

Bioinformática Estrutural Aplicada ao Estudo de Proteínas Alvo do Genoma do *Mycobacterium tuberculosis*.

Nelson José Freitas da Silveira

Tese apresentada para obtenção do título de Doutor em Biofísica Molecular, área de concentração em Biofísica Molecular do Departamento de Física do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista "Julio de Mesquita Filho" – UNESP.

Orientador: Prof. Dr. Walter Filgueira de Azevedo Júnior

São José do Rio Preto
Estado de São Paulo – Brasil
Agosto – 2005

Silveira, Nelson José Freitas da.

Bioinformática estrutural aplicada ao estudo de proteínas alvo do genoma do *Mycobacterium tuberculosis* / Nelson José Freitas da Silveira – São José do Rio Preto : [s.n.], 2005
117 f. : il. ; 30 cm.

Orientador: Walter Filgueira de Azevedo Júnior
Tese (doutorado) – Universidade Estadual Paulista. Instituto de Biociências, Letras e Ciências Exatas

1. Biologia molecular. 2. Bioinformática estrutural. 3. Proteínas Estrutura. I. Azevedo Júnior, Walter Filgueira de. II. Universidade Estadual Paulista. Instituto de Biociências, Letras e Ciências Exatas. III. Título.

CDU – 577.112

“Jamais considere seus estudos como uma obrigação, mas como uma oportunidade invejável para aprender a conhecer a influência libertadora da beleza do reino do espírito, para seu próprio prazer pessoal e para proveito da comunidade à qual seu futuro trabalho pertencer”.

Albert Einstein

**À toda minha família, em especial
minha mulher Eluy e minha filha
Isabela pelo amor e companheirismo.**

Agradecimentos

Ao Prof. Walter Filgueira de Azevedo Jr. pela orientação dedicada na realização deste trabalho, pela amizade e companheirismo na convivência diária e pela confiança em mim depositada.

À Profa. Eloiza Helena Tajara da Silva, pela sua dedicação a um treinamento pessoal e científico em bioinformática, proporcionado com seu contato com o Prof. Walter a orientação que resultou em um trabalho de doutoramento, parceria e amizade.

Aos Profs. Walter Filgueira de Azevedo Jr., José Márcio Machado, Valmir Fadel, Paula Rahal Liberatore e Fernanda Canduri pelas sugestões para complementação deste trabalho durante o exame geral de qualificação.

Aos membros da banca de defesa da tese Profs. Walter Filgueira de Azevedo Jr., Paulo Sérgio Lopes de Oliveira, Paula Regina Kuser Falcão, Jorge Chahine, José Roberto Ruggiero, Andréia Machado Leopoldino, Osmar Norberto de Souza e Fernanda Canduri por aceitarem a colaborar para o enriquecimento do trabalho e meu crescimento pessoal, fortalecendo o caráter científico desenvolvido.

A todo o corpo docente do Departamento de Física, pelos ensinamentos e dedicação transmitidos nas disciplinas cursadas, pela amizade, pelo tempo cedido ao esclarecimento de dúvidas e solução de problemas e pela receptividade acolhedora e harmoniosa.

Aos colegas da pós-graduação pela convivência, pelas conversas no café e churrascos e principalmente pela integração científica e disseminação de conhecimentos.

Aos colegas do grupo do Laboratório de Sistemas Biomoleculares, os quais foram de importância relevante para a conclusão deste trabalho, fazendo de cada seminário um ambiente de discussão em grupo favorecendo o aprendizado e pelas conversas, piadas que sempre descontraíram nos momentos difíceis.

Aos funcionários Barbosa, Paulinho e Ilva pelos favores e prestação de serviços de suporte ao bom funcionamento do departamento.

Aos meus pais Nelson (*in memoriam*) e Sebastiana, meus irmãos Luiz e Patrícia, pelo apoio e incentivo que sempre recebi, tendo sempre como conselho, a calma necessária e a obstinação por aquilo que deseja.

À minha mulher Eluy e minha filha Isabella pelo apoio, incentivo, alegria e suporte, gerando estrutura e ambiente familiar favorecendo o sucesso de meu trabalho.

Ao assessor *ad hoc* da FAPESP, o qual desde a concessão da bolsa tem emitido pareceres sobre os relatórios sempre com informações e sugestões de grande importância para o enriquecimento do trabalho.

À FAPESP (processo nº 02/10239-6) pela bolsa concedida e pelo apoio financeiro, possibilitando a aquisição de material necessário ao desenvolvimento e conclusão do trabalho, além das viagens a congressos.

À Deus por me conceder serenidade, saúde e a possibilidade de realizar tudo o que desejo, sempre em direção à paz.

Índice Geral

Índice de Figuras.....	3
Índice de Tabelas.....	5
Glossário de Termos e Abreviaturas.....	6
Resumo.....	9
Abstract.....	11
1. Introdução.....	13
1.1 Genoma do <i>Mycobacterium tuberculosis</i>	13
1.2 <i>Mycobacterium tuberculosis</i> multidroga resistente.....	18
1.3 Vias metabólicas.....	21
1.4 Alvos para desenho de drogas baseado em estrutura.....	23
1.5 Bioinformática estrutural aplicada ao estudo de proteínas alvo.....	26
2. Objetivos.....	29
3. Materiais e Métodos.....	30
3.1 Modelagem molecular.....	30
3.1.1 Procura e seleção de <i>templates</i>	31
3.1.2 Alinhamento <i>template/alvo</i>	32
3.1.3 Construção do modelo.....	33
3.1.4 Avaliação dos modelos.....	34
3.2 Aplicações da modelagem molecular comparativa.....	35
3.3 Possíveis erros em modelagem molecular comparativa.....	38
3.4 Modelagem em larga escala do genoma do <i>M. tuberculosis</i>	40
3.5 Busca por <i>templates</i> e algoritmo de alinhamento.....	44
3.6 Modelagem molecular comparativa usando um <i>cluster Beowulf</i>	46
3.7 Softwares de análise estrutural e validação de modelos.....	48
3.8 Perl/CGI.....	50
3.9 Banco de dados MySQL.....	53
3.10 Programas, servidores e <i>links</i> no DBMODELING.....	56
4. Resultados e Discussão.....	59
4.1 Conteúdo de dados no DBMODELING.....	59
4.2 Dados para referência sobre estruturas de <i>M. tuberculosis</i>	61
4.3 Acesso e interface do banco de dados.....	64
4.4 Precisão dos modelos gerados.....	71
4.5 Análises realizadas para uma estrutura contida no DBMODELING.....	76
4.5.1 Alinhamento das seqüências primárias e qualidade dos modelos.....	77
5. Conclusões.....	85
6. Desenvolvimentos Futuros.....	87
7. Bibliografia.....	88
APÊNDICE A – Descrição dos softwares utilizados.....	98
I. MODELAGEM MOLECULAR.....	98
II. ALINHAMENTO (NEEDLEMAN-WUNSCH).....	103
III. PROCHECK.....	108
IV. VERIFY 3D.....	113
V. RMSD.....	114

APÊNDICE B – Produção bibliográfica..... 117

Índice de Figuras

Figura 1. Distribuição geográfica da localização dos níveis de mortalidade causados pelo <i>M. tuberculosis</i> no mundo.....	14
Figura 2. Via do ácido chiquímico na seqüência de sete passos metabólicos.....	22
Figura 3. Interação das áreas que têm contribuído para a formação e o desenvolvimento da bioinformática.....	27
Figura 4. Precisão e aplicação de modelos estruturais de proteínas.....	37
Figura 5. Possíveis erros em modelagem molecular comparativa.....	40
Figura 6. Fluxograma do algoritmo criado para automatizar a modelagem comparativa de estruturas de proteínas.....	43
Figura 7. Arquivo de entrada para gerar o alinhamento pelo MODELLER.....	45
Figura 8. Arquivo de entrada da modelagem, indicando o número de modelos a serem gerados e a semente aleatória.....	45
Figura 9. Arquitetura do <i>cluster</i>	46
Figura 10. <i>Cluster</i> utilizado para executar a ferramenta desenvolvida para modelagem molecular e análise em larga escala.....	46
Figura 11. Diagrama esquemático de como a programação CGI interage com o banco de dados de <i>M. tuberculosis</i>	53
Figura 12. Relação entidade-relacionamento para as tabelas do banco de dados DBMODELING.....	56
Figura 13. Dados estatísticos sobre a modelagem.....	60
Figura 14. Gráfico representando a estimativa de dados que serão acrescentados ao DBMODELING.....	61
Figura 15. Interface de entrada para as ferramentas do grupo do Laboratório de Sistemas Biomoleculares (BMSys).....	65
Figura 16. Interface de entrada para o DBMODELING.....	66
Figura 17. Visualização da interface após a seleção do organismo.....	67
Figura 18. Interface de busca por uma via metabólica ou enzima específica.....	68
Figura 19. Links direcionando as enzimas de interesse para as informações estruturais.....	69
Figura 20. Dados estruturais da enzima selecionada para pesquisa.....	70
Figura 21. Análise dos resultados de uma proteína alvo para os 1000 modelos gerados.....	71
Figura 22. Histograma representando as regiões mais favoráveis do gráfico de Ramachandran.....	73
Figura 23. Histograma mostrando a freqüência de proteínas com relação aos intervalos dos valores de RMSD de sobreposição $C_{\alpha} - C_{\alpha}$	74
Figura 24. Gráfico de dispersão dos dados do Procheck e RMSD da geometria ideal.....	75
Figura 25. Reação catalisada pela Glucose-1-fosfato timidilil-transferase (RmlA).....	77
Figura 26. Alinhamento das seqüências de aminoácidos da <i>MtRmlA</i> e da <i>PaRmlA</i>	78
Figura 27. Gráfico de Ramachandran da modelagem da enzima <i>MtRmlA</i>	79
Figura 28. Estrutura 3D da enzima <i>MtRmlA</i> e <i>PaRmlA</i> , respectivamente.....	80
Figura 29. Gráfico gerado pelo programa VERIFY 3D da enzima <i>MtRmlA</i>	82
Figura 30. Sobreposição do modelo da <i>MtRmlA</i> com o <i>template</i> 1FXO_A mostrando as diferenças conformacionais entre as estruturas.....	83
Figura 31. Passos para construção de um modelo utilizando modelagem comparativa por satisfação de restrições espaciais.....	98
Figura 32. Mudança conformacional pela minimização da função objetivo.....	100
Figura 33. Alinhamento de seqüências por algoritmo de programação dinâmica.....	104
Figura 34. Método para calcular o escore ótimo no algoritmo de programação dinâmica.....	107

Figura 35. Matriz de valores de alinhamento BLOSUM62.....	108
Figura 36. Representação dos ângulos de torção em uma cadeia polipeptídica.....	110
Figura 37. Diagrama de Ramachandran.....	112
Figura 38. Representação estrutural da molécula de Glicina (a) e de Prolina (b).....	113

Índice de Tabelas

Tabela 1. Alvos moleculares para o diagnóstico de resistência do <i>M. tuberculosis</i>	19
Tabela 2. Programas e servidores <i>web</i> usados nos alinhamentos, construção e avaliação dos modelos.....	57
Tabela 3. Qualidade dos modelos estruturais usando as análises do gráfico de Ramachandran.....	58
Tabela 4. Cálculo do RMSD de sobreposição C _α -C _α	63
Tabela 5. Dados estatísticos para a região mais favorável do gráfico de Ramachandran.....	73
Tabela 6. Análises da estrutura do <i>template</i> e do modelo.....	81
Tabela 7. Dados gerais apresentados no banco de dados sobre o modelo e o <i>template</i> . análises.....	81

Glossário de Termos e Abreviaturas

3D – Tridimensional.

Ângulos diedros – Um ângulo formado por quatro pontos *i, j, k, l* (por exemplo, átomos). Ele é definido como o ângulo entre os planos normais *ijk* e *jkl*.

BMSys – Laboratório de Sistemas Biomoleculares (<http://www.biocristalografia.df.ibilce.unesp.br>).

CCP4 – *Collaborative Computational Project* N° 4.

Constrições e Restrições – Constrição restringe uma característica espacial, tal como a distância entre dois átomos, para um simples valor em particular. A restrição permite um intervalo maior de valores, possível com a variação da probabilidade.

CGI – *Common Gateway Interface*.

Docking – Um método utilizado para detectar sítios de ligação em proteína e avaliar interações proteína-proteína ou proteína-ligante, utilizando para cálculo a energia livre de ligação entre as moléculas. É dividido em *docking* rígido e flexível.

DBMS – *Database Management System*.

DBMODELING – Banco de dados relacional que utiliza a plataforma SQL/MySQL e programação Perl/CGI com o objetivo de disponibilizar modelos moleculares de proteínas alvo de genomas como *Mycobacterium tuberculosis*, *Xylella fastidiosa*, etc.

dTDP – desoxi-timidina di-fosfato.

dTMP – desoxi-timidina mono-fosfato.

dTTP – desoxi-timidina tri-fosfato.

Glc – Glucose.

G-1-P – Glucose-1-fosfato.

HTML – *Hypertext Markup Language*.

KEGG – (*Kyoto Encyclopedia of Genes and Genomes*) Banco de dados de vias metabólicas.

Minimização de energia – Técnica que muda a conformação de uma molécula, no sentido de diminuir sua energia tanto quanto possível.

Método da função alvo variável – Uma técnica de otimização determinística que envolve a otimização de uma função objetivo, iniciando com um pequeno subconjunto de restrições para otimizar, e finaliza com a função objetivo total, incluindo todas as restrições.

MPI – *Message Passing Interface*. Protocolo utilizado na paralelização de softwares.

MetaCyc – Banco de dados de vias metabólicas e reações químicas.

MtRmlA - Glucose-1-fosfato timidilil-transferase de *Mycobacterium tuberculosis*

ORF – (*Open Reading Frame*). Forma de leitura para estimar as possíveis seqüências codificantes de cada gene extraído do genoma.

Pdf (Função Densidade de Probabilidade) – Uma função que especifica a probabilidade para cada valor possível da característica restrita (por exemplo, a distância entre dois átomos), dado alguma característica conhecida (por exemplo, uma distância equivalente em uma estrutura relacionada). Esta é a formulação matemática mais geral de uma restrição espacial.

PaRmlA – Glucose-1-fosfato timidilil-transferase de *Pseudomonas aeruginosa*.

PDB – (Protein Data Bank) É uma coleção de estruturas 3D de proteínas determinadas principalmente por cristalografia de raios X ou ressonância magnética nuclear. É acessível em <http://www.rcsb.gov/pdb>.

Potencial de Lennard-Jones – Um termo de energia que é freqüentemente usado para descrever uma interação entre um par de partículas: $E = A/d^{12} - B/d^6$, onde A e B são constantes positivas e d é a distância entre as partículas. Após cada modelo gerado pelo MODELLER este potencial é usado em conjunto com o CHARMM para minimização da energia do modelo.

Programação dinâmica – Método que busca solução ótima dividindo o problema original em problemas menores, por isso ideal na utilização em computadores de arquiteturas paralelas.

Perl – (*Practical Extraction and Reporting Language*). Linguagem de programação.

RmlA – Glucose-1-fosfato timidilil-transferase

RMSD – Mede a diferença estrutural entre duas estruturas sobrepostas. É definido como:

$$\sqrt{\frac{\sum_i d_i^2}{N}}$$
, onde a soma é executada sobre os N pares de átomos equivalentes, um de cada estrutura, e d_i é a distância entre os dois átomos no i -ésimo par.

Restrições estereoquímicas – Estas restrições espaciais estão implicadas pela topologia covalente da molécula. Elas incluem restrições sobre os comprimentos de ligação, ângulos diedros, ângulos de ligação, planaridade de anéis e quiralidade de centros quirais.

RDBMS – *Relational Database Management System*.

RMN – Ressonância Magnética Nuclear.

Seqüência Alvo – Seqüência primária de proteína utilizada na modelagem para determinação de sua estrutura.

Screening virtual – Método que contribui para o processo de descobrimento de novas drogas. Utiliza biblioteca de ligantes com o objetivo de selecionar novos compostos candidatos a drogas contra uma determinada proteína utilizando simulações de *docking*.

SQL – *System Query Language*.

Swiss-Prot – Banco de dados de seqüências primárias.

Threading – Método sensível para se detectar relação remota seqüência/estrutura e para alinhar uma seqüência com uma estrutura.

Template – Estrutura resolvida experimentalmente presente no PDB utilizada como molde na modelagem comparativa.

