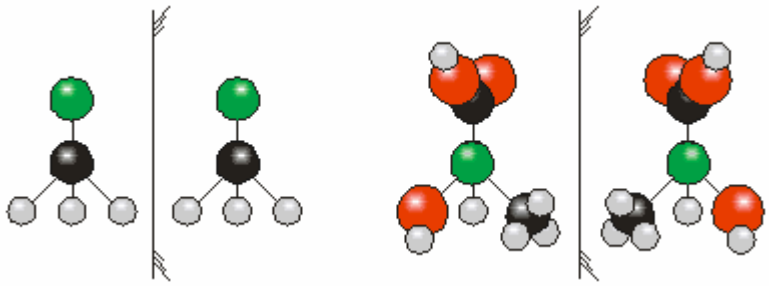


**Figura 2.3.IV: Moléculas Especulares** - A primeira molécula é aquiral, por não apresentar um átomo ligado a quatro grupos distintos. A segunda é quiral, sendo o átomo na cor verde o centro de quiralidade.

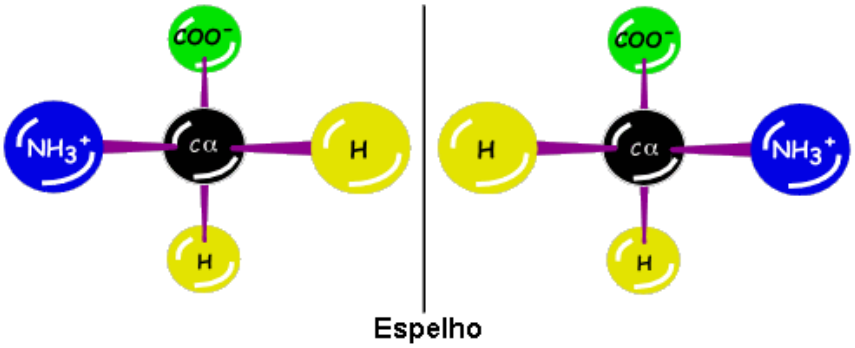



**Figura 2.3.V: Glicina** – A molécula da Glicina e sua imagem especular, caso em que o aminoácido não apresenta quiralidade. As imagens se superpõem, pois apesar de estar conectado a quatro grupos, dois deles são idênticos. Em ambas as imagens o grupo $NH_3^+$ e o H (esfera maior) estão acima do plano do papel, o Cα está no plano do papel e o H (esfera menor) e o grupo carboxila abaixo do plano do papel.

As duas possibilidades para um aminoácido quiral são nomeadas *Laevus* (L-) e *Dexter* (D-), denominações originadas do latim que significam esquerda e direita, respectivamente. No primeiro, o grupo hidroxila (OH) se encontra à esquerda da molécula, enquanto na configuração D, à direita. Os aminoácidos das proteínas são todos do tipo L. A **Figura 2.3.VI: Alanina L-D** apresenta as duas possíveis configurações da Alanina, onde os triângulos preenchidos representam um plano acima do papel e o tracejado um plano abaixo.

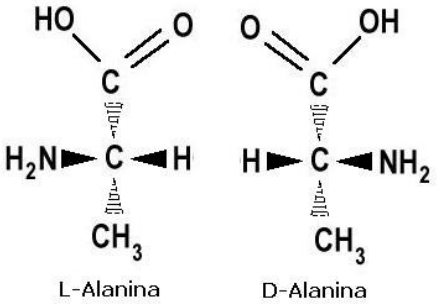



**Figura 2.3.VI: Alanina L-D** - A estrutura espacial da molécula da Alanina com as configurações L e D.

## Seção 2.4. Proteínas

As proteínas são as macromoléculas mais abundantes na célula **[LNC02]**, além de serem as mais diversificadas quanto à forma e à função, exercendo atividade tanto em termos estruturais (regula e organiza a estrutura arquitetônica das células e dos tecidos, como o citoesqueleto; proporciona uma matriz para o osso, tecido conjuntivo e músculos) quanto funcionais (os hormônios, os catalisadores, as moléculas receptoras e de transporte; no mecanismo de defesa), sendo importantes para quase todos os processos biológicos.

### 2.4.1. Formação da Proteína

As proteínas são constituídas por cadeias de aminoácidos, normalmente mais de 100, unidos por uma ligação covalente chamada de ligação peptídica onde o grupo amina de um aminoácido interage com a carboxila do outro, liberando uma molécula de água **[KC99] (Figura 2.4.1.I: Ligação Peptídica)**. Os compostos resultantes deste tipo de ligação são os peptídeos, motivo pelo qual as proteínas são também mencionadas como cadeias polipeptídicas.

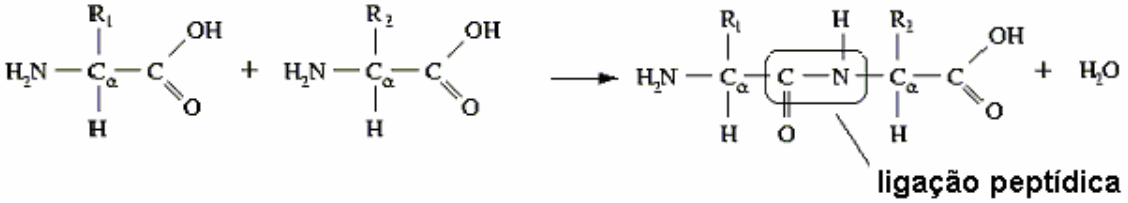



**Figura 2.4.1.I: Ligação Peptídica -** A ligação peptídica se forma entre o átomo de carbono (C) do grupo carboxila e o átomo de nitrogênio (N) do grupo amina. A água ($H_2O$) se forma a partir do -OH do grupo carboxila de um dos aminoácidos e de um átomo de hidrogênio do grupo -$NH_2$ do outro aminoácido.

As ligações peptídicas apresentam duas estruturas de ressonância: $^-$O-C=$NH^+$ e O=C-NH. Essas estruturas representam a mesma ligação, mas diferem pelo posicionamento dos elétrons **(Figura 2.4.1.II: Ressonância)**. Isso faz com que este tipo de ligação tenha um caráter parcial de ligação dupla, sendo mais forte que uma ligação simples comum, e como conseqüência, o grupo peptídico (O=C-N-H) responsável pela união dos dois aminoácidos é planar **(Figura 2.4.1.III: Ligação Planar)**, adotando conformação *syn* (0º) ou *anti* (180º), sendo esta última a mais comum, por apresentar uma menor probabilidade de repulsão entre átomos **(Figura 2.4.1.IV: Anti-Syn)**.

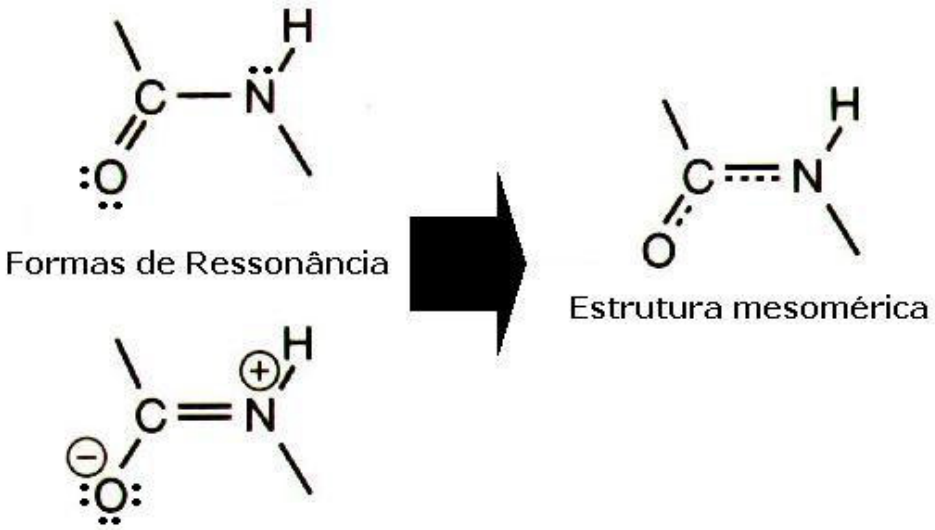



**Figura 2.4.1.II: Ressonância -** Do lado esquerdo têm-se as duas estruturas de ressonância possíveis para a ligação peptídica. Na primeira, O=C-NH, um único par de elétrons é compartilhado entre o C e o N; em $^{-}$O-C=NH$^{+}$ há uma ligação dupla entre eles. Do lado direito está representada uma estrutura intermediária, ou de equilíbrio mesomérico, dos modelos de ressonância.

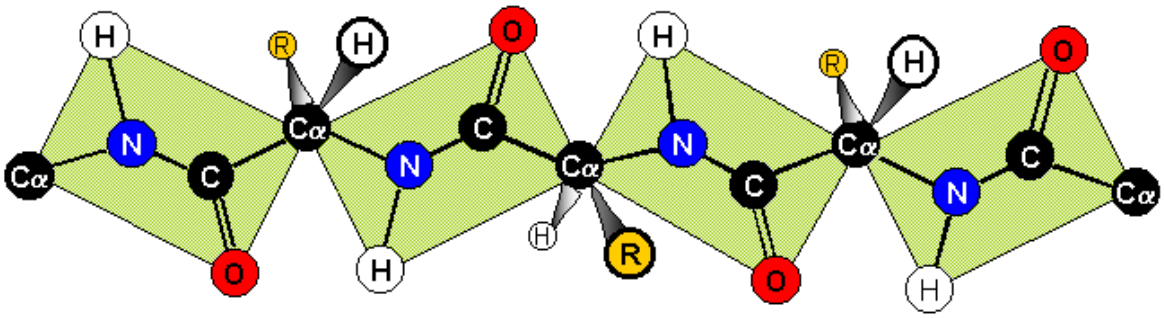


**Figura 2.4.1.III: Ligação Planar –** Seqüência polipeptídica com quatro ligações peptídicas, onde em cada um dos planos encontra-se o grupo peptídeo que é responsável pela união de dois aminoácidos.

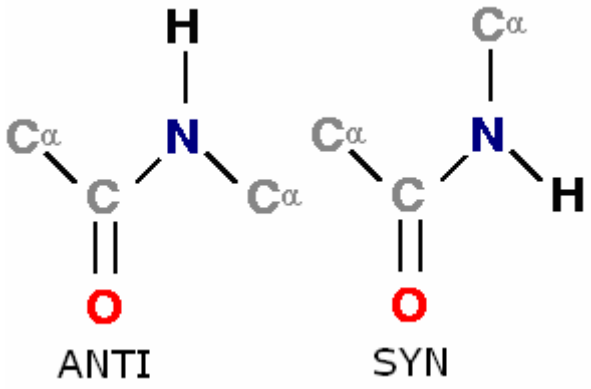



**Figura 2.4.1.IV: Anti-Syn -** As duas possíveis orientações para as ligações peptídicas. Na conformação *anti* os Cα se encontram em posições opostas no plano em relação a ligação C-N. Na *syn*, eles se apresentam do mesmo lado.

## 2.4.2. Do DNA à proteína

Nas células dos seres vivos, as proteínas são sintetizadas a partir do processo de transcrição-tradução das seqüências de nucleotídeos presentes nos genes. Inicialmente, o DNA sofre a transcrição, o que produz moléculas de RNA-mensageiro (RNAm) que, diferente do DNA, apresentam uma única cadeia de nucleotídeos, moldados a partir de um único filamento do DNA. Neste processo, cada nucleotídeo do DNA corresponde a um nucleotídeo complementar no RNAm, seguindo a correspondência citosina-guanina e adenina-uracila **(Figura 2.4.2.I: Transcrição-DNA)**.

Em seguida, o RNAm sofre o processo de tradução, onde o código da seqüência de nucleotídeos é lido em trincas de bases nitrogenadas, denominadas de códons. Cada um desses tercetos corresponde obrigatoriamente a um aminoácido da proteína a ser formada ou a um códon que indica a finalização da tradução para uma proteína **(Figura 2.4.2.II: Tradução-RNAm) (Tabela 2.4.2.I: Códon-Nucleotídeo)**.

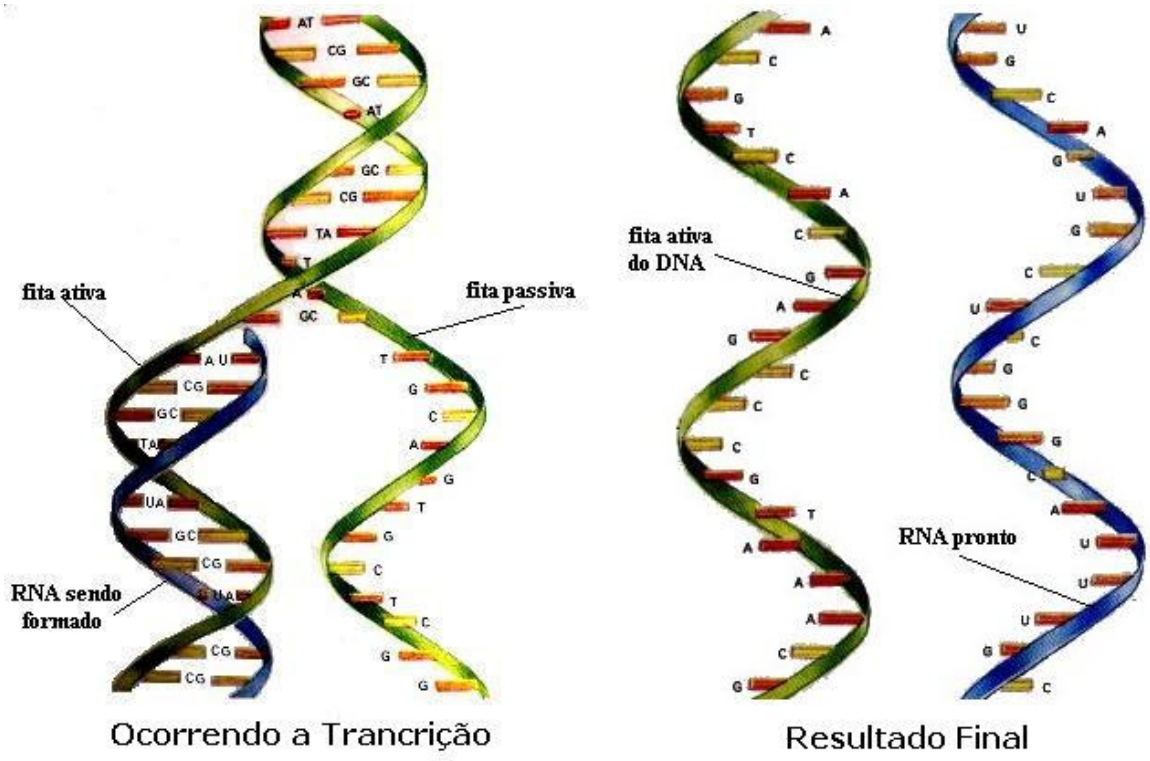



**Figura 2.4.2.I: Transcrição-DNA -** Na imagem da esquerda, as fitas de DNA estão se afastando e o RNAm está sendo formado em apenas uma delas, a chamada fita ativa, pelo processo de transcrição do código genético. Do lado direito, o resultado final é apresentado, quando a fita de RNAm se destaca da fita de DNA, que torna a parear com a fita passiva.

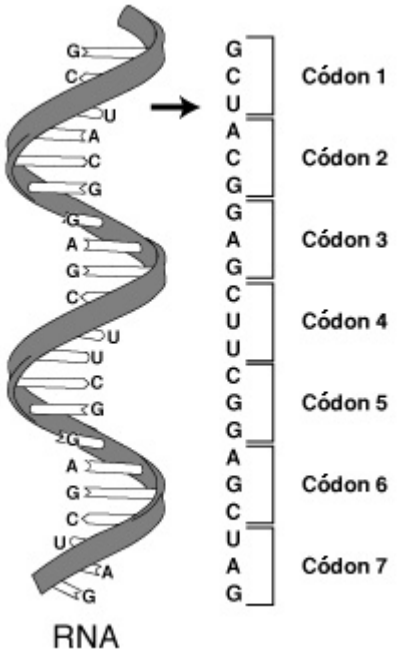



**Figura 2.4.2.II: Tradução-RNAm -** A imagem apresenta uma fita de RNA e a organização das trincas em códons.

| primeira posição | segunda posição: U | | | | segunda posição: C | | | | segunda posição: A | | | | segunda posição: G | | | | terceira posição |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | códon | Aminoácido | | | códon | Aminoácido | | | códon | Aminoácido | | | códon | Aminoácido | | | |
| U | UUU | phe | | | UCU | ser | | | UAU | tyr | | | UGU | cys | | | U |
| | UUC | | | | UCC | | | | UAC | | | | UGC | | | | C |
| | UUA | leu | | | UCA | | | | UAA | Terminal | | | UGA | Terminal | | | A |
| | UUG | | | | UCG | | | | UAG | Terminal | | | UGG | trp | | | G |
| C | CUU | leu | | | CCU | pro | | | CAU | his | | | CGU | arg | | | U |
| | CUC | | | | CCC | | | | CAC | | | | CGC | | | | C |
| | CUA | | | | CCA | | | | CAA | gln | | | CGA | | | | A |
| | CUG | | | | CCG | | | | CAG | | | | CGG | | | | G |
| A | AUU | ile | | | ACU | thr | | | AAU | asn | | | AGU | ser | | | U |
| | AUC | | | | ACC | | | | AAC | | | | AGC | | | | C |
| | AUA | | | | ACA | | | | AAA | lys | | | AGA | arg | | | A |
| | AUG | met | | | ACG | | | | AAG | | | | AGG | | | | G |
| G | GUU | val | | | GCU | ala | | | GAU | asp | | | GGU | gly | | | U |
| | GUC | | | | GCC | | | | GAC | | | | GGC | | | | C |
| | GUA | | | | GCA | | | | GAA | glu | | | GGA | | | | A |
| | GUG | | | | GCG | | | | GAG | | | | GGG | | | | G |

**Tabela 2.4.2.I: Códon-Nucleotídeo –** A tabela determina as 64 possíveis combinações com três bases nitrogenadas, formando os códons, e os aminoácidos. Três códons (UGA, UAG, UAA) indicam o final da cadeia polipeptídica.

## 2.4.3. Estruturas das Proteínas

Como conseqüência da complexidade das proteínas, definiu-se quatro níveis de estruturas para seu estudo **[Mou99]**:

- **Estrutura primária:** é a representação estrutural mais simples, formada pela seqüência dos aminoácidos, representando as ligações peptídicas. Neste tipo de estrutura, o aminoácido N-terminal (com o grupo amina livre) é representado como o primeiro à esquerda. Informações sobre os posicionamentos das pontes dissulfídicas (ligação entre os átomos de enxofre) podem ser combinadas com a estrutura primária, fornecendo a *estrutura covalente completa* da proteína **(Figura 2.4.3.I: Estrutura Covalente)**;

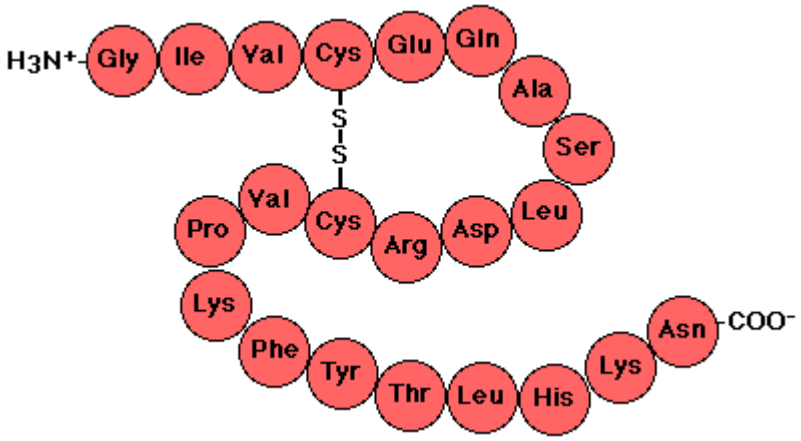



**Figura 2.4.3.I: Estrutura Covalente -** Representação da estrutura primária de uma proteína iniciando, do lado esquerdo, com um aminoácido N-terminal (com o grupo $H_3N^+$ livre), e finalizando com um C-terminal (com o grupo $COO^-$ livre). O 4º e o 11º aminoácidos, ambos a Cisteína, apresentam uma ligação dissulfídica.

- **Estrutura secundária**: representa interações intermoleculares entre os elementos dos aminoácidos que formam padrões regulares e repetitivos, ocorrendo localmente ao longo do esqueleto da proteína (cadeia principal). As ligações não-covalentes são as responsáveis pela estabilização dessas conformações, sendo as ligações de hidrogênio as mais determinantes. Os grupos laterais não são levados em consideração. As formas mais comuns de estrutura secundária são:
    - α-hélice: estrutura secundária mais presente nas proteínas, formada por uma única cadeia polipeptídica e com conformação helicoidal. Apresenta uma média de dez aminoácidos, mas pode variar de cinco até quarenta. São mais presentes nas superfícies das proteínas, com aminoácidos hidrofóbicos voltados para a face interna. Cada grupo C=O do aminoácido na posição *n* está ligado, através de ligação de hidrogênio, com o grupo NH do aminoácido localizado na posição *n+4*. As cadeias

laterais situam-se fora da hélice **(Figura 2.4.3.II: Estrutura Secundária A)**;

- folha-β: segmento da cadeia da proteína formado por cinco a dez aminoácidos, que interage através de ligações de hidrogênio com alguma outra seqüência de aminoácidos presente mais adiante na proteína. Estas seqüências podem ser adjacentes, com apenas alguns aminoácidos separando-as, ou distantes, com outra estrutura entre elas. Podem estar dispostas em três direções: paralelas, quando os aminoácidos de duas cadeias consecutivas estão conectados numa mesma direção (cadeias alinhadas em termos de suas extremidades N-terminais e C-terminais); antiparalelas, quando duas cadeias consecutivas apresentam os aminoácidos conectados em direções opostas; ou numa mistura das duas, sendo o padrão de disposição das ligações de hidrogênio diferentes em cada caso **(Figura 2.4.3.II: Estrutura Secundária B)**;
- alça-β: região de estrutura secundária, formada por quatro aminoácidos, que se dobra sobre ela mesma, sendo responsável pela reversão da direção da cadeia de proteína **(Figura 2.4.3.III: Alça-β)**;
- *coil*: região de estrutura secundária que não é hélice, nem folha nem alça.

- **Estrutura terciária**: é a estrutura espacial da cadeia polipeptídica, incluindo todos os átomos da molécula, constituída pelas estruturas secundárias, cadeias laterais e qualquer outro grupo prostético (átomos não pertencentes aos aminoácidos, mas presentes em algumas proteínas), sendo um resultado de todas as forças estabilizadoras, como interações hidrofóbicas, atração eletrostática, ligações de hidrogênio e pontes dissulfeto. Neste tipo de estrutura, definem-se dois novos elementos:
  - laços ou *loops*: regiões presentes entre estruturas secundárias e localizadas na superfície da proteína, apresentando diversas configurações tridimensionais e comprimentos, e que pode variar de dois a trinta aminoácidos. São as regiões da proteína mais suscetíveis aos processos de substituição, remoção e inserção de aminoácidos, pois não constituem partes das estruturas secundárias e não têm uma relação forte com o ambiente em que a proteína está imersa, como o apresentado pela região central **(Figura 2.4.3.IV: Estrutura Terciária)**.

- dobras ou *folds*: combinações de várias estruturas secundárias conectadas por laços, formando configurações que se repetem em diversas proteínas. Podem ser chamados de *motifs* ou estruturas supersecundárias **(Figura 2.4.3.V: Dobras)**.

A grande importância da estrutura terciária está no fato da funcionalidade de uma proteína estar relacionada com sua conformação tridimensional.

- **Estrutura quaternária**: é o arranjo espacial da estrutura da proteína no caso em que há mais de uma cadeia polipeptídica independente, podendo estar presente entre 2 a 12 cadeias, idênticas ou diferentes **(Figura 2.4.3.VI: Estrutura Quaternária)**. As cadeias são mantidas pelas mesmas forças que determinam as estruturas secundárias e terciárias.

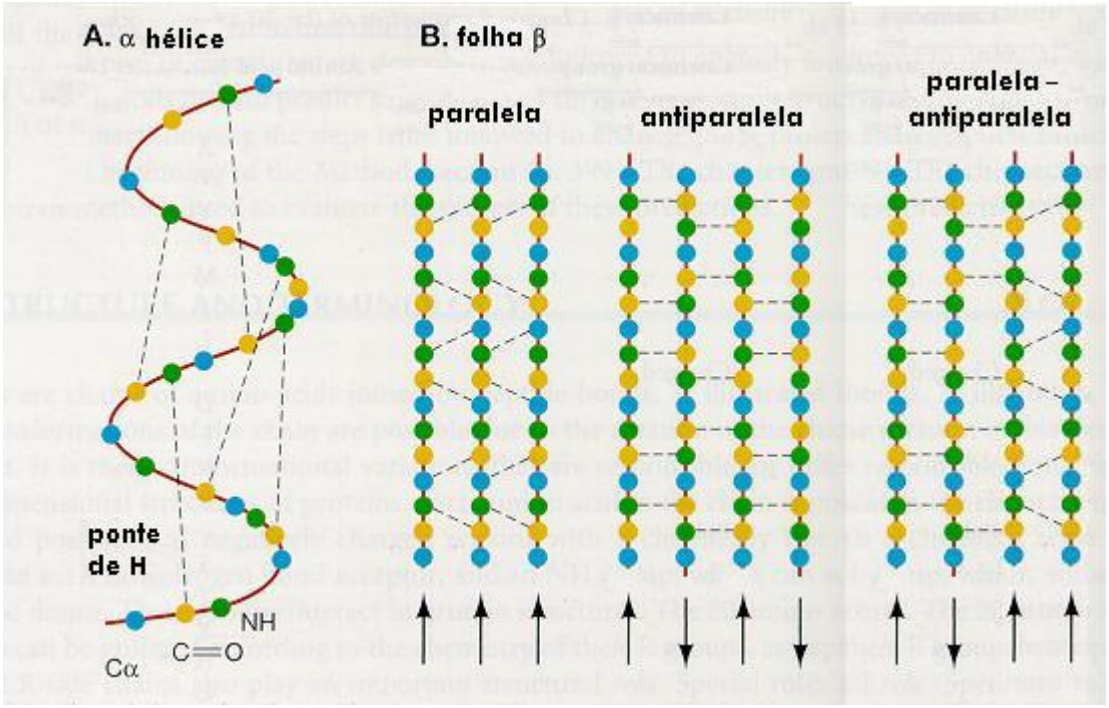



**Figura 2.4.3.II: Estrutura Secundária –** Representação dos dois principais tipos de estruturas secundárias: α-hélice e folha-β, com os Cα em azul, C=O em amarelo e o NH em verde. A) Reprodução da estrutura secundária α-hélice, onde cada grupo C=O do aminoácido na posição *n* está conectado, por ligação de hidrogênio com o aminoácido na posição *n+4*. B) Reprodução das estruturas secundárias folha-β paralela, antiparalela e paralela-antiparalela. No primeiro caso, os aminoácidos de duas cadeias consecutivas estão com o grupo NH na parte inferior do aminoácido ao qual pertencem; no segundo caso, a posição do grupo NH para duas cadeias consecutivas se encontra, em uma delas, na parte inferior do aminoácido ao qual pertence, e na outra, numa posição superior. As conformações antiparalelas são mais estáveis, devido ao melhor alinhamento das ligações de hidrogênio. C) Terceiro caso em que na folha-β há a presença de cadeias paralelas e antiparalelas.

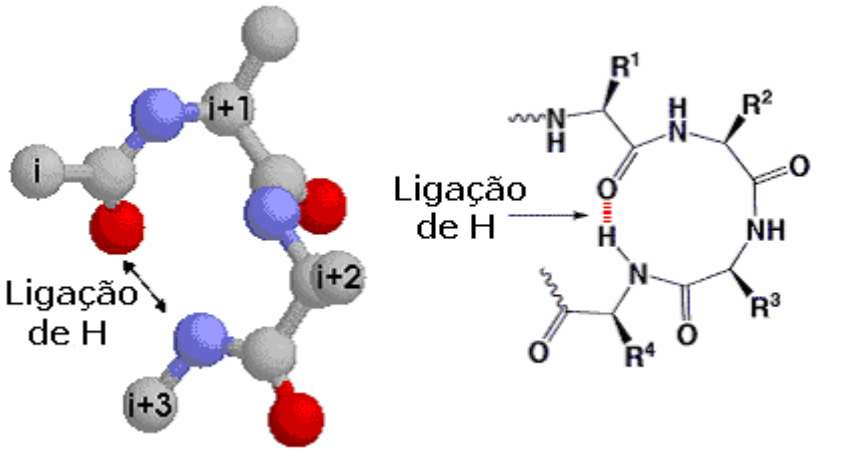



**Figura 2.4.3.III: Alça-β –** Representação da estrutura alça-β. Na figura da esquerda, uma das sete conformações tridimensionais possíveis: em vermelho os Oxigênios, em lilás os Nitrogênios e em cinza os Carbonos. Na imagem da direita, a representação bidimensional genérica desse tipo de estrutura.

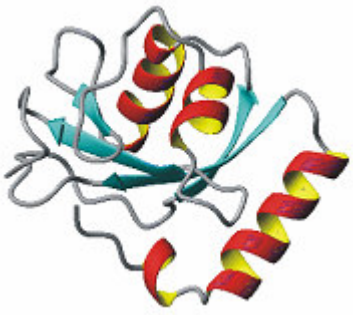


**Figura 2.4.3.IV: Estrutura Terciária –** Representação da estrutura terciária de uma proteína, com destaque para as estruturas secundárias, como α-hélice (vermelho e amarelo) e as folhas-β (azul); os laços estão em cinza.

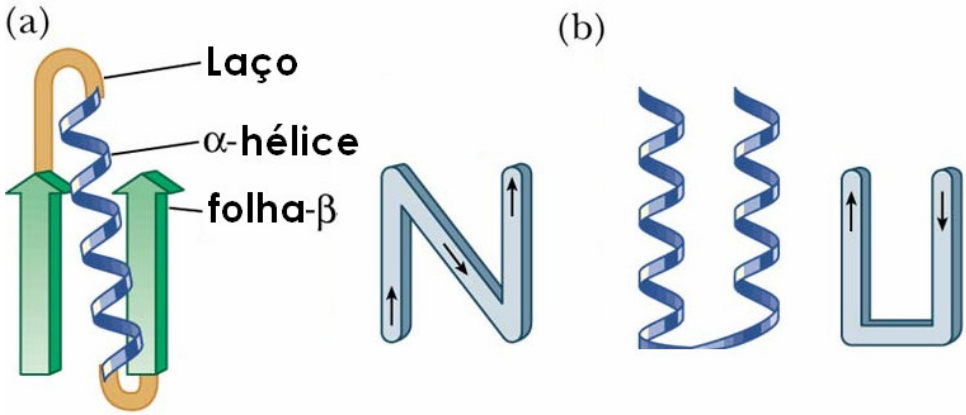



**Figura 2.4.3.V: Dobras –** Representação de duas possíveis dobras pertencentes a uma estrutura terciária. A) Dobra folhas-β - α-hélice - folhas-β e B) Dobra α-hélice - α-hélice.

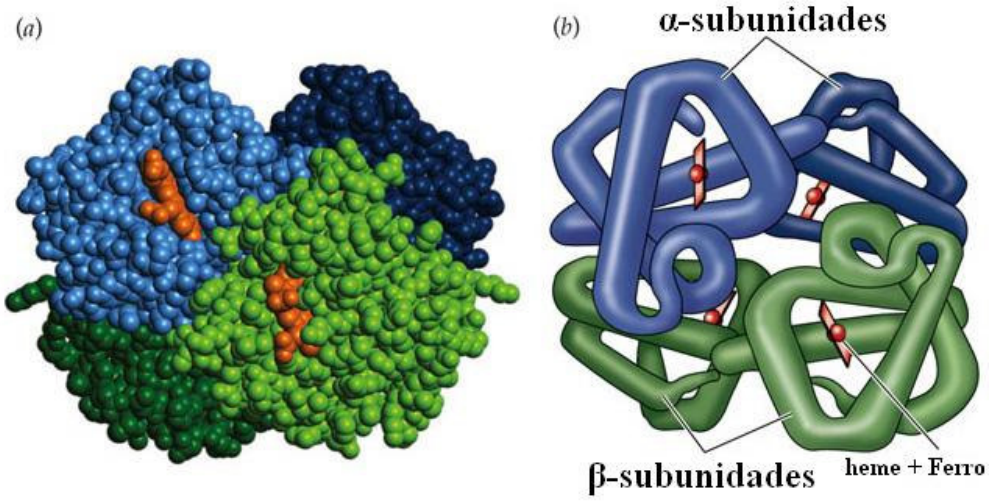



**Figura 2.4.3.VI: Estrutura Quaternária –** Representação da hemoglobina do sangue humano, composta por quatro cadeias protéicas, quatro heme (pigmento que dá cor ao sangue) e quatro átomos de Ferro. As subunidades α (cadeias em azul) são semelhantes entre si, assim como as subunidades β (cadeias em verde).

## 2.4.4. Ângulos diedros das proteínas

Este tópico apresenta a definição dos *ângulos diedros* ou *ângulos de torção* presentes na conformação tridimensional das proteínas, originários da intersecção entre dois planos determinados por aminoácidos consecutivos.

Para a cadeia principal de uma proteína, três são os tipos de ângulos diedros presentes: Φ (phi), ψ (psi) e ω (ômega). O primeiro é formado a partir da intersecção dos planos determinados pelos átomos $C_{n-1}$ - $N_n$ - $C\alpha_n$ e $N_n$ - $C\alpha_n$ - $C_n$ , onde ***n*** é a posição do aminoácido na seqüência da proteína e $C_{n-1}$ e $C_n$ referem-se aos carbonos do grupo carbonila **(Figura 2.4.4.I: Φ)**.

O ψ é determinado pelos átomos de $N_n$ - $C\alpha_n$ - $C_n$ - $N_{n+1}$ **(Figura 2.4.4.II: ψ)**, enquanto que o ω é definido a partir dos átomos de $C\alpha_n$ - $C_n$ - $N_{n+1}$ - $C\alpha_{n+1}$ **(Figura 2.4.4.III: ω)**.

Para as cadeias laterais, os ângulos diedros são representados pela letra grega χ (chi), e as coordenadas de quatro átomos também são necessárias. A **Tabela 2.4.4.I: Ângulos Diedros das Cadeias Laterais** indica os possíveis ângulos diedros para cada um dos grupos laterais.