Resumo

O seqüenciamento de genomas em larga escala estão nos munindo com várias informações biológicas sobre centenas de organismos. O entendimento das diferentes funções de proteínas expressas por genes obtidos nos projetos de seqüenciamento, nos leva à era pós-genômica, com a caracterização das estruturas 3D de proteínas. A determinação de estruturas de proteínas nem sempre é possível devido a limitações nas técnicas de cristalografia de raios X e RMN, tornando a utilização da modelagem molecular comparativa muito útil. O principal interesse no estudo de vias metabólicas identificadas em genomas de patógenos, é o fato de que algumas destas vias não estão presentes em humanos, o que as tornam alvos seletivos para desenho de drogas, diminuindo o impacto das drogas em humanos. O DBMODELING é um banco de dados relacional, criado para evidenciar a importância dos métodos de modelagem molecular aplicadas ao genoma do *Mycobacterium tuberculosis*. A motivação deste trabalho é o fato de que o *M. tuberculosis* é a causa de morte de milhões de pessoas no mundo, assim a caracterização estrutural de proteínas alvo para propor novas drogas tornou-se essencial. Há atualmente no banco de dados mais de 260 modelos de proteínas do genoma do *M. tuberculosis* e outros genomas de interesse também serão acrescentados. Este banco de dados contém uma descrição detalhada da reação catalisada, do gene e da qualidade estrutural de cada proteína, e suas coordenadas atômicas estão disponíveis para *download*, podendo ser acessadas em <http://www.biocristalografia.df.ibilce.unesp.br/tools>. Este trabalho aumenta a

certeza de que a modelagem comparativa é uma ferramenta útil em bioinformática estrutural, uma vez que não se tem acesso às estruturas determinadas experimentalmente, podendo ser valiosa na anotação de seqüências genômicas, contribuindo para a genômica estrutural e funcional, e simulações de *docking* proteína-ligante.

Abstract

The large-scale genome sequencing are providing us with several information about hundreds of organisms. Understanding of different protein functions expressed by genes obtained in the sequencing projects, lead us to the post-genomic era, with characterization of protein 3D structure. The determination of protein structure, is not always possible, due to limitations in X-ray crystallography and NMR techniques. This fact makes the utilization of comparative modeling very useful. The main interest in the study of metabolic pathways is the fact that some of these pathways are not present in human, which make them selective targets for drug design, decreasing the impact of drugs in humans. DBMODELING is a relational database, created to highlight the importance of molecular modeling methods applied in the *Mycobacterium tuberculosis* genome. The motivation of this work is the fact that *M. tuberculosis* is the cause of the deaths of millions of people in the world, thus the protein target structural characterization to propose new drugs became essential. There are currently in the database more than 260 protein models of the *M. tuberculosis* genome, and other genomes of interest will be added. This database contains a detailed description of reaction catalized, of gene and structural quality of each protein, and their atomic coordinates are available to download, can be accessed at <http://www.biocristalografia.df.ibilce.unesp.br/tools>. This work increase the conviction that comparative modeling is an useful tool in structural bioinformatics, once that we have not access to the experimentally structures

determined, can be valuable in annotating genome sequence, contributing to the structural and functional genomics, and protein-ligand docking simulations.

1. Introdução

1.1 Genoma do *Mycobacterium tuberculosis*

Projetos de genoma estrutural têm como um de seus objetivos finais fornecer estruturas tridimensionais, determinadas experimentalmente ou por modelagem molecular comparativa de proteínas, para todas as proteínas passíveis de estudo, que estão codificadas no genoma. Espera-se que as estruturas das proteínas, determinadas experimentalmente e os modelos computacionais, produzam avanços no entendimento da função molecular e no mecanismo de milhares de proteínas (BRENNER, 2001; TAYLOR, 2002; LIU *et al.*, 2002).

A tuberculose é uma doença infecciosa crônica que vem afligindo a humanidade há mais de 5 milênios. Seu agente etiológico, o *Mycobacterium tuberculosis*, ou bacilo de Koch, é o patógeno que, provavelmente, mais mortes causou até o momento (ISEMAN, 1994). Nos países mais desenvolvidos, o impacto da doença sobre a população foi reduzido pelas melhorias radicais nas condições de vida que ocorreram em meados do século XIX, e tornou-se ainda melhor pela implementação da quimioterapia efetiva nos últimos 50 anos. Nos países ainda em desenvolvimento, ao contrário, a tuberculose manteve-se como um sério problema de saúde pública (KOCHI, 1991).

Apesar da disponibilidade da efetiva quimioterapia de caminho curto (DOTS) e a vacina Bacilo Calmette-Guérin (BCG), o bacilo da tuberculose continua a reclamar mais vidas que outros agentes infecciosos (SNIDER *et al.*, 1994).

Recentemente, tem aumentado a incidência da tuberculose nos países em desenvolvimento (Figura 1), aumentando a emergência generalizada de resistência às drogas e criando e uma sinergia mortal com o vírus da imunodeficiência humana (HIV), provocando a morte de milhares de pessoas no mundo. Em 1993, a gravidade da situação levou a Organização Mundial de Saúde (OMS) a declarar a tuberculose uma emergência global em uma tentativa de intensificar a consciência pública e política.

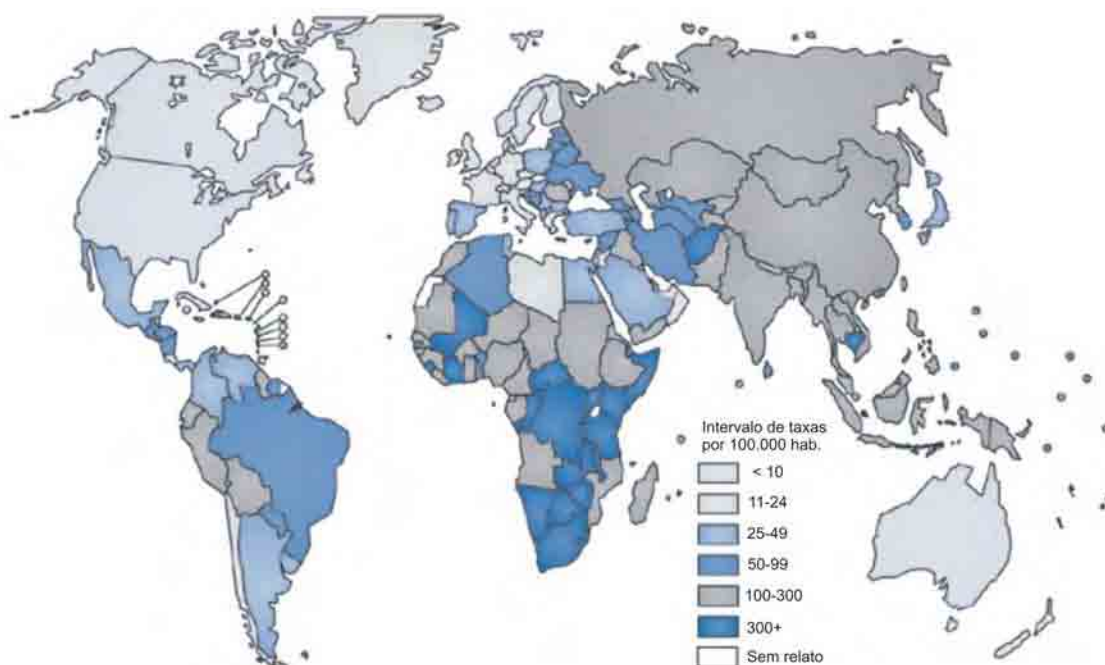


Figura 1. Distribuição geográfica das taxas de incidência de tuberculose estimadas pela OMS. No leste da Europa e África, estão aumentando as mortes após quase 40 anos de declínio. Outro fato importante é a presença da bactéria em níveis alarmantes nos países em desenvolvimento (OMS).

O perfil característico do bacilo inclui um crescimento lento, dormência, complexo envelope celular, patogênese intracelular e homogeneidade genética (WHEELER & RATLEDGE, 1994).

O envelope celular do *M. tuberculosis*, uma bactéria Gram-positiva, contém uma camada adicional mais afastada do peptidoglicano que é excepcionalmente rico em lipídios incomuns, glicolipídios e polissacarídeos (BRENNAN & DRAPER, 1994). Vias biosintéticas originais geradas a partir de componentes da parede celular tais como ácidos micólicos, fenoltiocerol, arabinogalactano, e vários destes devem contribuir para a longevidade micobacteriana, causando reações inflamatórias ao hospedeiro.

Com o seqüenciamento genético e a identificação e anotação das *ORFs* (*Open Reading Frames*) do *M. tuberculosis* (COLE *et al.*, 1998) abre-se uma nova linha de frente com o advento da bioinformática, na identificação das estruturas 3D das proteínas codificadas pelos genes seqüenciados. Foram codificadas 3.924 *ORFs* em todo o genoma do *M. tuberculosis*, dentre as quais, podemos identificar alvos moleculares para desenho de drogas baseado em estrutura, utilizando como seleção de alvos, vias metabólicas presentes na bactéria, porém ausentes em humanos. Esta seleção possibilitará (por métodos de modelagem molecular comparativa de proteínas), realizar simulações de *docking* contra possíveis inibidores de enzimas pertencentes às vias específicas selecionadas. A combinação de genômica e bioinformática tem o potencial para gerar a informação e conhecimento que possibilitará a concepção e o desenvolvimento de novas terapias e intervenções necessárias para lidar com esta doença aerotransportada e para elucidar a biologia incomum deste agente etiológico, *M. tuberculosis* (COLE *et al.*, 1998). A combinação das inovações nos campos da biologia estrutural e bioinformática,

fornece uma sinergia para o descobrimento de novos alvos para desenho de drogas. Com isso, foi formado o *TB Structural Genomics Consortium* (<http://www.doe-mbi.ucla.edu/TB>).

O principal objetivo do consórcio é determinar as estruturas de mais de 400 alvos de drogas do genoma do *M. tuberculosis* e analisar suas estruturas no contexto da informação funcional. Os alvos potenciais para drogas foram selecionados utilizando uma variedade de métodos de bioinformática. Os métodos para determinação dos alvos incluem perfil filogenético de proteínas e o uso de vias bioquímicas para selecionar genes relacionados de procariotos essenciais (GOULDING *et al.*, 2003). Foi realizada a re-anotação completa do genoma do *M. tuberculosis* da linhagem H37Rv quatro anos após a primeira submissão (CAMUS *et al.*, 2002). As informações sobre a nova anotação do genoma foram incorporadas no banco de dados público TubercuList (<http://genolist.pasteur.fr/TubercuList>). Na anotação da seqüência original do *M. tuberculosis*, da linhagem H37Rv, foram identificados 3.924 genes (COLE *et al.*, 1998) e na re-anotação foram incluídos 82 genes.

A nova anotação genômica do *M. tuberculosis* incorporou muitas mudanças à classificação funcional de proteínas preditas. Atualmente, está predita a função para 2058 proteínas (52% do proteoma) e mais de 150 destas proteínas foram experimentalmente provadas em pesquisa micobacterial. O número de proteínas hipotéticas conservadas foi mudado de 910 em 1998 para 1051 atualmente (um total de 376 possíveis proteínas não mostrou similaridade com proteínas conhecidas de

outros organismos e algumas delas devem ser específicas de *M. tuberculosis*). Atualmente, mais de 400 proteínas de *M. tuberculosis* foram detectadas experimentalmente, a maioria por estudos de proteômica (WELDINGH *et al.*, 1998; JUNGBLUT *et al.*, 1999; MOLLENKOPF *et al.*, 1999; ROSENKRANDS *et al.*, 2000; BETTS *et al.*, 2000).

Segundo o relatório anual da *World Health Organization* de 2001, estima-se que ocorreram cerca de 8,4 milhões de novos casos de tuberculose no mundo em 1999, o que representa um aumento de cerca de 20% em relação ao ano de 1997. Este aumento é devido à ocorrência da tuberculose em pacientes co-infectados com o vírus da AIDS. Outro fator que está relacionado com o aumento de casos de tuberculose é a emergência de cepas resistentes aos antimicrobianos utilizados para o seu tratamento. O abandono do tratamento ou a prescrição de regimes inapropriados para o tratamento da tuberculose resulta na seleção de cepas resistentes aos fármacos de primeira linha utilizados no seu tratamento.

O principal agente anti-tuberculose entre outros é a Isoniazida, ou hidrazida do ácido isonicotínico (INH) e é, provavelmente, o mais antigo fármaco sintético efetivo contra o *M. tuberculosis* (WHO, 1998). Foi descrita pela primeira vez em 1912 (MEYER & MALLY, 1912), mas só foi reconhecida como potente agente contra o *M. tuberculosis* em 1951 (FOX, 1951). Sua concentração inibitória mínima muito baixa (0,02 – 0,05 mg/ml) indubitavelmente contribui para sua eficácia. Um outro fator responsável pela sua potência pode ser o fato de que a droga age em diversos alvos na célula micobacteriana. A inibição da síntese de ácidos micólicos

(WINDER, 1982), enfraquecendo a parede bacteriana, foi uma das primeiras ações descritas da INH sobre o bacilo causador da tuberculose. Esses ácidos gordurosos e insaturados de cadeia longa contribuem para a impermeabilidade do envelope celular e, por serem restritos as micobactérias, configuram um alvo seletivo para os fármacos (CAMPOS, 1999).

1.2 *Mycobacterium tuberculosis* multidroga resistente

Pouco depois da introdução da INH no arsenal terapêutico contra a tuberculose, observou-se que algumas cepas isoladas altamente resistentes a ela não continham a enzima catalase-peroxidase, e que eram frequentemente não virulentos para a cobaia (MIDDLEBROOK *et al.*, 1954). Sabe-se hoje que a toxicidade da INH ao bacilo resulta de uma reação peroxidativa catalisada pela enzima catalase-peroxidase, a qual é codificada pelo gene *KatG* (HEYM *et al.*, 1995; YOUNG, 1994). A ausência desse gene em isolados de *M. tuberculosis* altamente resistentes a INH pode ser uma evidência de uma ligação entre a enzima catalase-peroxidase e a resistência a INH (ZHANG *et al.*, 1992). Uma outra forma de desenvolvimento da resistência a INH pode se dar por mutações que levem à expressão reduzida do gene ou à redução da atividade peroxidativa. A tabela 1 traz alguns alvos moleculares para diagnóstico de resistência do *M. tuberculosis*.

O *M. tuberculosis* resistente é um sério problema por dois motivos principais: 1) como há poucos fármacos efetivos disponíveis, uma infecção pelo bacilo resistente pode levar a uma doença potencialmente intratável; 2) embora a menor

parte dos infectados venha a adoecer (5-10%), a doença é altamente contagiosa. Portanto, se houver um número elevado de doentes tuberculosos portadores de germes resistentes a duas ou mais drogas potentes do arsenal terapêutico contra a doença, a probabilidade desse número aumentar exponencialmente é grande, e estaremos de frente a um sério problema com poucas possibilidades de solução.

Tabela 1. Alvos moleculares para o diagnóstico de resistência do *M. tuberculosis*.

Fármaco	Gene	Produto do gene	Frequência de mutações associadas à resistência
Isoniazida	<i>KatG</i>	Catalase-peroxidase	47-58%
	<i>InhA</i>	Biosíntese de ácidos graxos	21-28%
	<i>mabA</i>	Enzimas EnvM e FabG	21-28%
	<i>ahpC</i>	Alquil-hidroperóxido redutase C	10%
Rifampicina	<i>rpoB</i>	Subunidade da RNA polimerase	96-98%
Estreptomicina	<i>RpsL</i>	Proteína ribossômica S12	52-59%
	<i>Rrs</i>	RRNA 16S	8-21%
Etambutol	<i>EmbA</i>	-	-
	<i>EmbB</i>	-	-

Bactérias possuem diferentes mecanismos de defesa, provocando resistência a alguns antibióticos. De um modo geral, esses mecanismos de defesa podem ser divididos em três grupos: 1) mecanismos de “barreira” (a parede celular tem a capacidade de variar sua permeabilidade a diferentes compostos) (NIKAIDO, 1994); 2) a degradação ou inativação de enzimas (produzem enzimas que degradam ou modificam fármacos) (KWON *et al.*, 1995); 3) a modificação do “alvo” do fármaco (mutações pontuais em genes específicos modificam a especificidade da droga pela

enzima codificada). A resistência aos fármacos usados no tratamento da tuberculose depende desse terceiro mecanismo de resistência. A tuberculose multidroga resistente reflete a acumulação de etapas de mutações individuais de diversos genes independentes (HEYM *et al.*, 1994), e não a aquisição em bloco de resistência a múltiplas drogas.

Os mecanismos de resistência identificados até o momento são resultantes de mutações pontuais em genes codificadores das proteínas que são os alvos destes agentes anti-tuberculose (BASSO & BLANCHARD, 1998; BASSO *et al.*, 1998). Cepas de *M. tuberculosis* resistentes às drogas anti-tuberculose de primeira linha têm sido identificadas globalmente. A estimativa anual da tuberculose no Brasil é de 120.000 novos casos. Um aspecto preocupante da situação brasileira é que taxas superiores a 45% de pacientes previamente tratados apresentam multi-resistência (definida como resistente a isoniazida e rifampicina) adquirida, tornando imperiosa a busca de novos alvos para o desenvolvimento de novas drogas.

Dentre as prioridades para o combate à tuberculose, o desenvolvimento de novas drogas para substituírem àquelas comprometidas pela resistência é premente para o desenvolvimento de um tratamento quimioterápico eficaz. O principal objetivo da quimioterapia é atacar alvos peculiares aos microrganismos como, por exemplo, vias metabólicas ausentes no organismo humano. Tal fato, teoricamente, minimizaria o efeito tóxico destas drogas antimicrobianas para a espécie humana. Sob este aspecto, as enzimas da via do ácido chiquímico representam bons exemplos da utilidade de tal abordagem aplicada aos constituintes enzimáticos de uma rota

biossintética presente em microrganismos e plantas, e inexistente no organismo humano.

Um dos principais fatores que pode levar à resistência do *M. tuberculosis* é o tratamento irregular. Durante a quimioterapia, ciclos de destruição bacteriana (durante a administração das drogas) se alternam com ciclos de crescimento bacilar (quando a droga é suspensa). Em cada um desses ciclos ocorre seleção, favorecendo os mutantes resistentes em detrimento dos sensíveis. O recrudescimento da população bacteriana ao tamanho da população inicial, pré-início da quimioterapia, pode ocorrer com a presença de proporções crescentes de bacilos resistentes ao início de cada ciclo. Diferentes mecanismos, incluindo o efeito bactericida precoce das drogas usadas, a “monoterapia” durante a esterilização de populações bacterianas especiais (bacilos semi-dormentes) e inatividade metabólica da micobactéria pós-exposição ao fármaco favoreceriam a seleção de mutantes resistentes.

Diante deste cenário, incluímos política de saúde pública inadequada, o que provocou ao aumento do número de casos de tuberculose em escala mundial. Para combater a tuberculose faz-se necessário o desenvolvimento de novas drogas, preferencialmente usando-se alvos moleculares ausentes em humanos.

1.3 Vias metabólicas

Nas células as reações enzimáticas não ocorrem isoladamente, mas são organizadas em seqüências de múltiplas etapas denominadas rotas ou vias, nas quais

o produto de uma reação serve como substrato da reação subsequente. Por sua vez, diferentes vias se inter-relacionam, formando uma rede integrada e objetiva de reações químicas, coletivamente denominada metabolismo. É conveniente investigar o metabolismo examinando suas vias componentes. Cada via é composta de seqüências multienzimáticas e, cada enzima, por sua vez, pode exibir importantes características catalíticas ou regulatórias. A figura 2 mostra a via metabólica do ácido chiquímico, utilizada como alvo potencial para desenho de drogas baseado em estrutura, devido esta via estar presente em *M. tuberculosis*, porém ausente em humanos.

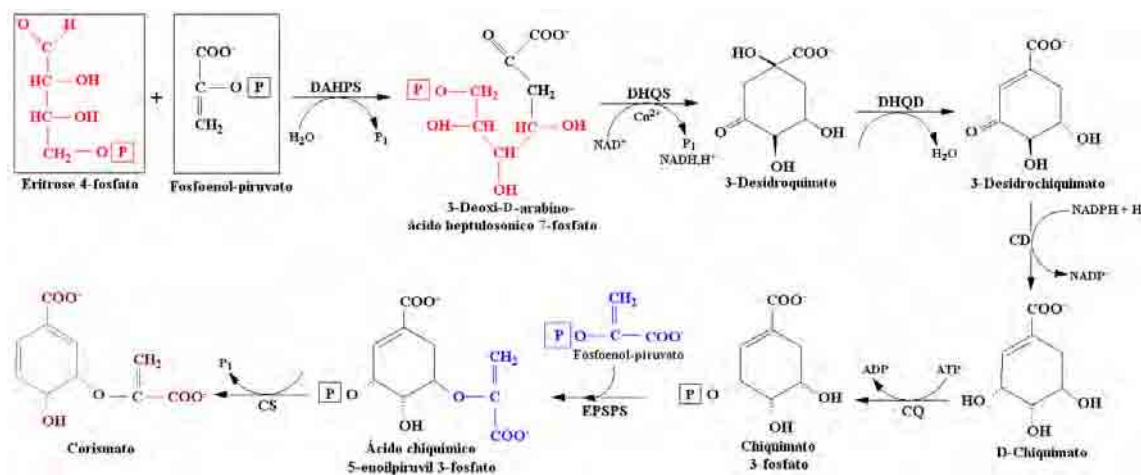


Figura 2. Via do ácido chiquímico na seqüência de sete passos metabólicos, iniciando no fosfoenolpiruvato e eritrose-4-fosfato até a conversão para corismato. A via é composta de 7 enzimas, as quais são: 3-deoxi-D-arabino-ácido heptulosônico 7-fosfato Sintase (DAHPS), 3-Desidroquinato Sintase (DHQS), 3-Desidroquinato Desidratase (DHQD), Chiquimato-5-Desidrogenase (CD), Chiquimato Quinase (CQ), 5-Enoilpiruvilchiquimato 3-fosfato Sintase (EPSPS) and Corismato Sintase (CS).

Da seqüência genômica, está claro que o bacilo da tuberculose tem o potencial de sintetizar todos os aminoácidos essenciais, vitaminas e cofatores de

enzimas, embora algumas das vias envolvidas devem diferir daquelas estabelecidas em outras bactérias. O *M. tuberculosis* pode metabolizar uma variedade de carboidratos, hidrocarbonetos, álcoois e ácidos carboxílicos (COLE *et al.*, 1998).

Desta forma, a tuberculose ou qualquer outra doença causada por um microrganismo que contém, por exemplo, a via do ácido chiquímico, poderá em princípio, ser tratada com inibidores das enzimas da rota do ácido chiquímico que impossibilitarão a produção do ácido corísmico - precursor chave para a biossíntese de PABA (ácido *p*-aminobenzóico, precursor do tetraidrofolato), ácido *p*-hidroxibenzóico (precursor da coenzima Q ou *ubiquinona*), micobactinas e dos aminoácidos aromáticos essenciais para a vida do bacilo.

1.4 Alvos para desenho de drogas baseado em estrutura

Uma extraordinária característica do *M. tuberculosis* que complica muito o tratamento é a “persistência”, a habilidade do organismo para ir a um estado de semidormência por muitos anos, fazendo com que durante este tempo mais drogas tenham sua eficácia limitada (PASCOPELLA *et al.*, 1994). Também, diferentes variedades de *M. tuberculosis* mostram diferentes virulências clínicas devido às variações genéticas no “fator de virulência” de proteínas que estão ainda somente identificadas parcialmente. Entender a persistência, a reativação de um estado persistente e a virulência, são os maiores desafios para os quais um entendimento fundamental do metabolismo do organismo pode ter importância direta para desenho de drogas baseado em estrutura.

O desenvolvimento de novas drogas deve ser facilitado pela identificação de genes essenciais para a viabilidade do bacilo, bem como fatores de persistência e fatores de virulência que contribuem para a patogênese. Outro alvo atrativo para desenho de drogas envolve produtos de genes de vias metabólicas importantes tal como a via do ácido chiquímico presente no *M. tuberculosis*.

O desenho de drogas baseado em estrutura tornou-se uma tecnologia altamente desenvolvida e utilizada nas maiores empresas farmacêuticas. A modelagem molecular comparativa ou por homologia é uma chave característica de um esforço integrado no descobrimento de novas drogas, porque ela permite que estas informações genômicas sejam utilizadas no desenvolvimento de ligantes alvos ou na engenharia de especificidade de ligantes (VEERAPANDIAN, 1997).

Uma das mais importantes técnicas utilizadas em conjunto com a modelagem molecular em biologia estrutural é o *docking* de um ligante a um receptor, tal como uma proteína. Se a estrutura do receptor é conhecida, então a aplicação é essencialmente de um desenho de droga baseado em estrutura. Estes métodos têm alguns objetivos relacionados, tais como: procurar identificar a localização do sítio ativo do ligante e talvez a geometria do ligante no sítio ativo. Uma outra meta é a classificação de uma série de ligantes relacionados em termos de sua afinidade ou avaliar a energia livre de ligação absoluta tão precisamente quanto possível (FOSTER, 2002).

Atualmente, modelos comparativos estão sendo usados em conjunto com *screening* virtual para identificar novos inibidores. Uma série de trabalhos

demonstra o sucesso no uso de modelos estruturais para auxiliar no desenho racional de drogas contra parasitas. Modelos comparativos foram usados em simulações de *docking*, identificando uma baixa constante de inibição para inibidores não peptídicos de proteases em malária e Schistosoma (RING *et al.*, 1993). Adicionalmente, modelos comparativos foram usados para justificar a afinidade de ligantes pelo sítio de ligação em *E. histolytica* (QUE *et al.*, 2002). A única maneira prática de explorar interações proteína-ligante para um número maior de sistemas é o uso de modelos moleculares de estruturas de proteínas, estabelecendo um limite mínimo de identidade com o *template* de 40%, podendo variar de acordo com a aplicação a que o modelo será submetido (Figura 4). Uma aproximação alternativa, implementada no programa MODELLER (ŠALI & BLUNDELL, 1993), procura satisfazer restrições estruturais expressas como função densidade de probabilidade (f.d.p.), as quais são derivadas de outras proteínas homólogas.

Computadores rápidos e a disponibilidade de computadores configurados como *clusters* de custo relativamente baixo tem aumentado a velocidade na qual as drogas podem ser identificadas e avaliadas *in silico*. O primeiro ciclo para o desenho de novas drogas inclui a determinação da estrutura da proteína alvo por um dos três principais métodos usados para desenho de drogas: cristalografia de raios X, RMN ou modelagem molecular comparativa de estruturas de proteínas. Uma vez que o alvo foi identificado, é necessário se obter informações sobre a precisão estrutural. Todas as estruturas devem ser avaliadas por vários programas para determinar desvios do comprimento de ligação com relação à geometria ideal, (as quais não

devem ser maiores que 0,015 Å ou 3° para ângulos de ligação. Átomos planares não devem estar mais que 0,015 Å fora do plano e não deve haver centros quirais incorretos. Finalmente, no mínimo 90% dos ângulos ϕ e ψ da cadeia principal devem cair na região mais favorável do gráfico de Ramachandran) garantindo maior precisão aos modelos (ANDERSON, 2003).

As estruturas 3D de novas proteínas alvo relevantes terapeuticamente estão se tornando disponíveis numa razão dramaticamente crescente através da determinação de estruturas por cristalografia de raios X, RMN ou por modelagem molecular comparativa. Devido ao crescimento do conhecimento estrutural, os experimentos de *docking* estão se tornando essenciais no desenho racional de drogas. Este interesse é atribuído ao *screening* virtual a bancos de dados de ligantes por métodos computacionais levando à identificação de novos alvos terapêuticos.

1.5 Bioinformática estrutural aplicada ao estudo de proteínas alvo

A bioinformática vem sendo utilizada bem antes dos grandes projetos genoma e das tecnologias que a tornaram uma área tão importante atualmente. A partir da década de 80, com o aprimoramento das técnicas de seqüenciamento e de novas tecnologias, o termo bioinformática foi lançado como uma nova área do conhecimento científico, representando a interação da biologia com a informática. Esta nova área passou a fazer parte de todos os projetos biológicos como forma de analisar grandes quantidades de dados, possibilidade de armazenamento em bancos de dados e apresentação de tais resultados em interfaces acessíveis via *web*,

tornando a pesquisa mais interativa e dinâmica. Uma definição mais ampla da bioinformática seria a aplicação de ferramentas de computação para análise, captura e interpretação de dados biológicos. É uma área interdisciplinar e absorve a ciência da computação, matemática, biologia, física e medicina (Figura 3) (BAYAT, 2002).

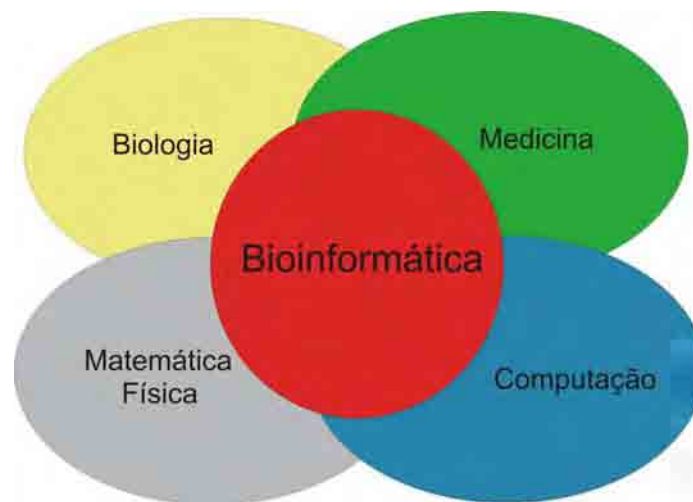


Figura 3. Interação das áreas que têm contribuído para a formação e o desenvolvimento da bioinformática (BAYAT, 2002).

Um dos grandes desafios da bioinformática começa a aparecer na era pós-genômica, na qual a proteômica se torna um dos principais alvos de estudo juntamente com o entendimento estrutural e funcional de proteínas.

A pesquisa de proteínas em bioinformática utiliza-se de anotações de proteínas e bancos de dados de eletroforese bidimensional. Após a separação, identificação e caracterização de uma proteína, o próximo desafio na bioinformática é a predição de sua estrutura. Biólogos estruturais usam a bioinformática para manusear o vasto e complexo conjunto de dados de cristalografia de raios X e RMN

para prever modelos 3D de moléculas de proteínas por modelagem molecular comparativa (BURLEY *et al.*, 1999).

O rápido aumento no número de estruturas 3D disponíveis em bancos de dados como o PDB (*Protein Data Bank*) (BERMAN *et al.*, 2000), levou à criação de uma sub-disciplina da bioinformática: a bioinformática estrutural. O principal foco desta sub-disciplina é a representação, armazenamento, recuperação, análise e visualização da informação estrutural a níveis atômicos. Assim, a predição de estruturas 3D de proteínas permanece uma área de grande interesse, sendo que a principal categoria de predições de estruturas de proteínas tem sido a modelagem molecular comparativa, baseada na alta homologia de uma seqüência por uma estrutura conhecida (SÁNCHEZ & ŠALI, 1997).

Os projetos de seqüenciamento de genomas completos têm nos fornecido uma enorme quantidade de dados, possibilitando a análise em larga escala das estruturas 3D obtidas através do método de modelagem molecular comparativa daquelas estruturas não determinadas por cristalografia de raios X e RMN, tal como o genoma do *M. tuberculosis*. Contudo, a bioinformática atuando em diversas áreas, nos dá opções de análise de dados de genômica e proteômica, munindo-nos de grande quantidade de dados armazenados em bancos de dados públicos, podendo ser utilizados no cruzamento de dados, obtendo inúmeras informações relevantes ao avanço em pesquisas biológicas e tecnológicas.

2. Objetivos

O objetivo do presente trabalho foi desenvolver ferramentas computacionais com o auxílio da computação de alto desempenho, integrando softwares de modelagem molecular comparativa e de análise de estruturas 3D de proteínas para o estudo estrutural do genoma completo do *M. tuberculosis*, disponibilizando os resultados estruturais em um banco de dados público, o DBMODELING. O banco de dados contém informações sobre vias metabólicas, enzimas, anotações, genes, coordenadas atômicas, seqüências primárias e dados sobre as análises e a modelagem. Todos os modelos foram checados com *softwares* de análises químicas de proteínas e *softwares* que avaliam a geometria da proteína, garantindo a precisão com dados apresentados na interface do banco de dados. O DBMODELING pode ser acessado no site: <http://www.biocristalografia.df.ibilce.unesp.br/tools>.

3. Materiais e Métodos

3.1 Modelagem molecular

A seqüência primária de uma proteína determina sua estrutura tridimensional, contudo o algoritmo que permita, com precisão absoluta, determinar a estrutura tridimensional de uma proteína partindo-se de sua seqüência ainda está por ser determinado. A modelagem molecular comparativa tem o potencial de gerar modelos confiáveis. A condição necessária é que a semelhança entre a seqüência designada e as estruturas do modelo sejam detectáveis e que o alinhamento correto entre elas possa ser construído. Esta aproximação para a modelagem da estrutura é possível porque uma pequena mudança na seqüência de uma proteína normalmente resulta em uma pequena mudança em sua estrutura tridimensional (LESK, 2001). Todas as aproximações baseadas nas restrições para modelagem molecular comparativa de proteínas, extraem as distâncias e as restrições dos ângulos diedros a partir do alinhamento da seqüência alvo com as estruturas relacionadas, adicionando restrições implícitas pela topologia covalente (restrições estereoquímicas) e calculam o modelo pela minimização das violações de todas as restrições. Desta forma, as duas principais diferenças entre as várias aproximações estão na derivação e satisfação das restrições espaciais (ŠALI, 1995). A precisão do método está em assumir que se há semelhança detectável entre duas seqüências lineares, a semelhança estrutural pode ser assumida e a função potencial guiará o modelo no caminho dos *templates* em direção à estrutura correta. A modelagem molecular

comparativa é composta de quatro passos sequenciais descritos a seguir (SÁNCHEZ & ŠALI, 1997).

3.1.1 Procura e seleção de *templates*

A modelagem inicia-se pela procura de *templates* em um banco de dados de estruturas de proteínas (PDB) (<http://www.rcsb.org/pdb>), usando-se como parâmetro de entrada uma seqüência primária de estrutura não determinada experimentalmente (alvo) para que esta seja alinhada com possíveis seqüências homólogas de estruturas conhecidas depositadas no PDB (*templates*). Nesta etapa, foram adquiridos os bancos de dados de seqüências primárias de proteínas contidas no PDB e o banco de seqüências primárias de *M. tuberculosis* (http://www.sanger.ac.uk/Projects/M_tuberculosis/) para que fossem feitos os alinhamentos por pares do proteoma do *M. tuberculosis* com o banco de dados extraído do PDB.