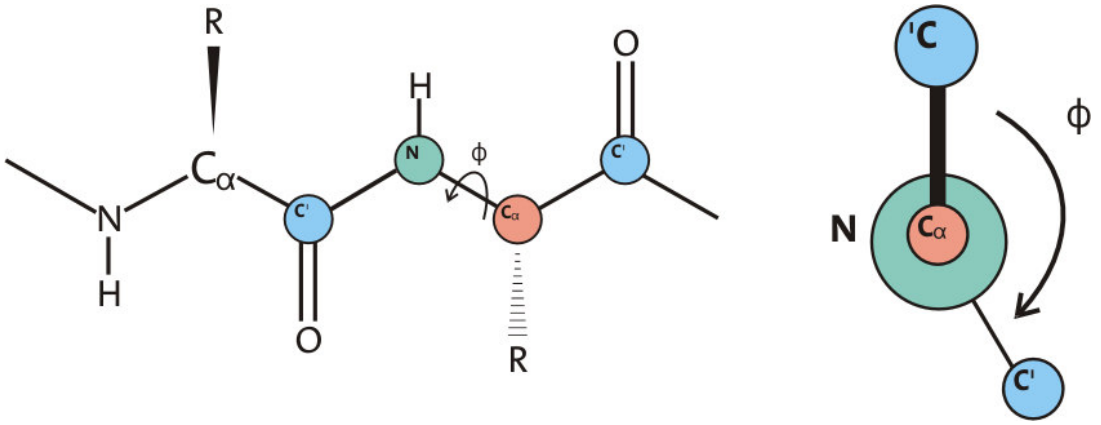



**Figura 2.4.4.I: Φ –** Representação do ângulo diedro Φ. Na imagem à esquerda, uma perspectiva bidimensional. Na segunda figura, uma visualização ao longo do eixo da ligação entre o Cα e o N .

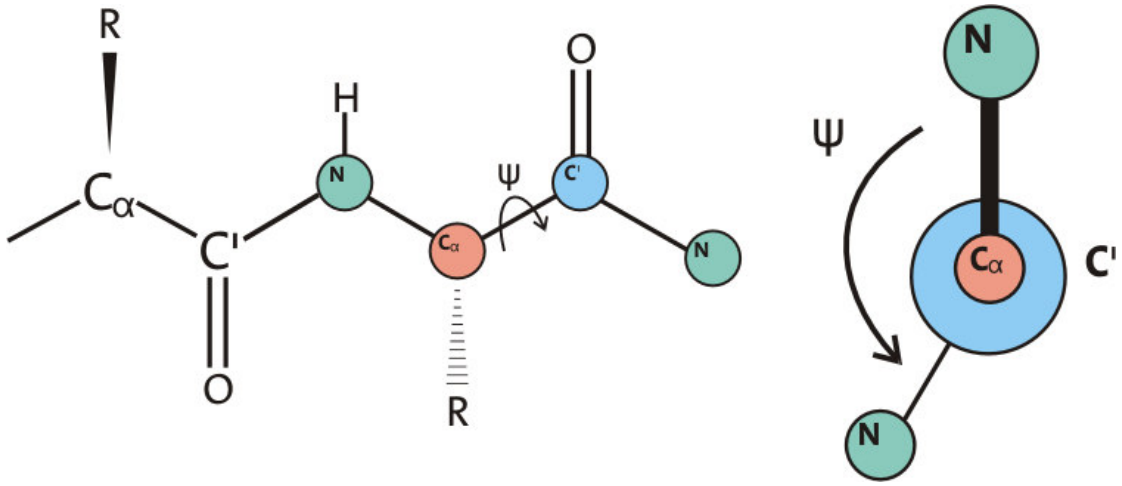



**Figura 2.4.4.II: ψ –** Representação do ângulo diedro ψ. Na imagem à esquerda, uma perspectiva bidimensional. Na segunda figura, uma visualização ao longo do eixo da ligação entre o Cα e o C do grupo carboxila.

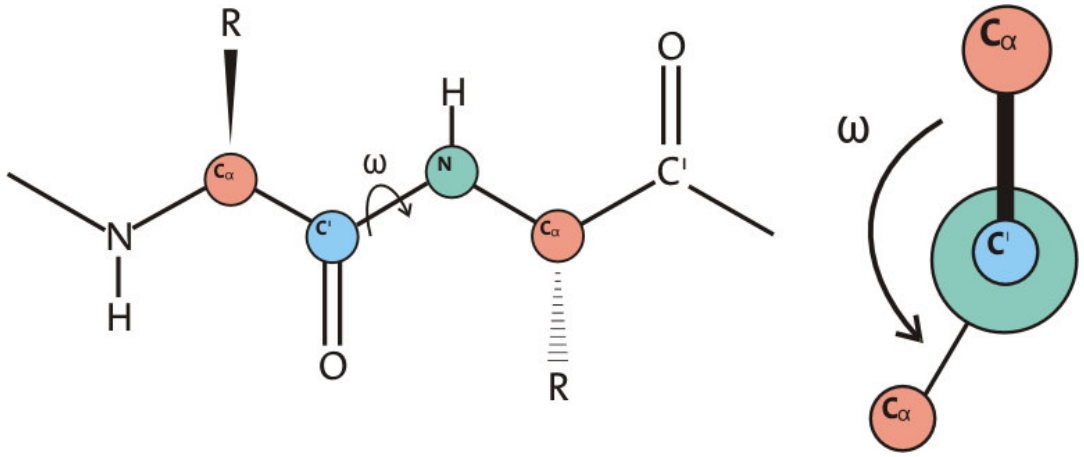



**Figura 2.4.4.III: ω –** Representação do ângulo diedro ω. Na imagem à esquerda, uma perspectiva bidimensional. Na segunda figura, uma visualização ao longo do eixo da ligação entre o C da carboxila e o N.

| $\chi_1$ | | $\chi_2$ | |
|---|---|---|---|
| Aminoácido | Átomos Usados | Aminoácido | Átomos Usados |
| ARG | N-CA-CB-CG | ARG | CA-CB-CG-CD |
| ASN | N-CA-CB-CG | ASN | CA-CB-CG-OD1 |
| ASP | N-CA-CB-CG | ASP | CA-CB-CG-OD1 |
| CYS | N-CA-CB-SG | GLN | CA-CB-CG-CD |
| GLN | N-CA-CB-CG | GLU | CA-CB-CG-CD |
| GLU | N-CA-CB-CG | HIS | CA-CB-CG-ND1 |
| HIS | N-CA-CB-CG | ILE | CA-CB-CG1-CD |
| ILE | N-CA-CB-CG1 | LEU | CA-CB-CG-CD1 |
| LEU | N-CA-CB-CG | LYS | CA-CB-CG-CD |
| LYS | N-CA-CB-CG | MET | CA-CB-CG-SD |
| MET | N-CA-CB-CG | PHE | CA-CB-CG-CD1 |
| PHE | N-CA-CB-CG | PRO | CA-CB-CG-CD |
| PRO | N-CA-CB-CG | TRP | CA-CB-CG-CD1 |
| SER | N-CA-CB-OG | TYR | CA-CB-CG-CD1 |
| THR | N-CA-CB-OG1 | $\chi_3$ | |
| TRP | N-CA-CB-CG | ARG | CB-CG-CD-NE |
| TYR | N-CA-CB-CG | GLN | CB-CG-CD-OE1 |
| VAL | N-CA-CB-CG1 | GLU | CB-CG-CD-OE1 |
| $\chi_4$ | | LYS | CB-CG-CD-CE |
| ARG | CG-CD-NE-CZ | MET | CB-CG-SD-CE |
| LYS | CG-CD-CE-NZ | $\chi_5$ | |
| | | ARG | CD-NE-CZ-NH1 |

**Tabela 2.4.4.I: Ângulos Diedros das Cadeias Laterais** – Apresenta os cinco diferentes tipos de ângulos diedros das cadeias laterais. Uma das colunas é preenchida com o tipo de aminoácido que contém o diedro, a outra contém os átomos que determinam os planos de intersecção. Nesta tabela, A = α; B = β; G = γ, D = δ, E = ε , Z = ζ e E = η;

## Seção 2.5. Banco de Proteínas

Como já foi dito anteriormente, umas das preocupações iniciais da Bioinformática era a construção de formas de armazenamento dos principais dados utilizados na área: as seqüências de nucleotídeos e aminoácidos. As estruturas dos bancos de dados foram aplicadas justamente com o objetivo de dar suporte à resolução dos problemas de armazenamento e acesso aos dados obtidos. As principais bases utilizadas nos estudos da Biologia Computacional são:

- *GenBank*: base de dados do *National Institutes of Health* **[NIH]**, construída pelo *National Center for Biotechnology Information* (*NCBI*), e que compreende seqüências de nucleotídeos e estruturas primárias de proteínas. Juntamente com outros dois bancos, *DNA Data Bank of Japan* **[DDBJ]** e *European Molecular Biology Laboratory* **[EMBL]**, forma o *International Nucleotide Sequence*

*Database Collaboration*, trocando informações diariamente, fazendo com que contenham as mesmas seqüências. Cada entrada do *GenBank* contém a descrição da seqüência, o nome científico e taxonômico do organismo de origem, entre outras informações **(Apêndice GenBank)**.

- *SWISS-PROT* **[Swiss-Prot]**: mantido pelo *Swiss Institute of Bioinformatics* **[SIB]** e pelo *The European Bioinformatics Institute* **[EBI]**, diferencia-se dos bancos acima citados por apresentar, além das informações centrais disponibilizadas por eles, um conteúdo relacionado à anotação da proteína (função, estrutura secundária, similaridade com outras proteínas, entre outras), sendo, portanto, considerado um banco de dados de estruturas secundárias.
- *Protein Sequence Database* (PSD): originado do Atlas de Seqüência e Estruturas de Proteínas criado por Margaret Dayhoff **[Day72]** e mantido pelo *Protein Information Resource* **[PIR]**, é outro banco de estruturas secundárias, apresentando dados sobre a funcionalidade e classificação das proteínas.
- *Protein DataBank* **[PDB]**: mantido atualmente pela *Research Collaboratory for Structural Bioinformatics* (*RCSB*), disponibiliza as estruturas tridimensionais das proteínas modeladas e publicadas, com informações relacionadas às coordenadas cartesianas dos átomos, à qualidade do modelo, ao organismo de origem, entre outras **(Apêndice PDB)**.

## Seção 2.6. Predição de Estruturas de Proteínas

As seqüências das proteínas podem ser obtidas a partir do processo de translação do DNA complementar, o cDNA, o qual é obtido a partir do processo de transcrição reversa do RNAm. Infelizmente, este tipo de seqüência está pouco presente nos bancos de dados.

Assim, a obtenção da maioria das seqüências de proteínas é alcançada considerando dois casos: os organismos com poucos *introns*, regiões do genoma que não sofrem a transcrição, com a obtenção das estruturas primárias das proteínas sendo feita a partir da transcrição de DNA genômico; e organismos com muitos *introns*, onde métodos para a identificação dos genes e, conseqüentemente, dos trechos que produzem proteínas, precisam ser aplicados. Entretanto, esses mecanismos são rápidos e acessíveis, dado que a maioria deles já existe desde a década de 70, gerando um grande número de seqüências de proteínas.

Com o armazenamento de grande quantidade de dados, outros estudos do conjunto de proteínas dos organismos foram considerados pela Biologia Computacional. Um dos principais focos é a compreensão do relacionamento existente entre as seqüências de aminoácidos e as estruturas tridimensionais das proteínas, pois é a partir das conformações obtidas que:

- Planejam-se experimentos;
- Associam-se funcionalidades a uma proteína;
- Formulam-se novas drogas;
- Descobrem-se estruturas de outras proteínas.

Há mais de 30 anos, busca-se uma solução rápida e precisa para o problema de predição das conformações tridimensionais das proteínas a partir das estruturas primárias. Para se ter uma idéia da dificuldade envolvida nesse problema, a cada seis seqüências do banco de dados de proteínas *Swiss-Prot* (170.140 seqüências submetidas até 15.02.05) apenas uma tem sua estrutura tridimensional conhecida no *PDB* (29.636 estruturas submetidas até 15.02.05). E os empecilhos variam desde a lentidão e a baixa qualidade das respostas à existência de ferramentas criadas para resolver apenas algumas situações específicas.

As várias técnicas de predição são classificadas de acordo com a informação predominante usada nos cálculos **[SB93]**:

- Métodos experimentais: incluem cristalografia de raios-X e técnicas de *RMN* - ressonância magnética nuclear.
- Aplicações teóricas: dividem-se em métodos físicos, os quais se baseiam na interação entre átomos, incluindo dinâmica molecular e minimização da energia; e em métodos empíricos ou baseados em conhecimento prévio, que dependem de estruturas de proteínas já determinadas.

A cristalografia de raios-X consiste em formar um cristal da proteína e submetê-lo a exposição de raios-X para estudar seu padrão de difração. A dificuldade desta técnica é a própria formação dos cristais, que pode durar dias. Uma das soluções para acelerar esse processo de formação é a construção deles num ambiente de gravidade nula, ou seja, no espaço. O modelo originário deste processo é estático, por apresentar a conformação da proteína no cristal.

A ressonância magnética nuclear funciona submetendo a proteína a um campo magnético, fazendo com que os spins dos núcleos dos átomos interajam com o campo, havendo uma separação (desdobramento) dos seus níveis energéticos proporcional ao campo. Em seguida, ondas de rádio são emitidas por um transmissor posicionado perto da proteína. Cada vez que os núcleos são atingidos, eles ganham energia. Entretanto, há a tendência deles retornarem ao seu estado de energia mínima, emitindo a energia excedente por radiação eletromagnética. Esta energia liberada é detectada e decodificada, originando informações que, após tratamento detalhado, fornece relações espaciais entre os núcleos, a partir dos quais pode-se obter a estrutura da proteína. A ressonância é classificada como um processo dinâmico por permitir a determinação das conformações ao longo de um intervalo de tempo.

As técnicas experimentais, apesar de apresentarem uma precisão superior aos demais métodos, são lentas e caras, pois necessitam de equipamentos, laboratórios e pessoas especializados. Por outro lado, as aplicações teóricas são mais acessíveis, sendo realizadas em qualquer computador. Dentre elas, o método físico mais conhecido é o *ab initio* **[BS00, Pil02]**, enquanto no campo empírico encontra-se a modelagem por homologia **[MSF$^{+}$00]** e o *threading* **[HT00]**.

- ***Ab initio***: com base na teoria termodinâmica de *Anfinsens* **[Anf72]**, que sugere que a estrutura nativa de uma proteína é aquela com a menor energia global, esse método busca a estrutura de menor energia, a partir da otimização ou minimização global de uma função de energia potencial, que representa o quanto a proteína precisa se esforçar para assumir a conformação. O procedimento de obtenção do modelo é dividido em três etapas:
    - Busca do mínimo global em uma representação simplificada da seqüência;
    - Conversão da estrutura de menor energia para uma representação com todos os átomos;
    - Busca da conformação de menor energia da nova representação.

  Esse método apresenta melhores resultados para pequenas seqüências (de até 150 aminoácidos), sendo a precisão reduzida com o aumento da quantidade de aminoácidos, além de ser mais lento que as demais aplicações teóricas. A função de energia pode ser composta por potenciais derivados de mecanismos moleculares (atração entre átomos, ângulos de ligação, etc.), presentes em ferramentas como *CHARMM* (*Chemistry at HARvard Macromolecular*

*Mechanics*) **[CHARMM]** e *TINKER* **[TINKER]**, assim como por potenciais de energia originados da análise estatística de estruturas conhecidas; a otimização é aplicada através de simulação de dinâmica molecular ou de Monte Carlo. Em termos de precisão, essa metodologia apresenta os piores resultados quando comparada aos outros métodos descritos a seguir.

- **Homologia ou comparação**: com base nas observações obtidas nos laboratórios em que proteínas com seqüências similares apresentam estruturas similares, determina-se a estrutura de uma proteína não modelada a partir de outras conhecidas, explorando-se o grau de similaridade entre as seqüências obtido a partir do processo de alinhamento entre as seqüências. As estruturas conhecidas são obtidas nos bancos de dados públicos.
- ***Threading* ou reconhecimento de dobras**: este método de modelagem foi desenvolvido considerando a característica de que uma grande quantidade de proteínas apresenta um conjunto limitado de possíveis dobra, sendo bastante semelhante ao processo da homologia. De uma maneira breve, os passos a serem seguidos na construção do modelo são:
    - A partir de uma base de dados de estruturas de proteínas é selecionado um conjunto representativo;
    - Determinam-se as dobras relevantes;
    - Trechos da seqüência alvo são alinhados com as dobras, permitindo inserções e remoções de laços, em um procedimento de otimização da função que determina a pontuação para o alinhamento entre seqüência e estrutura (função objetiva);
    - A saída do processo é a pontuação das dobras em função dos valores da função objetiva.

  Um caso especial é quando a seqüência alvo tem uma semelhança significativa com alguma estrutura do conjunto representativo, nesse caso, o problema recai em uma modelagem por homologia.

## Seção 2.7. Modelagem por homologia

Por se tratar, dentre as aplicações teóricas, a metodologia de determinação de estrutura terciárias mais rápida e precisa, a modelagem por homologia foi selecionada como o mecanismo para a construção dos novos modelos das proteínas.

Como já foi dito anteriormente, a idéia central dessa metodologia é obter o modelo de uma proteína de conformação desconhecida (seqüência alvo) a partir de proteínas com estruturas conhecidas (seqüências moldes ou *templates*) **[MSF$^{+}$00]**. Esta abordagem surge do fato de proteínas com seqüências similares apresentam estruturas tridimensionais também similares e, em alguns casos, proteínas com seqüências diferentes podem ter estruturas semelhantes, sugerindo que as estruturas 3-D são mais conservadas que as seqüências. Com a aplicação deste método é possível predizer 1/3 de todas as seqüências de proteínas armazenadas no *GenBank*, o que representa mais estruturas do que as já construídas e registradas no *PDB*.

Alguns pontos contribuem para o bom desempenho desta metodologia: a quantidade de possíveis configurações (dobras) adotadas pela proteína é limitada; o número de famílias de proteínas tende a não passar de 1.000 **[Mou01, HT00]**, o que limita a quantidade de conformações existentes; e o número de novas estruturas conhecidas vem crescendo continuamente.

São quatro os passos necessários para a construção do modelo gerado por homologia: busca de dobras e seleção dos moldes, alinhamento entre alvo e molde, construção do modelo e avaliação do modelo **(Figura 2.7.I: Modelagem por Homologia)**.

- **Busca das dobras e seleção dos moldes**: tem o objetivo de identificar proteínas de estruturas conhecidas relacionadas com a seqüência alvo. A busca pode ser realizada a partir dos bancos de dados de estruturas e seqüências, tais como *PDB*, *SCOP* **[SCOP]**, *DALI* **[DALI]** e o *CATH* **[CATH]**. Nessa busca, pode-se usar a comparação seqüência-seqüência (alinhamento entre pares), onde a seqüência alvo é comparada com cada uma das seqüências do banco de dados individualmente; ou uma comparação múltipla (alinhamento múltiplo de seqüências) envolvendo três ou mais proteínas, sendo esta última mais sensível por indicar domínios estruturais e regiões que representam conservação da função. Em seguida, faz-se uma seleção daquelas proteínas que serão utilizadas na modelagem, ou seja, os *templates*. Na maioria dos casos, o critério de classificação se baseia apenas na similaridade entre as seqüências. No entanto, outros aspectos podem ser considerados, como a divisão de família em subfamílias e a formação de uma árvore filogenética para a seleção dos moldes pertencentes a subfamílias próximas a seqüência alvo; o ambiente de imersão do molde (tipo de solvente, pH, etc.); e a qualidade experimental das estruturas

(precisão, fator-R). Programas como o *BLAST* **[BLAST]** e o *FASTA* **[FASTA]** podem ser empregados no alinhamento entre pares. Para alinhamentos múltiplos usa-se o *PSI-BLAST* **[BLAST]**.

- **Alinhamento alvo-molde**: diferentemente dos alinhamentos executados na etapa anterior, feitos a partir das seqüências e com o objetivo de detectar qualquer relação fraca existente entre elas, nesse estágio é construído o alinhamento ótimo entre a seqüência alvo e as estruturas dos *templates*. É preciso ter bastante atenção nesta fase, já que muitos erros de modelagem ocorrem devido a equívocos de alinhamento. O uso de conhecimentos sobre as estruturas secundárias e alterações manuais contribuem para um melhor resultado.
- **Construção do modelo**: cria-se o modelo da seqüência alvo a partir do alinhamento obtido no passo anterior, podendo-se utilizar uma dentre as três técnicas seguintes:
    - **Montagem por corpos rígidos**: baseada na divisão das estruturas das proteínas em três conjuntos: as regiões centrais conservadas, os laços e as cadeias laterais. Num primeiro estágio, sobrepõe-se os moldes e constrói-se uma moldura formada pela média das coordenadas dos Cα's das regiões conservadas dos moldes. Em seguida, o modelo da cadeia principal da região central da seqüência alvo é gerado pela superposição da moldura com os segmentos de regiões centrais dos moldes cuja seqüência é semelhante à alvo. Os laços são obtidos a partir de bancos de dados e de acordo com a região central, e as cadeias laterais são baseadas nas conformações das cadeias laterais presentes nos moldes.
    - **Reconstrução de coordenada**: hexapeptídeos de um conjunto representativo das estruturas padrões das proteínas podem ser organizados em aproximadamente 100 conjuntos distintos **[UHW$^+$89]**. Assim, adotando os moldes como guias e tendo como base os alinhamentos da seqüência alvo com as estruturas dos moldes, associam-se trechos da seqüência alvo a conjuntos de hexapeptídeos.
    - **Satisfação das restrições espaciais**: com base em análises muitas vezes estatísticas das várias características presentes nas estruturas das proteínas (ângulos diedros, comprimentos da ligação, ângulos de ligação, contatos entre átomos não ligados) obtém-se os valores médios

correspondentes a cada uma delas, baseados num conjunto representativo de estruturas de proteínas conhecidas. Desta forma, com base nos moldes, determinam-se os valores das restrições que deverão ser respeitados, ou ser minimamente infringidos, durante a construção do modelo da seqüência alvo.

A modelagem dos laços se apresenta como um ponto de destaque na determinação do modelo de uma proteína, pois a diferença funcional entre proteínas de uma mesma família está normalmente associada à diferença estrutural da superfície da proteína. Tais mudanças freqüentemente ocorrem nos laços, indicando que eles têm uma forte relação com a funcionalidade de uma proteína.

O principal problema para a boa determinação da conformação dessas regiões se deve ao fato de que os laços da seqüência alvo são regiões estruturalmente diferentes daquelas correspondentes nos moldes devido à variação conformacional para uma mesma seqüência de aminoácidos. Além disso, por serem regiões pequenas, apenas sua composição não é suficiente na determinação do modelo, sendo necessário o conhecimento do solvente em que se encontra a proteína e dos segmentos vizinhos ao laço na cadeia principal, ou seja, dos aminoácidos que precedem e sucedem o laço.

Para a solução do problema, duas técnicas, ou uma combinação delas, normalmente são aplicadas:

- **Busca em base de dados**: encontra, em bancos de dados de estruturas conhecidas, segmentos de cadeia principal que coincidam com a seqüência do laço e com as estruturas vizinhas a ele na proteína alvo. A limitação desse método se encontra no próprio tamanho das bases de dados, pois o número de conformações retornadas para um laço diminui com o aumento de seu comprimento, não disponibilizando um conjunto suficientemente grande para a escolha da melhor conformação para o laço.
- **Busca conformacional**: também conhecida por predição *ab initio*, faz a enumeração de possíveis conformações para a seqüência num determinado ambiente, guiado por uma função de energia ou de pontuação para a seleção das melhores.

- **Avaliação do modelo**: tem o objetivo de verificar a precisão da modelagem. Esta pode ser feita dos seguintes modos: pela similaridade entre os dobramentos dos *templates* e do modelo, e entre as seqüências *templates* e alvo, através da avaliação das propriedades estereoquímicas e por características espaciais (acessibilidade do solvente, formação das regiões hidrofóbicas, distribuição espacial de grupos de carga, volume atômico). A qualidade do modelo está associada ao tipo de informação que será retirada da nova estrutura. Por exemplo, se as identidades entre a seqüência alvo e os *templates* ficam abaixo de 30% (baixa precisão), o modelo serve para predizer, aproximadamente, algumas funções bioquímicas; entre 30-50% (média precisão), é possível determinar funcionalidades importantes da seqüência alvo; e acima de 50% (alta precisão), é possível ser aplicado no processo de *docking*. Caso não se tenha um modelo muito bom, retorna-se ao início, selecionando-se novos *templates*.

Os principais erros no processo de modelagem por homologia acontecem na determinação das cadeias laterais, quando uma mesma seqüência apresenta conformações distintas; quando os moldes são localmente diferentes da seqüência alvo; caso regiões da seqüência alvo não apresente um correspondente nos *templates*; e devido a erros de alinhamento e escolhas de moldes muito distantes entre si, com menos de 25% de identidade.

### 2.7.1. MODELLER

Dentre os vários programas que propõem a modelagem das estruturas das proteínas por homologia **(Tabela 2.7.1.I: Modelagem)**, o selecionado para o desenvolvimento do projeto foi o *MODELLER* **[MODELLER]**, por se tratar de uma ferramenta gratuita para pesquisadores e ser a mais utilizada no processo de construção de modelos.

O *MODELLER* foi desenvolvido para a predição da estrutura tridimensional de uma proteína por modelagem comparativa, sem considerar os átomos de hidrogênio e utilizando a técnica da satisfação das restrições espaciais. A ferramenta também oferece outras funcionalidades, como: comparação de seqüências e estruturas, *clustering* de proteínas, alinhamento entre seqüência e estrutura, busca em base de dados de seqüências, etc. Um material de avaliação da ferramenta, para três proteínas diferentes, feito por Andrej Šali *et al.* pode ser encontrado em **[SPY+95]**.

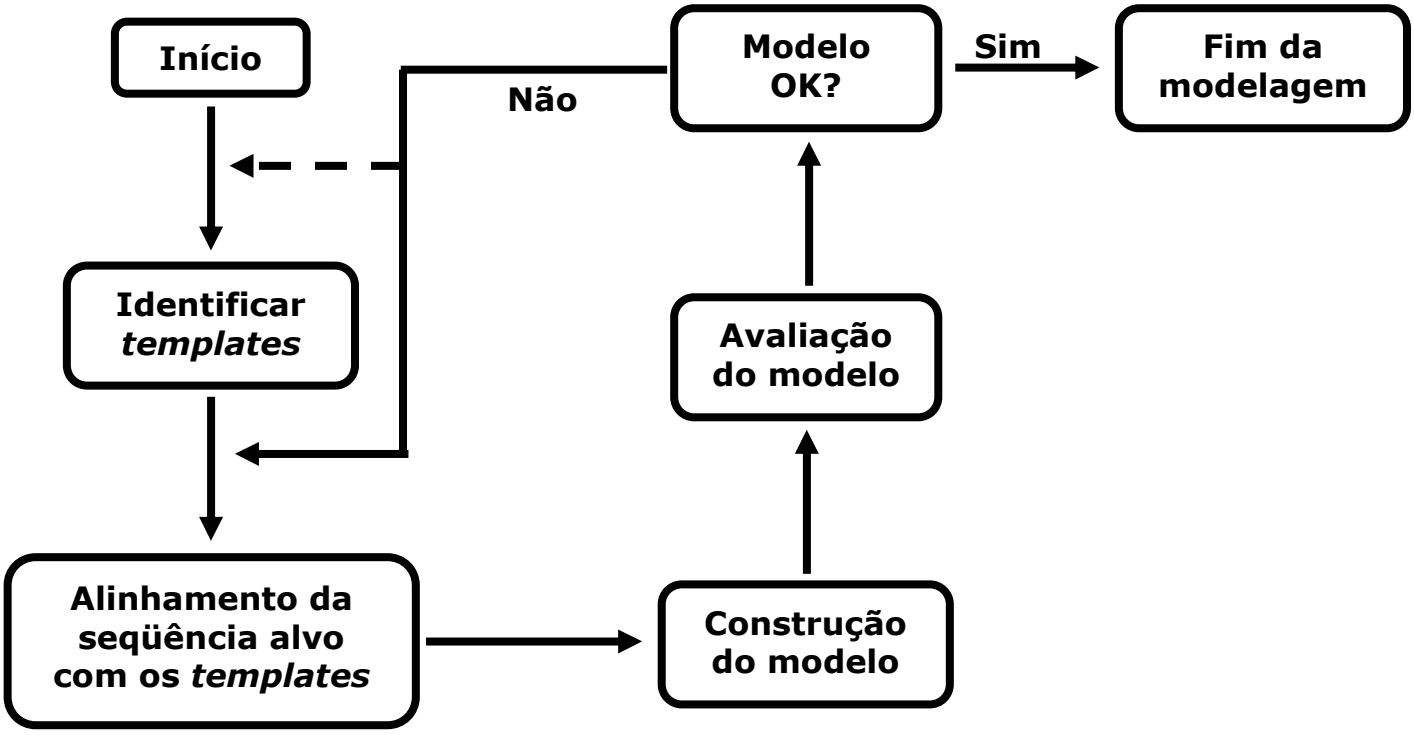



**Figura 2.7.I: Modelagem por Homologia** – Fluxo da modelagem por homologia

| Programa | Site |
|---|---|
| COMPOSER | http://www.tripos.com/sciTech/inSilicoDisc/bioInformatics/composer.html |
| CONGEN | http://www.congenomics.com/congen/congen.html |
| CPH models | http://www.cbs.dtu.dk/services/CPHmodels/ |
| Insight II | http://www.accelrys.com/insight/ |
| MODELLER | http://salilab.org/modeller/ |
| QUANTA | http://www.accelrys.com/quanta/ |
| SWISS-MOD | http://swissmodel.expasy.org/ |
| WHAT-IF | http://swift.cmbi.kun.nl/whatif/ |

**Tabela 2.7.1.I: Modelagem** – Algumas ferramentas aplicadas na modelagem de proteínas.

Por se tratar de uma ferramenta de modelagem por comparação, o *MODELLER* está dividido em vários blocos, de acordo com as etapas descritas no início desta seção. Para cada um deles será feito a seguir um breve comentário sobre os algoritmos aplicados, dos formatos de entrada e das saídas esperadas.

- **Busca das estruturas e seqüências:** a partir de uma base de dados de estruturas, procuram-se aquelas que apresentam similaridade com a seqüência alvo. Uma pontuação, que pode ser calculada de três formas distintas, define o grau de similaridade. A busca ocorre numa base de dados formada por um subconjunto representativo do *PDB*, onde estão presentes seqüências com menos de 40% de identidade entre si, constituindo cerca de 3.000 estruturas. Isto garante uma

busca mais rápida, pois evita a comparação da proteína alvo com seqüências muito semelhantes. A busca é baseada no algoritmo de programação dinâmica de alinhamento de seqüência de *Needleman & Wunsch* **[NW70]**; a entrada é um arquivo texto contento a seqüência alvo, no formato *PIR,* e a saída é outro arquivo texto, com os alinhamentos e as respectivas pontuações. Neste arquivo de saída, cada uma das seqüências similares à proteína alvo está associada a uma lista com as demais estruturas do *PDB* semelhantes a elas que não estavam incluídas na base de dados de busca. **(Apêndice MODELLER).**

- **Alinhamento múltiplo:** considerando-se apenas as proteínas selecionadas no passo anterior, constrói-se o alinhamento múltiplo de seqüências e estruturas. Esse alinhamento pode ser executado de quatro formas distintas:
    - Alinhamento múltiplo entre estruturas;
    - Alinhamento múltiplo entre seqüências;
    - Alinhamento entre dois blocos de seqüências já alinhadas;
    - Alinhamento entre um bloco de seqüências e um bloco de estruturas ambos alinhados.

  Em alguns casos, o alinhamento deve ser editado manualmente para otimizar sua qualidade.
- **Melhorando o alinhamento:** o seu objetivo é aperfeiçoar o alinhamento obtido no passo anterior, de forma manual ou automática. No processo manual, através de uma análise visual, determinam-se regiões mais estáveis/variáveis e *gaps* que podem ser excluídos ou movidos. No automático, para cada par de estruturas dos moldes é feita uma superposição da estrutura dos carbonos-α a partir do alinhamento, verificando a distância entre os carbonos-α equivalentes, sendo os erros de alinhamento indicados por valores acima de 6 Å (Âmgström).
- **Selecionando os moldes:** são escolhidas as estruturas mais representativas e mais próximas à seqüência alvo, a partir de um dendrograma **(Apêndice MODELLER)** ou utilizando a técnica de análise de componentes principais. Ambos os casos são construídos com base nas matrizes de distâncias dos alinhamentos dos pares das estruturas selecionadas na primeira etapa. Essas matrizes são criadas em função da similaridade dos resíduos ou da raiz quadrada da média da distância (*RMS*) entre as posições dos átomos.

  Antes do próximo passo, um alinhamento ótimo entre as seqüências moldes e a alvo é realizado e corrigido a partir da seguinte avaliação: para cada dois átomos

consecutivos de Cα da seqüência alvo, calcula-se a distância entre os átomos correspondentes nas estruturas presentes no alinhamento. Resultados acima de 8 Å indicam erros de alinhamento entre alvo e molde.

- **Construindo o modelo:** para o desenvolvimento da nova conformação da proteína alvo, a ferramenta MODELLER se baseia no cálculo da satisfação das restrições espaciais **[SB93]**. Com base em observações retiradas de uma base de dados ou por cálculo analítico das restrições estereoquímicas, propriedades das proteínas foram representadas por funções de densidade de probabilidade (pdf), que, combinadas, formam a pdf molecular. E a partir da otimização dessa expressão, por técnicas de gradiente conjugado e *simulated annealing*, o modelo tridimensional é formado.

  No MODELLER, as informações retiradas a partir de banco de dados para a construção das pdfs estão relacionadas a: distância entre dois átomos de Cα equivalentes entre estruturas semelhantes, distância entre os átomos de oxigênio e nitrogênio da cadeia principal e valores dos ângulos diedros ψ, Φ, $\chi_1$, $\chi_2$, $\chi_3$ e $\chi_4$. As informações com origem no cálculo analítico das restrições estereoquímicas para o desenvolvimento das pdfs, com base na estatística e mecânica clássica, são: a distância entre as ligações, os ângulos das ligações, as pontes dissulfídicas, a planaridade das cadeias laterais e a interação de van der Waals, entre outros valores.

  Após o alinhamento entre a seqüência alvo e as estruturas moldes, determina-se a probabilidade do modelo por cálculo da pdf molecular, formada tanto por termos que expressam o relacionamento entre a estrutura alvo e a dos moldes, como por elementos que limitam as características estereoquímicas do alvo. Um processo de otimização reduz as violações apresentadas pelo modelo, aumentando sua probabilidade baseada na fpd.

- **Modelagem dos laços:** com o modelo construído, a ferramenta MODELLER permite que o usuário modifique regiões classificadas como laços **[FDS00]**. Da mesma forma que para a modelagem de toda a estrutura da proteína, a alteração das conformações dos laços também não considera os átomos de hidrogênio. Durante o processo de atualização, são apreciadas as interações entre os átomos dos laços e os átomos presentes no restante da proteína que não se movimentam durante a execução desta etapa.

Durante a modelagem dos laços, uma função de pontuação baseada nas preferências estatísticas dos átomos de uma base de dados de estruturas conhecidas é otimizada. Inicialmente o laço apresenta-se com uma conformação onde os átomos são posicionados numa linha. Essa linha conecta o oxigênio da carboxila da cadeia principal do aminoácido anterior ao laço, com o nitrogênio do aminoácido posicionado logo após o laço. As coordenadas cartesianas dos átomos são modificadas randomicamente sobre valores uniformemente distribuídos entre -5 e 5 Å.

A função de pontuação inclui: propriedades estereoquímicas, como ângulos de ligação, distância de ligação; potenciais relacionados à interação entre os átomos não ligados e preferências estatísticas para os ângulos diedros das cadeias principais e laterais. Cada um dos termos representado por uma função de densidade de probabilidade.

- **Avaliando os modelos:** apesar da ferramenta apresentar rotinas que permitem comparar os valores das características espaciais do modelo com os valores padrão das restrições aplicadas, é comum avaliar a nova estrutura a partir de outras ferramentas desvinculadas do *MODELLER*, como: *PROCHECK*, *ANOLEA*, *PROSA*, *PROVE, WHAT-IF*, entre outras.