Uma vez obtida uma lista de *templates* potenciais usando-se um ou mais métodos de busca, é necessário selecionar os *templates* que são apropriados para o problema de modelagem em particular. Normalmente selecionamos os modelos que possuem identidade mais elevada, isto é, porcentagem mais alta de resíduos idênticos e um menor número de *gaps* no alinhamento. Para a construção de um complexo proteína-ligante, a escolha do *template* que contém um ligante semelhante é provavelmente mais importante que a resolução do modelo. Por outro lado, se o

modelo será usado para analisar a geometria do sítio ativo de uma enzima, é preferível usarmos um modelo de alta resolução.

3.1.2 Alinhamento *template/alvo*

Uma vez selecionado o *template*, um método deve ser utilizado para executar o alinhamento *template/alvo*. O alinhamento é um dos principais passos na modelagem, pois é dele que são extraídas as restrições espaciais para a construção do modelo. Portanto, usuários de métodos de modelagem molecular comparativa podem utilizar variadas faixas de identidade, sempre relacionando o modelo gerado a partir de uma identidade seqüencial com sua utilização. Para seqüências de proteínas proximamente relacionadas com identidade superior a 40% de identidade residual, o alinhamento será mais preciso. Regiões de baixa similaridade local de seqüências, são comuns quando a identidade total da seqüência está abaixo de 40% (SAQI *et al.*, 1998), podendo o modelo gerado a partir deste alinhamento ser utilizado para outros fins que não o *docking* ou a inferência de características evolutivas comuns. Alinhamentos abaixo de 30% começam a apresentar muitas falhas com grandes extensões de *gaps* e erros nos alinhamentos.

No alinhamento executado pelo MODELLER é utilizado o comando ALIGN2D, o qual é baseado no algoritmo de programação dinâmica, proposto por Needleman e Wunsch para alinhamento global de seqüências (NEEDLEMAN & WUNSCH, 1970).

3.1.3 Construção do modelo

Uma vez realizado o alinhamento entre a seqüência do alvo e do *template*, o modelo é construído utilizando-se a modelagem molecular comparativa por satisfação das restrições espaciais implementadas no programa MODELLER (ŠALI & BLUNDELL, 1993) e usa distância geométrica e técnicas de otimização para satisfazer as restrições espaciais obtidas do alinhamento. O programa MODELLER deriva muitas distâncias e restrições de ângulos diedros no alinhamento da seqüência alvo com o modelo da estrutura 3D. As restrições espaciais na seqüência alvo são obtidas da análise estatística das relações entre várias características da estrutura da proteína (pdf). Um banco de dados com 105 famílias incluindo alinhamentos de 416 proteínas com estrutura 3D conhecida foi construído para obter as tabelas quantificando as relações, tais como distâncias equivalentes entre $C_\alpha - C_\alpha$, ou entre ângulos diedros equivalentes da cadeia principal de duas proteínas relacionadas. Estas relações são expressas pela distribuição densidade de probabilidade condicional e podem ser usadas diretamente como restrição espacial. As restrições derivadas do *template* para a composição do conjunto de restrições total do modelo, violando a própria estereoquímica, compõem a função objetivo. Finalmente, o modelo é obtido pela otimização da função objetivo no espaço cartesiano. Vários modelos ligeiramente diferentes podem ser calculados variando a estrutura inicial. Outros fatores como seleção de *template* e um alinhamento preciso, têm um grande impacto na construção do modelo e em sua precisão, especialmente para modelos baseados em uma identidade seqüencial abaixo de 40% (Apêndice A.I).

3.1.4 Avaliação dos modelos

A qualidade do modelo predito determina a informação que pode ser extraída dele. Assim, estimar a precisão do modelo 3D da proteína é essencial para interpreta-lo. O modelo pode ser avaliado como um todo bem como em regiões individuais, com base na similaridade entre as seqüências do *template* e do alvo, observando resíduos importantes em regiões da proteína como o sítio ativo e sua conservação (SÁNCHEZ & ŠALI, 1998). Um requerimento básico para um modelo é ter uma boa qualidade estereoquímica. Os programas mais utilizados são o PROCHECK (LASKOWSKI *et al.*, 1998) e WHATCHECK (HOOFT *et al.*, 1996). As características de um modelo que são checadas por estes programas incluem comprimento de ligação, ângulo de ligação, ligação peptídica e planaridade de anéis da cadeia lateral, quiralidade, ângulos de torção da cadeia principal e cadeia lateral e choques entre pares de átomos não ligados.

Há também métodos para testar modelos 3D que implicitamente carregam muitas características espaciais compiladas de estruturas de proteínas a alta resolução. Estes métodos são baseados nos perfis 3D e potenciais estatísticos de força (SIPPL, 1990; LUTHY *et al.*, 1992). Os programas que implementam estas aproximações incluem o VERIFY3D (LUTHY *et al.*, 1992), PROSAIL (SIPPL, 1993), HARMONY (TOPHAM *et al.*, 1994) e ANOELA (MELO & FEYTMANS, 1998). Os programas avaliam o ambiente químico de cada resíduo em um modelo com respeito ao ambiente químico esperado como encontrado em estruturas de raios X à alta resolução.

3.2 Aplicações da modelagem molecular comparativa

A necessidade da modelagem molecular comparativa de estruturas de proteínas se encaixa nos mais variados tipos de pesquisa. Por exemplo, modelos comparativos podem ser úteis em desenhos para testes de hipótese sobre função de proteínas mutantes (BOISSEL *et al.*, 1993; WU *et al.*, 1999), identificar sítio ativo e ligações (RING *et al.*, 1993), modelar um substrato específico (XU *et al.*, 1996), simular *docking* de proteína-proteína ou proteína-ligante (VAKSER, 1997), facilitar a substituição molecular na determinação de estruturas de raios X (HOWELL *et al.*, 1992), refinar modelos baseados em restrições de RMN (MODI *et al.*, 1996) e confirmar uma relação estrutural remota (GUENTHER *et al.*, 1997; MIWA *et al.*, 1999).

As aplicações de modelos moleculares determinados por modelagem molecular comparativa estão diretamente relacionadas à precisão dos modelos com relação à identidade entre o alvo e o *template*, estabelecendo uma escala que varia de acordo com sua identidade e o r.m.s.d. determinado (Figura 4). Alta precisão em modelos comparativos é baseada na identidade seqüencial acima de 50% com relação aos seus *templates*. Tais modelos tendem a ter um r.m.s.d. de aproximadamente 1 Å para átomos da cadeia principal, o qual é comparável à precisão de estruturas determinadas por RMN e estruturas obtidas por difração de raios X a média ou a baixa resolução. Precisão média em modelos comparativos é baseada em uma identidade de 30-50%. Estes modelos tendem a ter aproximadamente 90% da cadeia principal modelada com um r.m.s.d. de 1,5 Å. Há

um empacotamento de cadeias laterais mais freqüentes, erros de distorção de *core* e modelagem de *loop* e há ocasionalmente erros nos alinhamentos. Finalmente, modelos de baixa precisão são aqueles obtidos com identidade inferior a 30%. Os erros nos alinhamentos aumentam rapidamente quanto menor a identidade e tornam-se mais significantes, originando erros nos modelos gerados. Assim, quando um modelo é gerado com um alinhamento insignificante com relação a uma estrutura conhecida, ele deve ter um enovelamento totalmente incorreto. Outros fatores como seleção do *template* e alinhamento preciso normalmente tem um grande impacto na precisão dos modelos, especialmente para modelos gerados com identidade acima de 40% (PIEPER *et al.*, 2004).

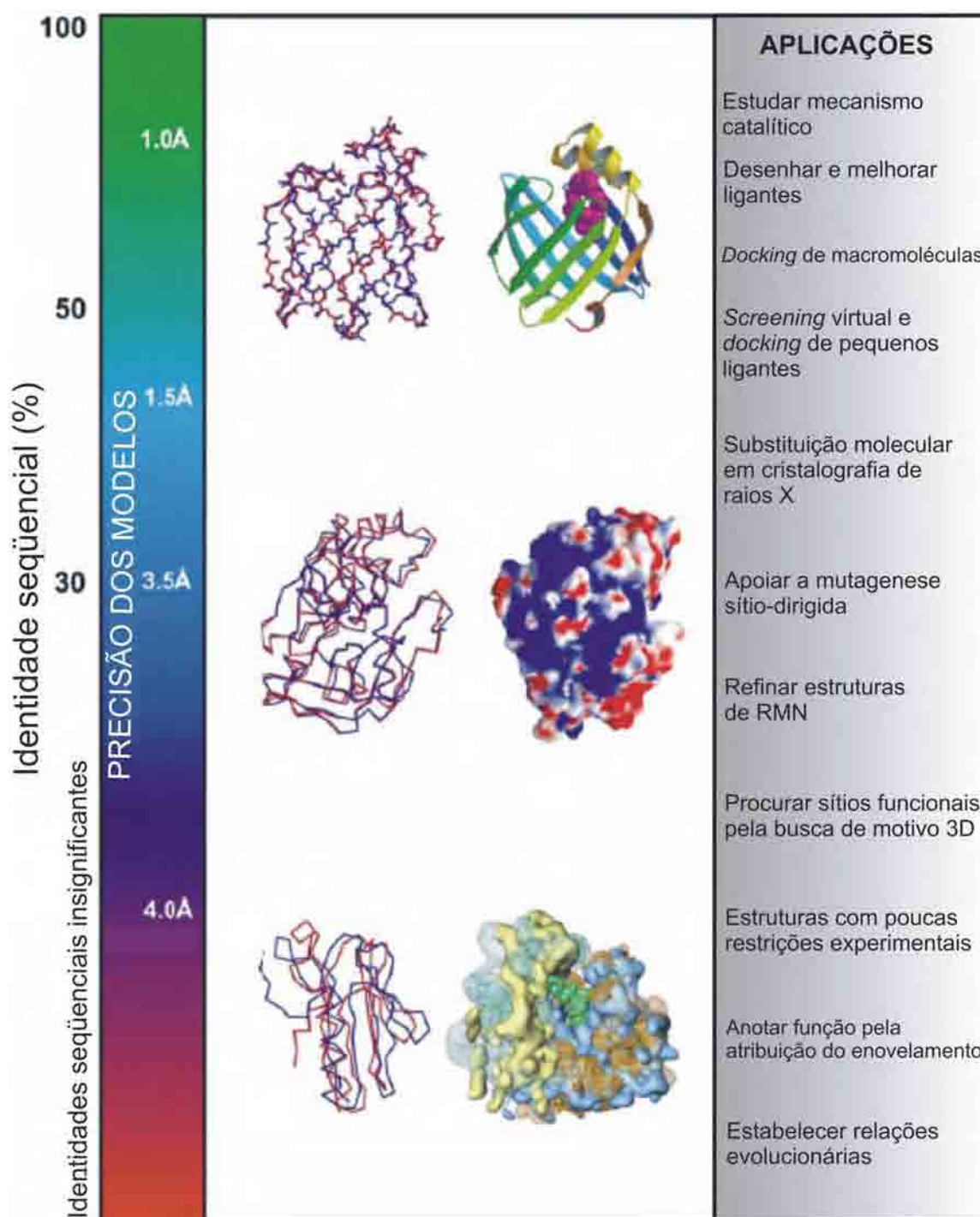


Figura 4. O diagrama acima descreve a precisão e aplicação de modelos estruturais de proteínas. O eixo vertical indica os diferentes intervalos da aplicabilidade da modelagem molecular comparativa de estruturas de proteínas, a precisão correspondente dos modelos estruturais e suas aplicações de acordo com a porcentagem de identidade relacionada ao r.m.s.d. (MARTIRENOM *et al.*, 2002)

3.3 Possíveis erros em modelagem comparativa

Com a diminuição da identidade entre a seqüência alvo e o *template*, os erros na modelagem aumentam, podendo ser divididos em cinco categorias (SÁNCHEZ & ŠALI, 1997) (Figura 5). Um caminho informativo para testar métodos de modelagem de estrutura de proteínas é fornecido pelo EVA-CM (EYRICH *et al.*, 2001) e LiveBench (BUJNICKI *et al.*, 2001).

a) Erros no empacotamento das cadeias laterais. Como as seqüências divergem, o empacotamento das cadeias laterais muda a estrutura da proteína. Erros em cadeias laterais são críticos se ocorrem em regiões que estão envolvidas na função da proteína, tais como sítios ativos e sítios de interação com ligantes.

b) Distorções e mudanças em regiões corretamente alinhadas. Como uma consequência da divergência de seqüências, há mudanças na conformação da cadeia principal, mesmo que o enovelamento geral permaneça o mesmo. Portanto, é possível que em alguns segmentos de um modelo alinhados corretamente, o *template* seja localmente diferente do alvo, resultando em erros naquela região. As diferenças estruturais são algumas vezes não devido a diferenças na seqüência, mas sim uma consequência de artefatos na determinação da estrutura em diferentes ambientes (ex. empacotamento de subunidades em um cristal). O uso de vários *templates* pode minimizar esta variedade de erros (SRINIVASAN & BLUNDELL, 1993; SÁNCHEZ & ŠALI, 1997).

c) Erros em regiões sem *template*. Segmentos da seqüência alvo que não têm região equivalente na estrutura do *template* (ex. inserções e *loops*) são as regiões mais

difíceis de modelar. Se a inserção é relativamente curta, menor que 9 resíduos, alguns métodos podem prever corretamente a conformação da cadeia principal (VAN VLIJMEN & KARPLUS, 1997; FISER *et al.*, 2000). As condições para o sucesso na predição são o alinhamento correto e um ambiente precisamente modelado em torno da inserção.

d) Erros devido a alinhamentos ruins. A maior fonte de erros em modelagem molecular comparativa são os alinhamentos ruins, especialmente quando a identidade seqüencial entre o *template* e o alvo está abaixo de 30%. Uma forma de minimizar estes erros é utilizar várias seqüências para construir um alinhamento múltiplo, atribuindo identidade em regiões da seqüência onde o alinhamento com apenas um *template* gerava *gaps*. (SÁNCHEZ & ŠALI, 1997).

e) Modelos incorretos. Este é um problema potencial quando proteínas distantemente relacionadas são usadas como *templates* (< 25% de identidade seqüencial). Distinguir entre um modelo baseado em um *template* incorreto e um modelo baseado em um alinhamento incorreto com um *template* correto é difícil. Em ambos os casos, os métodos de avaliação irão prever um modelo irreal. A conservação da chave funcional ou estrutural de resíduos na seqüência alvo aumenta a confiança em um dado enovelamento.

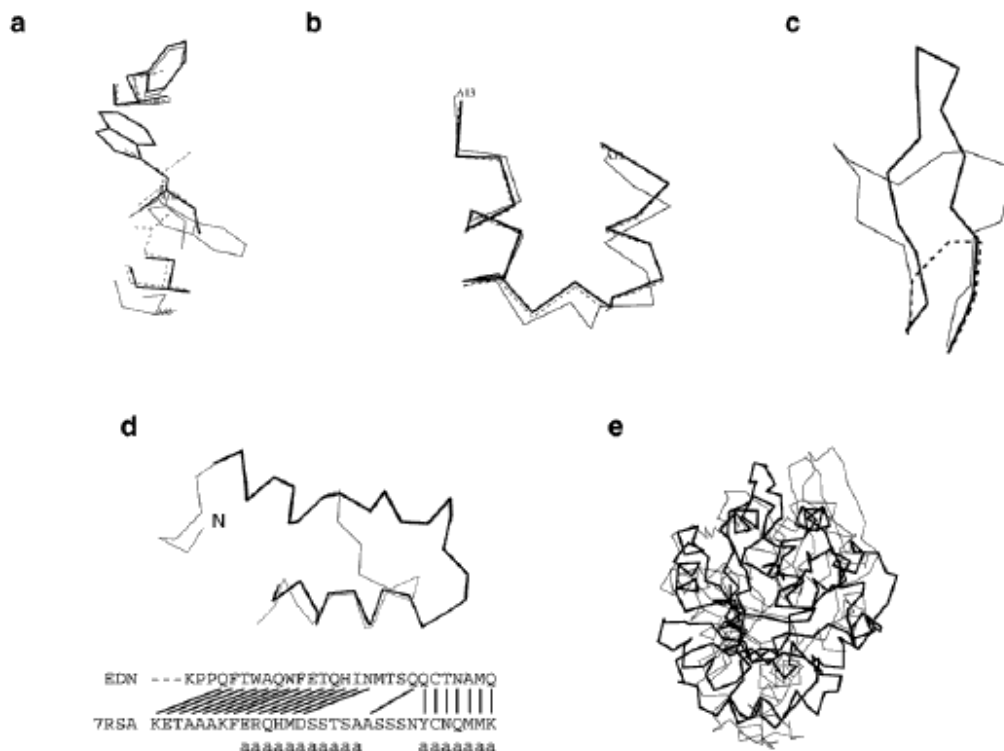


Figura 5. Possíveis erros em modelagem molecular comparativa. **a)** Erros no empacotamento das cadeias laterais. **b)** Distorções e mudanças em regiões corretamente alinhadas. **c)** Erros em regiões sem *template*. **d)** Erros devido a alinhamentos ruins. **e)** Modelos incorretos (Marti-Renom *et al.*, 2002).

3.4 Modelagem em larga escala do genoma do *M. tuberculosis*

Devido ao excelente progresso na biologia, existe a necessidade de descrever e entender a função de muitas proteínas em mais detalhes. Embora funções de proteínas sejam melhores determinadas experimentalmente (OLIVER, 1996), algumas vezes podem ser preditas pela comparação da seqüência de uma proteína com proteínas de funções conhecidas (OLIVER, 1996; KOONIN & MUSHEGIAN, 1996; DUJON, 1996). Isto é possível porque seqüências de proteínas similares tendem a ter funções similares, embora exceções também ocorram (ORENGO *et al.*, 1994). O sucesso e a utilidade da assinatura computacional de função de proteínas,

recentemente aumentou dramaticamente, devido ao grande número de projetos de seqüenciamento de genomas (MIKLOS & RUBIN, 1996), procurando atribuir estruturas 3D e inferir funções às proteínas identificadas nestes genomas.

Devido à importância do genoma do *M. tuberculosis* para a saúde pública e por ser uma doença negligenciada, houve a necessidade de se criar uma ferramenta computacional automatizada que pudesse, com a utilização da modelagem molecular comparativa de proteínas, determinar todos os modelos possíveis para este genoma. Como o genoma de *M. tuberculosis* é composto de aproximadamente 3.924 *ORFs*, a automatização da modelagem foi executada em um *cluster Beowulf*, com o objetivo de se minimizar o tempo de busca por *templates* e da modelagem. Os modelos depositados no DBMODELING apresentaram uma boa qualidade estereoquímica (mais de 85% na região mais favorável do gráfico de Ramachandran).

O método de modelagem molecular comparativa de proteínas gera modelos de estruturas de proteínas mais precisos e detalhados, maximizando sua utilidade em aplicações tais como interpretação da existência de dados funcionais, desenho de ligantes e construção de proteínas mutantes para teste de novas hipóteses funcionais (JOHNSON *et al.*, 1994). O fluxograma apresentado na figura 6 foi implementado em um *cluster Beowulf* com sistema operacional UNIX/Linux Conectiva 9.0 e configuração de *hardware* composta de 16 nós com Athlon XP 2100+, 1Gb de RAM, 80Gb HD e placas de rede 3Com de 100 Mbits conectadas a um Switch 3Com SuperStack 3300 10/100 Mbits. O programa *Parmodel* (UCHÔA *et al.*, 2004) foi utilizado para distribuir eficientemente as tarefas para todos os nós do

cluster, sem ter que adaptar os programas individuais para execução em paralelo (<http://www.biocristalografia.df.ibilce.unesp.br/tools/parmodel>). Todos os modelos estão acessíveis no DBMODELING (DA SILVEIRA *et al.*, 2005) (Apêndice B) no site <http://www.biocristalografia.df.ibilce.unesp.br/tools>. Os passos do fluxograma da figura 6 têm como objetivo, otimizar o tempo e generalizar o processo de modelagem em larga escala para genomas diversos.

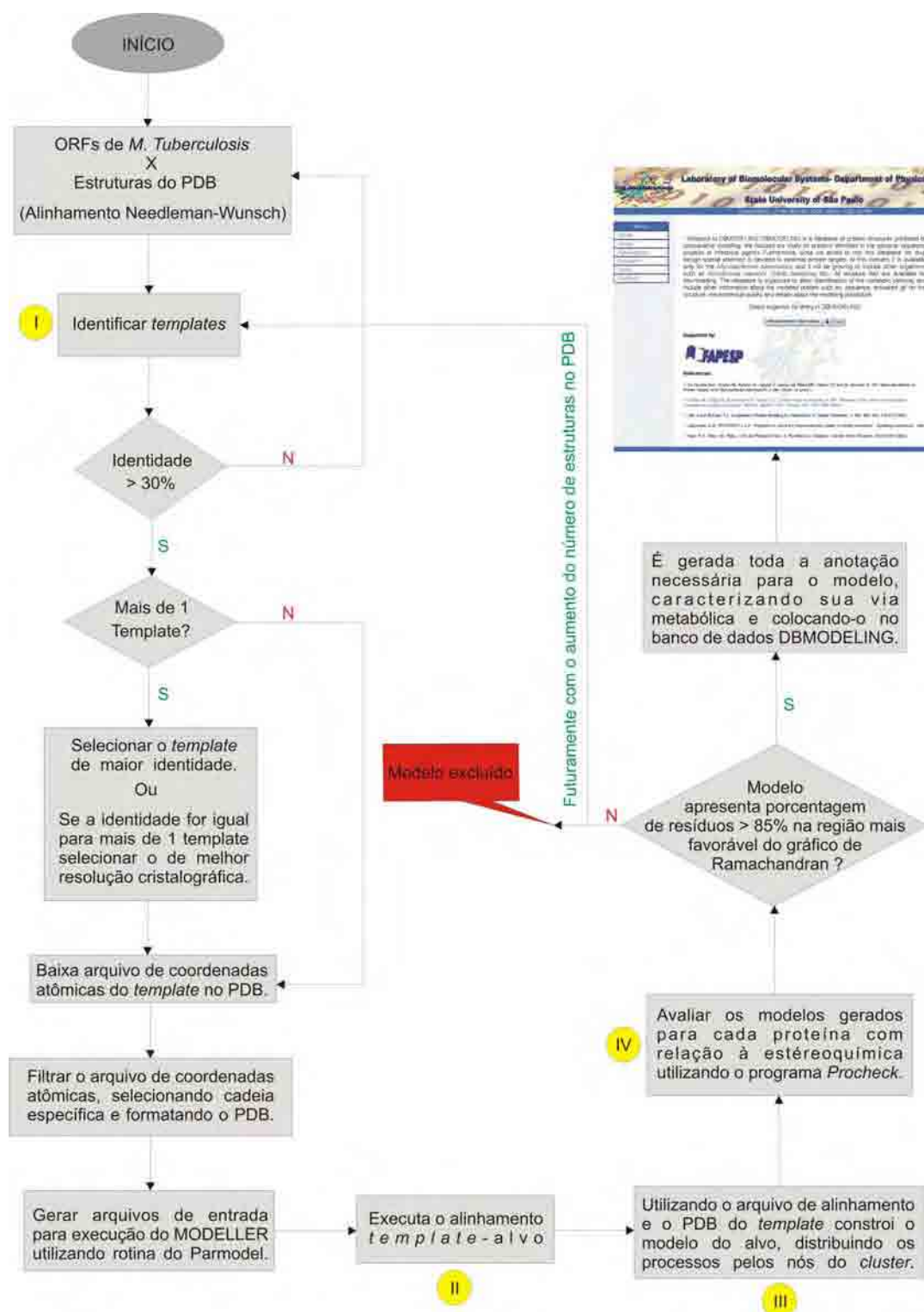


Figura 6. Fluxograma do algoritmo criado para automatizar a modelagem comparativa de estruturas de proteínas, onde os círculos amarelos representam as etapas de modelagem utilizadas por MODELLER (ŠALI & BLUNDELL, 1993).

3.5 Busca por *templates* e algoritmo de alinhamento

Para a busca por estruturas *template*, descrita no passo I da figura 6, para a modelagem molecular comparativa das *ORFs* de *M. tuberculosis*, foi utilizado o algoritmo de programação dinâmica para alinhamento de seqüências proposto por Needleman e Wunsch em 1970 (NEEDLEMAN & WUNSCH, 1970) (<http://emboss.sourceforge.net/download/>) (Apêndice A.II). Cada uma das 3924 *ORFs* foi submetida a um *script* desenvolvido em Perl, o qual utiliza as seqüências primárias de todas as estruturas depositadas no PDB (*Protein Data Bank*) e as alinham contra cada seqüência do proteoma de *M. tuberculosis*, extraíndo todos os *templates* que possuïrem identidade residual acima de 30% (limite inferior atribuído no *script* para estabelecer relação de homologia mínima entre a seqüência alvo e o *template*). Todo o processo de busca e seleção de *templates*, foi executado particionando o arquivo de *ORFs* pelo número de nós do *cluster* e disparando processos para iniciar o alinhamento de cada parte do arquivo com todas as seqüências de estruturas do PDB. Após a realização do alinhamento, o programa identifica automaticamente a identidade de todos os *templates* (> 30%), extraíndo apenas o de maior identidade. Posteriormente, outro *script* é acionado automaticamente executando o acesso ao *site* do PDB e baixando a estrutura do *template* a ser utilizado. No momento em que o *template* é baixado, é feita uma filtragem do arquivo de coordenadas atômicas, excluindo dados desnecessários e formatando-o para ser utilizado como entrada no programa MODELLER (ŠALI & BLUNDELL, 1993).

Nesta etapa, após a busca por *templates*, são geradas as entradas utilizadas no programa de modelagem molecular comparativa na forma necessária para a paralelização da modelagem (Figuras 7 e 8), dividindo o arquivo de entrada que estabelece o número de modelos a ser gerado e implantando uma semente aleatória em cada arquivo para que os modelos não se repitam ao serem executados em nós diferentes (Figura 9).

```
READ_MODEL FILE = '1HMS.pdb'  
SEQUENCE_TO_ALI ALIGN_CODES = '1HMS'  
READ_ALIGNMENT FILE = 'blbp.seq', ALIGN_CODES = 'blbp', ADD_SEQUENCE = on  
ALIGN2D  
WRITE_ALIGNMENT FILE = 'blbp-1HMS.ali', ALIGNMET_FORMAT = 'PIR'  
WRITE_ALIGNMENT FILE = 'blbp-1HMS.pap', ALIGNMET_FORMAT = 'PAP'
```

Figura 7. Arquivo de entrada para gerar o alinhamento pelo MODELLER

```
INCLUDE  
SET ALNFILE = 'blbp-1HMS.ali'  
SET KNOWN = '1HMS'  
SET SEQUENCE = 'blbp'  
SET STARTING_MODEL = 1  
SET ENDING_MODEL = 1000/nº nós  
RAND SEED = -1247  
CALL ROUTINE = 'model'
```

Figura 8. Arquivo de entrada da modelagem, indicando o número de modelos a serem gerados e a semente aleatória.

A partir deste passo há a integração com o programa *Parmodel*, que utilizará todos os dados gerados inicialmente para executar a modelagem em paralelo em um *cluster Beowulf* com 16 nós.

3.6 Modelagem molecular comparativa usando um *cluster Beowulf*

A ferramenta desenvolvida para modelagem molecular comparativa em larga escala só foi possível com a implementação dos processos de modelagem utilizando a tecnologia de *clusters*, que integrou a capacidade de gerar dados e analisá-los com a rapidez necessária ao desenvolvimento da pesquisa. A figura 9 mostra a arquitetura do *cluster Beowulf* utilizada no Laboratório de Sistemas Biomoleculares e uma foto do mesmo no laboratório onde o projeto foi desenvolvido (Figura 10).

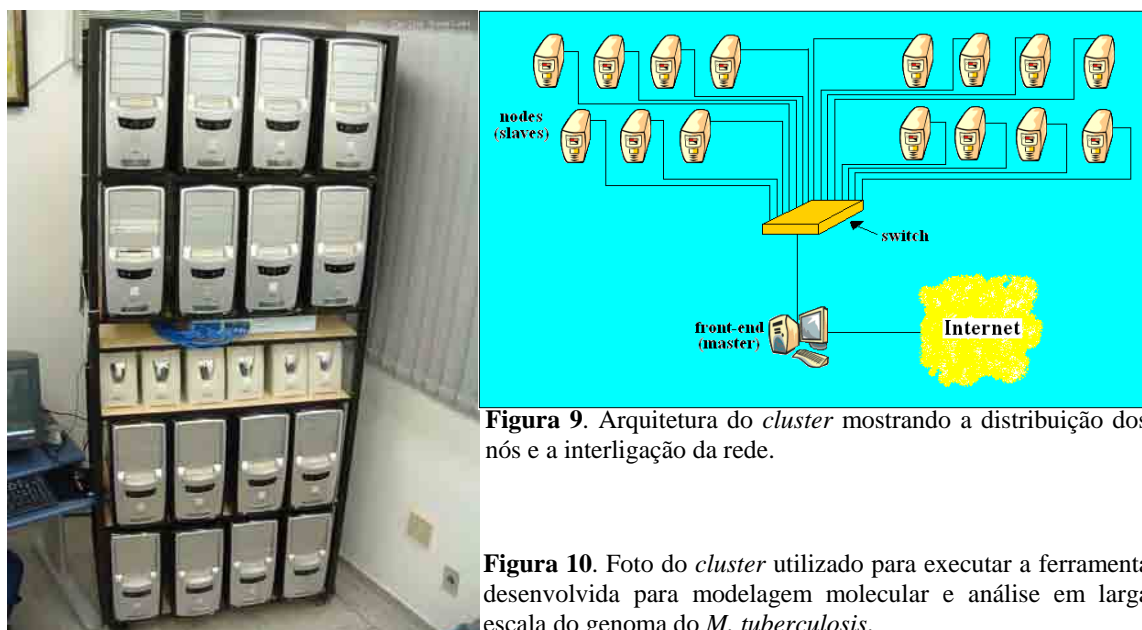


Figura 9. Arquitetura do *cluster* mostrando a distribuição dos nós e a interligação da rede.

Figura 10. Foto do *cluster* utilizado para executar a ferramenta desenvolvida para modelagem molecular e análise em larga escala do genoma do *M. tuberculosis*.

A partir do segundo passo do fluxograma da figura 6 para a modelagem em larga escala de genomas completos, os processos são executados utilizando as rotinas de modelagem do *Parmodel* (UCHÔA *et al.*, 2004). Estas rotinas visam agilizar todo o processo de modelagem molecular e análise de estruturas 3D de

proteínas. Esta agilidade é obtida de duas formas: integrar automaticamente todas as etapas do processo de modelagem molecular para que não haja nenhuma intervenção do usuário no decorrer deste processo e paralelizar a etapa de modelagem para que possa se obter uma diminuição significativa no seu tempo de processamento. O *Parmodel* está acessível publicamente, utilizando uma interface amigável ao usuário no site <http://www.biocristalografia.df.ibilce.unesp.br/tools/parmodel>.

A paralelização da modelagem molecular comparativa utilizando o MODELLER, foi realizada utilizando uma biblioteca de linguagem C, o MPI (*Message Passing Interface*), que controla a distribuição dos processos de modelagem pela divisão dos arquivos de entrada do MODELLER no *cluster Beowulf*. Isto permite paralelizar a execução do MODELLER e diminuir o tempo de processamento das rotinas de modelagem. A arquitetura do *cluster Beowulf* utilizada no Laboratório de Sistemas Biomoleculares foi inteiramente projetada aos interesses da bioinformática estrutural, adaptando todos os programas e construindo ferramentas que tornassem o *cluster* o mais específico possível para a pesquisa estrutural de proteínas.

Após a seleção dos *templates* e a criação das entradas para realizar a modelagem, é executado o alinhamento *template-alvo*, do qual é obtido o alinhamento a ser utilizado no terceiro passo da modelagem como arquivo de entrada para o MODELLER. Antes de iniciar o passo que executará a construção dos modelos, um *script* verifica se há alguma modelagem sendo realizada no momento. Isto porque, caso esteja sendo realizada alguma modelagem e outra

modelagem for submetida, ocorrerá uma perda no desempenho do programa. Assim, é verificada a existência da execução de alguma modelagem. Caso haja alguma modelagem sendo executada, então os parâmetros de entrada necessários à execução do programa serão gravados em uma fila de espera e os arquivos que o usuário submeteu permanecerão no diretório que foi criado. Ao término de cada modelagem, o programa verificará a existência de alguma modelagem nesta fila. Caso não haja modelagens sendo feitas, o *script* executará um programa em C implementado com rotinas MPI, iniciando a execução da modelagem distribuindo o número total de modelos solicitados pelos 16 nós para o próximo elemento da fila. Para cada proteína de *M. tuberculosis*, foi gerado e analisado um total de 1000 modelos, ampliando o espectro de análise com o objetivo de se obter melhores modelos.

3.7 Softwares de análise estrutural e validação de modelos

Logo que a modelagem é finalizada, a avaliação de cada modelo é feita automaticamente utilizando programas como: PROCHECK (LASKOWSKI *et al.*, 1993) (Apêndice A.III), WHATCHECK (HOOFT *et al.*, 1996), VERIFY3D (BOWIE *et al.*, 1991; LUTHY *et al.*, 1992) (Apêndice A.IV) e X-PLOR (BRÜNGER, 1992) para avaliarmos o RMSD da geometria ideal de cada proteína (Apêndice A.V). Algumas das propostas dos programas PROCHECK e WHATCHECK são (i) determinar erros grosseiros nas estruturas, tais como cadeias laterais deslocadas, (ii) checar anormalidades locais da estereoquímica e (iii) produzir critérios para a qualidade estereoquímica global (EU 3-D VALIDATION

NETWORK, 1998). O WHATCHECK (HOOFT *et al.*, 1996) oferece informações sobre a formação de regiões centrais hidrofóbicas, a acessibilidade de resíduos e átomos a moléculas de solvente (água), a distribuição espacial de grupos iônicos, a distribuição das distâncias atômicas e das ligações de hidrogênio da cadeia principal para cada modelo no banco de dados. Portanto, ele retrata a estereoquímica, comprimentos de ligações, ângulos diedros, entre outras quantidades na forma de um relatório gerado no formato “pdf”, descrevendo todas as análises executadas com a enzima pesquisada e há um *link* para o PROCHECK com as porcentagens da região mais favorável até a região não permitida, além da figura do gráfico de Ramachandran. As características de um modelo que são checadas por estes programas incluem comprimento de ligação, ângulo de ligação, ligações peptídicas e planaridade dos anéis das cadeias laterais, quiralidade, ângulos de torção de cadeias laterais e cadeia principal e choques entre pares de átomos não ligados na estrutura.