Todo o ciclo descrito pode ser repetido até que a conformação seja boa o suficiente, ou nenhuma melhora a mais seja possível. Isto pode ser feito através da adição ou remoção de moldes ou, ainda, de modificação dos alinhamentos, podendo-se utilizar algumas regiões do modelo intermediário como informações para o próximo ciclo.

## Seção 2.8. Avaliação de Estruturas de Proteínas

Uma vez finalizado o processo de modelagem, seja por métodos experimentais ou teóricos, é necessário descobrir se a construção atingiu os resultados esperados ou se será preciso melhorá-lo, e conseqüentemente, repetir parte do experimento. Assim, para se desvendar possíveis problemas das novas estruturas, é preciso avaliá-las através de mecanismos de campo de força, de métodos baseados em conhecimento ou de análises das propriedades estereoquímicas.

Campos de força são funções matemáticas que descrevem o comportamento estrutural de moléculas através de funções de potencial, associadas às interações entre

átomos não-ligados (como interações de Van der Waals e Coulomb), entre átomos ligados (distância das ligações, ângulos de ligação, ângulos diedros, etc.) e restrições especiais (restrições de posição e de distância). Parâmetros estão associados a cada um dos termos variando de acordo com os tipos de átomos. A energia da molécula é dada pela soma de cada um dos termos de interação e o menor valor indica a conformação mais provável. Por exemplo, o potencial entre dois átomos ligados numa molécula é dado pela expressão:

$$\sum_{i=1}^{N} \sum_{j=i+1}^{N} K_{ij} (r_{ij} - r_0)^2 \qquad \text{( Eq. 2.8.I )}$$

onde **i** e **j** são dois átomos conectados da molécula, **N** é o número de átomos da molécula, $\mathbf{K_{ij}}$ é a constante de força que varia de acordo com os tipos dos átomos, $\mathbf{r_{ij}}$ é a distância entre esses átomos na molécula e $\mathbf{r_0}$ corresponde ao valor ideal da distância.

Os diferentes tipos de campos de força existentes variam de acordo com os termos utilizados e os valores das constantes, assim como os tipos de estruturas a que devem ser submetidos. Por exemplo, o campo de força AMBER (*Assisted Model Building with Energy Refinement*) **(Figura 2.8.I: AMBER)** é aplicado em estruturas de proteínas e ácidos nucléicos, apresentando termos referentes aos comprimentos das ligações entre os átomos, aos valores dos ângulos de ligação (o somatório **ligantes** da **Figura 2.8.I: AMBER**), dos ângulos de torção (o somatório **ângulos**), de uma combinação de interação de Van der Waals e eletrostática dos átomos (o somatório **não-ligantes**) e das ligações de hidrogênio (o somatório **ligações de H**).

Os métodos baseados em conhecimento, ou potenciais estatísticos, são caracterizados por uma análise estatística de estruturas conhecidas e bem formuladas, modelando a energia potencial das interações intramoleculares entre pares de átomos em função da distância de separação do tipo de aminoácido **[Sip90]**. A grande vantagem dos métodos baseados em conhecimento é a inclusão de conhecimentos e/ou propriedades que ainda não foram totalmente entendidos.

A análise das propriedades estereoquímicas determina valores padrão para algumas das propriedades e compara-os a valores encontrados para as mesmas propriedades na estrutura de proteínas que está sendo submetida à avaliação.

Além destas três técnicas, a avaliação pode ter dois objetivos distintos: classificar um conjunto de conformações ou selecionar regiões de uma estrutura.

$$E_{\text{total}} = \sum_{\text{ligantes}} K_r(r - r_0)^2 + \sum_{\text{ângulos}} K_\theta(\theta - \theta_0)^2$$
$$+ \sum_{\text{diedros}} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] + \sum_{\text{não-ligantes}} \left[ \frac{\sigma}{R_{ij}^{12}} - \frac{\sigma}{R_{ij}^{6}} + \frac{q_i q_j}{\epsilon R_{ij}} \right]$$
$$+ \sum_{\text{ligações de H}} \left[ \frac{C_{ij}}{R_{ij}^{12}} - \frac{D_{ij}}{R_{ij}^{10}} \right]$$

**Figura 2.8.I: AMBER** – Expressão para o cálculo do campo de força AMBER de uma molécula, onde: **r** é a distância entre dois átomos conectados e $\mathbf{r_0}$ a distância ideal entre eles; **θ** é o ângulo de ligação e $\mathbf{\theta_0}$ o valor ideal; **Ø** é um ângulo diedro da cadeia principal (Φ, ψ, ω) ou lateral (χ), **n** é a periodicidade da função potencial e **γ** o deslocamento da fase de acordo com o tipo de diedro considerado; $\mathbf{R_{ij}}$ a distância entre dois átomos não conectados; $\mathbf{K_r}$, **σ**, **ϵ** representam as constantes para os potenciais.

Classificar um conjunto de conformações, seja de uma proteína completa ou de um fragmento de polipeptídio, tem a finalidade de selecionar as conformações mais próximas a um modelo real num grupo formato por estrutura nativas e não-nativas. Para isso, associa-se a cada seqüência uma pontuação originada de uma função de energia, onde os valores mais altos correspondem a estruturas distantes da nativa e os baixos resultados a modelos próximos do desejado. Para uma classificação aplicando campos de força, pode-se calcular a pontuação através de ferramentas como: *CHARMM*, *GROMACS* **[GROMACS]** e *TINKER*, sendo esta última de uso gratuito. Por potenciais estatísticos, pode-se citar o *RAPDF* criado por Ram Samudrala e John Moult **[SM98]**, que calcula a probabilidade de uma seqüência de aminoácidos apresentar uma conformação próxima da nativa.

Selecionar regiões tem o objetivo de, numa única estrutura, determinar trechos da seqüência que não estão de acordo com uma propriedade específica. Nesta situação, pode-se aplicar as seguintes ferramentas:

- *ANoLEA* (*Atomic Non-Local Environment Assessment*) **[ANoLEA]**: calcula a energia da estrutura com base nas interações entre os pares de átomos. A função aplicada no cálculo da energia, derivada do trabalho desenvolvido por Sippl **[Sip95]**, depende da distância de separação dos átomos, sendo derivada de um banco de dados de 147 estruturas de proteínas não-redundantes;
- *AQUA* (*Analyzing the QUAlity of biomolecular structures*) **[LRM$^+$96; AQUA]**: conjunto de programas que avalia a qualidade de estruturas originadas do processo de ressonância magnética nuclear através de informações originadas

dos *NOEs*, das distâncias entre os átomos, das ligações de hidrogênio, das ligações dissulfídicas e dos ângulos de torção;

- *ERRAT* **[ERRAT; CY93]**: direcionada para modelos originados da cristalografia de Raios-X, esboça os erros de uma estrutura baseado numa análise estatística do número de interações entre átomos não-ligados de C, N e O, ou seja, seis tipos de interação (CC, CO, CN, ON, OO, NN), separados por uma distância de no máximo 3,5 Å e de resíduos diferentes. Compara a freqüência do número de interações da estrutura de entrada com os valores obtidos de um banco de dados de 96 estruturas de alta resolução;
- *PROCHECK* **[PROCHECK]**: esboça os valores das propriedades estereoquímicas das estruturas por meio de arquivos textos ou gráficos, permitindo uma comparação com os valores padrão, originados dos trabalhos de Morris et al **[MMH$^+$92]** e Engh & Huber **[EH91]**;
- *ProSa* **[ProSa, Sip93]**: a partir de potenciais criados com base nas propriedades das proteínas globulares, originadas da cristalografia e da RMN, em função dos pares de átomos e das interações entre o solvente e a estrutura, calcula-se a energia dos modelos, medindo-se o desvio em relação a conformações globulares mais comuns;
- *PROVE* (*PROtein Volume Evaluation*) **[PROVE]**: avalia os modelos com base na comparação do volume dos átomos da estrutura de entrada com resultados originados de um conjunto de proteínas do *PDB*;
- *Verify3D* **[Verify3D; BLE91; LBE92]**: compara a estrutura de entrada com sua própria seqüência, através da técnica de *3D profile*. Essa técnica constrói para cada posição de aminoácido da estrutura uma tabela de 20 linhas representando as preferências estatísticas dessa posição para cada um dos diferentes tipos de aminoácidos. Esta tabela está em função do ambiente no qual a proteína esta imersa. Aplicando os valores da tabela de acordo com a seqüência de aminoácidos da estrutura, obtêm-se pontuações, onde modelos corretos produzem altos valores **[EBL$^+$92]**;
- *WHAT_CHECK* **[WHATCHECK]**: avaliador de proteínas, presente no programa *WHAT IF*, que calcula propriedades estruturais (fatores-B, ligações de hidrogênio, etc.) e geométricas (quiralidade, planaridade, comprimentos de

ligação, ângulos de ligação e de torção, entre outras), e compara esses valores com os obtidos a partir de um conjunto representativo do *PDB*.

No projeto, apenas quatro ferramentas foram selecionas (*ANoLEA*, *ProSa*, *PROVE* e *PROCHECK*), sendo a escolha baseada no fato de que as mesmas:

- São utilizadas em outros projetos de grande importância;
- Apresentam um material explicativo da teoria em que se fundamentaram;
- São executadas em ambientes *Linux* ou *Unix*, com saída em formato de texto;
- A interpretação dos resultados não requer qualquer conhecimento das propriedades das proteínas utilizadas como entrada;
- Diferenciam-se nas informações utilizadas para as verificações.

### 2.8.1. ANoLEA

O ANoLEA é uma ferramenta aplicada na determinação de regiões defeituosas de uma estrutura de acordo com a energia de interação entre os átomos. Esse cálculo de energia considera apenas as interações não-locais entre os átomos de C, O, N, S da estrutura, denominados por átomos pesados, e gera um perfil da energia, onde fragmentos com altos valores correspondem aos prováveis erros de modelagem **[MDD+97]**.

A partir dos 20 aminoácidos mais comuns presentes nas proteínas, definem-se 40 tipos diferentes de átomos **(Figura 2.8.1.I: 40 Átomos)** Em seguida, calculam-se dois tipos de força média potencial de nível atômico (*FMP-AT*), ou seja, potenciais estatísticos derivados de um banco de dados de proteínas não-redundantes, um dependente da distância entre os átomos e envolvendo apenas interações não-locais; o outro referente a superfície de acesso do átomo **[MF97]**. Todos os potenciais estatísticos são derivados do trabalho desenvolvido por Sippl **[Sip90]**.

A região não-local de um átomo compreende os átomos pesados da estrutura localizados a uma distância menor que 7 Å e separados por pelo menos 11 aminoácidos, quando presentes na mesma cadeia. No caso de estarem em outra cadeia, apenas a primeira condição é necessária. Para o cálculo da *FMP*, o raio de 7 Å é dividido em 35 partes **(Figura 2.8.1.II: Região não-local)**. A superfície de acesso é o número total de átomos pesados incluídos em um raio de 7 Å de uma esfera centrada no átomo dado. O raio é dividido em 20 classes de freqüência em um intervalo de 0 a 99 átomos **[MF98]**.

A energia é calculada inicialmente para cada átomo, com a soma das *FMP-AT* dependente da distância e da superfície de acesso. Para cada aminoácido, o potencial é obtido pelo somatório das energias dos seus respectivos átomos.

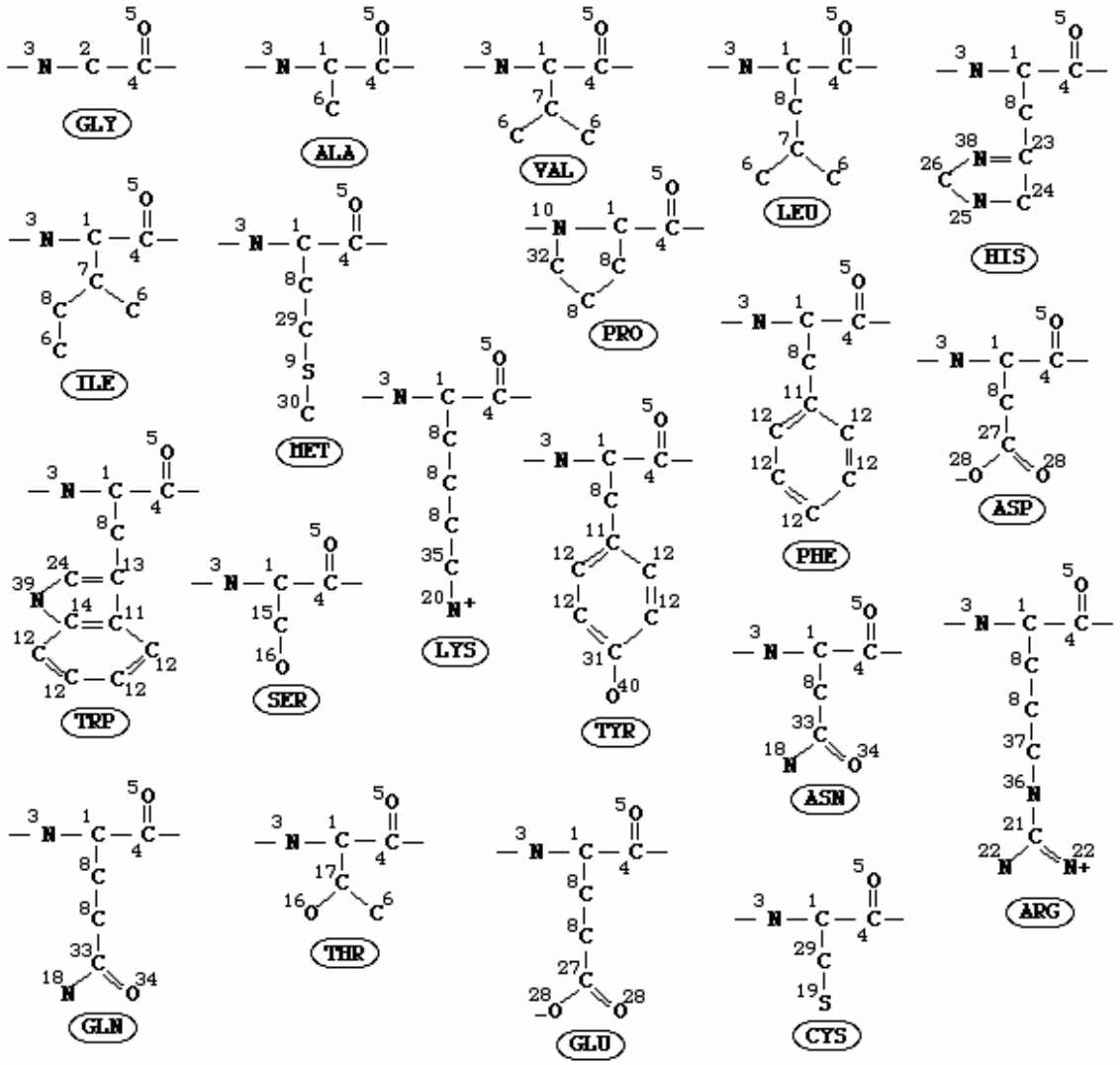



**Figura 2.8.1.I: 40 Átomos -** Apresenta os 20 tipos de aminoácidos e a classificação para os átomos, de acordo com a natureza química, localização (cadeia lateral ou principal) e as ligações com os outros elementos.

## 2.8.2. ProSa

Esta ferramenta emprega um banco de dados de proteínas, o *PDB*, com a finalidade de obter valores padrão para as forças que estabilizam as dobras de uma proteína nativa em solução. Aplicando o cálculo de potencial estatístico derivado de **[Sip90]**, as forças descrevem as interações entre os átomos das proteínas em função da distância de separação entre eles e dos tipos de aminoácidos aos quais os átomos pertencem.

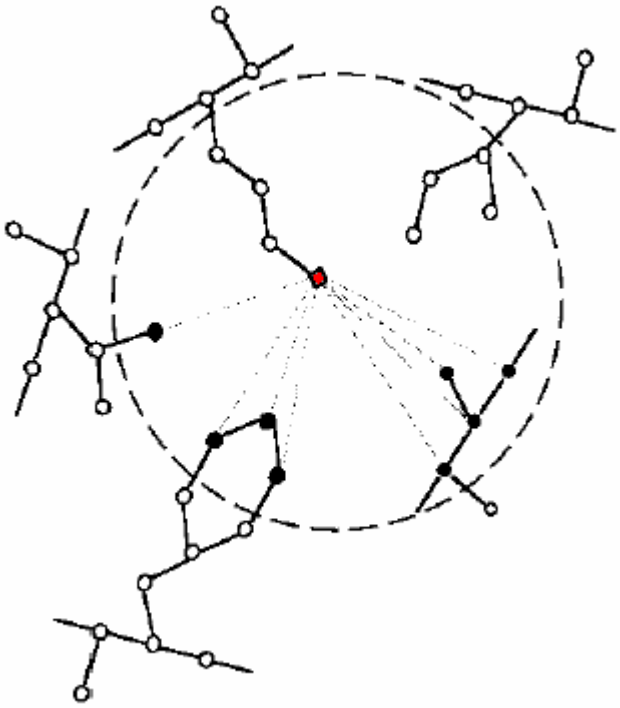


**Figura 2.8.1.II: Região não-local -** em vermelho está o átomo do qual define-se a região não-local. A circunferência de raio 7 Å encontra-se tracejada e em seu interior alguns átomos preenchidos, que correspondem àqueles que se encontram a uma distância maior que 11 resíduos em relação ao átomo central.

Assim, a energia de uma estrutura é determinada pela recombinação desses potenciais em função da seqüência de aminoácidos e da distância entre os átomos. Regiões com alto valor de energia são consideradas problemáticas. Diferente do *ANoLEA*, que aplica o mesmo princípio de cálculo de potencial estatístico, o ProSa considera apenas os átomos de Cα e Cβ, mas divide-os em grupos de acordo com os aminoácidos aos quais pertencem, caracterizando uma força média potencial aplicada apenas em nível de aminoácido (*FMP-AA*).

## 2.8.3. PROVE

Como já citado anteriormente, a ferramenta PROVE avalia os modelos com base no volume dos átomos pesados. O cálculo do volume é feito segundo o método de *Voronoi*, e os valores padrão são determinados a partir de um conjunto de 64 estruturas de proteínas com alto valor de resolução (precisão menor ou igual a 2 Å e fator-R menor que 0,20).

O volume correspondente a cada átomo se resume ao menor poliedro limitado pelos planos que interceptam os vetores de conexão do átomo central e seus vizinhos perpendicularmente na mediana **(Figura 2.8.3.I: Voronoi)**. Apenas os átomos que se

encontram no interior da proteína, ou seja, que apresentam área de contato com o solvente igual a zero, são empregados.

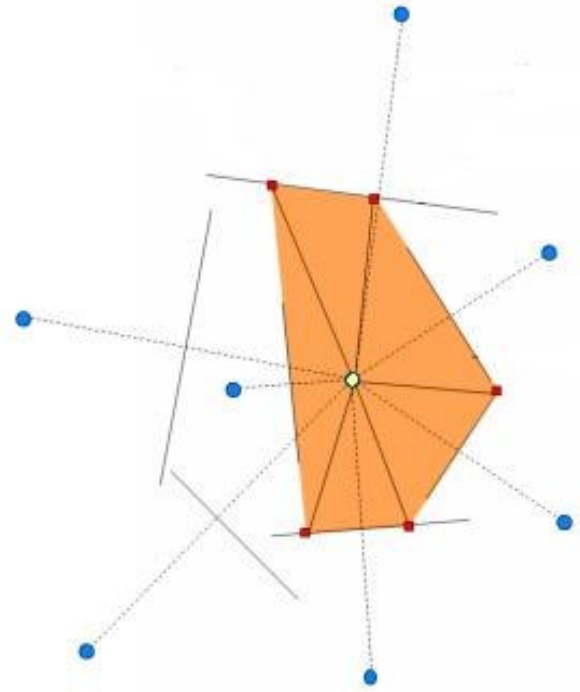


**Figura 2.8.3.I: Voronoi -** a ilustração em 2D representa o polígono de Voronoi (em laranja) considerando o átomo central, em amarelo, e seus vizinhos em azul. As linhas tracejadas são os vetores de conexão entre dois átomos e as linhas cheias correspondem aos planos que interceptam os vetores perpendicularmente na mediana.

Para o *PROVE*, os átomos foram divididos em 23 grupos (**Tabela 2.8.3.I: Prove**), de acordo com o tipo químico e com os ligantes. O volume médio e o desvio padrão para cada um dos grupos foram calculados e estabelecidos como padrão.

Dessa forma, para se descobrir irregularidades no volume de uma proteína modelada, efetua-se a operação de *Voronoi* para cada um de seus átomos e analisa-se cada um deles de acordo com a pontuação-Z **[PRW96]**:

$$\text{Pontuação-Z} = [V_{i,k} - MV_k] / \sigma_k \quad \text{( Eq. 2.8.3.I )}$$

onde $V_{i,k}$ é o volume do i-ésimo átomo da proteína, sendo este do tipo k. $MV_k$ é a média dos volumes do tipo k, originados da base de dados, e $\sigma_k$ é o desvio padrão. Para as pontuações maiores que três, é provável a presença de erros.

Infelizmente, quando se comparam as regiões defeituosas indicadas pelo *PROVE* com as recomendadas pelo *PROCHECK*, há algumas diferenças, provavelmente devido à pequena quantidade de proteínas da base de dados da primeira ferramenta. Apesar disso, o *PROVE* é utilizado no projeto por considerar um parâmetro, o volume do

átomo, que não é aplicado na maioria dos procedimentos de construção de modelos, ou seja, um termo independente do processo de modelagem.

| | |
|---|---|
| **C** | Carbono sem H (dos grupos amida e carboxila) |
| **CH1** | Carbono conectado a 1 H (Cα da maioria dos aminoácidos) |
| **CH2** | Carbono conectado a 2 H (Cβ da maioria dos aminoácidos) |
| **CH3** | Carbono conectado a 3 H (Cδ1 e Cδ2 da leucina) |
| **CR15** | Carbono conectado a 1 H em um anel de 5 átomos (Cη1 da histidina) |
| **CR16** | Carbono conectado a 1 H em um anel de 6 átomos (Cη1 da fenilalanina) |
| **CR5** | Carbono sem H em um anel de 5 átomos (Cγ1 da histidina) |
| **CR56** | Carbono entre dois anéis, um de 5 átomos e outro de 6 (Cδ2 do triptofano) |
| **CR6** | Carbono sem H em um anel de 6 átomos (Cγ da tirosina) |
| **N** | Nitrogênio sem H (N da cadeia principal da prolina) |
| **NC1** | Nitrogênio conectado a 1 H em um grupo de carga (Nε da arginina) |
| **NC2** | Nitrogênio conectado a 2 H em um grupo de carga (Nη2 da arginina) |
| **NC3** | Nitrogênio conectado a 3 H em um grupo de carga (Nζ2 da lisina) |
| **NH1** | Nitrogênio conectado a 1 H (N da cadeia principal) |
| **NH2** | Nitrogênio conectado a 2 H (Nε2 da glutamina) |
| **NRD5** | Nitrogênio sem H em um anel de 5 átomos (Nε2 da histidina) |
| **NR15** | Nitrogênio conectado a 1 H em um anel de 5 átomos (Nδ1 da histidina) |
| **O** | Oxigênio fora de uma cadeia com carga (O da cadeia principal) |
| **OC** | Oxigênio em uma cadeia com carga (Oε1 e Oε2 do ácido glutâmico) |
| **OH1** | Oxigênio de grupo alcoólico (Oγ1 da treonina) |
| **OH2** | Oxigênio da água |
| **S** | Enxofre sem H |
| **SH1** | Enxofre com 1 H |

**Tabela 2.8.3.I: PROVE -** os grupos de átomos considerados pela ferramenta *PROVE*.

## 2.8.4. PROCHECK

O programa gera um relatório em formato texto que discrimina valores para diversas propriedades estereoquímicas da estrutura de entrada **[LMM$^+$93]**. Os resultados também podem ser obtidos nos formatos de gráficos e de tabelas, denominados de *plots,* não utilizados neste projeto.

Os valores para cada um dos resíduos são comparados com outros, considerados ideais, originados a partir dos estudos de Morris *et al.* **[MMH$^+$92]** e Engh & Huber **[EH91]**, sobre estruturas de proteínas de alta resolução. As propriedades foram escolhidas para o projeto, exatamente as duas aplicadas na comparação do programa *PROVE*, são mostradas abaixo:

- Ângulo diedro $\chi_1$: ângulo diedro formado pelos átomos de N, Cα, Cβ e Cγ, podendo estar em 3 conformações – *trans*, *gauche*$^-$ e *gauche*$^+$ **(Figura 2.8.4.I: $\chi_1$).**

- Quiralidade do Cα: valor do ângulo de torção, definido pelos átomos Cα, N, C, e Cβ.

Para os aminoácidos serem considerados defeituosos, é necessário que ao menos umas das duas propriedades apresentem uma diferença em relação ao valor ideal de pelo menos duas vezes o desvio padrão.

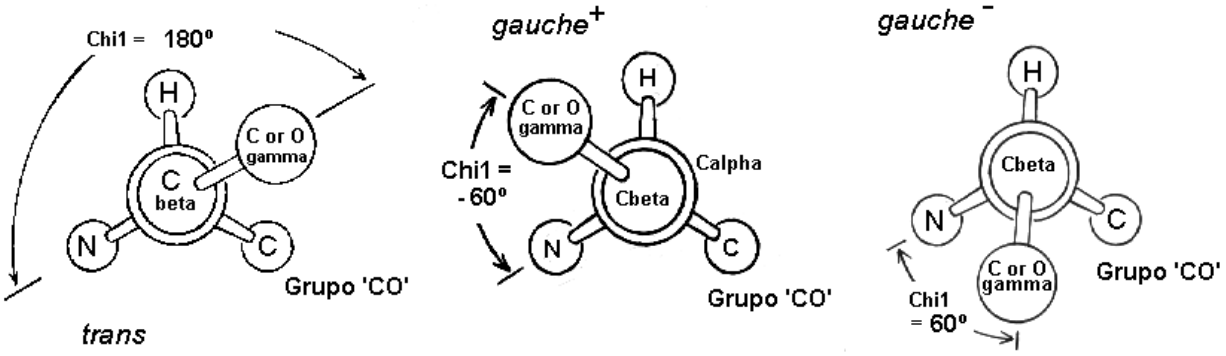



**Figura2.8.4.I: $\chi_1$ -** as diversas conformações apresentadas pelo ângulo diedro $\chi 1$.

## Seção 2.9. Reconstrução de uma cadeia polipeptídica

A reconstrução da cadeia principal, ou *backbone*, de uma proteína, pode ser implementada a partir da trajetória espacial dos carbonos-α, ou seja, com as coordenadas dos carbonos-α informadas, ou por *Ramachandran*.

O primeiro método é aplicado, sobretudo, nos estágios iniciais da difração por Raios-X, quando apenas um mapeamento dos carbonos-α é gerado; nas seqüências do *PDB* que apresentam apenas os carbonos-α; ou em processos de modelagem. As principais metodologias para o desenvolvimento desse primeiro método são:

- Analítica: aplica relações trigonométricas para posicionar os demais átomos da cadeia principal da proteína (Cβ, N, O) a partir de valores padrão dos ângulos e das distâncias entre os átomos da cadeia principal **[RS92],** obtidos da análise de um conjunto de estruturas bem formuladas;
- Baseada em conhecimento: a seqüência de Cα é dividida em vários segmentos. Em seguida, um banco de dados é percorrido na busca por modelos desses trechos baseados nas posições dos Cα's e nos tipos de aminoácidos. As seqüências retornadas da busca são ordenadas e usadas como moldes para o posicionamento dos demais átomos da estrutura submetida à reconstrução

**[Adc04]**. Dificuldades podem aparecer quando o banco não apresenta uma boa cobertura de seqüências semelhantes à cadeia de entrada;

- Minimização de energia: posiciona os átomos numa conformação inicial para aplicar, em seguida, a otimização de uma função de energia, seja por dinâmica molecular, *simulated annealing*, Monte Carlo **[KLS02]** ou minimização local, submetendo a proteína a diversas reconstruções. Infelizmente, o tempo computacional gasto pode ser muito grande quando comparado com as demais técnicas.

O método de *Ramachandran* é bastante semelhante ao processo analítico descrito acima, contando com informações relacionadas à distância entre os átomos e os ângulos de ligação, obtidas da análise de banco de dados de estruturas bem formuladas, e aplicação de relações trigonométricas. A diferença é que as posições dos Cα não são fornecidas, apenas os valores dos ângulos diedros (Φ, ψ, ω) entre os átomos são dados.

## Seção 2.10. Gerando conformações de polipeptídios

Esta seção descreve algumas técnicas aplicadas na construção de um conjunto de diversas conformações para uma mesma seqüência polipeptídica, sendo as estruturas formadas aplicadas em processos de predição de proteínas (RMN, cristalografia por Raios-X, *ab initio* ou homologia), na análise do poder de seleção da melhor estrutura das ferramentas que trabalham com potencial estatístico e força molecular, entre outras aplicaçoes. As metodologias desenvolvidas podem ser divididas em três categorias:

- **Método analítico:** o objetivo é a resolução de um polinômio de alta ordem formado por termos dependentes de restrições geométricas, como ângulos e distâncias de ligação, e da posição fixa de aminoácidos situados nos extremos da cadeia polipeptídica, denominados aminoácidos âncoras. Apesar de apresentar uma boa eficiência, essa metodologia é limitada pelo tamanho da conformação que consegue resolver: dois a três aminoácidos. Acima dessa quantidade, o grau de liberdade das equações, e conseqüentemente a aplicação de novas restrições, cresce, fazendo com que o processo se torne bastante demorado. Entretanto, alguns trabalhos buscam conformações para cadeias pela combinação das soluções das cadeias menores **[WS99]**;

- **Métodos baseados em conhecimento:** a partir da proteína alvo, busca-se, num banco de dados, estruturas ou seqüências compatíveis **[DB01]**. Em seguida, extrai-se as regiões similares com o fragmento desejado da proteína em estudo, construindo um conjunto de conformações. As dificuldades enfrentadas estão relacionadas com o tamanho limite das seqüências (entre três e oito resíduos), além do tamanho e da diversificação do conjunto, devido aos limites de cobertura das bases de dados.
- ***Ab initio*:** apresenta-se de duas formas: através da busca discreta **[DB00]**, com enumeração exaustiva de novas conformações e aplicando uma função de energia simplificada, como campo de força ou potencial estatístico, para excluir modelos de baixa energia, funcionando apenas como um filtro de seleção; ou com modelagem a partir da otimização da função de energia, onde esta é responsável pela construção dos modelos, aplicando dinâmica molecular **[FDS00]**, Monte Carlo ou *simulated annealing*. Em ambas as formas, há a modificação dos ângulos de ligação e das distâncias entre os átomos.

É claro que podem existir mecanismos que combinem essas técnicas, associados de forma independente, como o *CODA*, que executa, separadamente, programas que geram conformações por *ab initio* (*PETRA*) e baseadas em conhecimento prévio (*FREAD*) para em seguida combinar os resultados; ou intercalar os algoritmos, como o realizado por Martin *et al.* **[MCR89]**.

## Seção 2.11. Construção das cadeias laterais

A maioria das metodologias que se dedicam à construção de novas conformações não levam em consideração os elementos das cadeias laterais. Mas, para a obtenção de um modelo final é necessário que ele esteja completo. Em outras palavras, em algum momento, o restante dos átomos deverá ser inserido.

Muitos métodos se dedicam à inserção das cadeias laterais nos *backbones*, a maioria aplicado via bibliotecas de rotâmeros, conformações de cadeias laterais de baixa energia, obtidas por análises estatísticas em bancos de dados como o *PDB*, podendo estar organizados em função dos ângulos diedros $\Phi$ e $\psi$ da cadeia principal ou não. Para esse método, são desenvolvidos um algoritmo de busca, para selecionar as conformações da cadeia lateral de acordo com a seqüência de aminoácidos do *backbone*, e uma função de energia para guiar essa busca. Caso o algoritmo de busca garanta

encontrar a energia mínima global para uma biblioteca, ele é dito exato. Um exemplo de algoritmo exato é o *dead-end elimination,* que elimina os rotâmeros que não podem ser membros do processo de minimização, reduzindo o espaço de busca, até que uma procura exaustiva possa ser efetuada. Outros algoritmos, como métodos por Monte Carlo, buscam apenas uma conformação de baixa energia, não necessariamente a mínima global.

Durante o projeto, foi aplicado o programa *SCWRL* **[CSD03]** na construção de cadeias laterais, que utiliza uma biblioteca de rotâmeros dependente dos ângulos diédricos da cadeia principal. Os passos aplicados na execução dessa ferramenta são:

- Leitura da cadeia principal e medição dos ângulos diedros Φ e ψ;
- Busca dos rotâmeros para cada um dos aminoácidos de acordo com os ângulos Φ e ψ e o tipo do aminoácido, até que a densidade cumulativa (soma das probabilidades) dos rotâmeros para o aminoácido atinja 90%. A leitura da biblioteca é feita da mais alta para a mais baixa probabilidade dos rotâmeros associados ao aminoácido em questão;
- Cálculo das coordenadas cartesianas para todos os rotâmeros e da energia associada a cada um. A função de enegia **(Figura 2.11.I: Energia SCWRL)** é composta por dois termos: o primeiro depende da probabilidade do rotâmero em função dos ângulos diedros Φ e ψ; o segundo consiste da interação do rotâmero do i-ésimo aminoácido com a cadeia principal;

$$E_{bb}(r_i) = -K \log \frac{p(r_i|\phi,\psi)}{p(r_i = 1|\phi,\psi)} + \sum_{\substack{j<i-1 \\ j>i+1}} E_{sb}(r_i, j)$$

**Figura 2.11.I: Energia SCWRL –** Energia associada a um rotâmero $\mathbf{r_i}$ do **i-ésimo** aminoácido da seqüência da cadeia principal. **K** é uma constante, normalmente com o valor 3,0, baseado na otimização da função de energia para um conjunto de 180 proteínas, e **j** é o j-ésimo aminoácido da cadeia principal.

- Seleção dos melhores rotâmeros para cada aminoácido, onde o rotâmero $\mathbf{s_i}$ para o i-ésimo aminoácido é descartado se existir outro rotâmero $\mathbf{r_i}$ que sempre apresente uma energia de interação menor com todos as outras cadeias laterais e com a principal;
- Aminoácidos que, após o último passo, apresentarem mais de um rotâmero são denominados "ativos", sendo associados a vértices de um grafo. Caso a energia

de interação entre rotâmeros de aminoácidos ativos seja diferente de zero, uma aresta conecta esses dois aminoácidos.

- Determinação dos componentes biconexos e dos pontos de articulação, que conectam dois componentes biconexos;
- Busca da energia mínima para os componentes com um único ponto de articulação, considerando-se todas as possíveis combinações dos rotâmeros do componente para cada um dos rotâmeros do ponto de articulação. Em seguida, agrega-se o componente no ponto de articulação para resolver o próximo componente biconexo. Por exemplo, a **Figura 2.11.II: Resolvendo Componente Biconexo** apresenta cinco componentes biconexos. O primeiro a ser resolvido é o A com a formação do super-aminoácido K17-V10-R15-Y21. Caso o ponto de articulação K17 apresente oito rotâmeros, o novo super-aminoácido terá oito possíveis rotâmeros, a serem usados na resolução do componente B. Em seguida, os componentes E, D e B são solucionados.