O VERIFY3D mede a compatibilidade da estrutura 3D com sua seqüência primária, usando um perfil 3D. Cada posição do resíduo na estrutura é caracterizada pelo seu ambiente químico e é representado por uma fileira de 20 números no perfil. Estes números são as preferências estatísticas (chamadas 3D-1D scores) de cada um dos 20 aminoácidos para este ambiente químico (MARTI-RENOM *et al.*, 2004). Os ambientes dos resíduos são definidos por três parâmetros: a área do resíduo que está no interior da proteína, a fração de área de cadeia lateral que está ocupado por átomos polares (O e N) e a estrutura secundária local. O *link* no banco de dados de *M. tuberculosis* possibilita gerar um gráfico para a enzima de interesse com a análise

do escore 3D-1D para cada aminoácido, o escore do perfil 3D S para sua seqüência de aminoácidos e o escore ideal S_{ideal} , que é calculado a partir do comprimento da proteína. Logo, estruturas que possuem erros em seus enovelamentos têm tipicamente escores menores que $0,45 S_{ideal}$. Um escore próximo ou acima do S_{ideal} indica uma estrutura confiável (LUTHY *et al.*, 1992). Esses softwares de avaliação química de modelos gerados por modelagem molecular comparativa, nos dão maior confiabilidade nas estruturas geradas, possibilitando propor simulações de *docking* contra bibliotecas de ligantes (*screening* virtual) com o objetivo de selecionar alvos terapêuticos para desenho de drogas baseado em estrutura.

3.8 Perl/CGI

A programação e a bioinformática estão relacionadas, tanto na obtenção de dados, quanto no desenvolvimento de ferramentas que resolvam os problemas e obstáculos encontrados na pesquisa em desenvolvimento. A elaboração de ferramentas para acessar bancos de dados e realizar tarefas importantes para a otimização da modelagem e análise de proteínas, tornou-se possível com a utilização de programação em linguagem Perl (*Practical Extraction and Reporting Language*), a qual é uma linguagem muito utilizada em bioinformática por sua facilidade na manipulação de *strings*, conexão a bancos de dados e acesso via *web*.

Existem inúmeras vantagens em se utilizar esta linguagem, principalmente devido a sua eficiência, pois é necessário menos tempo para extrair dados e manipula-los, comparada ao C ou Java. Os dados biológicos são armazenados em

bancos de dados e arquivos de texto enormes. É possível analisar e classificar esses dados manualmente, mas levaria muito tempo, por isso, cientistas e programadores desenvolvem ferramentas para automatizar o processo. A Perl, com sua capacidade altamente desenvolvida para detectar padrões em dados, e especialmente seqüências de caracteres de texto, é a melhor opção para desenvolvimento dos *scripts*. A riqueza dos códigos Perl existentes para a bioinformática, a integração sem problemas do código com sistemas baseados em Unix, a portabilidade para várias plataformas e uma comunidade de usuários incrivelmente entusiástica tornam a *Perl* a linguagem de *scripts* preferida para aplicativos de bioinformática.

A Perl é gratuita e já vem incorporada ao Linux quando instalado, além de ser uma das linguagens de programação que mais favorece a criatividade. Esta linguagem de programação é muito rica, onde os tipos e as estruturas são simples de usar e compreender, tendo diversos recursos que enriquecem a linguagem, tais como: depuradores, perfis, referências cruzadas, compiladores, interpretadores, bibliotecas e editores direcionados para a sintaxe. Assim, a conexão com bancos de dados fica facilitada e o desenvolvimento de programas simplificado, mas com objetividade e utilidade (WALL *et al.*, 2000).

A finalidade de bancos de dados biológicos públicos é permitir que a comunidade científica compartilhe dados com simplicidade. Nada é mais simples e direto do que a *Web*. Portanto, é quase um pré-requisito ao desenvolver um banco de dados, pensar em como tornar os dados disponíveis na *Web*. Há muitas tecnologias que permitem a comunicação entre páginas da *Web* e bancos de dados.

A mais antiga é denominada programação CGI (*Common Gateway Interface*). Um programa ou *script* CGI é um aplicativo de software que reside em um servidor da *Web*. Quando o programa CGI é chamado por um usuário remoto do servidor da *Web*, o aplicativo é executado no servidor e, em seguida, passa as informações de entrada do formulário de volta ao usuário remoto na forma de uma página da *Web*, conforme mostrado na figura 11. Os programas CGI são acessados utilizando-se o protocolo HTTP (*Hypertext Transport Protocol*), exatamente como as páginas *Web* em HTML (*Hypertext Markup Language*). Quando o servidor recebe uma solicitação de HTTP, em vez de apenas exibir o código CGI em seu navegador, como faria com uma página da *Web* normal, o servidor executa o programa CGI. A CGI é uma tecnologia relativamente madura e é suportada por todos os principais servidores da *Web*.

Os programas CGI geralmente consistem de algumas seções (Figura 11). Primeiro há uma seção do programa que coleta entradas do usuário a partir de um formulário *Web*, tal como selecionar o banco de dados a ser pesquisado. Isso é seguido por uma seção do programa que carrega a entrada do usuário em uma variável e executa algo com base nesta entrada. O programa CGI pode conter o código completo para fazer o processamento da entrada, mas é mais provável que o programa formate a entrada apropriadamente e a repasse para um programa separado residente no servidor, depois colete a saída daquele programa quando a execução terminar. A função final do programa CGI é retornar a saída do processo que foi executado no servidor para o usuário, na forma de uma página *Web*, que

pode conter saída textual ou *links* para arquivos de resultados disponíveis para *downloads* ou ambos.

A figura a seguir mostra um desenho esquemático da integração entre CGI e o banco de dados, utilizando a linguagem de programação Perl para interação com o banco de dados criado para o genoma do *M. tuberculosis* e a forma com a qual o CGI solicita as informações ao servidor, organizando suas diferentes vias metabólicas e suas respectivas enzimas.

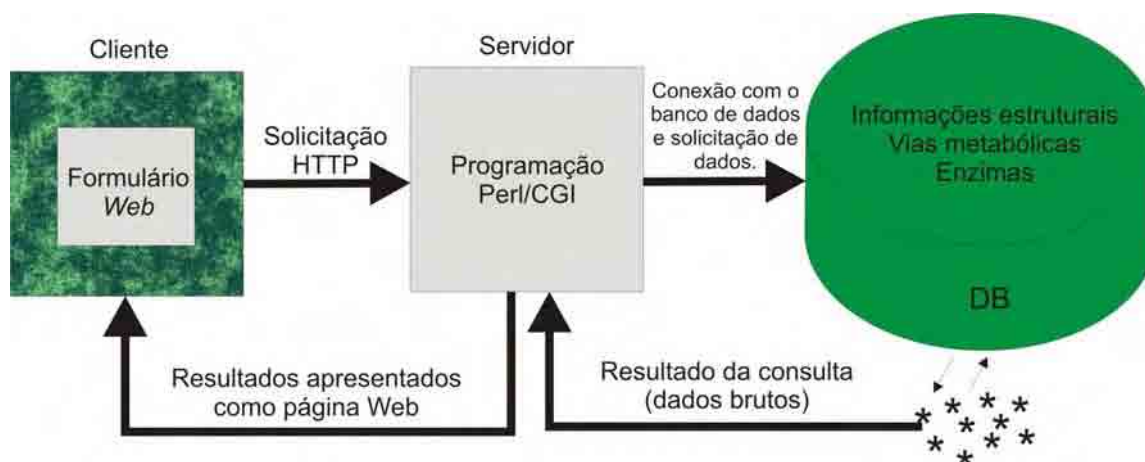


Figura 11. Diagrama esquemático de como a programação CGI interage com o banco de dados de *M. tuberculosis*, organizando e filtrando os dados devolvidos ao usuário.

3.9 Banco de dados MySQL

O banco de dados é uma ferramenta de fundamental importância na bioinformática, tanto na busca como no armazenamento de informações biológicas. Há dois tipos de sistemas de gerenciamento de banco de dados: sistemas de indexação de arquivos simples e DBMSs (*Database Management Systems*)

relacionais. Optar por um sistema de indexação de arquivos simples ou um sistema de banco de dados relacional é uma decisão importante que terá implicações de longo prazo para a capacidade e utilidade do banco de dados. Um banco de dados de arquivos simples é simplesmente uma coleção ordenada de arquivos semelhantes, geralmente em conformidade com um formato padrão de conteúdo. Os bancos de dados de arquivos simples podem ser pesquisados devido à indexação. Um índice extrai um atributo específico de um arquivo e alinha o valor do atributo no índice com um nome de arquivo e uma localização.

Os bancos de dados relacionais são apenas uma forma de coletar todas as informações sobre algo e armazená-las em um computador. Em um banco de dados de arquivos simples, todas as informações sobre o objeto de estudo são armazenadas em um grande arquivo de texto estruturado. Em um banco de dados relacional, as informações são armazenadas em um conjunto de tabelas. Os dados em uma tabela de banco de dados relacional são organizados em linhas, onde cada linha representa um registro no banco de dados. Uma linha pode conter várias informações separadas (campos). Cada campo no banco de dados pode conter uma informação distinta. Não pode consistir em um conjunto ou lista que possam ser divididos em partes menores. A função do RDBMS (*Relational Database Management System*) é fazer conexão entre tabelas relacionadas, localizando rapidamente os elementos comuns que estabelecem esses relacionamentos. A rede de tabelas e relacionamentos que compõe um banco de dados é denominada esquema de banco de dados. Para que um banco de dados mantenha sua utilidade com o passar do tempo, é melhor

desenvolver o esquema com cuidado antes mesmo de pensar em começar a popular o banco de dados. O MySQL é um DBMS relacional de médio porte que possibilita ao usuário criar, manter e gerenciar bancos de dados eletrônicos, além de ser gratuito e estar disponível para *download* no *site* do MySQL (<http://www.mysql.com/download>). O banco de dados foi instalado em um sistema operacional Linux-Conectiva 9.0, em um AthlonXP 2100+ com 80 Gb de HD, 1 Gb RAM e placa de rede 3Com de 100 Mbits, tendo grande capacidade para armazenamento de dados e rapidez nos processos de busca e aquisição das informações.

Desde o advento da *World Wide Web*, os bancos de dados biológicos se tornaram uma parte vital da literatura biológica. Saber localiza-los e fazer *download* de informações dos repositórios centrais de dados biológicos é atualmente uma habilidade tão importante para os pesquisadores quanto à pesquisa na literatura tradicional. A Figura 12 mostra o relacionamento entre as tabelas do banco de dados de vias metabólicas e enzimas.

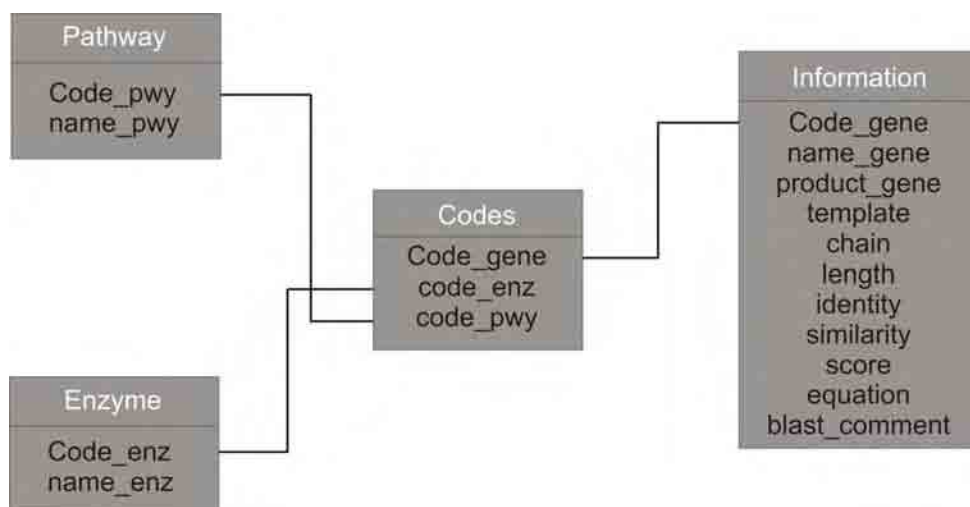


Figura 12. O diagrama descreve a relação entidade-relacionamento para as tabelas do banco de dados DBMODELING. Cada tabela é relacionada a um código (code_pwy, code_enz e code_gene). Para cada via metabólica selecionada há somente um código code_pwy, o qual selecionará sua enzima de forma a não cometer redundâncias linkando à tabela de informações e resultados sobre cada enzima. O diagrama descreve o relacionamento das tabelas para o conjunto de dados apresentados na interface web sem redundâncias.

3.10 Programas, servidores e *links* no DBMODELING

Alguns dados como vias metabólicas, foram obtidos de *sites* confiáveis e de constante atualização, garantindo a precisão dos dados apresentados no DBMODELING. A tabela 2 mostra alguns *sites* de interesse em biologia estrutural para obtenção e análise de dados utilizados no presente banco de dados. A tabela se divide em 3 partes, as quais apresentam informações sobre a utilidade, a caracterização em S ou P (Servidor ou Programa) e o endereço eletrônico, respectivamente.

Tabela 2. Programas e servidores *web* usados nos alinhamentos, construção e avaliação dos modelos. A primeira coluna indica para onde o *link* está direcionado ou sua função. A segunda coluna mostra se é programa (P) ou aplicativo em servidor (S) e a terceira coluna o endereço *web* de cada programa e servidor.

Laboratório de Sistemas Biomoleculares (BMSys)		
Grupo	S	http://www.biocristalografia.df.ibilce.unesp.br/node4_english.php
Publicações	S	http://www.biocristalografia.df.ibilce.unesp.br/publications.php
Pesquisas	S	http://www.biocristalografia.df.ibilce.unesp.br/pesquisa/pesquisa_english.php
Ferramentas	S	http://www.biocristalografia.df.ibilce.unesp.br/tools
Alinhamento		
Needleman-Wunsch	P	http://laboheme.df.ibilce.unesp.br/cgi-bin/db_modeling
Construção do modelo		
MODELLER	P	http://salilab.org/modeller
Avaliação e validação dos modelos		
Procheck	P	http://www.biochem.ucl.ac.uk/~roman/procheck/procheck.html
Verify3D	S	http://www.doe-mpi.ucla.edu/Service/Verify_3D
Whatcheck	S	http://www.sander.emb.-heidelberg.de/whatcheck
Links a outros bancos de dados e softwares para referência cruzada		
KEGG	S	http://www.genome.jp/kegg
MetaCyc	S	http://biocyc.org
TBdb	S	http://www.doe-mpi.ucla.edu/TB
TubercuList	S	http://genolist.pasteur.fr/TubercuList
Swiss-Prot	S	http://bo.expasy.org
PDB	S	http://www.rcsb.org/pdb
ProtGif	S	http://www.biocristalografia.df.ibilce.unesp.br/tools
Parmodel	S	http://www.biocristalografia.df.ibilce.unesp.br/tools/parmodel
Downloads		
Seqüência primária em fasta	S	http://www.biocristalografia.df.ibilce.unesp.br/tools
Coordenadas atômicas	S	http://www.biocristalografia.df.ibilce.unesp.br/tools
Imagem 3D da proteína	S	http://www.biocristalografia.df.ibilce.unesp.br/tools
Inputs para o MODELLER	S	http://www.biocristalografia.df.ibilce.unesp.br/tools

Todos os programas e servidores apresentados na tabela acima compõem o DBMODELING, garantindo sua utilidade e sua confiabilidade. A apresentação de resultados, requer a avaliação por diversos programas para fornecer ao usuário um

grau de confiabilidade maior possível quanto a estrutura 3D que está pesquisando, dando-lhe não só os resultados obtidos pelos programas mas o significado sobre a qualidade estrutural do modelo. Alguns softwares são utilizados dinamicamente, gerando resultados imediatos com relação à proteína de interesse, apresentando via *web* relatórios e gráficos das análises dos ambientes químicos das proteínas (WHATCHECK, VERIFY3D), além das análises dos ângulos ϕ e ψ estabelecendo a posição dos ângulos de torção dentro da cadeia protéica e verificando possíveis choques estereoquímicos entre átomos das cadeias laterais (PROCHECK). Todos os *links* no DBMODELING fornecem informações adicionais sobre as seqüências das proteínas, vias metabólicas, anotações funcionais e informações estruturais sobre proteínas já resolvidas de *M. tuberculosis*. Também estão incluídas informações sobre a qualidade da estrutura (Excelente, Bom e Regular) (Tabela 3), que é indispensável para a utilização da estrutura em simulações de *docking* a modelos obtidos por modelagem molecular comparativa, além da identidade residual e o RMSD da geometria ideal obtido pelo programa X-PLOR.

Tabela 3. Qualidade dos modelos estruturais usando as análises do gráfico de Ramachandran

Qualidade estrutural dos modelos	% de resíduos na região mais favorável do gráfico de Ramachandran
Excelente	>95
Bom	90-95
Regular	85-90

* Todos os modelos classificados abaixo de 85% não foram depositados na atual versão deste banco de dados

4. Resultados e Discussão

4.1 Conteúdo de dados no DBMODELING

Com a execução do programa de alinhamento contra a base de dados de seqüências primárias de estruturas resolvidas depositadas no PDB (BERMAN *et al.*, 2000; WESTBROOK *et al.*, 2003), selecionamos todas as proteínas possíveis de serem modeladas por modelagem molecular comparativa, ou seja, todas as proteínas que apresentaram identidade residual acima de 30% entre o *template* e o alvo. A figura 13 representa os dados inseridos no banco de dados de *M. tuberculosis*, que estão presentes na atual configuração do banco de dados. Também é apresentando a quantidade de estruturas e vias metabólicas presentes no banco de dados, além de dados relacionados à exclusão de modelos devido à baixa qualidade estereoquímica. O DBMODELING está aumentando o número de estruturas 3D e de vias metabólicas identificadas, utilizando bancos de dados específicos como PDB (WESTBROOK *et al.*, 2003), KEGG (OGATA *et al.*, 1999) e MetaCyc (KARP *et al.*, 2002), respectivamente. As ferramentas construídas para identificação de *templates* e modelagem serão utilizadas na atualização estrutural para garantir o melhor *template* para cada proteína do banco de dados, além de identificar novas proteínas devido ao grande volume de estruturas depositadas no PDB. O número de estruturas neste banco de dados pode ser alterado freqüentemente pelo aumento do número de estruturas de proteínas de *M. tuberculosis* que são depositadas no PDB, as quais são excluídas do banco no momento da atualização.

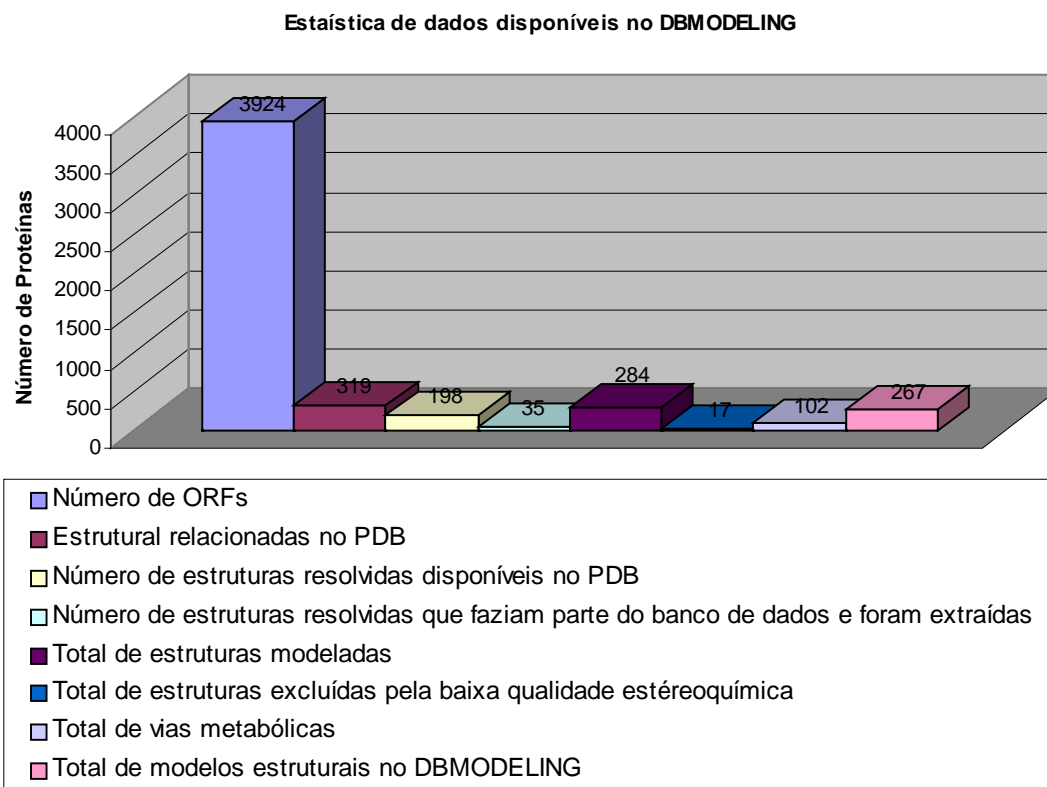


Figura 13. Dados estatísticos sobre a modelagem, mostrando a quantidade de enzimas inseridas no banco de dados, bem como a quantidade de excluídas pela qualidade estereoquímica e as já resolvidas experimentalmente.

Uma estimativa realizada (Figura 14), refletiu o aumento na identificação de proteínas relacionadas depositadas no PDB, as quais possibilitaram a seleção de novos *templates* para novos modelos que serão acrescentados no DBMODELING assim que passarem pelos processos de modelagem, análise, anotação e verificação quanto ao fato de suas estruturas estarem ou não resolvidas e depositadas no PDB.

O objetivo do DBMODELING é fornecer acesso a um conjunto de modelos de *M. tuberculosis* determinados por modelagem comparativa, de forma automatizada. Este banco de dados é atualizado freqüentemente para refletir o

aumento do número de seqüências e estruturas no banco de dados, bem como melhoras nos métodos, utilização de novos *softwares* usados na análise dos modelos, atualização de anotações funcionais e de vias metabólicas e agregar novas ferramentas de visualização e referências cruzadas para outros bancos de dados.

Estimativa para atualização do DBMODELING

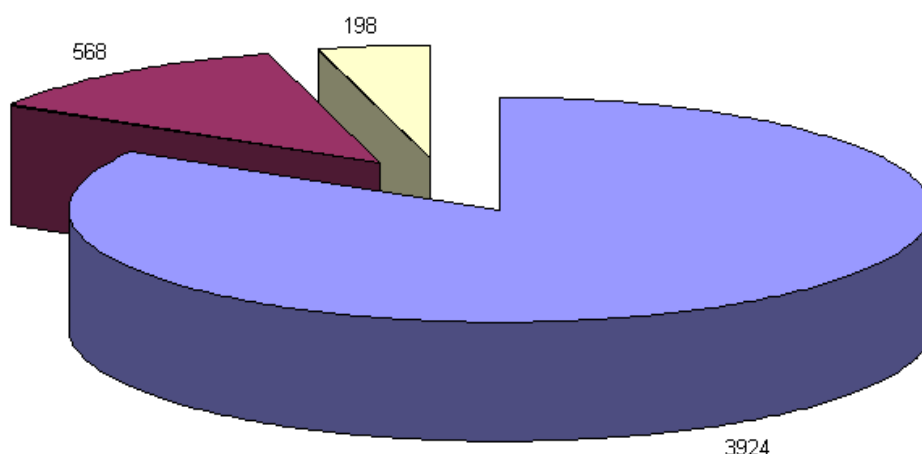


Figura 14. Gráfico representando a estimativa de dados que serão acrescentados ao DBMODELING. Em azul é representado o genoma total do *M. tuberculosis*, em vinho a quantidade de estruturas relacionadas no PDB e em amarelo o número de estruturas resolvidas de *M. tuberculosis* e depositadas no PDB.

4.2 Dados para referência sobre estruturas de *M. tuberculosis*

Há várias proteínas de *M. tuberculosis* as quais tiveram suas estruturas determinadas por difração de raios X ou RMN. As coordenadas atômicas destas proteínas estão disponíveis no *M. tuberculosis Structural Genomics Consortium*

(<http://www.doe-mbi.ucla.edu/TB>) (TERWILLIGER *et al.*, 2003). A estratégia do consórcio é determinar as estruturas 3D de proteínas do *M. tuberculosis* e colocá-las sob domínio público. Há atualmente 198 estruturas de *M. tuberculosis* com coordenadas atômicas depositadas no PDB e no *site* do consórcio para consultas sobre os grupos que as determinaram e publicações. O foco do DBMODELING é disponibilizar a maior quantidade possível de modelos estruturais e suas respectivas vias metabólicas, podendo utilizar os dados das estruturas resolvidas pelo consórcio para atualização e talvez como *templates* para modelagem de novas estruturas. Além disso, alvos atrativos para desenho de drogas envolvem produtos de genes pertencentes a vias metabólicas importantes, tais como a via do ácido chiquímico. Outro importante *link* para informações funcionais relacionadas ao genoma do *M. tuberculosis* é o TubercuList (<http://genolist.pasteur.fr/TubercuList>) (CAMUS *et al.*, 2002), também utilizado para atualizações do DBMODELING, devido à confiabilidade dos dados.

Com a estimativa de aumento do número de estruturas resolvidas de *M. tuberculosis*, há a necessidade de se refazer uma busca no DBMODELING com o objetivo de se identificar modelos tenham sido resolvidos por métodos experimentais e incluí-los em uma tabela com os cálculos dos valores de RMSD C_{α} - C_{α} para cada estrutura, estimando a precisão, validando protocolos utilizados. A tabela 4 representa o RMSD de sobreposição C_{α} - C_{α} para as estruturas contidas anteriormente no DBMODELING, excluídas por terem sido resolvidas experimentalmente.

Tabela 4. Cálculo do RMSD de sobreposição C_{α} - C_{α}

Códigos dos genes de <i>M.t.</i>	Códigos de acesso do PDB	Resolução Cristalográfica (Å)	RMSD C_{α}-C_{α} (Å)
Rv0009	1w74	2.60	0,55
Rv0137c	1nwa	1.50	1,49
Rv0467	1f8m	1.80	0,66
Rv0489	1rii	1.70	1,03
Rv0733	1p4s	RMN	0,77
Rv1379	1w30	1.90	1,17
Rv1389	1s4q	2.16	0,73
Rv1484	1enz	2.70	0,80
Rv1542c	1idr	1.90	1,03
Rv1837c	1n8i	2.10	0,96
Rv1886c	1f0p	1.90	1,15
Rv2002	1nff	1.80	1,12
Rv2150c	1rlu	2.08	0,93
Rv2445c	1k44	2.60	0,77
Rv2537c	1h05	1.50	0,55
Rv2539c	1l4u	1.80	0,66
Rv2697c	1mq7	1.95	1,21
Rv2711	1b1b	2.60	0,98
Rv2763c	1dg8	2.00	0,68
Rv2773c	1c3v	2.39	1,13
Rv2965c	1tfu	1.99	0,93
Rv2995c	1w0d	1.65	0,57
Rv3106	1lqt	1.05	1,57
Rv3247c	1g3u	1.95	0,13
Rv3307	1g2o	1.75	1,55
Rv3465	1upi	1.70	0,15
Rv3608c	1eye	1.70	0,44
Rv3803c	1r88	1.71	0,78
Rv3846	1gn4	2.50	0,89

4.3 Acesso e interface do banco de dados

O DBMODELING fornece uma interface com menus amigáveis, uma vez que todas as informações podem ser impressas em um único passo. Uma pequena representação da estrutura terciária de cada proteína está incluída para se obter uma

primeira impressão do modelo estrutural (Figura 20). Pode ser feito o *download* das coordenadas atômicas no formato PDB e de sua seqüência primária em formato fasta. O banco de dados pode ser pesquisado por enzimas ou vias metabólicas, como palavras chave, selecionando a tabela do banco e como opções “AND”, “OR” e “ONLY ONE KEYWORD”, para refinar a busca (Figura 18). A interface de busca permite combinar todas estas diferentes descrições para pesquisas mais complexas. Para cada modelo, o DBMODELING fornece interfaces *web*, sendo organizadas em forma de tabelas.

Os campos são definidos com *links* para a seqüência alvo e informações complementares no Swiss-Prot (BAIROCH & APWEILER, 1999), para o PDB (WESTBROOK *et al.*, 2003; ABOLA *et al.*, 1987) através do código do *template*, informação estrutural, análises e informações sobre a modelagem, tais como entradas usadas no MODELLER para cada um dos modelos. O DBMODELING inclui *links* para bancos de dados de vias metabólicas como o KEGG (OGATA *et al.*, 1999) e o MetaCyc (KARP *et al.*, 2002).

Todos os arquivos de entrada para a modelagem utilizando o programa MODELLER estão disponíveis na página, mostrando o alinhamento e as porcentagens de identidade e similaridade do alvo com relação ao *template*. Uma imagem inicial da proteína é apresentada e direcionada por um *link*, a um software de geração de imagens animadas executado pela ferramenta PROTGIF, desenvolvida no próprio laboratório, disponibilizando diversos recursos para visualização e animação de proteínas.

A figura 15 nos dá uma visão geral das ferramentas desenvolvidas no Laboratório de Sistemas Biomoleculares (BMSys) (<http://www.biocristalografia.df.ibilce.unesp.br/tools>), objetivando o desenvolvimento biotecnológico para auxiliar os projetos em andamento do grupo de pesquisa e de grupos externos, possuindo acesso público à comunidade científica. O DBMODELING é um banco de dados que compõe um conjunto de ferramentas de acesso aberto de interesse estrutural, dedicado ao genoma do *M. tuberculosis* e de futuros outros genomas que representem alvos potenciais para desenho de drogas baseado em estrutura como: *Xylella fastidiosa*, *Plasmodium falciparum*, etc., os quais poderão ser selecionados na página oficial do DBMODELING (Figura 16), promovendo acesso à pesquisa de genomas de interesse restrito a um ou mais grupos de pesquisa.

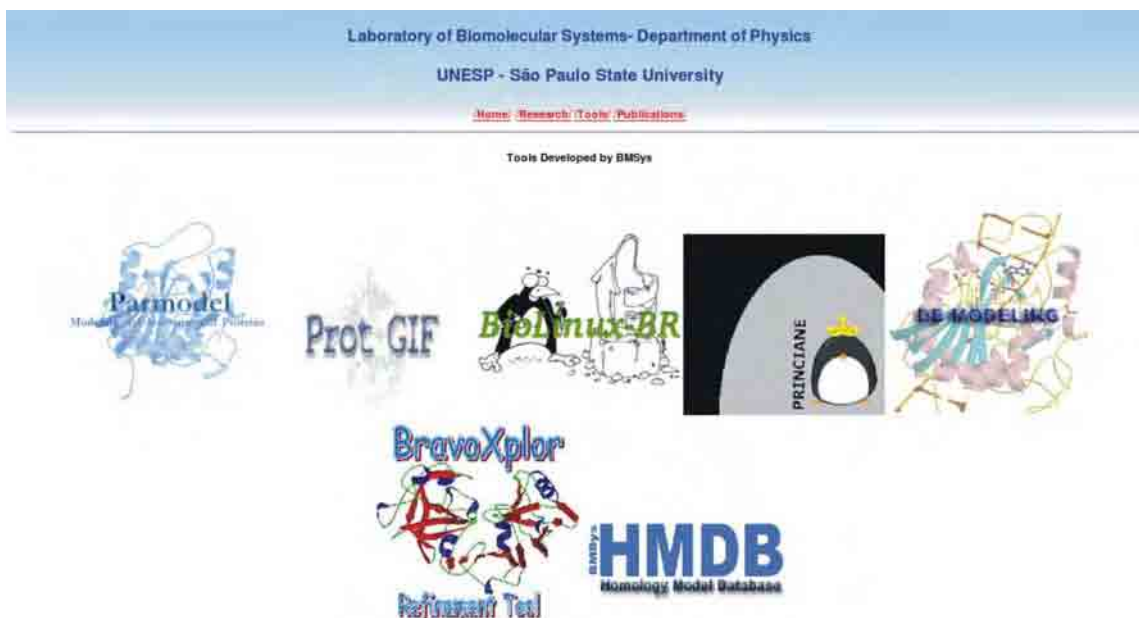


Figura 15. Interface de entrada para as ferramentas do grupo do Laboratório de Sistemas Biomoleculares (BMSys).

Logo após acessar a página de ferramentas e clicar no *link* de entrada para o DBMODELING o usuário irá para a interface oficial do banco de dados (Figura 16), podendo selecionar o organismo para pesquisa ou navegar no site do grupo em busca das publicações e pesquisas em andamento. Esta página dá ao usuário uma visão geral sobre os objetivos deste banco de dados e cita a inserção de futuros genomas completos em seu conteúdo, estabelecendo os mesmos protocolos de construção e análise de modelos citados no fluxograma da figura 6.