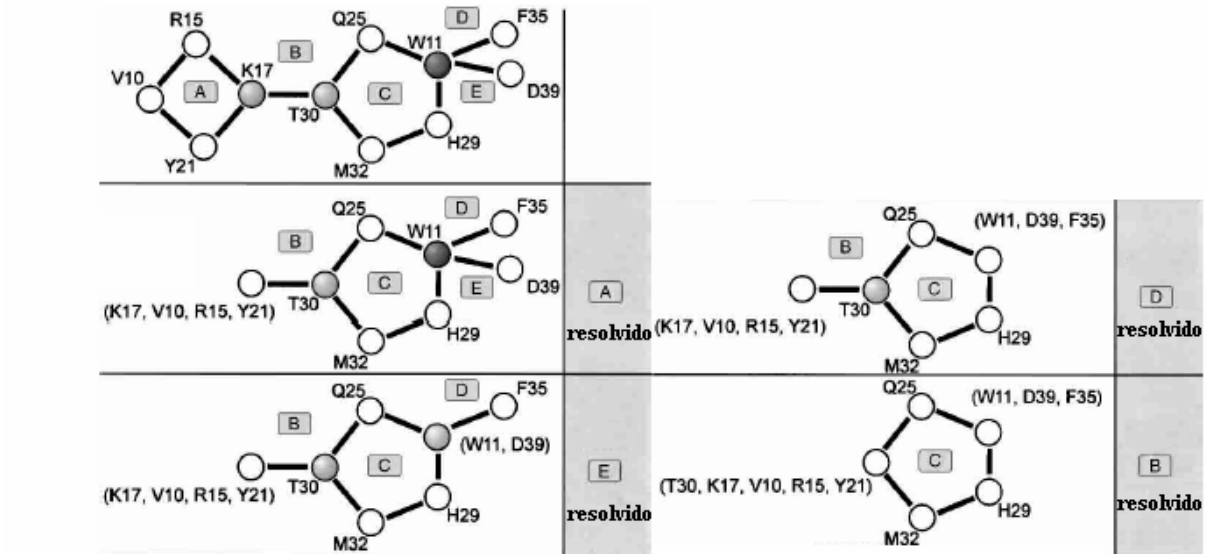



**Figura 2.11.II: Resolvendo Componente Biconexo** – Passos na resolução do problema de componentes biconexos.

# Capítulo 3. Metodologia

"That's probably a good Idea."
Jules – Pulp Fiction (1994)

Este capítulo descreve os procedimentos e algoritmos adotados no desenvolvimento do projeto que consiste de um ciclo dividido em três fases: a modelagem, a avaliação e a construção de novos modelos de proteínas. Tais fases podem ser executadas em conjunto ou de forma independente uma das outras

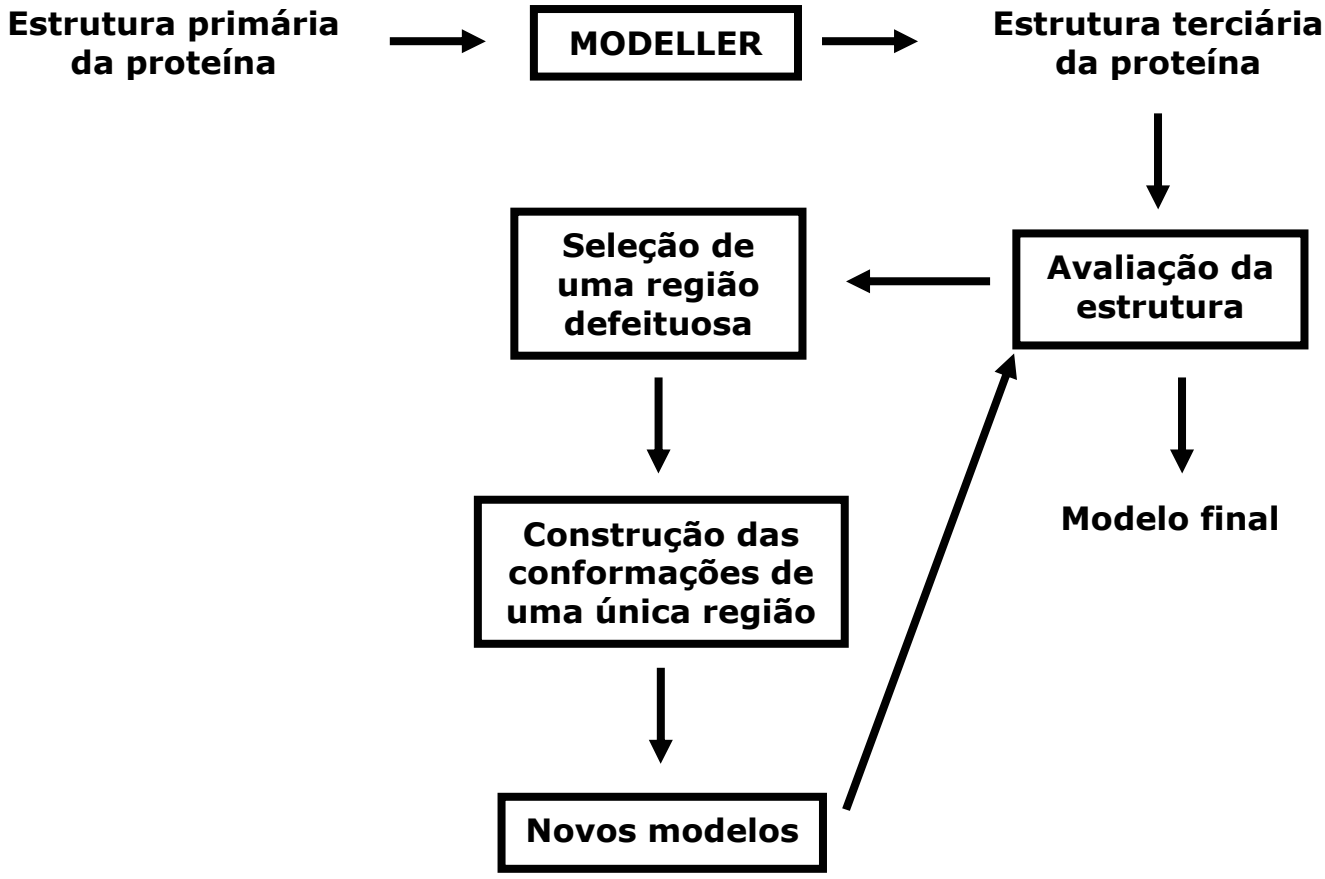



**Figura 3.I: Ciclo -** Representação do fluxo do projeto.

Para iniciar a execução de todo o ciclo, a estrutura primária da proteína alvo é fornecida, no formato *FASTA* **(Apêndice FASTA)**, como entrada para o programa, sendo sua estrutura modelada em seguida através do MODELLER. A partir do arquivo da seqüência alvo no formato *PDB*, construído na modelagem, faz-se uma busca e seleção por regiões defeituosas. Em seguida, seleciona-se uma única região defeituosa e constroem-se novas conformações para esta região. Essas novas conformações são substituídas no modelo original, gerando um conjunto de possíveis modelos. Estes

modelos são classificados e a estrutura com a melhor conformação é escolhida como o novo modelo. Nesse ponto, o procedimento pode ser repetido, selecionando-se novamente regiões defeituosas e seguindo os mesmos passos, ou ser finalizado, de acordo com a precisão do modelo desejada pelo usuário **(Figura 3.I: Ciclo)**.

A ferramenta foi construída para um ambiente Unix, podendo ser estendida a uma interface *Web*, com as linguagens de programação *Java* 1.4.2, sob a plataforma *Eclipse* 2.1.2, que permitiu a criação de um ambiente orientado a objeto (OO) na representação da estrutura das proteínas; e *Perl* 5.8.4, com o editor *UltraEdit*-32 5.0a, utilizada por facilitar a manipulação dos arquivos textos e execução de outros programas. O projeto é executado em uma máquina *Pentium* 4 2.40 *Ghz*, com 1 *GB* de *RAM*, rodando *Slackware Linux* versão 10.0 com *kernel* 2.6.6.

## Seção 3.1. Modelagem de proteínas

A primeira fase do programa executa a ferramenta *MODELLER* versão 7.7, com os parâmetros *default*. A entrada é um arquivo texto no formato *FASTA*, contendo a seqüência primária da proteína a ser modelada. A finalidade dessa etapa, além de criar a conformação para a proteína, é facilitar a aplicação da ferramenta de modelagem, pois esta apresenta diversos passos seqüenciais a serem executados e arquivos intermediários que precisam ser construídos. Para isto, os arquivos intermediários são criados com a mínima interferência dos usuários, além de executar todos os passos automaticamente, dando a impressão que toda a construção do modelo ocorre em um único processo. A execução é, na verdade, dividida em sete blocos:

- A partir da seqüência de entrada, constroem-se dois arquivos intermediários: o primeiro com as informações presentes no arquivo FASTA, e o segundo com os parâmetros aplicados na busca das seqüências similares, dentre eles o banco de dados representativo do *PDB*. Este processo é automático, sem a intervenção do usuário;
- Escolha das proteínas semelhantes à alvo: são mostrados ao usuário os códigos das proteínas mais similares com a alvo, obtidas no passo anterior e que pertencem ao conjunto representativo do *PDB*. As proteínas estão ordenadas de forma decrescente quanto à similaridade, assim como outras proteínas que não fazem parte do conjunto representativo, mas que apresentam uma similaridade de seqüência maior que 40% a pelo menos uma das seqüências similares à alvo.

O critério para a seleção das proteínas similares à alvo que devem ser mostradas ao usuário é baseado no trabalho de Fiser e Šali **[FS01]**. O usuário indica qual(is) quer utilizar na modelagem;

- Acesso automático ao banco de dados *PDB* para o *download* das proteínas selecionadas (ftp://ftp.rcsb.org/pub/pdb/data/structures/all/pdb/);
- Alinhamento entre as seqüências similares e a alvo: inicialmente, um alinhamento múltiplo estrutural entre as seqüências selecionadas no bloco anterior é construído, sendo, em seguida, utilizado para o alinhamento entre as estruturas e a seqüência alvo. Todos os alinhamentos são executados sem a intervenção do usuário;
- Construção do *dendograma*, uma árvore que permite a visualização da distância de similaridade entre as seqüências utilizadas no alinhamento;
- Seleção das estruturas moldes a partir do *dendograma*: é informada ao usuário a árvore de distância de similaridade construída anteriormente, e este escolhe quais estruturas serão aplicadas como moldes;
- Construção e seleção do melhor modelo automaticamente: nesta fase, cinco modelos são criados, sendo escolhido aquele que apresentar a menor energia de acordo com os cálculos do *MODELLER*.

## Seção 3.2. Avaliação das estruturas

Nesta fase, o usuário indica qual(is) ferramentas de avaliação de estruturas (*ANoLEA*, *ProSa*, *PROVE* ou *PROCHECK*) serão utilizada na análise do modelo da proteína, modelo este representado por um arquivo no formato *PDB*. A saída apresenta dois tipos de arquivo: um para cada ferramenta individualmente, indicando as regiões defeituosas **(Figura 3.2.I: Regiões Individuais)**; outro mostrando graficamente as regiões selecionadas por cada uma das ferramentas, e indicando a região que apresenta o maior consenso entre todos os avaliadores aplicados **(Figura 3.2.II: Todas as Regiões)**. Para determinar a região consenso, é associada uma pontuação (P) a cada posição da seqüência (i) de acordo com o peso das ferramentas (p) da seguinte forma:

$$P(i) = p_{ANoLEA} * f_{ANoLEA}(i) + p_{ProSa} * f_{ProSa}(i) + p_{PROVE} * f_{PROVE}(i) + p_{PROCHECK} * f_{PROCHECK}(i)$$

**( Eq. 3.2.I )**

onde o $p_{ferramenta}$ será zero caso ela não seja selecionada para a avaliação; $f_{ferramenta}(i)$ é igual a 1 se a posição **i** for indicada como defeituosa pela ferramenta e 0 caso contrário. A região consenso de tamanho **s** preterida é aquela de maior pontuação, o qual é dada pela Equação 3.2.II, onde **i** representa valores entre 1 e **L–s**, com **L** correspondendo ao tamanho máximo da seqüência. Os pesos apresentam um valor padrão igual a 1, e **s** um valor igual a 5, porém esses parâmetros podem ser alterados pelo usuário.

$$P_s(i) = \sum_{n=i}^{n=i+s} P(n) \qquad (\text{Eq. 3.2.II})$$

```
     ANOLEA          |      PROSA          |      PROVE
                     |                     |
Número de regiões: 5 | Número de regiões: 2 | Número de regiões: 5
INICIO FIM           | INICIO FIM           | INICIO FIM
     6-13            |      5-13            |     22-22
    20-30            |     22-49            |     29-30
    32-43            |                      |     37-37
    47-47            |                      |     43-43
    49-50            |                      |     47-47
```

**Figura 3.2.I: Regiões Individuais** - Exemplos de três arquivos de saída, cada um correspondendo individualmente a uma ferramenta de avaliação; nesse caso, ANoLEA, ProSa e PROVE.

```
          5         15        25        35        45
          ---------------------------------------------
anolea    -********-------**********-***********---*-**
prosa     *********--------***************************-
prove     -----------------*------**------*-----*---*---
          12222222200000011322222331222223222223111312 1

Região de maior pontuação: 26 30
```

**Figura 3.2.II: Todas as Regiões** – Cada aminoácido indicado por alguma das ferramentas é apresentado por um '*'. A linha abaixo da representação das regiões das ferramentas indica a pontuação de cada uma das posições da seqüência, onde os pesos para as ferramentas foi mantido com o valor *default* igual a 1.

Para que o processo continue é necessário que o usuário indique qual região deve ser considerada como problemática.

## Seção 3.3. Construção dos novos modelos

Neste último estágio, ocorre a implementação do processo de criação de novas conformações para o segmento de polipeptídio, selecionado na etapa anterior, pela metodologia *ab initio*; e a seleção do melhor modelo, por campos de força. Inicialmente,

constrói-se uma nova estrutura do fragmento da proteína por *Ramachandran*, selecionando-se, aleatoriamente, um par de valores para os ângulos diedros ψ e Φ para cada aminoácido. Caso o modelo gerado seja aprovado por todos os pré-requisitos presentes no algoritmo, a estrutura fará parte do conjunto de novas conformações do fragmento da proteína. As melhores conformações serão selecionadas e substituídas no modelo original, e os novos modelos que apresentarem os menores valores de energia serão apresentados ao usuário.

### 3.3.1. Obtenção do conjunto de pares de ângulos diedros ψ/Φ

Para que se possa construir uma conformação de boa qualidade de um segmento de proteína, pode-se fixar os ângulos e as distâncias de ligação, variando-se apenas os possíveis ângulos diedros ψ e Φ adotados pelos aminoácidos. No entanto, não é comum, durante a formação de modelo, permitir-se que os ângulos assumam qualquer valor no intervalo (–180º, +180º]. Normalmente monta-se um conjunto com os possíveis pares de ângulos a serem aplicados, a partir da aproximação dos valores obtidos de análises de banco de dados de estruturas.

Diversos estudos dedicaram-se a criar tais conjuntos de estados discretos dos ângulos diedros, sejam com a intenção de modelar apenas os laços ou toda a estrutura da proteína. Esses trabalhos resultaram na apresentação de poucos pares, como no projeto desenvolvido por Moult & James **[MJ86]**, que apresenta 11 elementos; ou até de 5184 pares **[DBL$^+$03]**. Neste último trabalho, fica evidente que os conjuntos maiores implicam na construção de melhores modelos, pois permitem atingir um maior espaço conformacional para o segmento da proteína.

O conjunto de pares de ângulos diedros aplicado no projeto aqui descrito foi o mesmo utilizado no trabalho desenvolvido por DePristo *et al.* **[DBL$^+$03]**, derivado da base de dados Top500 **[Top500]**, formada por seqüências originadas da cristalografia de Raios-X com resolução igual a 1,8 Å ou menor. Apenas as regiões que não apresentavam a conformação de folha-β ou α-hélice foram consideradas, fazendo com que o conjunto fosse voltado para a modelagem dos laços.

Os pares de valores dos ângulos diedros obtidos foram separados de acordo com o tipo do aminoácido. E para cada aminoácidos os pares de ângulos foram agrupados em conjuntos de precisão de 5º. Ou seja, esses pares foram organizados em conjuntos em que a diferença entre eles é de 5º, ou múltiplos de 5º, em, pelo menos, um dos dois ângulos, o que permite formar 5184 pares dos ângulos ψ e Φ para cada aminoácido. Os

conjuntos que não estavam associados a nenhum par foram descartados (**Tabela 3.3.1.I: Conjuntos ψ/Φ**).

| Aminoácido | Número de conjuntos associados |
|---|---|
| ALA | 3055 |
| ARG | 2582 |
| ASN | 3082 |
| ASP | 3275 |
| CYS | 2541 |
| GLN | 2573 |
| GLU | 2491 |
| GLY | 4264 |
| HIS | 2748 |
| ILE | 2229 |
| LEU | 2737 |
| LYS | 2658 |
| MET | 2268 |
| PHE | 2523 |
| PRO | 1207 |
| SER | 3203 |
| THR | 2870 |
| TRP | 2257 |
| TYR | 3497 |
| VAL | 2403 |

**Tabela 3.3.1.I: Conjuntos ψ/Φ** – O número de conjuntos associados a cada um dos aminoácidos que apresentaram pelo menos um par de ângulo ψ/Φ originados do Top500 com precisão de 5º.

Ainda decorrente do trabalho de DePristo, a análise do banco Top500 mostrou que o ângulo diedro ω normalmente se encontra na posição *trans*, sendo de apenas 0,05% os casos que apresentam a conformação *cis* para aminoácidos diferentes da Prolina, e de 5,0% para a Prolina. Assim, na construção das novas conformações, o valor adotado para o ω foi de 180º.

## 3.3.2. Representação da estrutura da proteína

Durante o processo de construção das novas conformações para um segmento da proteína, apenas os átomos pesados (C, O, N, S) foram utilizados. Para a região defeituosa, apenas os átomos da cadeia principal e o primeiro carbono do grupo lateral, caso exista, foram considerados, enquanto que para o restante da proteína, todos os átomos pesados foram utilizados, independentemente de sua localização, quer seja na cadeia principal ou lateral.

A representação computacional da região defeituosa se baseou nos padrões adotados pelo projeto BioJava **[BioJava]**, trabalho de código aberto que oferece estrutura em Java para processamento de dados biológicos. A implementação apresentou a seguinte organização:

- **Elemento Químico**: estrutura do programa que pode representar qualquer elemento químico da tabela periódica. Cada elemento apresenta o nome, o símbolo e o raio de Van der Walls, cujos valores foram obtidos do trabalho de Word *et al.* **[WLL+99]** (**Tabela 3.3.2.II: Raio de Van der Waals**).

| Elemento Químico | Raio (Å) |
|---|---|
| Carbono | 1,75 |
| Carbono da carboxila | 1,65 |
| Nitrogênio | 1,55 |
| Oxigênio | 1,4 |
| Enxofre | 1,8 |

**Tabela 3.3.2.II: Raio de Van der Waals** - Valores do raio de Van der Waals aplicados aos átomos durante o projeto.

- **Alfabeto**: conjunto de todos os tipos de **elementos químicos** utilizados na representação de um dos 20 diferentes aminoácidos presentes nas proteínas. Na definição do Alfabeto, adotou-se o padrão utilizado na nomenclatura aplicada aos átomos dos arquivos PDB, definindo-se 35 Elementos Químicos (**Tabela 3.3.2.II: Alfabeto**).
- **Átomo**: um **elemento químico** associado a coordenadas cartesianas, representando um átomo de um aminoácido e sua posição espacial. No projeto, o átomo é visto como uma esfera cujo centro está localizado nas coordenadas cartesianas do átomo, sendo o raio igual ao raio de Van der Waals.
- **Aminoácido**: conjunto de **átomos** representando um dos 20 tipos de aminoácidos encontrado nas proteínas. Os **elementos químicos** presentes nos **átomos** dos **aminoácidos** são apenas aqueles definidos no **alfabeto**.
- **Região**: formado por vários **aminoácidos**, simulando uma cadeia polipeptídica.

Uma caixa contendo toda a proteína foi construída e dividida em diversos cubos congruentes e não sobrepostos, que foram denominados de **vértice**s. Cada **átomo** da proteína não presente na região defeituosa foi associado às listas de átomos dos **vértices**

que interceptavam **(Figura 3.3.2.I: Grade2D)**. Assim, um mesmo átomo da proteína estar preso aos vértices próximos a ele.

| Elemento Químico | |
|---|---|
| **Símbolo** | **Nome** |
| CA | carbono α |
| CB | carbono β |
| CC | carbono carboxila |
| CG, CG1, CG2 | carbono γ, γ1, γ2 |
| CD, CD1, CD2 | carbono δ, δ1, δ2 |
| CE, CE1, CE2, CE3 | carbono ε, ε1, ε2, ε3 |
| CZ, CZ1, CZ2 | carbono ζ, ζ1, ζ2 |
| CH2 | carbono η2 |
| N | nitrogênio |
| NE, NE1, NE2 | nitrogênio ε, ε1, ε2 |
| NZ | nitrogênio ζ |
| NH1, NH2 | nitrogênio η1, η2 |
| O | oxigênio |
| OG, OG1 | oxigênio γ, γ 1 |
| OD1, OD2 | oxigênio δ1, δ2 |
| OE1, OE2 | oxigênio ε1, ε2 |
| OH | oxigênio η |
| S | enxofre |
| SG | enxofre γ |
| SD | enxofre δ |

**Tabela 3.3.2.II: Alfabeto** - Os símbolos e os nomes dos elementos químicos usados na representação dos aminoácidos.

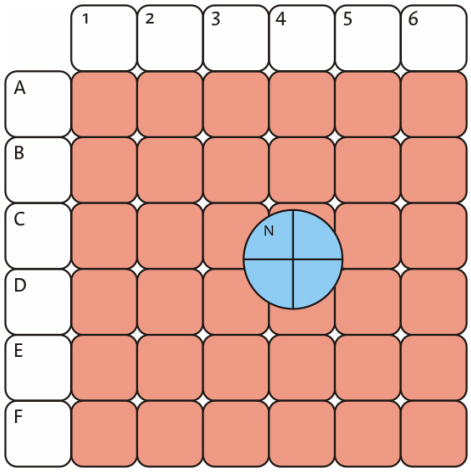



**Figura 3.3.2.I: Grade2D** - Nessa representação bidimensional da grade que cobre toda a proteína, o átomo N foi inserido na lista de átomos dos vértices (3C, 4C, 5C, 3D, 4D e 5D) que ele intercepta.

### 3.3.3. Técnica de construção das conformações

O processo aplicado neste projeto para a montagem dos novos modelos foi baseado na técnica de *Ramachandran*, metodologia que constrói uma conformação a partir de valores padrão de ângulos e distâncias entre os átomos. Os valores aqui aplicados foram obtidos no trabalho desenvolvido por Engh & Huber **[EH91] (Tabela 3.3.3.I: Engh & Huber)**.

| **A. Distâncias de ligação (Å)** | | |
|---|---|---|
| **Ligação** | **Aminoácido** | **Valor** |
| C – N | Pro | 1,341 |
| | os demais | 1,329 |
| C – O | todos | 1,231 |
| Cα – C | Gly | 1,516 |
| | os demais | 1,525 |
| Cα – Cβ | Ala | 1,521 |
| | Ile, Thr, Val | 1,540 |
| | os demais | 1,539 |
| N – Cα | Gly | 1,451 |
| | Pro | 1,466 |
| | os demais | 1,458 |
| **B. Ângulos de ligação (º)** | | |
| **Ângulo** | **Aminoácido** | **Valor** |
| C - N – Cα | Pro | 122,6 |
| | Gly | 120,6 |
| | os demais | 121,7 |
| Cα - C – N | Pro | 116,9 |
| | Gly | 116,4 |
| | os demais | 116,2 |
| Cα - C – O | todos | 120,8 |
| Cβ - Cα – C | Ala | 110,5 |
| | Ile, Thr, Val | 109,1 |
| | os demais | 110,1 |
| N - Cα – C | Gly | 112,5 |
| | Pro | 111,8 |
| | os demais | 111,2 |
| N - Cα – Cβ | Ala | 110,4 |
| | Ile, Thr, Val | 111,5 |
| | Pro | 103,0 |
| | os demais | 110,5 |
| O - C – N | Pro | 122,0 |
| | os demais | 123,0 |

**Tabela 3.3.3.I: Engh & Huber –** A) Os valores em Å para os comprimentos das ligações entre dois átomos pertencentes à cadeia principal ou o Cβ; B) Os valores em º (graus) para os ângulos entre os átomos da cadeia principal

Antes de iniciar a construção da nova conformação, a região é aumentada em quatro aminoácidos, colocando-se dois aminoácidos em cada extremidades, correspondentes aos aminoácidos vizinhos da região na proteína. Esses aminoácidos são denominados aminoácidos âncoras da região.

Para uma região de tamanho **n** após a inserção dos quatro aminoácidos âncoras, e sendo $\mathbf{A_i}$ um átomo do tipo **A** do **i-ésimo** aminoácido da seqüência, o processo se inicia considerando-se que as duas primeiras âncoras já estão presentes na nova conformação e na mesma posição em que se encontram na proteína, correspondendo aos dois primeiros aminoácidos do novo modelo. Em seguida, o $N_3$ tem sua posição determinada pelo ângulo diedro ψ e os átomos de $N_2$, $C\alpha_2$ e C da carboxila - da segunda âncora - $Cc_2$. O $O_2$ é re-posicionado, de acordo com a planaridade dos átomos $C\alpha_2$, $Cc_2$ e o $N_3$.

O $C\alpha_3$ é inserido em seguida, considerando-se o ângulo diedro ω e os átomos $C\alpha_2$, $Cc_2$ e $N_3$, assim como $Cc_3$, pelo diedro Φ e os átomos $Cc_2$, $N_3$ e $C\alpha_3$. Por fim, o $C\beta_3$ é ajustado, guiado pela quiralidade existente entre ele e os átomos $N_3$, $C\alpha_3$ e $C_3$. O $N_4$ do quarto resíduo é posicionado, assim como o $O_3$. Este processo continua até a inserção do aminoácido de ordem **n-1**, o chamado resíduo *dummy* **(Figura 3.3.3.I: Ramachandran)**.

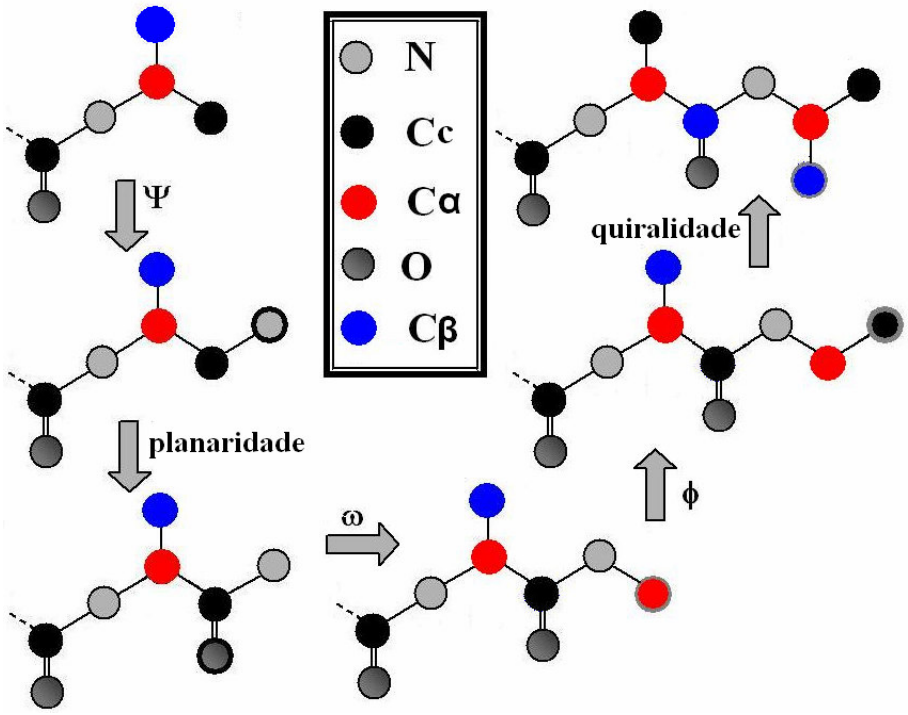



**Figura 3.3.3.I: Ramachandran -** Inicia com a inserção do N de acordo com o ψ, seguido pela atualização da posição do O pela planaridade. O Cα é determinado pelo diedro ω e o Cc pelo Φ. Por fim, posiciona o Cβ pela quiralidade.

A cada inserção de um novo aminoácido, um par de ângulos diedros Φ/ψ é selecionado aleatoriamente dos conjuntos descritos na sub-seção 3.3.1. Em contrapartida, o ângulo ω está associado a 180º em todo o processo.

### 3.3.4. Inserção de um átomo pelo ângulo diedro

Para inserir os átomos de N, Cα e C da carboxila (Cc) de um aminoácido durante a construção de uma nova conformação para uma região defeituosa pelo método de *Ramachandran* descrito acima, é preciso utilizar os ângulos diedros ψ, ω e Φ respectivamente. Assim, este tópico tem por objetivo explicar como inserir um átomo numa seqüência de aminoácidos pelo ângulo diedro.

Considere $\mathbf{A_4}$ um novo átomo a ser inserido numa conformação pelo ângulo diedro **φ**, formado pela intersecção dos planos determinados pelos átomos $\mathbf{A_1}$-$\mathbf{A_2}$-$\mathbf{A_3}$ e $\mathbf{A_2}$- $\mathbf{A_3}$-$\mathbf{A_4}$, onde ($X_n$, $Y_n$, $Z_n$) correspondem às coordenadas cartesianas do **n**-ésimo átomo (**Figura 3.3.4.I: Diedro $\mathbf{A_1}$-$\mathbf{A_2}$-$\mathbf{A_3}$-$\mathbf{A_4}$)**. O primeiro passo para se determinar a localização o $\mathbf{A_4}$ é deslocar $\mathbf{A_1}$, $\mathbf{A_2}$ e $\mathbf{A_3}$, de forma que $\mathbf{A_3}$ passe a estar na origem do sistema, $\mathbf{A_2}$ na orientação positiva do eixo Z e $\mathbf{A_1}$ no 1º quadrante do sistema de coordenadas XZ. Em seguida, posiciona-se o $\mathbf{A_4}$ de acordo com os ângulos **φ** e **θ.** O diedro **φ** corresponde ao ângulo entre o vetor $\mathbf{A_3A_4}$ e o vetor de orientação positiva do eixo X, com rotação no sentido horário em relação ao eixo Z, enquanto o **θ** corresponde ao ângulo entre os vetores $\mathbf{A_3A_4}$ e $\mathbf{A_3A_2}$ como mostrado na **Figura 3.3.4.II: Posicionando $\mathbf{A_4}$**. Em seguida, retornar-se $\mathbf{A_1}$, $\mathbf{A_2}$, e $\mathbf{A_3}$ às posições originais. Por estar agora conectado com o $\mathbf{A_3}$, o átomo $\mathbf{A_4}$ é levado a uma nova posição, tendo assim determinado as coordenadas de inserção do novo átomo.

Para deslocar $\mathbf{A_3}$ para a origem do sistema, inicialmente aplica-se uma translação $\mathbf{T}(-X_3, -Y_3, -Z_3)$ aos átomos $\mathbf{A_1}$, $\mathbf{A_2}$, e $\mathbf{A_3}$, como visto na **Figura 3.3.4.III: Translação.** As novas coordenadas para os átomos são:

$$\mathbf{A_3} = (X_3\text{-}X_3,\ Y_3\text{-}Y_3,\ Z_3\text{-}Z_3) = (0, 0, 0)$$
$$\mathbf{A_2} = (X_2\text{-}X_3,\ Y_2\text{-}Y_3,\ Z_2\text{-}Z_3) = (X_2', Y_2', Z_2')$$
$$\mathbf{A_1} = (X_1\text{-}X_3,\ Y_1\text{-}Y_3,\ Z_1\text{-}Z_3) = (X_1', Y_1', Z_1')$$

Determina-se o valor do menor ângulo $\mathbf{\theta_2}$ formado entre o vetor $\mathbf{A_3A_2}$ e o vetor de orientação positiva no eixo Z **(Figura 3.3.4.IV: $\mathbf{\theta_2/\varphi_2}$)**, onde $\mathbf{\theta_2}$ = Arc cos ($Z_2'$/ $R_2$), com $R_2$ representando a distância entre $\mathbf{A_2}$ e $\mathbf{A_3}$. Determina-se também $\mathbf{\varphi_2}$, ângulo entre

o vetor $\mathbf{A_3A_2}$ e o 1º quadrante do sistema ZX, sendo positivo na orientação anti-horária em relação a Z, como mostrado na **Figura 3.3.4.IV: $\theta_2/\varphi_2$**, onde:

Se $\theta_2 = \pi/2$ rad
$\cos(\varphi_2) = X_2'/(Z_2' * \text{sen}(\theta_2))$
$\text{sen}(\varphi_2) = Y_2'/(Z_2' * \text{sen}(\theta_2))$

Se $\theta_2 = \pi$ rad ou $\theta_2 = 0$
$\cos(\varphi_2) = 1.0$
$\text{sen}(\varphi_2) = 0.0$

Para qualquer outro valor de $\theta_2$
$\cos(\varphi_2) = X_2'/(Z_2' * \text{tg}(\theta_2))$
$\text{sen}(\varphi_2) = Y_2'/(Z_2' * \text{tg}(\theta_2))$

Nessa situação, $\theta_2$ varia em um intervalo de [0, π] radianos e $\varphi_2$ entre [0, 2π) radianos.

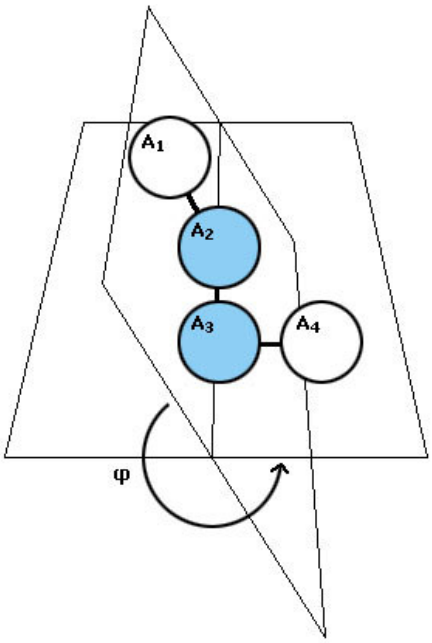



**Figura 3.3.4.I: Diedro $A_1$-$A_2$-$A_3$-$A_4$** - Representação do ângulo diedro **φ** para os átomos $A_1$-$A_2$-$A_3$-$A_4$.

Em seguida, move-se $\mathbf{A_1}$, $\mathbf{A_2}$ e $\mathbf{A_3}$ de forma que $\mathbf{A_2}$ se posicione no 1º quadrante do sistema ZY. Para isto, faz-se a rotação anti-horária desses três átomos em torno do eixo Z pelo ângulo de ($\varphi_2 - \pi$) **(Figura 3.3.4.V: Rotação $\varphi_2 - \pi$)**.

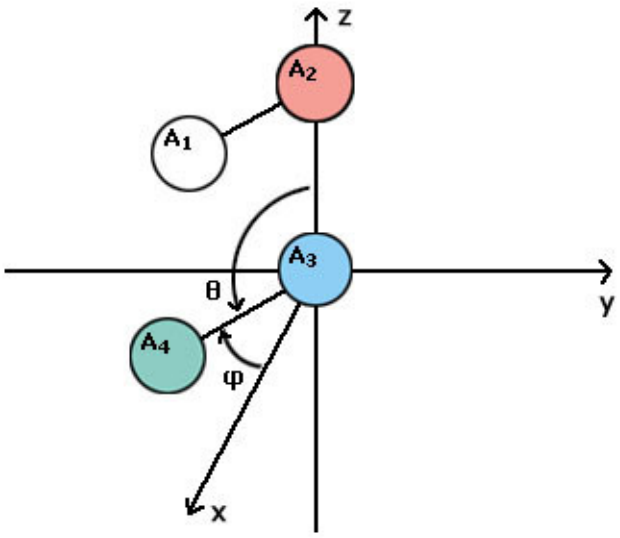



**Figura 3.3.4.II: Posicionando $A_4$** – Posicionamento do átomo $A_4$ no sistema XYZ após os deslocamentos de $A_1$, $A_2$ e $A_3$.

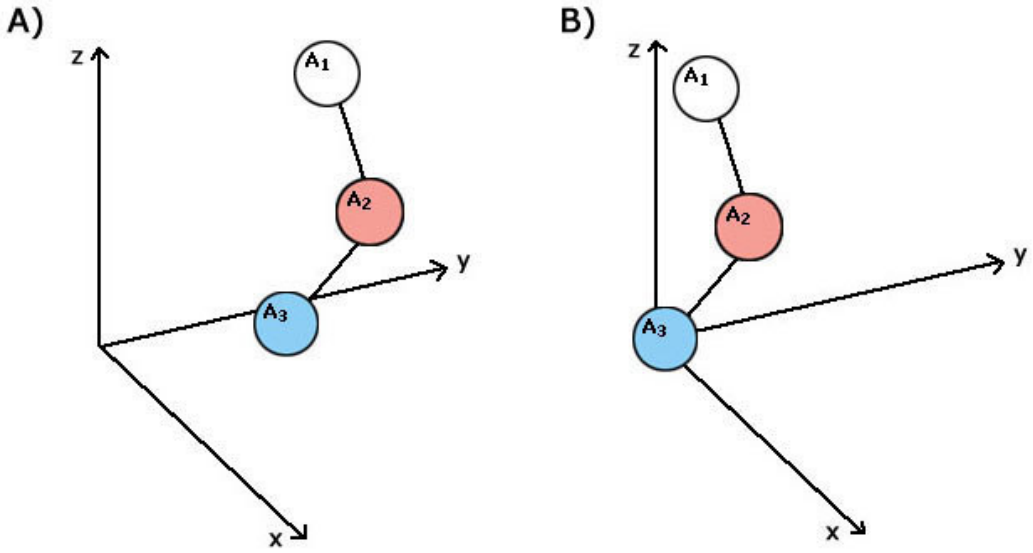



**Figura 3.3.4.III: Translação** - A) Representa os átomos na posição original. B) Átomos após a translação, com o $A_3$ posicionado na origem do sistema.

Para a rotação anti-horária no eixo Z de um átomo **A** por um ângulo **σ**, é preciso multiplicar as coordenadas cartesianas do átomo pela matriz de rotação em Z (**Figura 3.3.4.VI: Rotação Z**).

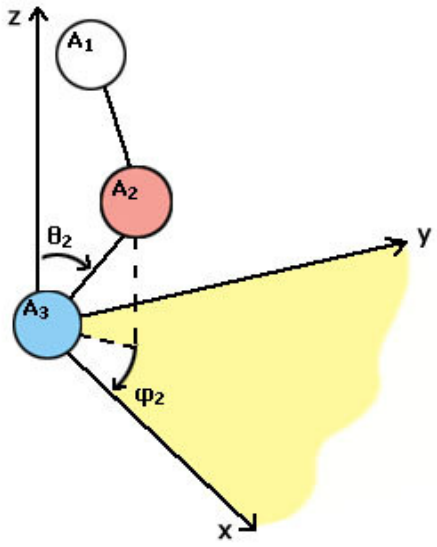



**Figura 3.3.4.IV: $\theta_2/\varphi_2$** - Representação dos ângulos $\theta_2$ e $\varphi_2$ utilizados para posicionar o $A_2$ segundo coordenadas esféricas.

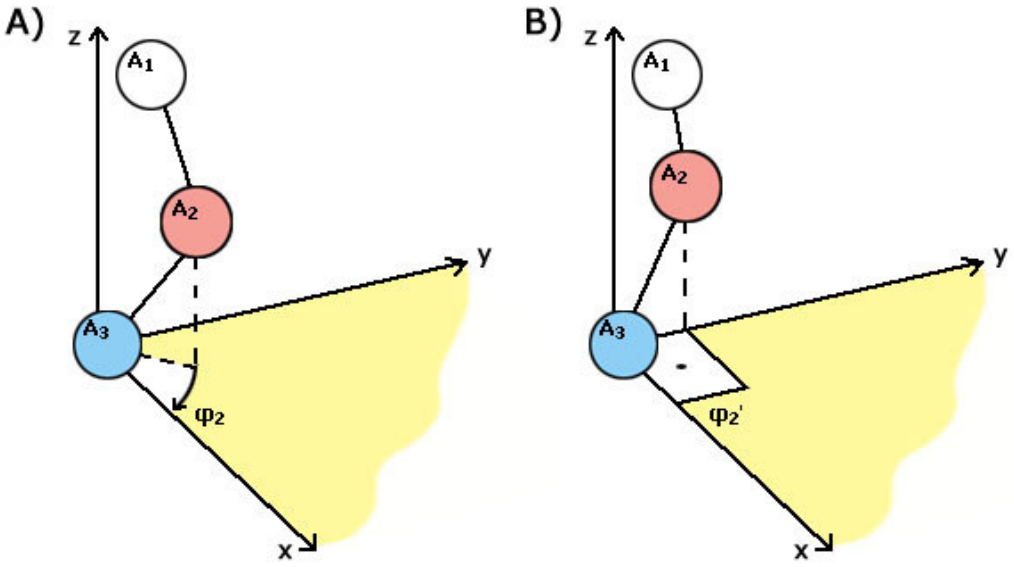



**Figura 3.3.4.V: Rotação $\varphi_2 - \pi$** – A) Posição dos átomos antes da rotação de $\varphi_2 - \pi$; B) Configuração dos átomos após a rotação anti-horária de $\varphi_2 - \pi$ em torno do eixo Z.

$$R_Z(\sigma) = \begin{pmatrix} \cos\sigma & -\text{sen}\sigma & 0 \\ \text{sen}\sigma & \cos\sigma & 0 \\ 0 & 0 & 1 \end{pmatrix} * \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}$$

**Figura 3.3.4.VI: Rotação Z** - Produto de matrizes que representa a rotação anti-horária de um átomo A, de coordenadas cartesianas (X, Y, Z), de $\sigma$ em torno do eixo Z.

Continuando com o re-posicionamento de $\mathbf{A_1}$, $\mathbf{A_2}$ e $\mathbf{A_3}$, aplica-se uma rotação anti-horária desses átomos em torno do eixo X de $\boldsymbol{\theta_2}$, a fim de posicionar o átomo $\mathbf{A_2}$ na orientação positiva do eixo Z **(Figura 3.3.4.VII: Rotação $\theta_2$)**.

A rotação anti-horária de um átomo **A** por um ângulo $\boldsymbol{\sigma}$ em torno do eixo X é dada pelo produto da matriz de rotação em X com as coordenadas cartesianas do átomo **(Figura 3.3.4.VIII: Rotação X)**.

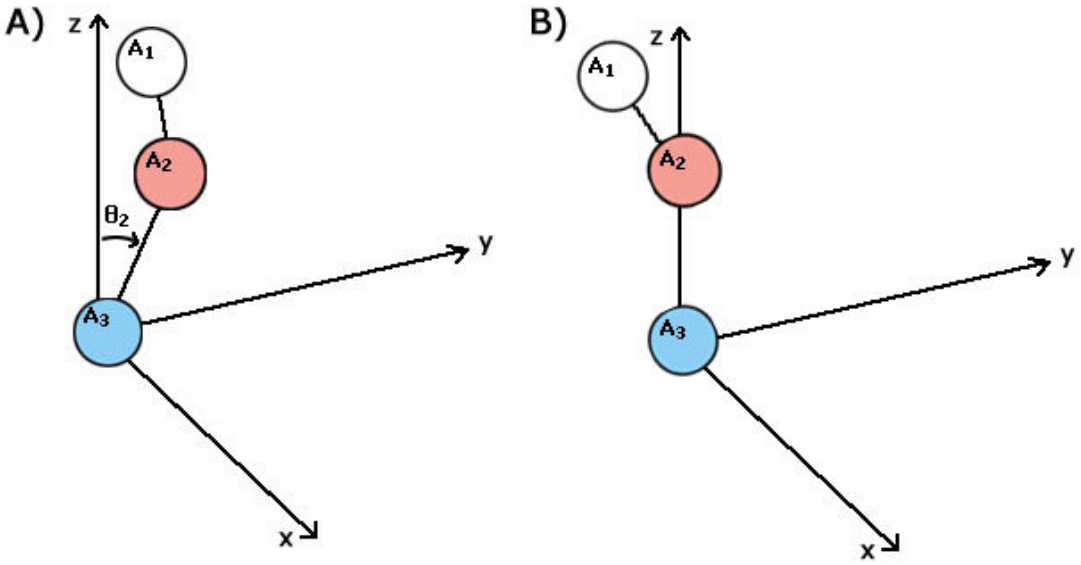



**Figura 3.3.4.VII: Rotação $\theta_2$** - A) Posição dos átomos antes da rotação de $\boldsymbol{\theta_2}$; B) Configuração dos átomos após a rotação anti-horária de $\boldsymbol{\theta_2}$ em torno do eixo X.

$$R_X(\sigma) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\sigma & -\text{sen}\sigma \\ 0 & \text{sen}\sigma & \cos\sigma \end{pmatrix} * \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}$$

**Figura 3.3.4.VIII: Rotação X** - Produto de matrizes que representa a rotação anti-horária de um átomo A, de coordenadas cartesianas (X, Y, Z), de $\boldsymbol{\sigma}$ em torno do eixo X.

Após as duas rotações, as novas coordenadas dos átomos são:

$$\mathbf{A_3} = (0, 0, 0)$$
$$\mathbf{A_2} = (0, 0, R_2)$$
$$\mathbf{A_1} = (X_1'', Y_1'', Z_1'')$$

onde:

$$X_1'' = X_1' * \text{sen}\ \boldsymbol{\varphi_2} - Y_1' * \cos\boldsymbol{\varphi_2}$$
$$Y_1'' = (X_1' * \cos\boldsymbol{\varphi_2} - Y_1' * \text{sen}\ \boldsymbol{\varphi_2}) * \cos\boldsymbol{\theta_2} - Z_1' * \text{sen}\ \boldsymbol{\theta_2}$$
$$Z_1'' = (X_1' * \cos\boldsymbol{\varphi_2} - Y_1' * \text{sen}\ \boldsymbol{\varphi_2}) * \text{sen}\ \boldsymbol{\theta_2} - Z_1' * \cos\boldsymbol{\theta_2}$$

Descobre-se $\varphi_1$, ângulo entre o vetor $\mathbf{A_1A_2}$ e o plano ZX, sendo positivo na orientação horária em relação a Z **(Figura 3.3.4.IX: $\varphi_1$ A)**, onde $\varphi_1$ = Arctg ($Y_1$"/ $X_1$").

Finalmente, por uma rotação anti-horária de $\varphi_1$ em torno do eixo Z dos átomos $\mathbf{A_1}$, $\mathbf{A_2}$ e $\mathbf{A_3}$, posiciona-se o $\mathbf{A_3}$ no 1º quadrante do sistema XZ **(Figura 3.3.4.IX: $\varphi_1$ B).** Como $\varphi_1$ varia num intervalo [-π/2, π/2], caso $X_1$" < 0, é preciso acrescentar a $\varphi_1$ o valor de π, para que $\mathbf{A_3}$ assuma a posição desejada.

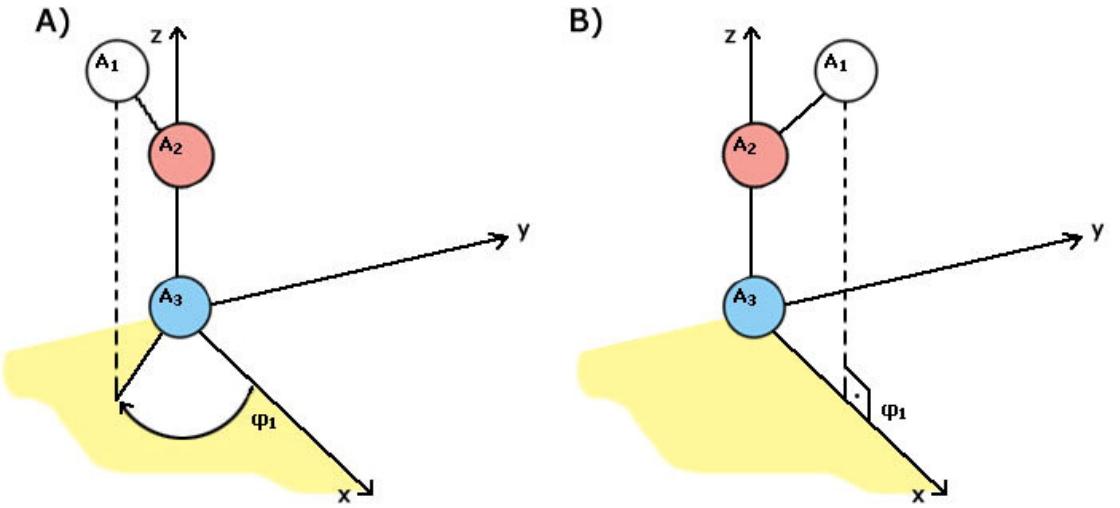



**Figura 3.3.4.IX: $\varphi_1$ -** A) Representação do ângulo $\varphi_1$ entre o vetor $\mathbf{A_1A_2}$ e o 1º quadrante do sistema ZX; B) Posição dos átomos $\mathbf{A_1}$, $\mathbf{A_2}$ e $\mathbf{A_3}$ após a rotação anti-horária de $\mathbf{A_1}$, $\mathbf{A_2}$ e $\mathbf{A_3}$, em torno de Z.

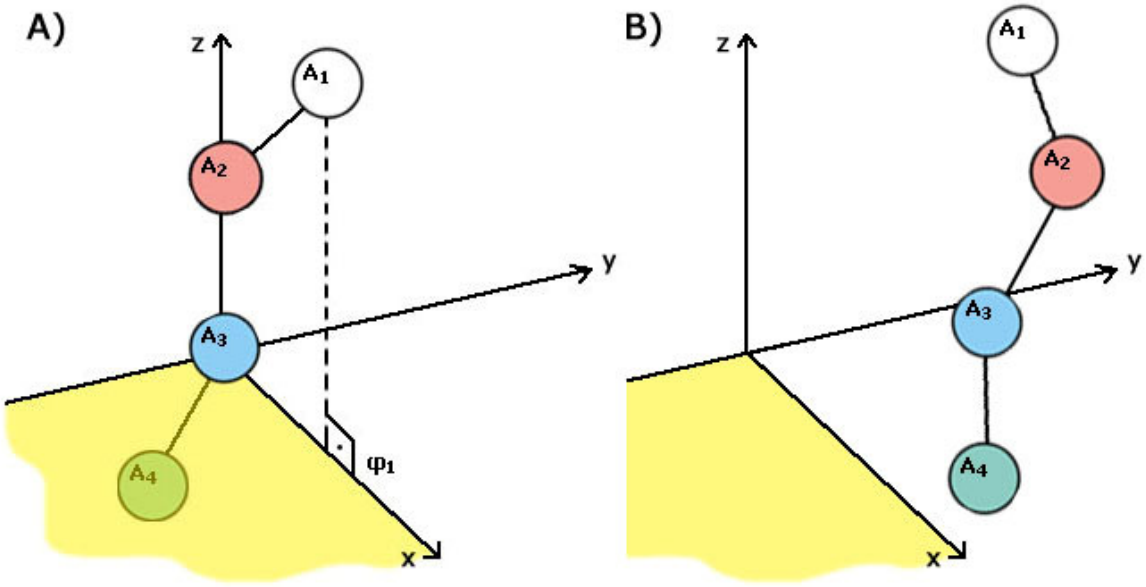



**Figura 3.3.4.X: Posição Final $\mathbf{A_4}$** - A) Posição de $\mathbf{A_4}$ em relação às novas localizações de $\mathbf{A_1}$, $\mathbf{A_2}$ e $\mathbf{A_3}$; B) Posição dos átomos após as transformações inversas na ordem inversa.

Para posicionar o $\mathbf{A_4}$ em relação às novas posições de $\mathbf{A_1}$, $\mathbf{A_2}$ e $\mathbf{A_3}$, usa-se a relação entre coordenadas cartesianas e esféricas:

$$X_4 = R_4 * \text{sen } \boldsymbol{\theta} * \cos \boldsymbol{\varphi}$$
$$Y_4 = R_4 * \text{sen } \boldsymbol{\theta} * \text{sen } \boldsymbol{\varphi}$$
$$Z_4 = R_4 * \cos \boldsymbol{\theta}$$

onde $R_4$ é a distância entre $\mathbf{A_4}$ e $\mathbf{A_3}$ e $\boldsymbol{\theta}$ é o menor ângulo entre os vetores $\mathbf{A_2A_3}$ e $\mathbf{A_4A_3}$.

Para retornar $\mathbf{A_1}$, $\mathbf{A_2}$ e $\mathbf{A_3}$ para as posições iniciais, juntamente com o $\mathbf{A_4}$, aplica-se a todos os quatro átomos as transformações inversas e na ordem inversa da descrita acima. Ou seja, deve-se aplicae $Rz(-\boldsymbol{\varphi_1}) * Rx(\boldsymbol{\theta_2}) * Rz(\pi - \boldsymbol{\varphi_2})$, e finalmente transladar todos eles por $\mathbf{T}(X_3, Y_3, Z_3)$ **(Figura 3.3.4.X: Posição Final $A_4$)**.

## 3.3.5. Inserção do átomo de oxigênio pela planaridade

Para inserir o átomo de oxigênio em uma nova conformação, aplica-se a propriedade da planaridade, segundo a qual dados dois aminoácidos $\mathbf{Aa_1}$ e $\mathbf{Aa_2}$ ligados peptidicamente, considera-se que os átomos de $C\alpha_1$, $Cc_1$ e $O_1$ estão no mesmo plano do átomo $N_2$ **(Figura 3.3.5.I: Planaridade)**. Isto é possível neste projeto porque adotou-se o ângulo diedro ω com o valor de 180º para todos os aminoácidos, permitindo que os vetores $C\alpha_1Cc_1$, $Cc_1N_2$ e $N_2C\alpha2$ estejam no mesmo plano.

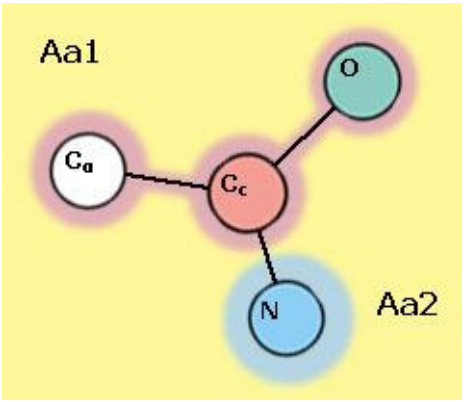



**Figura 3.3.5.I: Planaridade** – Os átomos $C\alpha_1$, $Cc_1$ e $O_1$ do $\mathbf{Aa_1}$ estão no mesmo plano do átomo $N_2$ do $\mathbf{Aa_2}$.

Assim, para determinar a posição do $O_1$, desloca-se inicialmente os átomos de $C\alpha_1$, $Cc_1$ e $N_2$, para que o $N_2$ esteja na origem do sistema, $Cc_1$ na orientação positiva do eixo Z, e o $C\alpha_1$ no 1º quadrante do sistema XZ **(Figura 3.3.5.II: Deslocamento da Planaridade)**. Esse deslocamento ocorre da mesma forma que a descrita no passo anterior, adotando-se $A_1 = C\alpha_1$, $A_2 = Cc_1$ e $A_3 = N_2$.

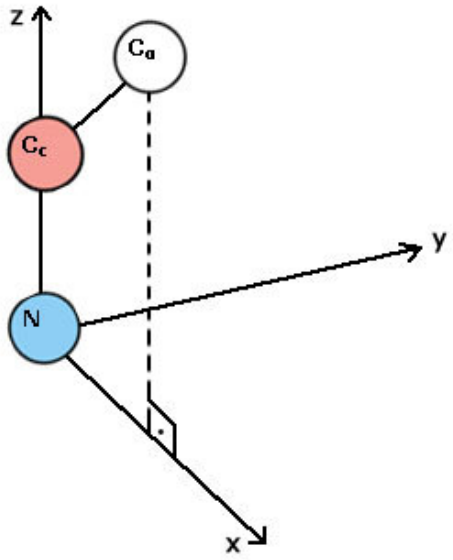



**Figura 3.3.5.II: Deslocamento da Planaridade** - Posição dos átomos $C\alpha_1$, $Cc_1$ e $N_2$ após as mesmas rotações e translações aplicadas no tópico 3.3.4.

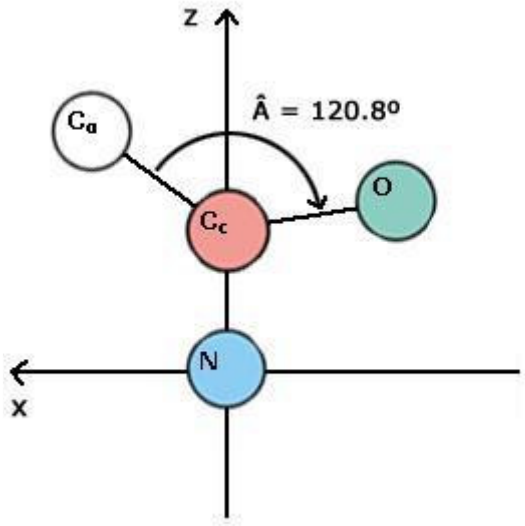



**Figura 3.3.5.III: Planaridade do O** – Posicionando o $O_1$ no 4º quadrante do sistema XZ, de acordo com as novas posições dos átomos $C\alpha_1$, $Cc_1$ e $N_2$. Nesta imagem, o eixo Y está perpendicular ao plano do papel, com a orientação positiva voltada para o leitor.

Em seguida, posiciona-se o $O_1$ no 4º quadrante do sistema XZ, de forma que os átomos $C\alpha_1$, $Cc_1$ e $O_1$ formem um ângulo de 120,8º, de acordo com a **Tabela 3.3.3.I: Engh & Huber (Figura 3.3.5.III: Planaridade do O)**, onde:

$$X_{O1} = -R_{O1_Cc1} * \cos(120{,}8º - \pi/2)$$
$$Y_{O1} = 0.0$$
$$Z_{O1} = R_{Cc1_N2} + ( R_{O1_Cc1} * \text{sen}(120{,}8º - \pi/2) )$$

com $R_{O1_Cc1}$ representando a distância entre o $O_1$ e o $Cc_1$, e $R_{Cc1_N2}$ a distância entre o $Cc_1$ e o $N_2$. Para retornar os átomos $C\alpha_1$, $Cc_1$ e $N_2$ para as posições originais, juntamente com o $O_1$, aplica-se o mesmo procedimento do tópico 3.3.4.

### 3.3.6 Inserção do átomo de Cβ pela quiralidade

Para inserir o átomo de Cβ é preciso considerar os átomos de Ca, Cc e N do mesmo aminoácido. Adotando N = $A_1$, Cα = $A_2$ e Cc = $A_3$, aplica-se a mesma transformação apresentada na subseção 3.3.4 **(Figura 3.3.6.I: Deslocamento da quiralidade)**.

Em seguida, translada-se esses três átomos de ($-X''_{C\alpha}$, $-Y''_{C\alpha}$, $-Z''_{C\alpha}$), a fim de posicionar o Cα na origem do sistema, o Cc na orientação negativa do eixo Z, e manter o N no 1º quadrante do sistema XZ **(Figura 3.3.6.II: Translação da quiralidade)**.

Para posicionar o Cβ é necessário resolver a seguinte transformação de coordenadas esféricas para cartesianas:

$$X_{C\beta} = R_{C\beta_C\alpha} * \cos(\theta_{C\beta} - \pi/2) * \cos(\varphi)$$
$$Y_{C\beta} = R_{C\beta_C\alpha} * \cos(\theta_{C\beta} - \pi/2) * \operatorname{sen}(\varphi)$$
$$Z_{C\beta} = R_{C\beta_C\alpha} * \operatorname{sen}(\theta_{C\beta} - \pi/2)$$

onde $R_{C\beta_C\alpha}$ é a distância entre os átomos de Cα e Cβ; $\boldsymbol{\theta_{C\beta}}$ é o ângulo entre os vetores CcCα e CαCβ; $\boldsymbol{\varphi}$ é o ângulo entre o vetor CβCα e o plano XZ, como visto na **Figura 3.3.6.III: Posicionando Cβ**.

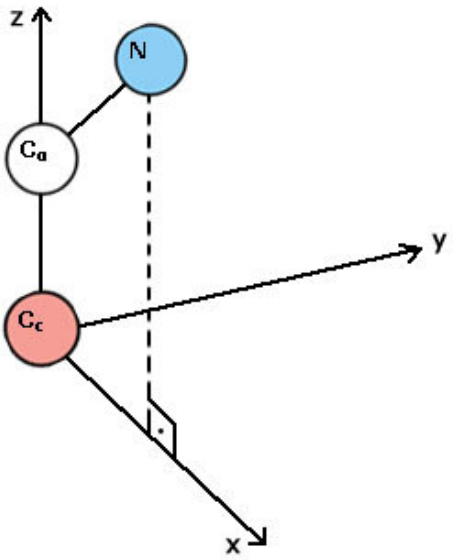



**Figura 3.3.6.I: Deslocamento da quiralidade -** Posição dos átomos de Ca, Cc e N após as mesmas transformações aplicadas no tópico 3.3.4.

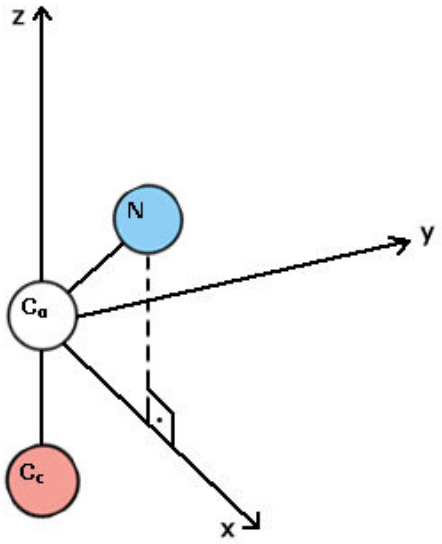



**Figura 3.3.6.II: Translação da quiralidade** – Posicionando o Cα na origem do sistema.

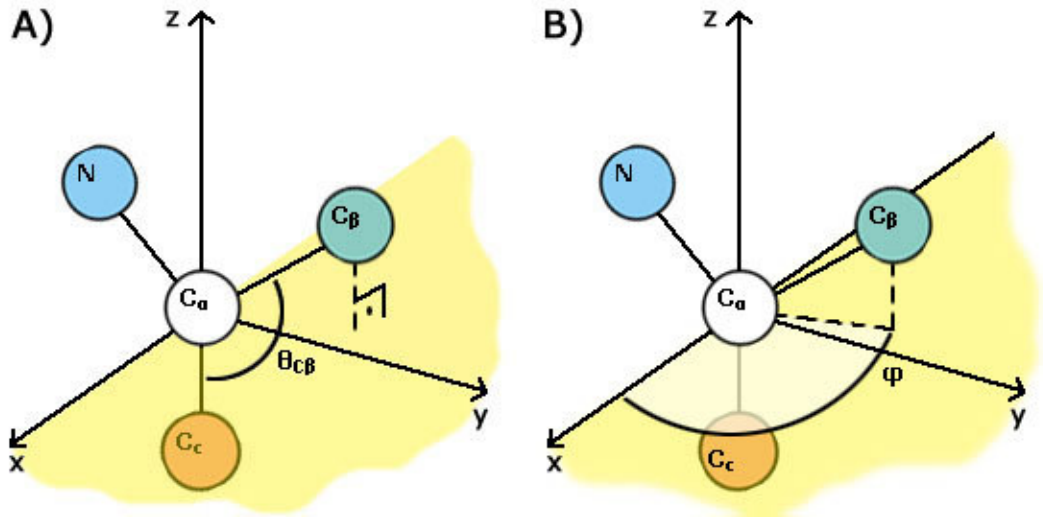



**Figura 3.3.6.III: Posicionando Cβ** – Representação dos dois ângulos usados pelas expressões de coordenadas cartesianas na determinação da posição do Cβ.

O problema na determinação da posição do Cβ seria descobrir se φ tem sentido positivo em uma orientação horária ou anti-horária em relação ao eixo Z. O fato de todos os aminoácidos presentes nas proteínas serem *L* faz com que o grupo carboxila se encontre logo à esquerda do Cα, fazendo com que φ apresente uma orientação anti-horária. Caso contrário, o grupo que estaria logo à esquerda do Cα seria o grupo lateral, como visto na **Figura 3.3.6.IV: Orientação φ**.

Isto faz com que a configuração do aminoácido apresente uma propriedade interessante: visualizando o aminoácido ao longo do eixo HCα, a palavra CORN é formada no sentido horário de rotação **(Figura 3.3.6.V: CORN)**.

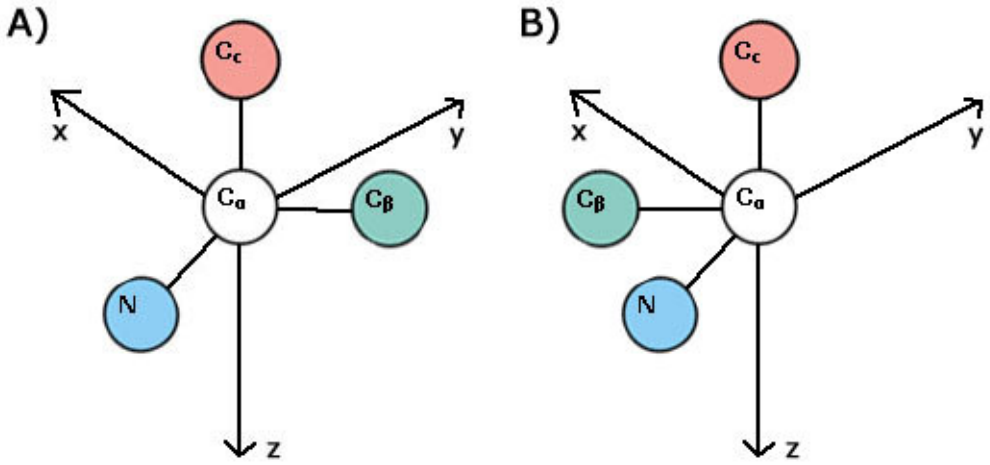



**Figura 3.3.6.IV: Orientação φ** – A) φ com orientação anti-horária, fazendo com que o Cc fique à esquerda do Cα; B) φ com orientação horária. Nesse caso, o Cβ é que está a esquerda do Cα.

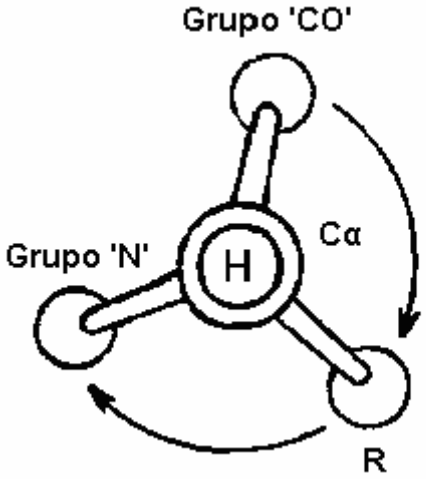



**Figura 3.3.6.V: CORN** - Formação da palavra CORN, visualizando-se o aminoácido ao longo do eixo HCα.

A obtenção do ângulo φ segue os seguintes passos:

- Muda-se a representação de coordenadas cartesianas do N e do Cβ para coordenadas esféricas:

$$X''_N = R_{N_C\alpha} * \cos(\theta_N - \pi/2)$$
$$Y''_N = 0.0$$
$$Z''_N = R_{N_C\alpha} * \text{sen}(\theta_N - \pi/2)$$

$$X''_{C\beta} = R_{C\beta_C\alpha} * \cos(\theta_{C\beta} - \pi/2) * \cos\varphi$$
$$Y''_{C\beta} = R_{C\beta_C\alpha} * \cos(\theta_{C\beta} - \pi/2) * \text{sen}\,\varphi$$
$$X''_{C\beta} = R_{C\beta_C\alpha} * \text{sen}\,\theta_{C\beta}$$

onde $R_{N_C\alpha}$ é a distância entre os átomos de N e Cα; $\theta_N$ é o ângulo entre o vetor NCα e o vetor CαCc; $R_{C\beta_C\alpha}$ é a distância entre o Cβ e o Cα; $\theta_{C\beta}$ é o ângulo entre o vetor CβCα e o vetor CαCc **(Figura 3.3.6.VI: Determinando φ A )**.

- Determina-se o ângulo entre os vetores NCα e CβCα ($\theta_{N-C\beta}$) por produto escalar:

$$\theta_{N-C\beta} = \text{Arc cos} ( (NC\alpha * C\beta C\alpha) / | NC\alpha | * | C\beta C\alpha |)$$

$$\cos(\theta_{N-C\beta}) = ( (X''_N * X''_{C\beta}) + (Y''_N * Y''_{C\beta}) + (Z''_N * Z''_{C\beta}) ) / R_{N_C\alpha} * R_{C\beta_C\alpha}$$

$$\cos(\theta_{N-C\beta}) = \cos(\theta_N - \pi/2) * \cos(\theta_{C\beta} - \pi/2) * \cos\varphi + \text{sen}(\theta_N - \pi/2) * \text{sen}\,\theta_{C\beta}$$

A partir dessa última equação, tem-se:

$$\varphi = \text{Arc cos}(( \cos(\theta_{N-C\beta}) - (\text{sen}(\theta_N - \pi/2) * \text{sen}\,\theta_{C\beta}) ) / \cos(\theta_N - \pi/2) * \cos(\theta_{C\beta} - \pi/2))$$

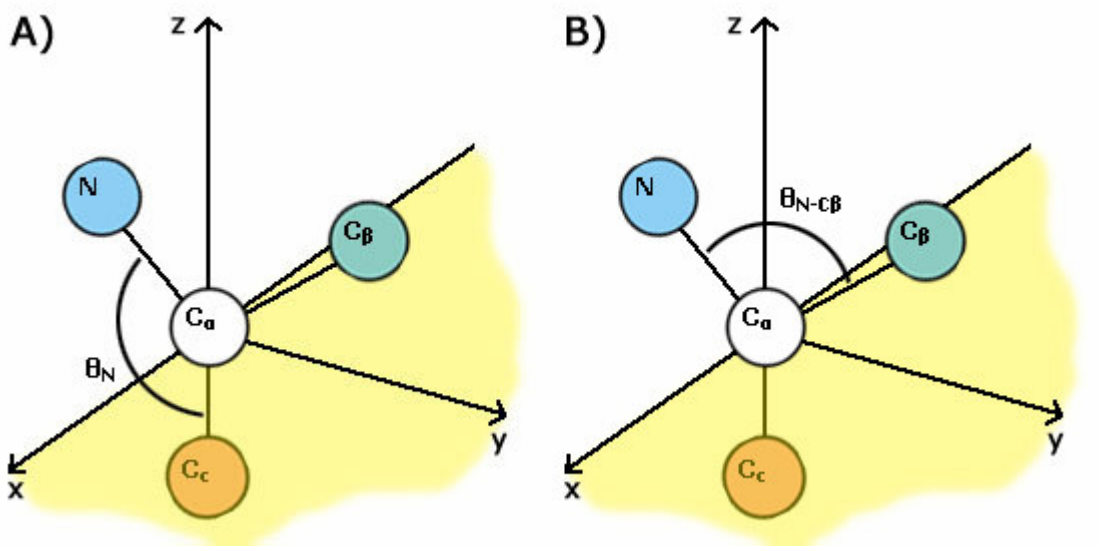



**Figura 3.3.6.VI: Determinando φ** – A) $\theta_N$, o menor ângulo determinado pelos átomos de N, Cα e Cc; B) $\theta_{N-C\beta}$, menor ângulo entre os átomos de N, Cα e Cβ.

Por fim, retorna-se o N, Cα e o Cc para as posições originais, juntamente com o Cβ. Inicialmente, uma translação de ($X''_{C\alpha}$, $Y''_{C\alpha}$, $Z''_{C\alpha}$) é feita para todos os quatro átomos. Em seguida, aplica-se a mesma transformação mostrada no fim da seção 3.3.4: $Rz(-\varphi_1) * Rx(\theta_2) * Rz(\pi - \varphi_2)$.

## 3.3.7. Validando as novas conformações

Após posicionar o penúltimo aminoácido na nova conformação da seqüência defeituosa, o aminoácido *dummy*, este é aproximado do posicionamento correspondente à primeira âncora final, por minimização do potencial:

$$\text{Potencial} = \sum_{i \in \{N, C\alpha\}} || \vec{d_i} - \vec{a_i} ||^2$$

onde, $\mathbf{d_i}$ e $\mathbf{a_i}$ são as coordenadas do átomo **i** para o aminoácido *dummy* e do átomo âncora, respectivamente. Durante a minimização, os ângulos Φ e ψ sofrem pequenas variações até ±5º, sendo o movimento aceito caso haja uma diminuição de potencial. Isto acontece durante 10000 passos ou enquanto a média das últimas 50 alterações for menor que –0,01. Após a aproximação, insere-se a última âncora, e atualiza-se a posição do O do aminoácido *dummy* pela planaridade.

Durante a construção e a minimização, algumas condições têm que ser respeitadas. A primeira é um filtro de choques entre átomos, em que nenhum dos átomos dos resíduos da nova conformação pode interceptar algum outro presente no restante da proteína ou na própria conformação. Para isto, é suficiente avaliar a seguinte propriedade: para que dois átomos não-ligados se interceptem, basta que a distância entre seus centros seja menor que a soma dos seus raios de van der Waals.

Para cada novo aminoácido inserido na nova conformação, faz-se dois tipos de verificação: se os átomos desse aminoácido se chocam com algum outro átomo da conformação ou com algum átomo do restante da proteína. No primeiro caso, para cada átomo do novo aminoácido calcula-se a distância para todos os átomos já presentes na nova conformação, assim como a soma dos raios, e compara os valores correspondentes. Para validar o aminoácido com relação ao restante da proteína, não se compara a distância e o raio de cada átomo do aminoácido com cada átomo da proteína, pois isso acarretaria uma grande demanda computacional. O que ocorre é que para cada átomo do aminoácido buscam-se os vértices que ele intercepta, e em seguida verifica-se a sobreposição com todos os átomos das listas desses vértices.

O outro filtro considera que a distância entre o Cα de um novo aminoácido da região e o Cα da penúltima âncora tem que ser, em Å, menor que **3,8 * (T - 1 - atual) + 1**, onde **T** é o comprimento total da seqüência defeituosa, incluindo as quatro âncoras, e **atual** é a posição do novo aminoácido. Esta fórmula se baseia no fato de que, numa geometria ideal, os átomos de Cα entre dois aminoácidos conectados estão a uma distância máxima de 3,8 Å. O fator final **+1** garante flexibilidade a esse valor, além de não forçar a superposição dos Cα's do aminoácido *dummy* e da penúltima âncora.

### 3.3.8. Algoritmo para construção do conjunto das conformações

O processo para obtenção do conjunto de 1000 conformações é bastante simples: inicialmente constrói-se uma fila (**F**), que armazenará as conformações intermediárias da região em estudo geradas durante a execução do algoritmo. A primeira conformação presente em **F** é formada pelos dois aminoácidos âncoras encontrados no início da região a ser alterada. Retira-se de **F** uma conformação (**C**). Nesta conformação, insere-se o próximo aminoácido (**Aa**), a partir da metodologia descrita na seção 3.3.3, com um par de ângulos diedros selecionado aleatoriamente do conjunto descrito na seção 3.3.1.

Se a nova conformação for recusada durante a validação das duas condições citadas no tópico anterior, escolhe-se outro par de ângulos, podendo-se repetir este processo no máximo 25 vezes. Caso nenhuma estrutura aceitável tenha sido construída, descarta-se **C** e o algoritmo recomeça, retirando uma nova conformação intermediária de **F**.

Caso o novo aminoácido tenha sido inserido com sucesso, verifica-se se a nova conformação intermediária (**C+Aa**) tem o tamanho correspondente à região sem a última âncora. Em caso negativo, insere-se **C+Aa** e **C** na fila **F**. Em caso positivo, faz-se a aproximação do resíduo *dummy* com a penúltima âncora e insere-se a última âncora na conformação, verificando-se se ocorrem os choques ou intersecções.

Se houver alguma intersecção, descarta-se a conformação **C+Aa** e busca-se outro par de ângulos diedros para a inserção de **Aa** em **C**, caso ainda seja possível, pois a escolha para cada novo resíduo só pode ser feita 25 vezes no máximo, como mencionado anteriormente.

Caso não haja interseção, **C+Aa** é adicionado ao conjunto de conformações finais e **C** retorna à fila **F**. O processo continua até que o conjunto final de conformações tenha o tamanho desejado. No caso desse projeto são 1000 conformações. A **Figura 3.3.8.I: Fluxo** traz o fluxo do algoritmo usado na criação do conjunto de conformações. Inicialmente as duas primeiras âncoras são consideradas como o primeiro modelo intermediário aceitável e que é inserido em **F**.

No algoritmo aplicado, a etapa de maior consumo computacional corresponde ao processo de verificação da interseção dos átomos do aminoácido inserido com os átomos do restante da proteína. Caso essa comparação ocorresse com os átomos do aminoácido com todos os outros átomos da estrutura, o tempo computacional seria da ordem de **O(r ( (p-r) + r') )**, onde **p** corresponde ao número de aminoácidos da

proteína, **r** ao tamanho de região a ser modelada e **r'** o numero de aminoácidos já inseridos na região alterada.

Mas devido à associação dos átomos da proteína aos vértices, como descrito no Tópico 3.3.2, a quantidade de comparações aplicadas sofre uma grande diminuição. Assim, ao invés de comparar cada átomo do aminoácido inserido com **(p-r)** átomos, correspondente ao restante da proteína, faz-se a verificação considerando apenas os átomos presentes nos vértices em que os átomos do novo aminoácido interceptam. Dessa forma, a nova ordem de grandeza é **O(r (q + r'))**, onde **q** corresponde ao somatório da quantidade de elementos encontrados nos vértices interceptados pelos átomos do novo aminoácido inserido quantidade esta bastante inferior a **(p-r).**

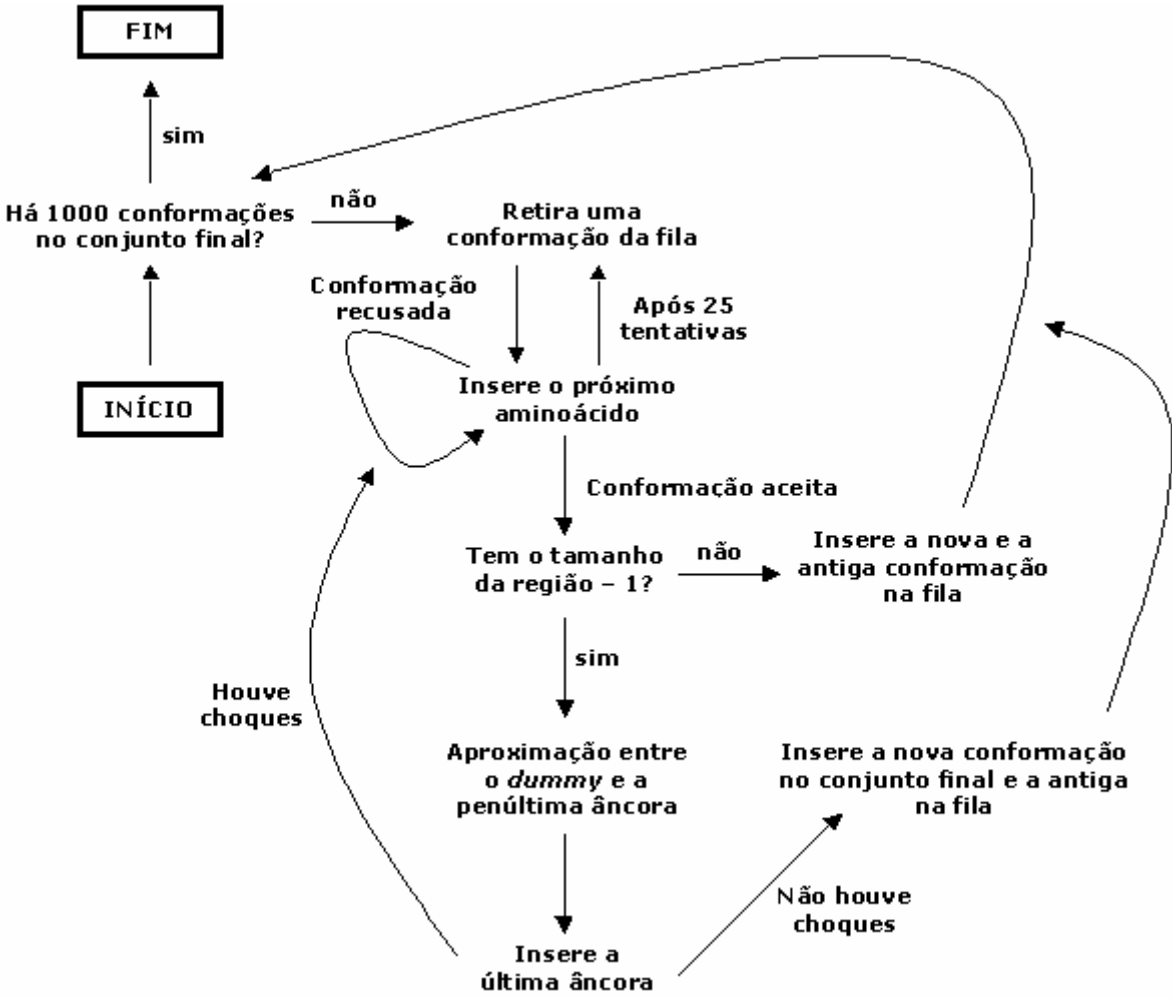



**Figura 3.3.8.I: Fluxo -** Fluxo do algoritmo para a construção das 1000 conformações

## 3.3.9. Seleção das melhores conformações

Com o objetivo de selecionar as melhores conformações, constrói-se a cadeia lateral correspondente para cada uma das 1000 conformações presentes no conjunto

criado pelo algoritmo do passo anterior. Para esta construção, executa-se o programa *SCWRL* levando em consideração o restante da proteína.

As estruturas nas quais algum átomo da cadeia lateral da região modificada apresentar intersecção com algum outro átomo, seja da região ou do restante da proteína, serão imediatamente descartadas. A verificação segue o mesmo procedimento do filtro de choques descrito na seção 3.3.7.

Com as demais estruturas que não apresentaram intersecção entre os átomos, efetua-se o cálculo da energia pelo campo de força *AMBER*, aplicado através da ferramenta *TINKER*. Por fim, as 50 estruturas com o menor valor de energia são selecionadas.

### 3.3.10. Seleção do modelo

Para cada umas das 50 seqüências obtidas na passagem anterior, aplica-se mais uma vez o cálculo da energia por campos de força *AMBER*, usando-se a ferramenta *TINKER*. A diferença é que, agora, a minimização da energia ocorre, ou seja, altera-se as posições dos átomos em busca de uma estrutura mais próxima do real.

O modelo com a menor energia é escolhido como a melhor estrutura até o momento. Caso o modelo apresente uma melhor energia antes da alteração da região defeituosa, o processo de melhoramento do modelo para. Caso contrário, o programa retorna à etapa de avaliação de estruturas, buscando uma nova região a ser modificada.

# Capítulo 4. Testes e discussão

"Com Grandes Poderes Vêm Grandes Responsabilidades"
Peter Parker – Spiderman I (2002)

Com o objetivo de avaliar a aplicabilidade deste projeto e demonstrar alguns dos resultados obtidos com a metodologia descrita, duas diferentes execuções foram realizadas. A primeira verificou-se a precisão obtida na geração do conjunto de conformações de pequenas seqüências de aminoácidos quando comparada com o projeto de DePristo *et al.* **[DBL+03]**. A segunda execução avaliou o comportamento de toda a ferramenta quando aplicada na determinação e melhoramento dos modelos de proteínas de estruturas desconhecidas, ou seja, sem nenhuma referência no *PDB*.

## Seção 4.1. Precisão do Conjunto de conformações

Neste primeiro teste será feita uma comparação dos conjuntos *decoys* obtidos a partir do projeto desenvolvido por Mark A. DePristo com os obtidos pelo projeto aqui descrito. Uma comparação entre os algoritmos também é necessária, a fim de justificar as diferenças existentes entre os conjuntos formados pelos dois projetos.

Os conjuntos *decoys* estudados foram criados a partir dos laços de proteínas utilizados no trabalho de Fiser et al. **[FDS00]**. Como alguns modelos de estruturas aplicados no projeto de Fiser se encontram obsoletos, foram substituídos por novos modelos presentes no PDB. Outras estruturas foram excluídas por apresentarem um alto valor para os fatores-B na região do laço estudado, por falta de cadeias laterais, pela presença de *gaps* na seqüência ou pela baixa qualidade da estrutura. Os códigos *PDB* das seqüências utilizadas, juntamente com as posições iniciais e comprimentos dos laços considerados para a estrutura, estão descritos na **Tabela 4.1.I: Laços Fiser**.

Para comparar as conformações criadas, foram utilizados os valores do RMSD-Global e Local obtidos entre os modelos desenvolvidos e a estrutura original. Esse tipo de comparação ajuda a identificar o quanto as novas conformações se assemelham à estrutura nativa. No caso do RMSD-Global, é avaliada a similaridade da orientação espacial entre duas estruturas. O RMSD-Local indica a similaridade interna das novas conformações com o laço nativo.

Na comparação dos resultados, representada na **Tabela 4.1.II: Média Dos Melhores RMSD**, é avaliado o RMSD, tanto Local como Global, das melhores

conformações geradas para cada um dos laços, divididos de acordo com comprimento, que varia entre dois e sete aminoácidos.

| | | |
|---|---|---|
| 154l (10, 2); (10, 3); | 1mba (97, 4); | 2cba (61, 7); |
| 1ads (149, 6); (14, 2); (14, 3); (274, 5); (99, 4); | 1pbe (117, 4); (194, 5); (254, 2); (254, 3); | 2cmd (163, 4); (168, 2); (168, 3); (188, 5); |
| 1aaj (82, 4); | 1nar (192, 8); (56, 5); | 2cpl (115, 5); (122, 6); |
| 1alc (110, 5); | 1nfp (117, 3); (118, 2); (37, 4); | 2ctc (53, 7); |
| 1amp (225, 7); | 1npk (35, 2); | 2cy3 (101, 4); |
| 1arb (249, 7); | 1onc (11, 6); (58, 5); | 2dri (161, 8); (30, 3); (31, 2); |
| 1arp (171, 2); (282, 6); (296, 3); | 1oyc (115, 3); | 2ebn (123, 2); (60, 3); |
| 1art (346, 5); (356, 7); | 1mbd (17, 7); | 2end (93, 5); |
| 1cbs (21, 4); (45, 7); (55, 8); (66, 6); | 1php (135, 7); (217, 6); (256, 3); (257, 2); | 2exo (161, 4); (262, 8); (73, 3); (74, 2); |
| 1byb (198, 7); (223, 5); | 1phf (17, 7); (268, 5); (341, 6); | 2fox (25, 5); |
| 1csh (98, 5); | 1pgs (223, 4); (33, 5); | 2hbg (37, 5); |
| 1ctf (113, 2); (113, 3); | 1plc (74, 4); (87, 7); | 2mnr (270, 7); (291, 2); (308, 6); |
| 1ctm (131, 2); (131, 3); | 1pmy (25, 7); | 2por (74, 2); |
| 1cus (31, 5); | 1poa (108, 5); | 2ran (26, 8); (40, 6); |
| 1cyo (24, 7); | 1poc (128, 3); | 2sga (16, 4); (78, 3); (79, 2); |
| 1dts (137, 3); (146, 6); (35, 2); (62, 5); (76, 7); | 1ppn (144, 6); (166, 7); (42, 4); (77, 3); (78, 2); | 1prn (150, 8); (160, 2); (187, 5); (57, 6); (66, 4); |
| 1eco (72, 5); (91, 2); | 2sil (176, 6); (220, 4); | 2tgi (40, 3); (41, 2); (72, 4); |
| 1ede (121, 7); (15, 5); (180, 6); | 1ptf (65, 7); | 3chy (12, 6); |
| 1fkf (10, 5); (42, 4); | 1rcf (111, 4); | 3cyr (25, 6); (69, 4); |
| 1flp (120, 2); (120, 3); | 1sbp (107, 8); (169, 7); (181, 5); | 3dni (132, 3); |
| 1fna (21, 2); (21, 3); | 1scs (19, 6); | 3grs (26, 7); (336, 6); |
| 1fnd (204, 6); | 1tca (287, 4); (94, 6); | 3hsc (227, 6); |
| 1tml (147, 5); (187, 8); (20, 7); (42, 4); | 1tib (176, 6); (211, 2); (211, 3); (253, 5); (46, 4); | 3tgl (159, 7); (207, 5); (213, 2); (213, 3); (82, 6); |
| 1gca (100, 6); (196, 7); | 1thw (158, 7); (18, 8); (194, 4); | 4enl (24, 8); (335, 4); |
| 1gky (31, 3); | 1iab (100, 4); (142, 7); | 4fgf (58, 7); |
| 1gpr (123, 4); (35, 7); (54, 5); (83, 6); | 5fd1 (77, 3); (78, 2); (81, 4); (87, 6); | 4gcr (116, 4); (150, 7); (169, 3); (170, 2); |
| 1gof (430, 5); | 1ubi (72, 2); | 4i1b (38, 3); |
| 1hbq (138, 7); (158, 5); (15, 6); | 1ukz (94, 5); | 1tys (131, 4); (66, 6); |
| 1hfc (142, 8); (151, 7); | 1xif (286, 3); (357, 6); (82, 4); | 5fx2 (27, 7); (41, 6); |
| 7rsa (14, 6); (20, 7); (47, 4); (75, 5); | 1frd (54, 2); (54, 3); (59, 4); (83, 5); | 5p21 (104, 6); (45, 8); (75, 4); (83, 7); |
| 1l58 (12, 2); | 1ycc (62, 7); (79, 5); (90, 6); | 5ptp (141, 2); (141, 3); |
| 1lif (24, 2); (35, 3); (64, 7); | 8abp (299, 5); (54, 4); (64, 6); | 1xnb (30, 4); |
| 1lst (109, 6); (117, 7); (80, 3); (81, 2); | 2ayh (123, 8); (142, 3); (160, 6); (54, 7); | 2alp (171, 5); (191, 3); (192, 2); (94, 7); |

**Tabela 4.1.I: Laços Fiser** - Todos os códigos *PDB* das proteínas utilizadas na execução dos testes de avaliação da precisão dos conjuntos de conformações, acompanhados de pares que indicam a posição inicial e o tamanho dos laços considerados para a estrutura.

Por exemplo, para cada um dos laços de tamanho três construiu-se um conjunto de 1000 conformações. Para cada uma delas calculou-se o RMSD-Global e o RMSD-

Local em relação à estrutura nativa correspondente. O melhor RMSD-Global e Local de cada conjunto foi utilizado no cálculo das médias do RMSD para os laços de tamanho três. O RMSD foi calculado aplicando a ferramenta MODELLER, considerando apenas os átomos pesados da cadeia principal (Cα, N, O e C da carboxila).

| Tamanho do laço | Número de seqüências de teste (DePristo) | Número de seqüências de teste (do autor) | RMSD-Global (DePristo) (Å) | RMSD-Global (do autor) (Å) | RMSD-Local (DePristo) (Å) | RMSD-Local (do autor) (Å) |
|---|---|---|---|---|---|---|
| 2 | 34 | 34 | 0,31 | 0,49 | 0,11 | 0,22 |
| 3 | 34 | 34 | 0,34 | 0,61 | 0,18 | 0,33 |
| 4 | 35 | 35 | 0,43 | 1,04 | 0,24 | 0,57 |
| 5 | 35 | 34 | 0,53 | 1,19 | 0,32 | 0,76 |
| 6 | 36 | 36 | 0,69 | 1,56 | 0,42 | 0,97 |
| 7 | 38 | 38 | 0,78 | 1,62 | 0,49 | 1,06 |

**Tabela 4.1.II: Média Dos Melhores RMSD** – Tabela com os valores de RMSD-Global e RMSD-Local para laços de comprimento entre dois e sete aminoácidos medido em Å.

Ao selecionar apenas a melhor conformação para cada um dos conjuntos criados, busca-se entender o comportamento da metodologia aplicada na construção dos novos modelos considerando o melhor caso.

Na Tabela 4.1.II, a primeira coluna indica o comprimento dos laços, seguida pela coluna que mostra o número de seqüências alvos aplicadas na construção dos modelos usadas por DePristo e pelo projeto aqui descrito. No caso de laços de tamanho cinco, uma seqüência a menos foi aplicada no teste deste projeto. Isso se deve ao fato de que na proteína 8abp, o laço é determinado na extremidade da estrutura (posição 299 até 303, sendo o comprimento da seqüência de 304 aminoácidos), que não permite a consideração dos dois últimos aminoácidos âncoras necessários no algoritmo deste projeto. As demais colunas representam os valores para o RMSD-Global e o RMSD-Local para os dois projetos.

Ao analisar os resultados, é evidente que, em termos relativos, os valores de ambos os tipos de RMSD são melhores no trabalho de DePristo para todos os tamanhos de aminoácidos, sendo em média duas vezes menor, conseqüentemente formando conformações mais próximas da estrutura nativa. Entretanto, em termos absolutos, a diferença entre os modelos gerados é pequena, sendo no RMSD-Local de no máximo 0.57 Å e no RMSD-Global de 0,84 Å.

O resultado de menor similaridade das conformações construídas pelo projeto aqui descrito era esperado, pois há uma grande diferença num dos principais pontos dos algoritmos: a seleção dos ângulos diedros. Durante a determinação dos possíveis valores

a serem atribuídos ao pares de ângulos ψ e Φ, cada par obtido da base de proteínas Top500 está associado a um aminoácido e a um elemento do conjunto de pares de precisão de 5º para este aminoácido (Seção 3.3.1). Os dois trabalhos usam apenas os elementos de cada conjunto que tenha ao menos um par do Top500 associado. A diferença é que para DePristo aqueles elementos que apresentam mais pares associados têm uma maior probabilidade de serem selecionados e aplicados na construção da nova conformação, enquanto que neste projeto todos os elementos têm a mesma probabilidade. Esta decisão faz com que o projeto de DePristo aplique os pares mais prováveis na construção das conformações e conseqüentemente com uma maior tendência de se aproximar da conformação nativa.

## Seção 4.2. Modelagem de estruturas

Nesta segunda parte dos testes do projeto, executaram-se todas as etapas disponibilizadas pela ferramenta desenvolvida para duas proteínas de estruturas desconhecidas, ou seja, não presentes no PDB. O objetivo é ter uma idéia de quanto a alteração de pequenos trechos pode melhorar os modelos das proteínas obtidas por homologia.

As proteínas aplicadas no teste foram obtidas a partir do conjunto utilizado na quinta edição do CASP (*Critical Assessment of Structure Prediction*) **[CASP]**. Este evento permite aos grupos que desenvolvem novas técnicas de modelagem de proteínas uma forma de avaliar suas metodologias. O funcionamento é simples: um conjunto de proteínas de estruturas desconhecidas é disponibilizado e os grupos interessados tentam modelar a proteína. Enquanto isso, as proteínas são modeladas por metodologias laboratoriais. No fim, as estruturas determinadas por cada um dos grupos são comparadas com aquelas construídas nos laboratórios.

No teste, cada proteína é modelada pela ferramenta MODELLER. O conjunto de proteínas moldes utilizado é formado pelas estruturas com maior similaridade em relação a seqüência alvo, considerando tantas quanto forem possíveis sem que o MODELLER apresente problemas com os alinhamentos dos moldes. Isso ocorre quando as estruturas são muito distintas entre si e não há a formação de um alinhamento.

Cinco modelos são construídos, sendo escolhido aquele que apresentar o menor valor de energia segundo os cálculos do MODELLER. Em seguida, o modelo selecionado é submetido às quatro ferramentas de avaliação disponíveis (PROCHECK,

ANoLEA, ProSa e PROVE), com o subseqüente cálculo da pontuação de cada aminoácido, com todas as ferramentas associadas ao mesmo peso. Em seguida seleciona-se a região defeituosa, a seqüência de cinco aminoácidos de maior pontuação contida em um laço.

As regiões que pertencem aos laços na proteína são todas aquelas não associadas às estruturas secundárias folha-β e α-hélice, sendo estas definidas a partir da execução do programa DSSP **[KS83]** e usando as mesmas considerações que as aplicadas no trabalho de Fiser et al. **[FDS00]**.

Para a região defeituosa, 1000 novas conformações são construídas, as 50 de mais baixa energia são minimizadas com base na função de energia do campo de força AMBER. Dessas, as cinco de mais baixa energia são comparadas com o modelo atual.

Para esta comparação é analisada a similaridade do modelo atual e das cinco novas conformações com a estrutura da proteína determinada em laboratório (o modelo real). Duas são as formas utilizadas na avaliação da similaridade: o RMSD-Local considerando os átomos pesados (Cα, N, O e C da carboxila), para toda a proteína e apenas para a região modificada. Dentre esses dois valores analisados, o mais importante corresponde ao RMSD da região considerando todos os átomos pesados, pois é ele que realmente indica se a nova conformação selecionada tem uma maior similaridade com o modelo laboratorial.

Os códigos das proteínas no conjunto do CASP5 aplicados nos testes foram T0129, T0130 e T0137.

### 4.2.1. T0129

Nesse teste aplicou-se a seqüência T0129, de comprimento de 182 aminoácidos e atualmente associada à estrutura no PDB de código 1IZM. Para que o resultado da ferramenta não apresentasse um comportamento tendencioso, o conjunto PDB utilizado em todas as execuções é formado apenas pelas estruturas submetidas antes do CASP5.

Neste exemplo, a construção dos cinco modelos da estrutura pelo MODELLER durou 43 minutos, a seleção da estrutura de menor energia e a subseqüente avaliação pelas ferramentas de validação duraram apenas 8 segundos. A configuração da pontuação do melhor modelo é mostrada na **Figura 4.2.1.I**.

A região de maior pontuação é limitada pelos aminoácidos da posição 90-94. Mas segundo a execução do DSSP, com os resultados na parte inferior da **Figura 4.2.1.I,** este é um trecho que se encontra em sua maior parte fora de um laço.

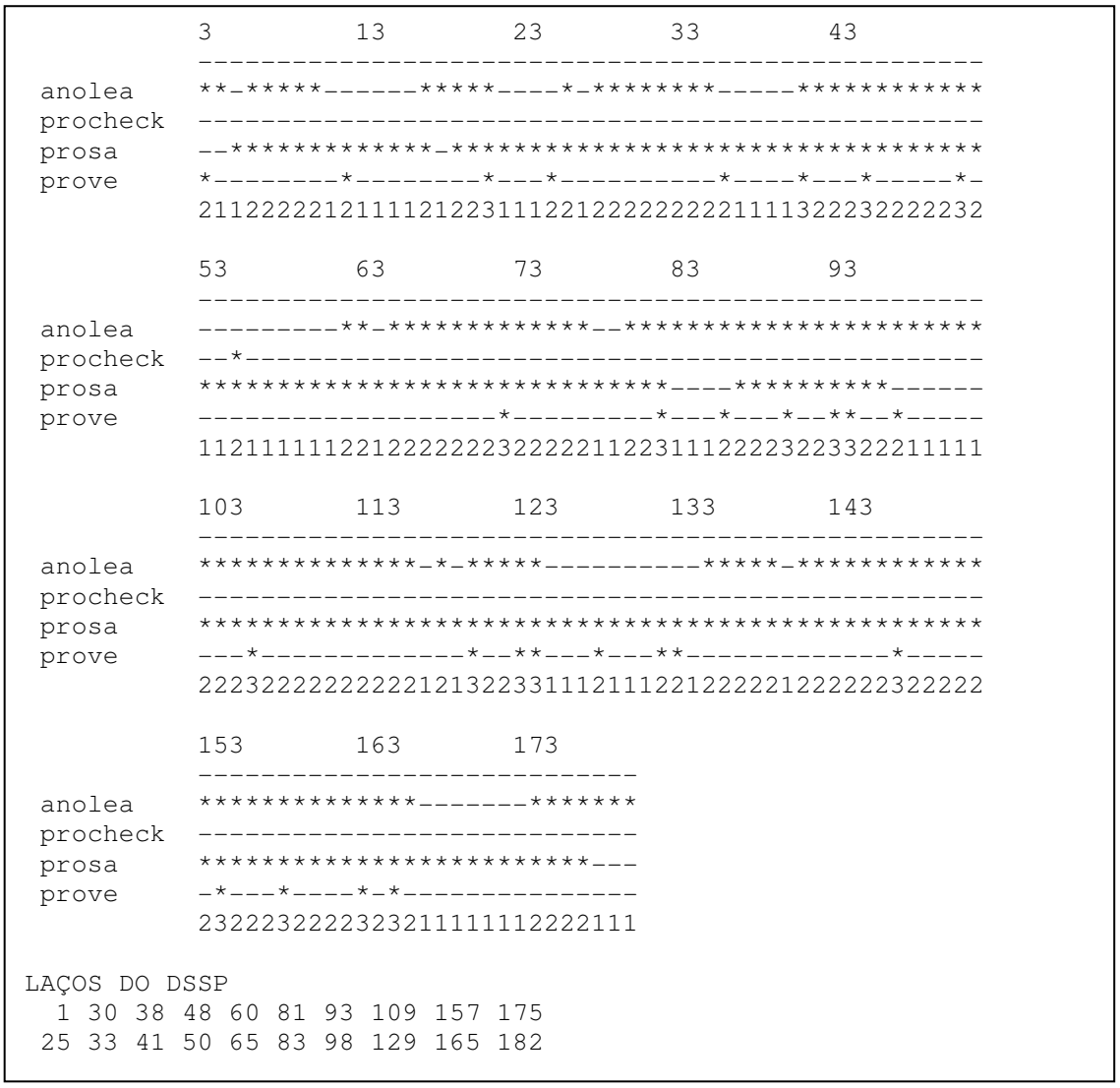



**Figura 4.2.1.I: Avaliação T0129-01** – Avaliação da estrutura originada do MODELLER e o resultado da execução da ferramenta de determinação de laços, o DSSP.

Assim, na primeira tentativa de melhora, selecionou-se para alteração uma região próxima à de maior pontuação, mas que pertencesse a um laço. Essa região foi a compreendida entre os aminoácisdos 93-97. Todo o processo de construção de novas conformações e cálculo da energia para todas as 1000 conformações durou 62 minutos e a minimização de energia das 50 melhores conformações, processo mais longo de toda a execução, durou 7 horas. Entretanto, nenhuma das cinco novas conformações apresentou valores de RMSD melhores que as do modelo originado pelo MODELLER.

Dessa forma, buscou-se alterar outras regiões, sendo selecionadas a 120-124 e 161-165. O tempo gasto para a construção das novas conformações e cálculo de suas energias, assim como o tempo de execução do processo de minimização foi o mesmo

apresentado para a região 93-97. E do mesmo modo que ocorreu anteriormente, nenhuma evolução em relação ao modelo criado pelo MODELLER foi indicada pelos valores de RMSD.

A explicação para os fracos resultados apresentados neste primeiro teste é a seguinte: o modelo originado do MODELLER está bastante próximo do modelo oriundo do laboratório; por exemplo, o RMSD-Local da região 93-97 apresenta um valor de 1.0894. Mas se for analisada a **Tabela 4.1.II** para o RMSD-Local para regiões de tamanho 5, encontra-se o valor 0.76 para os melhores casos, o que pode levar a concluir, erroneamente, que a execução deveria ser capaz de melhorar a região.

Na execução da ferramenta, não se escolhe a melhor conformação, já que não é possível determiná-la. O que se seleciona é o modelo de menor energia após a minimização de toda a estrutura, o que acarreta influência de outras regiões que não estão necessariamente corretas na homologia. Conseqüentemente, nem sempre a melhor conformação é escolhida. Ou seja, esse valor 0.76 é verdadeiro apenas quando o restante da estrutura está bem formulado, o que não é necessariamente o caso, já que o modelo de partida é originado da homologia.

### 4.2.2. T0130

O segundo caso de uso envolveu a proteína T0130, correspondente à estrutura 1NO5 no PDB, apresentando um resultado mais satisfatório na melhora das conformações, quando comparado com o teste anterior.

Neste exemplo, o tempo de execução do MODELLER para a construção dos cinco modelos foi de 25 minutos, inferior quando comparado ao T0129, devido ao seu menor comprimento, de apenas 114 aminoácidos. O tempo também é reduzido na seleção e pontuação dos aminoácidos do melhor modelo gerado pelo MODELLER, sendo de apenas 5 segundos. A pontuação para o modelo é descrita na **Figura 4.2.2.I**.

Assim como aconteceu com a proteína T0129, a região de maior pontuação da T0130, entre os aminoácidos 60-64, se encontra numa área que não é laço. Dessa forma, a região selecionada para a modificação foi a 71-75, pois pertencia a um trecho de laço, e era uma região de alta pontuação próxima à região indicada pelo programa.

O processo de construção das 1000 conformações durou 1hora e 13minutos. O cálculo da energia das 1000 novas conformações durou 1hora e 7minutos, enquanto que o processo de minimização dos modelos com as 50 melhores novas conformações foi reduzido para 2horas e 28minutos, quando comparado com o caso anterior.

```
           3         13        23        33        43
           ----------------------------------------------------
 anolea    *****_____*___*****_____*************_*****___***
 procheck  ----------------------------------------------------
 prosa     --*****************_*************_________*******
 prove     ____*_____*__*___*____*_____*__*__*__________*__*_
           11223111113112222320112122223223222111011112121232

           53        63        73        83        93
           ----------------------------------------------------
 anolea    --*********************************************
 procheck  ------------*--------------------------------------
 prosa     *********************************************
 prove     ____*__*_***________***_______*_*_*____*____*_____
           11223223233323222222333222222232323222322232222222

           103
           ----------
 anolea    **********
 procheck  ----------
 prosa     *******___
 prove     *_**______
           3233222111

LAÇOS DO DSSP
 1 12 25 46 55 63 79  87
 2 13 34 48 56 75 83 102
```

**Figura 4.2.2.I: Avaliação T0130-01** – Avaliação da estrutura originada do MODELLER e o resultado da execução da ferramenta de determinação de laços, o DSSP.

Cada uma das cinco estruturas selecionadas após a minimização foi comparada com o modelo original do MODELLER, e todas apresentaram valores de RMSD melhores. A estrutura que apresentou uma maior evolução de similaridade com relação ao modelo da T0130 obtida por laboratório foi escolhida para dar continuidade ao processo, sendo denominado por Modelo-01. A **Tabela 4.2.2.I** mostra os valores dos RMSD. A **Figura 4.2.2.II** mostra as conformações espaciais da região compreendida entre os aminoácidos da posição 71 e 75 do modelo originado do MODELLER e do novo modelo selecionado.

| | RMSD Total (Å) | RMSD-R Total (71-75) (Å) |
|---|---|---|
| **MODELLER** | 3.0939 | 2.3385 |
| **Modelo-01** | 3.061 | 2.229 |

**Tabela 4.2.2.I: Seleção T0130-01** – Os valores para o RMSD-Local considerando: todos os átomos pesados da proteína (RMSD Total), apenas os átomos pesados da região compreendida entre os aminoácidos 71-75 (RMSD-R Total).

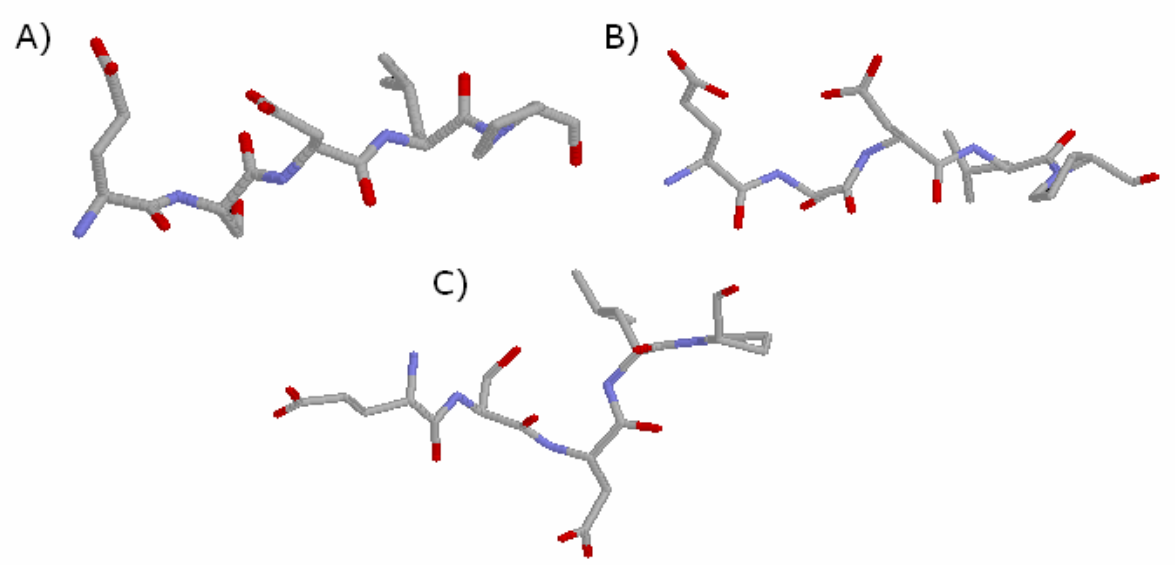



**Figura 4.2.2.II: Região 3D-01** – Representação da região 71-75 da estrutura: A) do MODELLER; B) do Modelo-01; C) Nativa.

Com o novo modelo, o processo é reiniciado, com a pontuação dos aminoácidos a partir das ferramentas de avaliação representado pela **Figura 4.2.2.III**. Mais uma vez a região de maior pontuação é formada por átomos que não fazem parte de um laço, nesse caso entre a posição 5 e a 9. Assim, a região selecionada a ser modificada foi alguma próxima a 5-9, como alta pontuação e pertencente a um laço. A região 9-13 se enquadrava nesse perfil.

A construção das 1000 conformações durou apenas 24 minutos, levando a crer que uma estrutura minimizada consome menos tempo na formação das conformações. O cálculo da energia dos 1000 novos modelos criados a partir das novas conformações ocorreu em 40 minutos. A minimização dos modelos com as 50 melhores novas conformações foi de apenas 7 minutos, pois parte do modelo já fora minimizado no primeiro ciclo.

As cinco melhores conformações foram comparadas com o modelo, e todas apresentaram uma evolução, considerando a análise pelos valores do RMSD. A **Tabela 4.2.2.II** mostra as medidas de similaridade da conformação de melhor resultado, Modelo-02, e do Modelo-01 quando comparados à estrutura 1NO5 presente no PDB. A análise dessa tabela mostra que houve uma boa aproximação da nova conformação com relação ao modelo do PDB. A **Figura 4.2.2.IV** representa as conformações da região entre os aminoácidos 9 e 13 do Modelo-01 e do Modelo-02 .

```
           3         13         23         33         43
            ------------------------------------------------
 anolea     ************************************************
 procheck   ------*--*-------------------------------*--*-*----
 prosa      --*****************_*********_*-----------*******
 prove      -*-*--*------*--*-*-*-*--*--*-----*---------------
            12232242232223223231323223223212112111112113232222

            53         63         73         83         93
            ------------------------------------------------
 anolea     ********____*********************_____***********
 procheck   ----------------*-----------------------------------*
 prosa      ***__*******************************************
 prove      *--*--*----*--**-------**---*-*-**-*--*-**-*------
            32221232111222333222222332223232321211323323222223

            103
            ----------
 anolea     *--*******
 procheck   *------*--
 prosa      *******---
 prove      --*-*-**--
            3122323311

LAÇOS DO DSSP
 1   8 21 45 62 80 86   113
 3 13 37 54 76 82 102 114
```

**Figura 4.2.2.III: Avaliação T0130-02** – Avaliação da estrutura originada do MODELLER e o resultado da execução da ferramenta de determinação de laços, o DSSP.

| | RMSD Total (Å) | RMSD-R Total (9-13) (Å) |
|---|---|---|
| **Modelo-01** | 3.061 | 2.1511 |
| **Modelo-02** | 2.8621 | 1.4183 |

**Tabela 4.2.2.II: Seleção T0130-02** – Os valores para o RMSD-Local considerando todos os átomos pesados da proteína (RMSD Total), aplicando apenas os átomos pesados da região (RMSD-R Total) e considerando apenas os Cα da região (RMSD-R Cα).

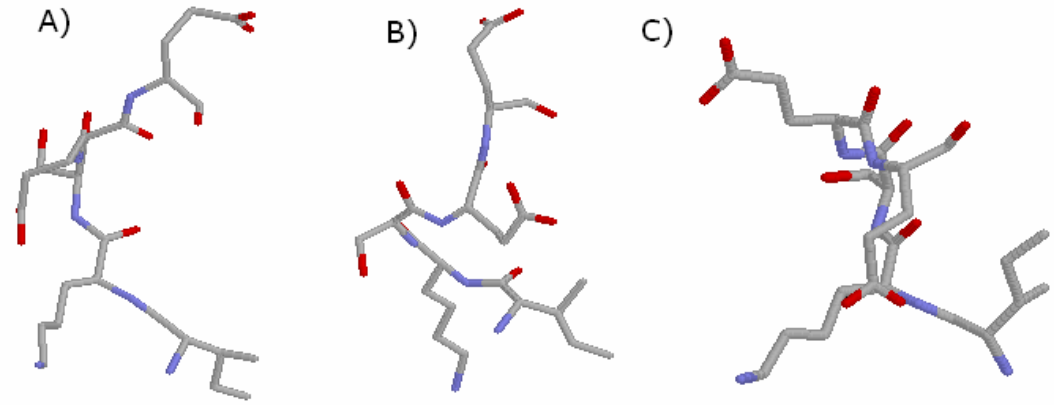



**Figura 4.2.2.IV: Região 3D-02** – Representação da região 9-13 da estrutura: A) do Modelo-01; B) do Modelo-02; C) Nativa.

Com um novo modelo selecionado, deu-se continuidade à execução do programa, com a determinação da pontuação dos aminoácidos, a construção das 1000 conformações, o cálculo da energia, a seleção das 50 melhores e a minimização.

A **Tabela 4.2.2.III: Análise T0130** mostra um resumo de todos os testes realizados com a proteína T0130. A primeira coluna indica os modelos, a coluna RMSD-Total corresponde ao RMSD-Local do modelo com a estrutura do PDB, considerando todos os átomos pesados. Com exceção do modelo originado pelo MODELLER e do último, todos os demais estão associados a duas linhas. Na primeira linha de cada modelo, a coluna regiões indica o trecho de proteína que foi alterado, enquanto que a última coluna indica o RMSD-Local dessa região em relação ao modelo do PDB, considerando os átomos pesados. A segunda linha para cada modelo é usada para indicar o RMSD-Local da nova região a ser alterada.

Em todos os casos houve uma melhora do RMSD considerando toda a estrutura ou apenas a região, aplicando os átomos pesados, mostrando que quando a ferramenta MODELLER não cria bons modelos o algoritmo de alteração das conformações ajuda bastante no processo de melhoramento da estrutura.

| | RMSD Total | Regiões | RMSD-R Total (Å) |
|---|---|---|---|
| **MODELLER** | 3.0939 | 71-75 | 2.3385 |
| **Modelo-01** | 3.061 | 71-75 | 2.229 |
| | | 09-13 | 2.1511 |
| **Modelo-02** | 2.8621 | 09-13 | 1.4183 |
| | | 45-49 | 1.8054 |
| **Modelo-03** | 2.7563 | 45-49 | 1.7664 |
| | | 88-92 | 1.6754 |
| **Modelo-04** | 2.5749 | 88-92 | 1.6567 |

**Tabela 4.2.2.III: Análise T0130** – Os valores para o RMSD-Local considerando todos os átomos pesados da proteína (RMSD Total), aplicando apenas os átomos pesados da região (RMSD-R Total) e considerando apenas os Cα da região (RMSD-R Cα) para as execuções realizadas com o T0130.

A **Figura 4.2.2.V** representa as conformações da região entre os aminoácidos 45 e 49 do Modelo-02 e do Modelo-03, enquanto que a **Figura 4.2.2.VI** representa as conformações dos aminoácidos entre as posições 88 e 92 do Modelo-03 e do Modelo-04.

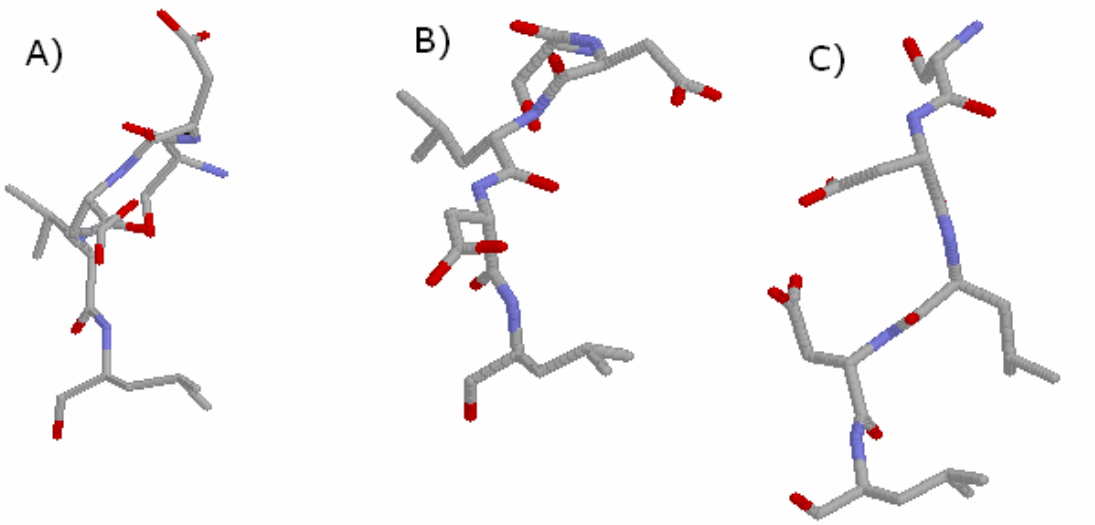



**Figura 4.2.2.V: Região 3D-03** – Representação da região 45-49 da estrutura: A) do Modelo-02; B) do Modelo-03; C) Nativa.

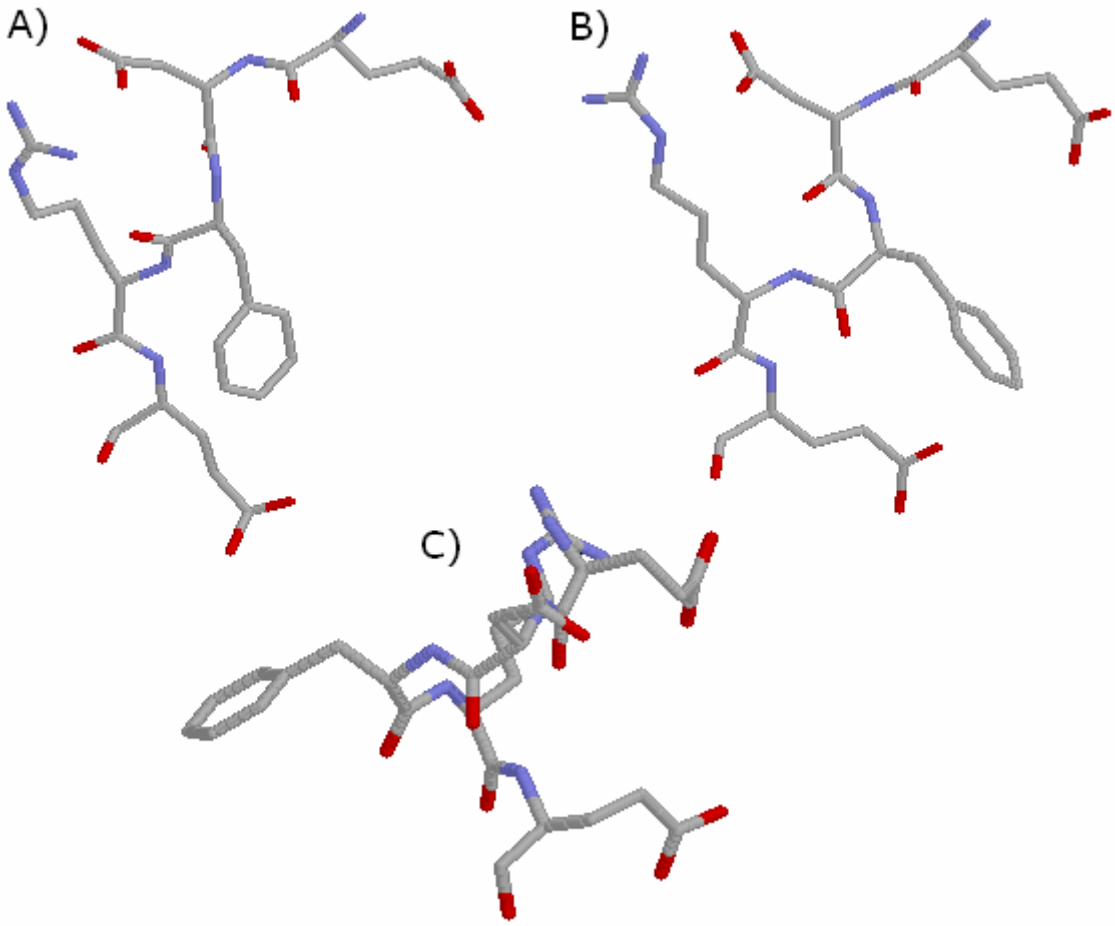



**Figura 4.2.2.VI: Região 3D-04** – Representação da região 88-92 da estrutura: A) do Modelo-03; B) do Modelo-04; C) Nativa.

Esses experimentos realizados mostram que apesar de alguns resultados indesejados, seja quando comparada ao trabalho de DePristo ou pela limitação de

aproximação a estruturas nativas (Seção 4.2.1), o projeto desenvolvido apresenta resultados promissores. Isto porque a ferramenta consegue aproximar conformações de regiões da estrutura nativa, mesmo no caso em que o modelo inicial é originado de um processo de homologia, que não garante certeza do posicionamento dos aminoácidos presentes no restante do modelo inicial.

Além desses, outros testes foram realizados, tendo por objetivo criar uma idéia do tempo necessário para a execução do primeiro ciclo de modelagem. A **Tabela 4.2.2.IV: Análise do Tempo** apresenta sete colunas: a primeira (Código) corresponde ao código da proteínas oriunda do conjunto do CASP5, a segunda (Tamanho) indica quantos aminoácidos estão presentes na proteína. A terceira coluna indica o tempo gasto para a execução do MODELLER e conseqüente criação do modelo inicial. A coluna Seleção mostra o tempo gasto para a execução das ferramentas de avaliação e para a indicação da região de maior pontuação. A quinta coluna indica o tempo decorrido para a criação das 1000 conformações, e a sexta coluna, o tempo gasto para o calculo da energia das 1000 conformações. Finalmente, a última coluna mostra o tempo gasto para a minimização de 50 conformações.

| Código | Tamanho | MODELLER | Seleção | 1000 conf | Energia | Minim 50 |
|---|---|---|---|---|---|---|
| **T0170** | 69 | 14min | 5seg | 37min | 58min | 1h56min |
| **T0150** | 100 | 41min | 6seg | 16min | 1h13min | 3h45min |
| **T0130** | 114 | 25min | 5seg | 1h13min | 1h7min | 2h28min |
| **T0137** | 133 | 32min | 5seg | 1h32min | 1h12min | 3h13min |
| **T0132** | 154 | 31min | 5seg | 1h54min | 1h23min | 5h52min |
| **T0129** | 182 | 43min | 8seg | 2h16min | 53min | 7h |
| **T0193** | 211 | 49min | 6seg | 24min | 1h55min | 7h52min |

**Tabela 4.2.2.IV: Análise de tempo** – Mostra os tempos gastos na execução do primeiro ciclo do processo de melhoramento da estrutura para seqüências extraídas do CASP5.

# Capítulo 5. Conclusão e Trabalhos futuros

Hasta La vista, Baby
T-101 – Terminator II: Judgment Day (1991)

Este documento descreveu o estudo realizado do processo de modelagem e melhora de estruturas de proteínas, assim como uma implementação inicial de um fluxo para a construção das conformações das proteínas, proporcionando a execução, em um mesmo programa, das etapas de modelagem por homologia, avaliação de regiões defeituosas das proteínas e modificação local das conformações das estruturas. Este projeto teve o intuito de dar o primeiro passo no desenvolvimento de uma ferramenta que torne a modelagem um processo mais rápido e acessível.

A execução de cada uma das etapas pode ser realizada de forma independente ou em conjunto, de acordo com os objetivos do usuário. Por exemplo, se apenas a seqüência da proteína está ao alcance do usuário, ele deve iniciar o processo pela modelagem. Caso a estrutura terciária e as regiões defeituosas já sejam de conhecimento do usuário, este pode pular as duas etapas iniciais e executar a alteração das conformações de regiões das proteínas.

No desenvolvimento deste projeto, o objetivo básico era facilitar o processo de modelagem das estruturas de proteínas para os principais profissionais interessados: os biólogos e químicos. Isto foi feito ao acomodar as principais ferramentas utilizadas na modelagem de uma forma mais prática, onde não houvesse a necessidade de executar cada um dos programas de forma independente e conseqüentemente trabalhar com diversos formatos de arquivos, tanto para a entrada como para a saída. E apesar de ainda estar apenas implementado como uma aplicação para ambiente Unix, o que não é o ideal, o projeto pode ser facilmente executado, com mínima interferência do usuário. Além disso, quando esta interferência é necessária, as decisões do usuário não dependem de todos os arquivos intermediários criados até aquele ponto do processo. As decisões podem ser baseadas em alguns pequenos trechos de alguns arquivos. O programa disponibiliza tais regiões de arquivos essenciais para guiar a decisão, não havendo a necessidade do usuário se preocupar em procurar as informações nos diversos arquivos.

Não só promover os trabalhos dos futuros usuários foi considerado durante a criação da ferramenta, também houve a preocupação de criar um código bastante modularizado a fim de facilitar futuras modificações e melhoras do projeto, sejam elas

feitas por analistas de sistemas ou químicos computacionais. Assim, a implementação foi feita de modo a permitir alterações em cada uma das etapas (modelagem, avaliação e alteração das estruturas) de forma independente, sem alterar o funcionamento do restante da ferramenta.

Durante o projeto, buscou-se a compreensão do processo da modelagem da melhor maneira possível, assimilando todos os detalhes bioquímicos envolvidos. Esta necessidade de formar a melhor base possível de todo o procedimento acabou por consumir um período considerável do projeto, fazendo com que a exploração para a implementação e melhora das etapas, seja com a inclusão de novas idéias ou otimização dos algoritmos aplicados, tivesse um tempo bastante reduzido.

Mesmo assim, com pouco trabalho envolvendo estudos direcionados a evolução da ferramenta, os testes indicam que a aplicação tem condições de ser melhorada, principalmente se considerar que nenhum conhecimento biológico ou bioquímico mais específico foi aplicado para a construção das conformações das regiões e mesmo assim a ferramenta conseguiu aproximações para a estrutura nativa.

Além disso, por ter havido a preocupação de construir uma base sólida dos principais conceitos e mecanismos envolvidos, surgiram idéias que podem ser aplicadas na melhora do projeto, podendo ser discutidas, implementadas e testadas em trabalhos futuros, sempre visando a evolução da ferramenta. Essas possíveis modificações são divididas de acordo com cada uma das etapas, como descrito a seguir:

**Modelagem:** deve-se incluir outros parâmetros do MODELLER não aplicados neste projeto, e permitir ao usuário a alteração dos valores pré-definidos. Além disso, uma interface gráfica que facilite o manuseio dos parâmetros e a visualização dos trechos dos arquivos intermediários deve ser disponibilizada.

Algumas áreas críticas do processo de modelagem precisam dar a flexibilidade de uma maior interferência do usuário, como o alinhamento das estruturas das proteínas semelhantes à seqüência alvo. Nesse estágio, através de uma interface que representasse as estruturas 3D das proteínas, seria possível alterar manualmente os alinhamentos, inserindo *gaps* e modificando as posições dos aminoácidos com o *mouse*. Uma função baseada em campos de força informaria a variação de energia do alinhamento à medida que fosse alterado, dando uma referência de melhora ou piora para o usuário.

Outra possível modificação seria criar uma pontuação que exprimisse o relacionamento entre a seqüência alvo e os possíveis moldes, levando em consideração termos como: a pontuação dos alinhamentos, as regiões alinhadas, a família da proteína

molde e a precisão de modelagem da estrutura molde. Isto permitiria uma seleção automática das proteínas a serem aplicadas na modelagem, ou no mínimo serviria como um guia para a seleção do usuário. Para que se construísse esta pontuação, uma análise preliminar, que considerasse todos os termos citados, seria necessária.

**Avaliação:** sempre buscando construir um ambiente mais amigável ao usuário final, seria importante criar uma interface onde fosse permitido selecionar as ferramentas a serem utilizadas, além de disponibilizar outras ferramentas de análise, diferentes daquelas quatro já implementadas.

Outro formato para a visualização das regiões defeituosas seria disponibilizado, semelhante ao citado na etapa de modelagem. Seria possível ter a visualização 3D da proteína em estudo, com os trechos defeituosos em destaque, fossem eles originados de um consenso entre ferramentas ou de uma única. O usuário teria a liberdade de selecionar outros aminoácidos que considerasse estarem em posições espaciais indesejáveis.

Uma inovação interessante seria atribuir o peso a cada uma das ferramentas de avaliação de acordo com a proximidade das estruturas moldes com as seqüências do banco de dados utilizados por cada uma das ferramentas. Um estudo prévio visando encontrar um padrão entre as famílias das proteínas e os bancos de dados das ferramentas precisaria ser concretizado.

Por exemplo, considere uma seqüência alvo modelada a partir de estruturas moldes que apresentam pouca representatividade no banco de dados que PROCHECK considerou na determinação de seus valores padrão, mas com grande similaridade com diversas estruturas utilizadas no banco do ANoLEA. Nesse caso, a seleção de uma região defeituosa sofreria uma maior influência das regiões determinadas pela ferramenta ANoLEA que das indicadas pelo PROCHECK.

**Construção e seleção de novas conformações:** para a construção de novas conformações, poder-se-ia combinar o método *ab initio,* já utilizado, com a metodologia baseada em conhecimento. A função de energia de campos de força utilizada na construção poderia ser determinada dinamicamente, a partir do relacionamento entre as famílias dos moldes, a região a ser alterada e os tipos de campos de força. Um estudo preliminar associando família de proteínas e campo de força precisa ser realizado.

Atualmente a seleção das melhores conformações ocorre através da aplicação de um campo de força sem minimização. A mais provável modificação neste trecho do projeto seria a formação de uma função descrevendo os principais termos de um campo

de força de forma aproximada. Esta função seria utilizada como um filtro negativo, eliminado as conformações realmente não aceitáveis, e na determinação de uma pontuação para a classificação das conformações. Esta função também poderia ser gerada dinamicamente, com base na família dos moldes e da região alterada.

Na escolha do modelo, uma função de energia de campo de força seria aplicada, assim como ocorre no atual projeto, mas da mesma forma que nos dois casos anteriores a função poderia ser determinada dinamicamente.

Em todos os casos acima citados, seria permitida a alteração da função por parte do usuário. Por fim, o modelo final disponibilizado estaria suscetível a alterações manuais, de acordo com o conhecimento e vontade do usuário.

Uma vez finalizado todo o processo de modelagem, avaliação e melhoramento da estrutura, pode-se considerar a possibilidade de estender o projeto para problemas de interação entre proteínas, através da:

- Determinação de prováveis sítios ativos de enzimas a partir das estruturas 3-D com a utilização de técnicas de busca de cavidades.
- Determinação de inibidores de enzimas a partir das técnicas de "*docking*";

No mais, o trabalho mostra que mesmo aplicando poucos conceitos bioquímicos é possível fazer melhoras no processo de modelagem, ou seja, a ferramenta apresenta um grande potencial. Mas para que isso seja possível, é interessante que a continuidade do projeto não seja apenas realizada por conhecedores da computação, mas também que possa contar com o apoio dos profissionais de outras áreas, como a química e a biologia.

# Bibliografia


**[Adc04]** Adcock, S. A.; *Peptide backbone reconstruction using dead-end elimination and a knowledge-based forcefield*; Journal of Computational Chemistry 25(1):16-27 (2004).

**[Anf72]** Anfinsen, C. B.; *Studies on the Principles That Govern the Folding of Protein Chains;* The Nobel Lecture (1972).

**[ANoLEA]** ANOLEA (Atomic Non-Local EnvironmentAssessment). Disponível: *Pontificia Universidad Católica de Chile – Laboratório de Bioinformática Molecular.* Última visita: 11 de Janeiro, 2005. URL: http://protein.bio.puc.cl/cardex/servers/anolea/.

**[AQUA]** About AQUA. Disponível: *Bijvoet Center for Biomolecular Research*. Última visita: 11 de Janeiro, 2005. URL: http://www.nmr.chem.uu.nl/users/rull/pcheck_dir/nmr_manual/manaquaint.html.

**[BioJava]** BioJava. Disponível: *BioJava site.* Última visita: 11 de Janeiro, 2005. URL: http://www.biojava.org/.

**[BLAST]** NCBI BLAST. Disponível: *The National Center for Biotechnology Information (NCBI)* Última visita: 11 de Janeiro, 2005. URL: http://www.ncbi.nlm.nih.gov/BLAST/.

**[BLE91]** Bowie, J. U., Luthy, R., Eisenberg, D.; A method to identify protein sequences that fold into a known three-dimensional structure; Science 12;253(5016):164-70 (1991).

**[BS00]** Bernasconi, A., Segre, A. M.; *Ab initio methods for protein structure prediction: a new technique based on Ramachandran Plots*. ERCIM News Nº 43 October 2000.

**[CASP]** Bourne, P. E.; *CASP and CAFASP Experiments and their findings*. Structural Bioinformatics by Wiley-Liss, Inc(2003).

**[CATH]** CATH Protein Structure Classification Database (UCL). Disponível: *CATH* site (05 de Maio, 2005). URL: http://cathwww.biochem.ucl.ac. uk/latest/index.html.

**[CHARMM]** CHARMM. Disponível: *CHARMM* site (11 de Janeiro, 2005). URL: http://www.charmm.org/.

**[CSD03]** Canutescu, A. A., Shelenkov, A. A., Dunbrack, R. L. Jr.; *A graph theory algorithm for protein side-chain prediction*; Protein Science 12: 2001-2014 (2003).

**[CY93]** Colovos, C., Yeates, T. O.; *Verification of protein structures: patterns of nonbonded atomic interactions*; Protein Sci. 2(9):1511-9 (1993).

**[DALI]** EMBL Dali: email server for 3-D protein structure database searches. Disponível: *European Bioinformatic Institute.* Última visita: 11 de Janeiro, 2005. URL: http://www.ebi.ac.uk/dali/.

**[Day72]** Dayhoff M. O. *Atlas of protein sequence and structure, vol.5.* National Biomedical Research Foundation, Georgetown University, Washington, D.C. (1972).

**[DB00]** Deane, C. M., Blundell, T. L.; *A novel exhaustive search algorithm for predicting the conformation of polypeptide segments in proteins*; Proteins 40:135-144 (2000).

**[DB01]** Deane, C. M.,Blundell, T. L.; *CODA a combined algorithm for predicting the sctructally variable regions of protein models*; Protein Science 10:599-612 (2001).

**[DBL+03]** DePristo, M. A., Bakker, P. I. W., Lovell, S. C., Blundell T.L.; *Ab Initio Construction of Polypeptide Fragments: Efficient Generation of Accurate, Representative Ensembles.* PROTEINS: Structure, Function, and Genetics (2003) 51:41–55.

**[DDBJ]** DDBJ Homepage. Disponível: *National Institute of Genetics* site (11 de Janeiro, 2005). URL: http://www.ddbj.nig.ac.jp/.

**[EBI]** European Bioinformatics Institute. Última visita: 20 de Fevereiro, 2005. URL:http://www.ebi.ac.uk/

**[EBL+92]** Eisenberg, D., Bowie, J. U., Luthy R., Choe S.; *Three-dimensional profiles for analysing protein sequence-structure relationships*. Faraday Discuss. 93:25-34 (1992).

**[EH91]** Engh, R. A., Huber R.; *Accurate bond and angle parameters for X-ray protein structure refinement*; Acta Cryst. A47: 392-400 (1991).

**[EMBL]** EMBL - The European Molecular Biology Laboratory. Disponível: *EMBL* site (11 de Janeiro, 2005). URL: http://www.embl.org/.

**[ERRAT]** ERRAT Structure Submission Page. Disponível: *UCLA-DOE Institute for Genomics and Proteomics.* Última visita: 11 de Janeiro, 2005. URL: http://www.doe-mbi.ucla.edu/Services/ERRATv2/.

**[FASTA]** FASTA Sequence Comparison. Disponível: *University of Virginia* . Última visita: 11 de Janeiro, 2005. URL: http://fasta.bioch.virginia.edu/fasta_www/cgi/search_frm.cgi?pgm=fa

**[FDS00]** Fiser, A., Do, R. K. G., Šali, A.; *Modeling of loops in protein structures*. Protein Science 9:1753-1773 (2000).

**[FS01]** Fiser, A., Šali, A.; *Comparative protein structure modeling with Modeller: A practical approach.* Laboratories of Molecular Biophysics (2001).

**[GROMACS]** Welcome to GROMACS. Disponível: *GROMACS* site. Última visita: 11 de Janeiro, 2005. URL: http://www.gromacs.org/.

**[HT00]** Higgins, D., Taylor, W.; *Bioinformatics: sequences, structure and databanks*; Oxford University Press (2000).

**[KC99]** Klug, W. S., Cummings R. M.; *Essential of Genetics*; Pretence Hall (1999).

**[KLS02]** Kamierkiewicz, R., Liwo, A., Scheraga, H. A.; *Energy-based reconstruction of a protein backbone from its alpha-carbon trace by a Monte-Carlo method*; Journal of Computational Chemistry 23(7):715-723 (2002).

**[KS83]** Kabsch, W., Sander, C.; *Adictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features.* Biopolymers (1983) 22:2577–2637.

**[LBE92]** Luthy, R., Bowie, J. U., Eisenberg, D.; Assessment of protein models with three-dimensional profiles; Nature 5;356(6364):83-5 (1992).

**[LMM+93]** Laskowski, R. A., MacArthur, M. W., Moss, D. S., Thornton, J. M.; PROCHECK: a program to check the stereochemical quality of protein structures; J. Appl. Cryst. 26, 283-291 (1993).

**[LNC02]** Lehninger, A. L., Nelson, D. L., Cox, M. M.; *Lehninger Princípios de Bioquímica*; Sarvier (2002).

**[LRM+96]** Laskowski, R. A., Rullmann, J. A. C., MacArthur, M. W., Kaptein, R., Thornton, J. M.; AQUA and PROCHECK-NMR: Programs for checking the quality of protein structures solved by NMR;J. Biomol. NMR 8(4):477-86 (1996).

**[MCR89]** Martin, A. C. R., Cheetham, J. C., Rees, A. R.; *Modeling Antibody Hypervariable Loops: A Combined Algorithm*; Proc Natl Acad Sci U S A. 86(23): 9268–9272 (1989).

**[MDD+97]** Melo, F., Devos, D., Depiereux, E., Feytmans, E.; *ANOLEA: a www server to assess protein structures*; Intelligent Systems for Molecular Biology, 110-113 (1997).

**[MF97]** Melo, F., Feytmans, E.; *Novel Knowledge-based Mean Force Potential at Atomic Leve*l; J. Mol. Biol. 267, 207-222 (1997).

**[MF98]** Melo, F., Feytmans, E.; *Assessing Protein Structures with a Non-local Atomic Interaction Energy*; J. Mol. Biol. (1998) 277, 1141-1152.

**[MG77]** Maxam A. M., Gilbert W. *A new method for sequencing DNA*. Proc. Natl. Acad. Sci. 74:560-564 (1977).

**[MJ86]** Moult, J., James, M. N.; An algorithm for determining the conformation of polypeptide segments in proteins by systematic search; Proteins 1:146-163 (1986).

**[MMH$^+$92]** Morris, A. L., MacArthur M. W., Hutchinson E. G., Thornton, J. M.; *Stereochemical quality of protein structure coordinates*; Proteins 12: 345-364 (1992).

**[MODELLER]** About MODELLER. Disponível: *Andrej Sali Lab.* Última visita: 11 de Janeiro, 2005. URL: http://salilab.org/modeller/.

**[Mou01]** Mount, D. W.; *Bioinformatics: sequence e genoma analysis*; Cold Spring Harbor Laboratory Press (2001).

**[MSF$^+$00]** Martí-Renom, M. A., Stuart, A. C., Fiser, A., Sánchez, R., Melo F., Šali, A.; *Comparative protein structure modeling of genes and genomes*. Annu. Rev. Biophys. Biomol. Struct. 29:291-325 (2000).

**[NIH]** National Institutes of Health (NIH). Disponível: *NIH* site (11 de Janeiro, 2005). URL: http://www.nih.gov/.

**[NW70]** Needleman, S. B., Wunsch, C. D. *A general method applicable to the search for similarities in the amino acid sequence of two proteins*; J. Mol. Biol.48:443-453 (1970).

**[PDB]** The RCSB Protein Data Bank. Disponível: *Research Collaboratory for Structural Bioinformatics* (11de Janeiro, 2005). URL: http://www.rcsb.org/pdb/.

**[Pil02]** Pillardy, J.; *Ab-initio protein structure prediction*; Computational Biology Service Unit (2002).

**[PIR]** Protein Information Resource Home Page - Protein Sequence Database - iProClass. Disponível: Georgetown University Medical Center. Última visita: 11 de Janeiro, 2003; URL: http://pir.georgetown.edu/.

**[PROCHECK]** PROCHECK – protein structure checks. Disponível: *UCL Department of Biochemistry & Molecular Biology.* Última visita: 11 de Janeiro, 2005. URL: http://www.biochem.ucl.ac.uk/~roman/

procheck/ procheck.html

**[ProSa]** PROSAII Page. Disponível: *University of Salzburg - Center of Applied Molecular Engineering site.* Última Visita: 11 de Janeiro, 2005. URL: http://www.came.sbg.ac.at/Services/prosa.html

**[PROVE]** ULB - SCMBB: Developed Software. Disponível: *SCMBB - Service de Conformation des Macromolécules Biologiques et de Bioinformatique, Université Libre de Bruxelles.* Última visita: 11 de Janeiro, 2005. URL: http://www.ucmb.ulb.ac.be/UCMB/PROVE/

**[PRW96]** Pontius J., Richelle J., Wodak S. J.; Deviations from standard atomic volumes as a quality measure for protein crystal structures; J. Mol. Biol. 264(1): 121-136 (1996).

**[RS92]** Rey, A., Skolnick, J.; Efficient Algorithm for the Reconstruction of a Protein Backbone from the alpha-carbon Coordinates. J.Comp.Chem. 13:443-456 (1992).

**[SB93]** Šali, A., Blundell, T. B.; *Comparative protein modelling by satisfaction of spatial restraints*; J. Mol. Biol. 243:779-815 (1993).

**[SCOP]** SCOP: Structural Classification of Proteins. Disponível: *Berkeley University of Califórnia.* Última visita: 11 de Janeiro, 2005. URL: http://scop.berkeley.edu/.

**[SIB]** Swiss Institute of Bioinformatics. Última visita: 20 de Fevereiro, 2005. URL: http://www.isb-sib.ch/

**[Sip90]** 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. J Mol Biol. 213:859-883 (1990).

**[Sip93]** Sippl, M. J.; Recognition of Errors in Three-Dimensional Structures of Proteins; Proteins (1993) 17: 355-362.

**[Sip95]** Sippl, M. J.; *Knowledge-based potentials for proteins*. Curr. Opin. Struct. Biol. 5:229-235 (1995).

**[SM98]** Samudrala, R., Moult, J.; *An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction*; J Mol Biol. 275:895-916 (1998).

**[SNC77]** Sanger F., Nicklen S., Coulson A. R. *DNA sequencing with chain terminating inhibitors.* Proc. Natl. Acad. Sci. 74:5463-5467 (1977);

**[SPY+95]** Šali, A., Potterton, L., Yuan, F., Vlijmen, H., Karplus, M.; *Evaluation on comparative protein modeling by MODELLER*; Proteins: Structure, Function, and Genetics 23:318-326 (1995).

**[ST51]** Sanger F., Tuppy H. *The amino acid sequence of the phenylalanylchain of insulin.* Biochem. J. 49:481-490 (1951).

**[Swiss-Prot]** EsPASy – Swiss-Protand TrEMBL. Disponível: *ExPASy Proteomics Server*. Última visita 11 de Janeiro, 2005. URL: http://au.expasy.org/sprot/

**[TINKER]** TINKER. Disponível: *Unilever Centre for Molecular Science Informatics from Cambridge University* (11de Janeiro, 2005). URL: http://www.ch.cam.ac.uk/cil/tinker/.

**[Top500]** Top500 DataBase. Disponível: *Richardson Lab & Kinemage Home Page.* Última visita: 11de Janeiro, 2005. URL: http://kinemage.biochem.duke.edu/databases/top500.php.

**[UHW+89]** Unger, R., Harel, D., Wherland, S., Sussman, J. L.; *A 3D building blocks approach to analyzing and predicting structure of proteins*; Proteins 5(4):355-73 (1989).

**[Verify3D]** Verify_3D. Disponível: *UCLA-DOE Institute for Genomics and Proteomics.* Última visita: 11 de Janeiro, 2005. URL: http://www.doe-mbi.ucla.edu/Services/Verify_3D/.

**[WC53]** Watson J. D., Crick F. H. *Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid.* Nature 171:737-8 (1953).

**[WHATCHECK]** WHAT_CHECK: free WHAT IF verification subset. Disponível: *University of Nijmegen – Centre for molecular and Biomolecular Informatics.* Última visita: 11 de Janeiro, 2005; URL: http://www.cmbi.kun.nl/swift/whatcheck/

**[WLL+99]** Word, J. M., Lovell, S. C., LaBean, T. H., Taylor, H. C., Zalis, M. E., Presley, B.K., Richardson, J. S., Richardson, D. C.; *Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms*; J Mol Biol 285:1711-1733 (1999).

**[WS99]** Wedemeyer, W. J., Scheraga, H. A.; *Exact Analytical Loop Closure in Proteins Using Polynomial Equations*; J. Comput. Chem. 20: 819-844 (1999).

# Glossário

## A

**Alinhamento**
Comparação entre duas ou mais seqüências procurando caracteres individuais ou padrões de caracteres que estão em uma mesma ordem nas seqüências.

**Análise de componentes principais**
Reduzir um grande conjunto de variáveis para um conjunto menor, mas que ainda represente a maioria da informação.

**Âmgström**
Unidade de medida de comprimento, representado pelo símbolo Å, igual a $10^{-10}$ m ou 10 nm.

**Árvore filogenética**
Representação gráfica em forma de árvore onde cada unidade é posicionada de acordo com as relações de parentesco, ou seja, a ordem filogenética.

## C

**cDNA**
Molécula similar ao DNA, porém sintetizada em laboratório com aplicação da enzima transcriptase reversa, que reproduz DNA a partir de uma cópia de molécula de RNA.

***Clustering***
Método para agrupar um conjunto de objetos que são mais similares em um grande grupo de objetos relacionados.

**Componente biconexo**
Grafos que não podem ser desconectados com a remoção de uma simples aresta, ou seja, após a remoção de uma aresta, ainda é possível atingir um vértice a partir de qualquer outro.

## D

**Dendrograma**
Também definido como diagrama de árvore ou árvore hierárquica, representa o agrupamento hierárquico de objetos, construído a partir de uma matriz de proximidade desses objetos. Os casos semelhantes são unidos em grupos cujas posições no dendrograma são determinadas pelo nível de (dis)similaridade entre eles.

**Dinâmica molecular**
Metodologia de simulação computacional da evolução temporal do arranjo estrutural de sistemas moleculares, compostos por átomos e moléculas, sob as leis da Mecânica Clássica.

**Distribuição de probabilidade**
Curva matemática que associa uma probabilidade a cada valor, ou faixa de valor, de uma variável.

***Docking***
Conformação resultante do processo de interação entre duas moléculas.

**Domínio (contexto estrutural)**
Segmento de uma cadeia de polipeptídio que apresenta uma conformação tridimensional independente de qualquer outro segmento da cadeia.

## E

**Estrutura Nativa**
A conformação tridimensional de uma proteína que apresenta funcionalidade. Qualquer outra conformação de uma proteína não apresenta funcionalidade.

## F

**Família de proteína (contexto estrutural)**
Duas estruturas que apresentam similaridade estrutural, mas não necessariamente similaridade de seqüência.

**Fator-B**
Também chamado fator de temperatura, descreve a flexibilidade da posição dos átomos na estrutura.

**Fator-R**
Fator originado no processo de cristalografia a partir de uma função de correlação entre as curvas experimentais e teóricas do processo de modelagem.

**Função de densidade de probabilidade densidade (pdf)**
É uma expressão que determina a probabilidade de uma variável assumir um valor x. Numa distribuição contínua, a probabilidade de um simples ponto é nula, e a de um intervalo compreendido entre os pontos a e b é dada por:

$$\int_a^b f(x)dx = Pr[a \leq X \leq b]$$

Para uma distribuição discreta, a fpd é determinada por:

$$f(x) = Pr[X = x]$$

## G

**Gene**
Seqüência de nucleotídeos que codificam seqüências de aminoácidos que irão formar proteínas ou envolvem regulação genética.

**Genoma**
Todo o material genético do organismo.

**Gradiente Conjugado**
Método interativo para a resolução de sistemas lineares. Aplicado quando se requer um armazenamento menor de informações, vantagem significativa quando o número de variáveis é muito grande, minimizando uma função quadrática *n*-dimensional, com no máximo *n* iterações.

## L

**Las Vegas**

Técnica semelhante ao processo de Monte Carlo, com a diferença de sempre retornar resultados corretos, mas com o tempo de execução indeterminado.

## M

**Monte Carlo**

Técnica que aplica números randômicos e probabilidade para a resolução de problemas matemáticos através da criação de múltiplos cenários para a solução, considerando a distribuição de probabilidade e o intervalo de valores aceitáveis como resposta do problema. Outra algoritmo de randomização é o Las Vegas.

## N

**NOEs - *Nuclear Overhauser Effects***

Quando dois átomos estão próximos, há uma influência dos *spins* de um com os do outro, gerada pelos campos magnéticos de cada um dos *spins*. Através de processos de ressonância magnética nuclear pode-se mudar a orientação dos *spins* ou parar um deles e verificar a influência dessa modificação. O NOEs é aplicado para determinar as restrições entre prótons para estruturas determinadas por ressonância magnética nuclear.

## P

**Propriedades estereoquímicas**

Propriedades relacionadas à disposição espacial dos átomos da molécula.

**Proteína globular**

Enovelamento de cadeias polipeptídicas em duas ou mais regiões compactas, semi-independentes, unidas por segmentos lineares e flexíveis de cadeia polipeptídica. Cerca de 50% da estrutura é associada à α-hélice ou à folha-β.

## R

**Raio de Van der Waals**

Metade da distância entre dois átomos equivalentes não ligados, em sua distribuição mais estável.

## S

***Simulated annealaing***

Aplicação de Monte Carlo para minimização de funções. O termo deriva de uma aproximação grosseira com o processo físico análogo de aquecer e esfriar uma substancia para obter a estrutura cristalizada. Na execução do *simulated annealing*, o sistema apresenta uma configuração inicial associada a uma temperatura inicial. Uma nova configuração é associada por um processo randômico. Se a energia desse novo modelo for menor do que a do anterior, a mudança é aceita e o sistema é atualizado. Caso a energia seja maior, uma probabilidade em função da temperatura atual determina se o novo modelo será aceito, dando ao sistema a flexibilidade do sistema de evitar mínimos locais. A cada nova atualização do sistema, a temperatura é reduzida. Quando a temperatura atinge o valor zero, o novo modelo só será aceito caso

a energia para o novo modelo seja menor que a do anterior. O critério de parada do processo é o número de ciclos realizados.

**Spins**

Momento magnético intrínseco que define a natureza das partículas, isto é, férmions têm *spin* semi-inteiro e bósons têm *spin* inteiro.

# Apêndice FASTA

Uma seqüência no formato FASTA, seja ela de nucleotídeos ou aminoácidos, é constituída por duas partes: a primeira seção corresponde a uma única linha iniciada pelo caractere “>” descrevendo o nome da seqüência, o código que a representa e até o organismo de origem; a segunda seção são linhas contendo a seqüência propriamente dita, recomendando que não apresentem mais que 80 caracteres de comprimento.

```
>gi|532319|pir|TVFV2E|TVFV2E envelope protein
ELRLRYCAPAGFALLKCNDADYDGFKTNCSNVSVVHCTNLMNTTVTTGLLLNGSYSENRT
QIWQKHRTSNDSALILLNKHYNLTVTCKRPGNKTVLPVTIMAGLVFHSQKYNLRLRQAWC
HFPSNWKGAWKEVKEEIVNLPKERYRGTNDPKRIFFQRQWGDPETANLWFNCHGEFFYCK
MDWFLNYLNNLTVDADHNECKNTSGTKSGNKRAPGPCVQRTYVACHIRSVIIWLETISKK
TYAPPREGHLECTSTVTGMTVELNYIPKNRTNVTLSPQIESIWAAELDRYKLVEITPIGF
APTEVRRYTGGHERQKRVPFVXXXXXXXXXXXXXXXXXXXXXXVQSQHLLAGILQQQKNL
LAAVEAQQQMLKLTIWGVK
```

Os códigos aplicados para os ácidos nucléicos seguem o padrão IUB/IUPAC.

```
A --> adenosina           M --> A C (amino)
C --> citosina            S --> G C (forte)
G --> guanina             W --> A T (fraca)
T --> timina              B --> G T C
U --> uridina             D --> G A T
R --> G A (purina)        H --> A C T
Y --> T C (pirimidina)    V --> G C A
K --> G T (keto)          N --> A G C T (qualquer uma das bases)
                          -  gap de comprimento indeterminado
```

Para as seqüências de aminoácidos, os códigos também estão no padrão determinado pela IUB/IUPAC.

```
A  alanina                    P  prolina
B  aspartato ou asparigina    Q  glutamina
C  cisteína                   R  arginina
D  aspartato                  S  serina
E  glutamato                  T  treonina
F  fenilalanina               U  selenocisteína
G  glicina                    V  valina
H  histidina                  W  triptofano
I  isoleucina                 Y  tirosina
K  lisina                     Z  glutamato or glutamina
L  leucina                    X  qualquer aminoácido
M  metionina                  *  códon de parada
N  asparagina                 -  gap de comprimento indeterminado
```

# Apêndice GenBank

A seguir é apresentado um exemplo do formato em que uma seqüência de proteína é armazenada no banco de dados GenBank. O exemplo utilizado corresponde à proteína de código 1FDN, a ferridoxina, originada da bactéria *Clostridium acidurici*.

```
LOCUS       1FDN                     55 aa            BCT 04-SEP-1998
DEFINITION  Ferredoxin.
ACCESSION   1FDN
VERSION     1FDN  GI:640397
KEYWORDS    .
SOURCE      Clostridium acidurici
  ORGANISM  Clostridium acidurici
            Bacteria; Firmicutes; Clostridia; Clostridiales; Clostridiaceae;
            Clostridium.
REFERENCE   1  (residues 1 to 55)
  AUTHORS   Murthy,H.M., Hendrickson,W.A., Orme-Johnson,W.H., Merritt,E.A. and
            Phizackerley,R.P.
  TITLE     Crystal structure of Clostridium acidi-urici ferredoxin at 5-A
            resolution based on measurements of anomalous X-ray scattering at
            multiple wavelengths
  JOURNAL   J. Biol. Chem. 263 (34), 18430-18436 (1988)
  MEDLINE   89054025
   PUBMED   3192542
FEATURES             Location/Qualifiers
     source          1..55
                     /organism="Clostridium acidurici"
                     /db_xref="taxon:1556"
     SecStr          1..5
                     /sec_str_type="sheet"
                     /note="strand 1"
     Het             join(bond(8),bond(11),bond(14),bond(47))
                     /heterogen="(FS4,  56 ) IronSULFUR CLUSTER"
     Het             join(bond(18),bond(37),bond(40),bond(43))
                     /heterogen="(FS4,  57 ) IronSULFUR CLUSTER"
     SecStr          22..27
                     /sec_str_type="sheet"
                     /note="strand 2"
     SecStr          28..33
                     /sec_str_type="sheet"
                     /note="strand 3"
     SecStr          51..55
                     /sec_str_type="sheet"
                     /note="strand 4"
ORIGIN
        1 ayvineacis cgacepecpv naissgddry vidadtcidc gacagvcpvd apvqa
```

O campo **LOCUS** corresponde ao código da proteína, o comprimento da seqüência e a data da última atualização. **DEFINITION** descreve resumidamente a seqüência. **ACCESSION** é o número de acesso primário, único e não passível de alteração, freqüentemente usado para citar seqüências nos jornais. **VERSION** indica o número de acesso primário mais uma chave associada à seqüência pelo NCBI. **KEYWORDS** são pequenas sentenças que descrevem a proteína. **SOURCE** é o nome do organismo ou a denominação mais freqüente na literatura da qual a seqüência em análise faz parte. **ORGANISM** é a denominação formal do organismo, mais a classificação taxonômica. **REFERENCE** são as

citações de todos os artigos contendo os dados registrados nesta estrada. **AUTHORS** é lista dos autores da citação. **TITLE** é o título da citação. **JOURNAL** lista o nome do jornal, volume, ano e número das páginas da citação. **MEDLINE** fornece o identificador único Medline (base de dados de bibliografia de citações e abstracts do NLM) para a citação. **PUBMED** corresponde ao identificador único da citação para o PubMed (serviço do NLM incluindo milhares de citações e links para diversos artigos com a finalidade de facilitar o acesso às informações, sendo o MedLine um dos subconjuntos). **FEATURES** contêm algumas propriedades com as respectivas posições. **ORIGIN** é a estrutura primária da proteína.

# Apêndice MODELLER

**Busca das estruturas e das seqüências:** nesta etapa, o arquivo de retorno apresenta uma tabela que indica as proteínas mais próximas com o alvo.

```
   # CODE_1     CODE_2   LEN1 LEN2 NID %ID   %ID    SCORE    SIGNI SIGNI2 SIGNI3
-------------------------------------------------------------------------------
   1 my         1azo     182  214  44  20.6  24.2   81472.    5.5 -999.0 -999.0
   2 my         1aj8A    182  371  57  15.4  31.3   21853.    4.3 -999.0 -999.0
   3 my         1ihoA    182  282  50  17.7  27.5   56064.    4.3 -999.0 -999.0
   4 my         1ldkA    182  358  64  17.9  35.2   27186.    4.3 -999.0 -999.0
   5 my         1h7dA    182   49  21  11.5  42.9  -26223.    4.2 -999.0 -999.0
   6 my         1earA    182  142  28  15.4  19.7   57440.    4.2 -999.0 -999.0
   7 my         1ituA    182  369  53  14.4  29.1   21962.    4.1 -999.0 -999.0
   8 my         1f9zA    182  128  30  16.5  23.4   45567.    4.1 -999.0 -999.0
   9 my         1ektA    182   53  25  13.7  47.2  -22316.    4.1 -999.0 -999.0
  10 my         1k8tA    182  498  60  12.0  33.0  -27765.    3.8 -999.0 -999.0
  11 my         1aa0     182  113  34  18.7  30.1   33403.    3.8 -999.0 -999.0
  12 my         1go3F    182  107  30  16.5  28.0   27331.    3.7 -999.0 -999.0
  13 my         1joyA    182   67  23  12.6  34.3   -9125.    3.6 -999.0 -999.0
  14 my         1bh9A    182   45  16   8.8  35.6  -30037.    3.6 -999.0 -999.0
  15 my         1vii     182   36  19  10.4  52.8  -38104.    3.5 -999.0 -999.0
  16 my         1h98A    182   77  26  14.3  33.8   -3171.    3.5 -999.0 -999.0
  17 my         1g7oA    182  215  33  15.3  18.1   79130.    3.4 -999.0 -999.0
  18 my         1ko7A    182  285  51  17.9  28.0   54044.    3.4 -999.0 -999.0
  19 my         1h3oA    182   50  16   8.8  32.0  -25482.    3.4 -999.0 -999.0
  20 my         1bh9B    182   89  24  13.2  27.0    9704.    3.4 -999.0 -999.0


Group added to ALIGNMENT:  1azo : 1azo 2azoA 2azoB
Group added to ALIGNMENT:  1aj8A : 1aj8A 1aj8B 1o7xA 1o7xB 1o7xC 1o7xD
Group added to ALIGNMENT:  1ihoA : 1ihoA 1ihoB
```

**CODE_1** é o código associado à seqüência alvo. A coluna **CODE_2** apresenta os códigos das proteínas selecionadas pela ferramenta MODELLER semelhantes a alvo. **LEN1** é o comprimento da seqüência alvo. **LEN2** corresponde ao comprimento das proteínas semelhantes. **NID** consiste no número de aminoácidos alinhados; **%ID's** representam o percentual de identidade, normalizados pelo comprimento da menor e da maior seqüência respectivamente. **SCORE** é a pontuação do alinhamento. **SGNI's** apresentam três formas distintas para se estabelecer a significância dos alinhamentos, sendo a primeira delas a mais relevante. Nela, para cada alinhamento, as duas seqüências são misturadas e alinhadas diversas vezes, calculando as pontuações desses alinhamentos. Por fim obtém-se a média e o desvio padrão das pontuações e associa com o **SCORE** do alinhamento original.

**Selecionando os moldes:** o *dendrograma* construído pela ferramenta MODELLER é formado a partir de uma matriz de distância entre as seqüências, baseado no método de *clustering* de pares pela média dos pesos, onde a distância entre dois agrupamentos é dada pela média das distâncias de cada grupo de um conjunto para

cada grupo do outro. Em outras palavras, a distância D(S, T) entre dois agrupamentos S={s1, s2, ..., sm} e T={t1, t2, ..., tn}, é dado pela média da distância de cada elemento S com cada elemento de T:

**D(S,T) = ( D(s1,t1) + D(s1,t2) + … + D(s1,tn) + D(s2,t1) + D(s2,t2) + … + D(s2,tn) + … + D(smt1) + D(smt2) + … + D(smtn) ) / m*n**

Por exemplo, na matriz de similaridade abaixo, a diagonal principal corresponde ao tamanho da seqüência, o triângulo superior à diagonal indica o número de resíduos equivalentes e a diagonal inferior ao percentual de resíduos equivalentes, normalizados pelo tamanho da menor. A partir dessa matriz de similaridade, deriva a matriz de distância, simétrica e com todos os elementos da diagonal principal nulos. Nessa matriz, cada campo corresponde ao percentual de elementos diferentes entre duas seqüências, normalizado pelo tamanho da menor seqüência.

**Matriz de Similaridade**

| | 1aj8A | 11aj8B | 11o7xA | 21ihoA | 11ihoB | 1my |
|---|---|---|---|---|---|---|
| 1aj8A | 371 | 370 | 129 | 2 | 6 | 18 |
| 1aj8B | 100 | 370 | 129 | 2 | 6 | 18 |
| 1o7xA | 35 | 35 | 367 | 4 | 7 | 13 |
| 1ihoA | 1 | 1 | 1 | 282 | 95 | 6 |
| 1ihoB | 2 | 2 | 2 | 34 | 282 | 9 |
| my | 10 | 10 | 7 | 3 | 5 | 182 |

**Matriz de distância**

| | 1aj8A | 11aj8B | 11o7xA | 21ihoA | 11ihoB | 1my |
|---|---|---|---|---|---|---|
| 1aj8A | 0.000 | 0.000 | 65.000 | 99.000 | 98.000 | 90.000 |
| 1aj8B | 0.000 | 0.000 | 65.000 | 99.000 | 98.000 | 90.000 |
| 1o7xA | 65.000 | 65.000 | 0.000 | 99.000 | 98.000 | 93.000 |
| 1ihoA | 99.000 | 99.000 | 99.000 | 0.000 | 66.000 | 97.000 |
| 1ihoB | 98.000 | 98.000 | 98.000 | 66.000 | 0.000 | 95.000 |
| my | 90.000 | 90.000 | 93.000 | 97.000 | 95.000 | 0.000 |

Em seguida, o dendrograma é criado a partir da matriz de distâncias. Inicialmente, as duas seqüências mais próximas são reunidas. No caso, as **1aj8A** e **1aj8B** são escolhidas, pois a distância entre elas é de 0.0% segundo a matriz, valor este indicado no dendrograma, ao lado do código **1aj8A**.

D(**1aj8A**, **1aj8B**) = D(**1aj8A**, **1aj8B**)/1 = 0.0%

Em seguida, o conjunto **1aj8A**-**1aj8B** é unido à seqüência **1o7xA**, sendo a distância entre o conjunto e a seqüência uma média da distância entre **1aj8A-1o7xA** e **1aj8B-1o7xa**.

D(**1aj8A**-**1aj8B, 1o7xa**) = (D(**1aj8A**, **1o7xA**) + D(**1aj8B**, **1o7xA**)) / 2
D(**1aj8A**-**1aj8B, 1o7xa**) = ((65.0% + 65.0%) / 2) = 65.0%

Ao mesmo tempo, forma-se um outro grupo com as seqüências **1ihoA** e **1ihoB**, cuja distância entre as seqüências é de 66.0%. Continuando o processo, o grupo **(1aj8A-1o7xa)-1o7xa** é unido à seqüência **my**, cuja distância entre o grupo e a seqüência é de 91.5%, dada pela média das distâncias dos grupos **(1aj8A-1o7xa)** e **1o7xa** com a seqüência **my**.

D((**1aj8A**-**1aj8B**)-**1o7xa, my**) = (D((**1aj8A**-**1aj8B**), **my**) + D(**1o7xA, my**)) / 2 =
((D(**1aj8A, my**) + D(**1aj8B, my**)) / 2 + D(**1o7xA, my**)) / 2 =
((90.0% + 90.0%)/2 + 93.0%) = 91.5%

Finalmente, o grupo **((1aj8A-1o7xa)- 1o7xa)-my** é agrupado com **1ihoA**-**1ihoB**, onde a distância é calcula pela média da distância de cada grupo do primeiro agrupamento, para cada seqüência do segundo agrupamento:

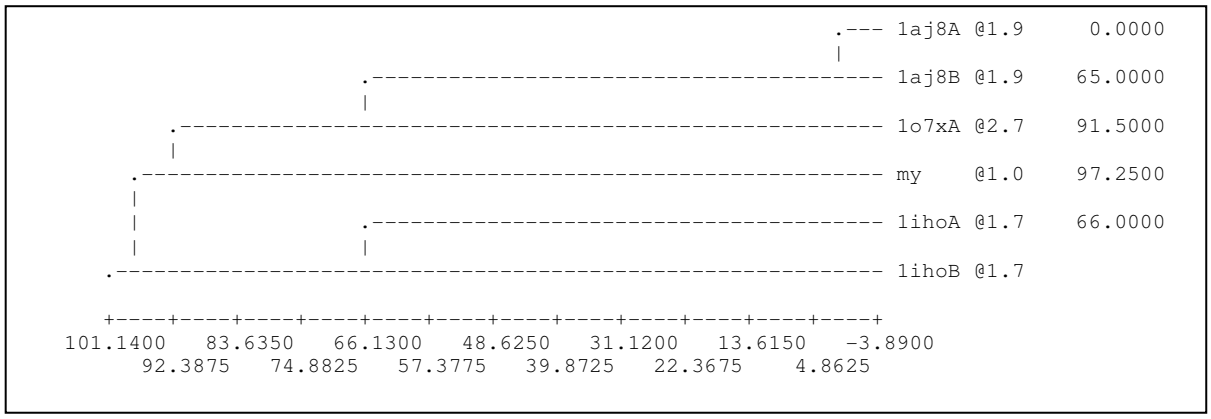
D( **((1aj8A-1o7xa)- 1o7xa)-my, 1ihoA-1ihoA** ) =
(D ( **((1aj8A-1o7xa)- 1o7xa)-my, 1ihoA** ) + D ( **((1aj8A-1o7xa)- 1o7xa)-my, 1ihoB** ))
/2

onde

D ( **((1aj8A-1o7xa)- 1o7xa)-my, 1ihoA** ) =
(D ( **(1aj8A-1o7xa)- 1o7xa, 1ihoA** ) + D ( **my, 1ihoA** ) ) /2 =
((D ( **1aj8A-1o7xa, 1ihoA** ) + D ( **1o7xa, 1ihoA** ) ) /2 + D ( **my, 1ihoA** ) ) /2 =
(((D ( **1aj8A, 1ihoA** ) + (D ( **1aj8A, 1ihoA** ))/2 + 99.0% ) /2 + 97% ) /2 =
(((99.0% + 99.0%)/2 + 99.0% ) /2 + 97% ) /2 = 98.0%

O mesmo acontece para **1ihoB**, caso em que a distância equivale a 96.5%. Dessa forma, tem-se que:

D( **((1aj8A-1o7xa)- 1o7xa)-my, 1ihoA-1ihoA** ) = (98.0% + 96.5%)/2 = 97.25%


1aj8A @1.9 0.0000
1aj8B @1.9 65.0000
1o7xA @2.7 91.5000
my @1.0 97.2500
1ihoA @1.7 66.0000
1ihoB @1.7
101.1400 83.6350 66.1300 48.6250 31.1200 13.6150 -3.8900
92.3875 74.8825 57.3775 39.8725 22.3675 4.8625

# Apêndice PDB

Uma simples descrição do formato utilizado no armazenamento das proteínas no *Protein Data Bank* é exemplificada com um trecho do arquivo da proteína de código 1FDN, a ferridoxina, originada da bactéria *Clostridium acidurici*.

```
HEADER    ELECTRON TRANSPORT                      31-MAR-94   1FDN
COMPND    FERREDOXIN
SOURCE    (CLOSTRIDIUM ACIDIURICI) ATCC: 7906
AUTHOR    E.DUEE,E.FANCHON,J.VICAT,L.C.SIEKER,J.MEYER,J-M.MOULIS
REVDAT   1   31-AUG-94 1FDN    0
JRNL        AUTH   E.DUEE,E.FANCHON,J.VICAT,L.C.SIEKER,J.MEYER,
JRNL        AUTH 2 J.M.MOULIS
JRNL        TITL   REFINED CRYSTAL STRUCTURE OF THE 2[4FE-4S]
JRNL        TITL 2 FERREDOXIN FROM CLOSTRIDIUM ACIDURICI AT 1.84
JRNL        TITL 3 ANGSTROMS RESOLUTION
JRNL        REF    TO BE PUBLISHED
JRNL        REFN                                                0353
REMARK   1
REMARK   1 REFERENCE 1
REMARK   1  AUTH   J.MEYER,J.-M.MOULIS,N.SCHERRER,J.GAGNON,J.ULRICH
REMARK   1  TITL   SEQUENCES OF CLOSTRIDIAL FERREDOXINS
REMARK   1  REF    BIOCHEM.J.                    V. 294   622 1993
REMARK   1  REFN   ASTM BIJOAK  UK ISSN 0306-3275                0043
REMARK   2
REMARK   2 RESOLUTION. 1.84 ANGSTROMS.
SEQRES   1     55  ALA TYR VAL ILE ASN GLU ALA CYS ILE SER CYS GLY ALA
SEQRES   2     55  CYS GLU PRO GLU CYS PRO VAL ASN ALA ILE SER SER GLY
SEQRES   3     55  ASP ASP ARG TYR VAL ILE ASP ALA ASP THR CYS ILE ASP
SEQRES   4     55  CYS GLY ALA CYS ALA GLY VAL CYS PRO VAL ASP ALA PRO
SEQRES   5     55  VAL GLN ALA
FTNOTE   1

                                      ...

ATOM     17  OH  TYR     2      12.652  20.633  70.674  1.00 14.24
ATOM     18  N   VAL     3       6.341  19.223  65.770  1.00  9.50
ATOM     19  CA  VAL     3       5.103  18.626  65.305  1.00  9.51
ATOM     20  C   VAL     3       4.839  17.445  66.248  1.00 10.00
ATOM     21  O   VAL     3       5.768  16.767  66.712  1.00  9.74
ATOM     22  CB  VAL     3       5.271  18.199  63.809  1.00  8.75
ATOM     23  CG1 VAL     3       6.379  17.179  63.634  1.00  9.61
ATOM     24  CG2 VAL     3       3.957  17.620  63.311  1.00  9.83
ATOM     25  N   ILE     4       3.587  17.263  66.657  1.00  9.05
ATOM     26  CA  ILE     4       3.204  16.165  67.531  1.00  9.65
ATOM     27  C   ILE     4       2.977  14.961  66.644  1.00 10.61
ATOM     28  O   ILE     4       2.171  15.043  65.716  1.00 11.05
ATOM     29  CB  ILE     4       1.941  16.587  68.304  1.00  8.61
ATOM     30  CG1 ILE     4       2.360  17.676  69.283  1.00  9.26
ATOM     31  CG2 ILE     4       1.323  15.427  69.064  1.00  7.79
ATOM     32  CD1 ILE     4       1.239  18.368  70.064  1.00 10.35
ATOM     33  N   ASN     5       3.717  13.878  66.858  1.00 10.52
ATOM     34  CA  ASN     5       3.551  12.729  65.991  1.00 11.91
ATOM     35  C   ASN     5       2.482  11.763  66.487  1.00 12.50
ATOM     36  O   ASN     5       1.794  12.018  67.477  1.00 11.77
ATOM     37  CB  ASN     5       4.928  12.033  65.826  1.00 13.38
ATOM     38  CG  ASN     5       5.545  11.386  67.055  1.00 13.93
ATOM     39  OD1 ASN     5       4.853  10.975  67.986  1.00 15.00
```

A primeira linha corresponde ao campo **HEADER**, informando a classificação da molécula, a data de recebimento das coordenadas dos átomos pelo PDB e o código identificando a proteína. **COMPND** dá uma maior descrição da molécula de entrada. **SOURCE** indica a origem biológica ou química da proteína. **AUTHOR** são os

responsáveis pelas informações contidas no arquivo. **REVDAT** informa o histórico das modificações do arquivo. **JRNL** registra a primeira citação da literatura que descreve o experimento que resultou nas coordenadas dos átomos. **REMARK** incluem anotações, comentários e outras referências não citadas. **SEQRES** informa a estrutura primária da proteína. **ATOM** indica cada um dos átomos que fazem parte da molécula, indicando o aminoácido ao qual pertence e sua posição na seqüência, as coordenadas cartesianas e parâmetros que indicam a qualidade do posicionamento do átomo. Existem outros campos que descrevem melhor a estrutura das proteínas e estão descritos no Guia do Formato de Descrição do PDB (http://www.rcsb.org/ pdb/docs/format/pdbguide2.2/guide2.2_frame.html).

Dissertação de Mestrado apresentada por **Erico Souza Teixeira** à Pós-Graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco, sob o título , **"Estudo de Processos para Modelagem**, **Avaliação, Alteração e Seleção da Estrutura de Proteínas",** orientada pela **Profa. Katia Silva Guimarães** e aprovada pela Banca Examinadora formada pelos professores:

Prof. Ruy José Guerra Barretto de Queiroz
Centro de Informática / UFPE

Prof. Ricardo Luiz Longo
Departamento de Química Fundamental / UFPE

Profa. Katia Silva Guimarães
Centro de Informática / UFPE

Visto e permitida a impressão.
Recife, 17 de junho de 2005.

**Prof. JAELSON FREIRE BRELAZ DE CASTRO**
Coordenador da Pós-Graduação em Ciência da Computação do
Centro de Informática da Universidade Federal de Pernambuco.