Laboratory of Biomolecular Systems- Department of Physics
State University of São Paulo

Sexta-feira, 06 de Maio de 2005, Hora: 9:54:08 AM

Menu

- [Home](#)
- [Group](#)
- [Publications](#)
- [Research](#)
- [Tools](#)
- [Contact](#)

Welcome to DBMODELING! DBMODELING is a database of protein structures predicted by comparative modeling. We focused our study on proteins identified in the genome sequencing projects of infectious agents. Furthermore, since we aimed to use this database for drug design special attention is devoted to potential protein targets. At this moment, it is available only for the *Mycobacterium tuberculosis*, and it will be growing to include other organisms such as *Xylolla fastidiosa*, etc... All structure files are available for downloading. The database is organized to allow identification of the metabolic pathway and include other information about the modeled protein such as, sequence, animated gif for the structure, stereochemical quality and details about the modeling procedure.

Select organism for entry in DBMODELING:

Mycobacterium tuberculosis

Supported by:

References:

- Da Silveira N.J.F., Uchôa H.B., Pereira J.H., Canduri F., Basso L.A., Palma M.S., Santos D.S. and De Azevedo Jr. W.F. Molecular Models of Protein Targets from *Mycobacterium tuberculosis*. *J. Mol. Model.* 11:160-166 (2005).
- Uchôa H.B., Josep G.E., da Silveira N.J.F., Carreira J.C., Canduri F. and de Azevedo Jr. W.F. Parmodel: a web server for automated comparative modeling of proteins. *Biochem. Biophys. Res. Commun.* 325: 1481-1486 (2004).
- Sali, A and Blundell, T.L. Comparative Protein Modelling by Satisfaction of Spatial Restraints. *J. Mol. Biol.* 234, 778-815 (1993).

Figura 16. Interface de entrada para o DBMODELING com a opção para selecionar o organismo de interesse contido no banco de dados e links para as principais publicações envolvidas no desenvolvimento do banco.

Após selecionar o organismo, o usuário terá uma apresentação completa de todas as vias metabólicas identificadas para os modelos gerados, tendo como principal ferramenta agregada ao banco, uma busca recursiva de dados específicos de interesse do pesquisador. Nesta interface (Figura 17), é possível utilizar a pesquisa dentro do banco de dados, fornecendo como entrada o nome da via metabólica ou da enzima de interesse, especificar onde a palavra-chave está inserida (Via metabólica ou Enzima) e selecionar AND, OR ou Only one keyword para se obter melhor desempenho na busca e filtrar os dados recebidos (Figura 18).

The image shows a web interface for the Laboratory of Biomolecular Systems. On the left, there is a sidebar with the title 'Homology models from Mycobacterium tuberculosis' and a 'SEARCH in database' section. Below this, a list of biosynthesis pathways is provided, each with a red link: folic acid biosynthesis, formylTHF biosynthesis, glycine biosynthesis I, serine and glycine biosynthesis, cobalamin biosynthesis, O-antigen biosynthesis, enterobacterial common antigen, dTDP-rhamnose biosynthesis, de novo biosynthesis of purine, his+purine+pyrimidine biosynthesis, homoserine methionine biosynthesis, sulfur amino acid biosynthesis, biosynthesis of proto- and siroh, fatty acid biosynthesis -- initial, peptidoglycan biosynthesis, UDP-N-acetylglucosamine biosynthesis, peptidoglycan and lipid A precursors, methionine and methyl-donor-r, and methionine biosynthesis I.

The main content area features a header for 'Laboratory of Biomolecular Systems- Department of Physics, State University of São Paulo'. Below this is a navigation bar with buttons for HOME, RESEARCH, TOOLS, PUBLICATIONS, and SEQ. MODELING. The main title is 'Structural Bioinformatics Applied to the Study of Protein Target from Mycobacterium tuberculosis Genome.' Below the title is a circular map of the chromosome of M. tuberculosis H37Rv. The map consists of several concentric rings: the outermost ring shows the scale in Mb with 0 at the origin of replication; the first ring shows stable RNA genes (tRNAs in blue, others in pink); the second ring shows the coding sequence by strand (clockwise in dark green, anticlockwise in light green); the third ring depicts repetitive DNA (insertion sequences in orange, 13E12 REP family in dark pink, prophage in blue); the fourth ring shows the positions of the PPE family members (green); the fifth ring shows the PE family members (purple, excluding PGRS); and the sixth ring shows the positions of the PGRS sequences (dark red). A central histogram represents G + C content, with < 65% G + C in yellow and > 65% G + C in red. The figure was generated with software from DNASTAR.

Figura 17. Visualização da interface após a seleção do organismo, relacionando todas as vias metabólicas identificadas para os modelos gerados com links para suas respectivas enzimas.

O resultado da busca será apresentado ao lado das vias metabólicas, citando a via relacionada à enzima consultada. A apresentação deve ser feita sem redundâncias, e estabelecendo *links* nas enzimas direcionados para os dados estruturais disponíveis (Figura 19).

Laboratory of Biomolecular Systems

Homology models from *Mycobacterium tuberculosis*

SEARCH in database

Biosynthesis pathway

- [folic acid biosynthesis](#)
- [formyl-THF biosynthesis](#)
- [glycine biosynthesis I](#)
- [serine and glycine biosynthesis](#)
- [cobalamin biosynthesis](#)
- [O-antigen biosynthesis](#)
- [enterobacterial common antigen](#)
- [dTDP-rhamnose biosynthesis](#)
- [de novo biosynthesis of purine](#)
- [his+purine+pyrimidine biosynthesis](#)
- [homoserine methionine biosynthesis](#)
- [sulfur amino acid biosynthesis](#)
- [biosynthesis of proto- and siro-](#)
- [fatty acid biosynthesis -- initial](#)
- [peptidoglycan biosynthesis](#)
- [UDP-N-acetylglucosamine biosynthesis](#)
- [peptidoglycan and lipid A precursors](#)
- [methionine and methyl-donor-*S*-adenosylmethionine biosynthesis I](#)

DB MODELING

Laboratory of Biomolecular Systems - Department of Physics
State University of São Paulo

HOME RESEARCH TOOLS PUBLICATIONS DB MODELING

Search for specific pathway or enzyme from *Mycobacterium tuberculosis*.

Enter with your keyword for search in DBMODELING

Table select
 Pathway Enzyme

Options for restrict your search
 AND OR Only one keyword

Go: Clear Help

Figura 18. Interface de busca por uma via metabólica ou enzima específica, com opções de refinamento da pesquisa a ser feita.

Laboratory of Biomolecular Systems

Homology models from *Mycobacterium tuberculosis*

SEARCH in database

Biosynthesis pathway

- [folic acid biosynthesis](#)
- [formyl-THF biosynthesis](#)
- [glycine biosynthesis I](#)
- [serine and glycine biosynthesis](#)
- [cobalamin biosynthesis](#)
- [O-antigen biosynthesis](#)
- [enterobacterial common antigen](#)
- [dTDP-rhamnose biosynthesis](#)
- [de novo biosynthesis of purine](#)
- [his+purine+pyrimidine biosynt](#)
- [homoserine methionine biosynt](#)
- [sulfur amino acid biosynthesis](#)
- [biosynthesis of proto- and siroh](#)
- [fatty acid biosynthesis -- initial](#)
- [peptidoglycan biosynthesis](#)
- [UDP-N-acetylglucosamine bios](#)
- [peptidoglycan and lipid A precu](#)
- [methionine and methyl-donor-r](#)
- [methionine biosynthesis I](#)

DB MODELING

Laboratory of Biomolecular Systems - Department of Physics
State University of São Paulo

HOME RESEARCH TOOLS PUBLICATIONS BIO MODELING

Results for query entries from *Mycobacterium tuberculosis* database.

phenylalanine, tyrosine and tryptophan biosynthesis, complete

- [3-dehydroquinate synthase](#)

chorismate biosynthesis

- [3-dehydroquinate synthase](#)

pathways of chorismate


- [3-dehydroquinate synthase](#)

Figura 19. Links direcionando as enzimas de interesse para as informações estruturais.

A figura 20 apresenta informações sobre a enzima a qual o usuário selecionou clicando no *link* com o nome da enzima. Todos os dados de análise estrutural estão disponíveis neste *site*, tais como *download* da seqüência primária da enzima e das coordenadas atômicas do modelo, resultados dinâmicos apresentados em tempo real na página utilizando os softwares VERIFY3D, PROCHECK E WHATCHECK, entradas de dados para modelagem, resultados das análises para todos os modelos gerados utilizando um robusto protocolo de modelagem (Figura 21), o método utilizado para alinhamento e busca por *templates*, além do mapa de Ramachandran determinado usando o PROCHECK e dados de RMSD da geometria ideal gerados com o programa X-PLOR, extraindo parâmetros de comparação estrutural

(comprimento de ligação, ângulos de ligação, ângulos diedros e ângulos impróprios).

Links para outros bancos como KEGG, MetaCyc, TBdb, TubercuList, Swiss-Prot e PDB também estão disponíveis para anotação, verificação estrutural e atualização, bem como identidade e similaridade entre as seqüências da proteína alvo e do *template*, escore do alinhamento global utilizando programação dinâmica, o nome da enzima, a reação catalisada pela enzima e a cadeia polipeptídica com a qual o modelo foi gerado.



>RV2538C
 MTDI GAPVTVQVAVDPPYPVVI GTGLLDELEDLLADRHKVAVVHQPLAETAEETRKPLA
 GKGVDAHRIEIPDAEAGKDLVVGFIWEVLGRI GIGRKDALVSLGGGAATDVAGFAAATW
 LRGVSI VHLPTLLGMVDAAVGGKTGINTDAGKNLVGAFHQPLAVLVDLATLQTLPRDEM
 ICGMAEVVKAGFIADPVI LDLIEADPAALDPAGDVLPELIRRAITVKAEVVADEKESE
 LREILNYGHTLGHAIERRERYFWRHGAAVSVGLVFAAELARLAGRLDDATAQRHRTILSS
 LGLPVSYPDALPQLLEIMAGDKKTRAGVLRVVL DGLAKPGRMVGPDPGLLVTAYAGVC
 AP

[Download fasta format \(RV2538C.fasta\)](#)
[Download PDB coordinates \(RV2538C.pdb\)](#)

Analysis report for wash model from *Mycobacterium tuberculosis*

[Procheck](#) | [Whatcheck](#) | [3DPlot](#) | [Anasoft](#) | [Alignment method](#) | [Modeler](#)

Links to related databases

[KEGG](#) | [MetaCyc](#) | [TBdb](#) | [TubercuList](#)

Structural model quality	Good
Gene code	RV2538C
Gene name	aroB
Gene product	3-dehydroquinate synthase
Template	1dqq
Chain	A
Length	426 aa
Identity with template	30.0%
Similarity with template	43.0%
Score (for the alignment)	432.5
Catalysed reaction	3-deoxy-D-ambino-heptulosonate-7-phosphate + phosphate → 3-dehydroquinate
Blast comment	(MTCY159.18), len: 362, aroB, almost identical to AROB_MYCTU_P38919

Figura 20. Dados estruturais da enzima selecionada para pesquisa.

Information About Models Evaluation

PROCHECK is used to check the model's stereochemistry. Before doing any external evaluation of the model, one should check the log file from the modeling run for errors and restraint violations.

The stereochemistry of the model can be checked by program PROCHECK. The output of PROCHECK is a series of POSTSCRIPT files with evaluations of different aspects of the model's stereochemistry. One of the most important charts is the Ramachandran plot which points out those residues that have anomalous combinations of ϕ and ψ angles. A few deviations of this type are usual even in experimentally determined structures. The table below describes the analysis for 1,000 models generated with MODELLER and analysed with PROCHECK for each enzyme from *Mycobacterium tuberculosis*.

Evaluated plots	Best	Good	Bad	Excluded	Average G-factor	Energy MODELLER
Rv2538c_0571.out	(93.4)%	(5.3)%	(0.7)%	(0.7)%	(-0.08)	(1886.55410)
Rv2538c_0976.out	(93.0)%	(5.6)%	(0.7)%	(0.7)%	(-0.09)	(1884.32700)
Rv2538c_0971.out	(93.0)%	(6.0)%	(0.3)%	(0.7)%	(-0.10)	(1907.50600)
Rv2538c_0924.out	(93.0)%	(5.0)%	(1.3)%	(0.7)%	(-0.10)	(1942.04300)
Rv2538c_0882.out	(93.0)%	(5.6)%	(0.7)%	(0.7)%	(-0.08)	(1760.08410)
Rv2538c_0862.out	(93.0)%	(6.0)%	(0.7)%	(0.3)%	(-0.12)	(1905.79540)
Rv2538c_0835.out	(93.0)%	(6.3)%	(0.3)%	(0.3)%	(-0.07)	(1765.15440)
Rv2538c_0761.out	(93.0)%	(6.0)%	(0.3)%	(0.7)%	(-0.07)	(1771.33520)
Rv2538c_0717.out	(93.0)%	(5.6)%	(1.0)%	(0.3)%	(-0.07)	(1782.80200)
Rv2538c_0640.out	(93.0)%	(5.3)%	(1.0)%	(0.7)%	(-0.11)	(1973.12400)
Rv2538c_0634.out	(93.0)%	(5.3)%	(1.0)%	(0.7)%	(-0.10)	(1856.45500)
Rv2538c_0627.out	(93.0)%	(6.0)%	(0.3)%	(0.7)%	(-0.08)	(1784.69400)
Rv2538c_0537.out	(93.0)%	(6.0)%	(0.3)%	(0.7)%	(-0.10)	(1821.13000)
Rv2538c_0531.out	(93.0)%	(5.3)%	(1.0)%	(0.7)%	(-0.12)	(2045.90740)
Rv2538c_0467.out	(93.0)%	(5.6)%	(1.0)%	(0.3)%	(-0.10)	(1996.16710)
Rv2538c_0350.out	(93.0)%	(6.0)%	(0.7)%	(0.3)%	(-0.08)	(1832.53030)
Rv2538c_0337.out	(93.0)%	(6.0)%	(0.7)%	(0.3)%	(-0.11)	(1899.11700)
Rv2538c_0033.out	(93.0)%	(6.0)%	(0.7)%	(0.3)%	(-0.09)	(1754.69100)
Rv2538c_0315.out	(93.0)%	(5.3)%	(0.7)%	(1.0)%	(-0.10)	(1930.67110)
Rv2538c_0246.out	(93.0)%	(6.0)%	(0.7)%	(0.3)%	(-0.10)	(1806.58700)
Rv2538c_0243.out	(93.0)%	(6.0)%	(0.7)%	(0.3)%	(-0.08)	(1751.32110)
Rv2538c_0202.out	(93.0)%	(6.0)%	(0.7)%	(0.3)%	(-0.07)	(1797.84940)
Rv2538c_0144.out	(93.0)%	(6.0)%	(0.3)%	(0.7)%	(-0.10)	(1836.37800)
Rv2538c_0133.out	(93.0)%	(5.6)%	(0.7)%	(0.7)%	(-0.07)	(1766.59800)
Rv2538c_0980.out	(92.7)%	(6.3)%	(0.7)%	(0.3)%	(-0.11)	(1869.71700)
Rv2538c_0098.out	(92.7)%	(6.0)%	(0.7)%	(0.7)%	(-0.09)	(1916.45400)
Rv2538c_0967.out	(92.7)%	(6.0)%	(0.7)%	(0.7)%	(-0.10)	(1916.55100)
Rv2538c_0958.out	(92.7)%	(6.3)%	(0.7)%	(0.3)%	(-0.08)	(1836.62100)

Figura 21. Análise dos resultados de uma proteína alvo para os 1000 modelos gerados, selecionando aquele que obteve 85% ou mais de resíduos na região mais favorável do gráfico de Ramachandran. A sexta e a sétima coluna representa o Fator-G e a função objetiva obtidos da modelagem, respectivamente.

4.4 Precisão dos modelos gerados

Para facilitar a avaliação da qualidade das estruturas foi criado um simples esquema de classificação para os modelos depositados, como indicado anteriormente na tabela 3. A precisão da modelagem comparativa de proteínas está relacionada à porcentagem de identidade na qual o modelo é baseado, estabelecendo uma correlação entre as similaridades estrutural e seqüencial das duas proteínas (MARTIRENOM *et al.*, 2000; SÁNCHEZ & ŠALI, 1998; KOEHL & LEVITT, 1999). Todos os modelos no banco de dados foram construídos usando alinhamentos que apresentaram uma identidade maior que 30%, a qual gerou modelos com média e alta precisão.

Como descrito anteriormente, o principal objetivo deste banco de dados é fornecer modelos estruturais para serem usados em simulações de *docking* e desenho de drogas baseado em estruturas. Sendo a precisão dos modelos altamente dependente da identidade entre as seqüências do alvo e do *template*, é recomendado fortemente que qualquer simulação de *docking* seja focada em modelos estruturais os quais apresentarem maior identidade possível e forem classificados como sendo de excelente qualidade estereoquímica. A figura 4 citada anteriormente, descreve uma escala para utilização de modelos comparativos de acordo com sua identidade com o *template* e a melhora gradativa no r.m.s.d. C_{α} - C_{α} entre a seqüência alvo e o *template* de acordo com o aumento da identidade. O histograma da figura 22 mostra a freqüência dos dados obtidos das estruturas modeladas com relação à região mais favorável do gráfico de Ramachandran, com base nos intervalos da tabela 3.

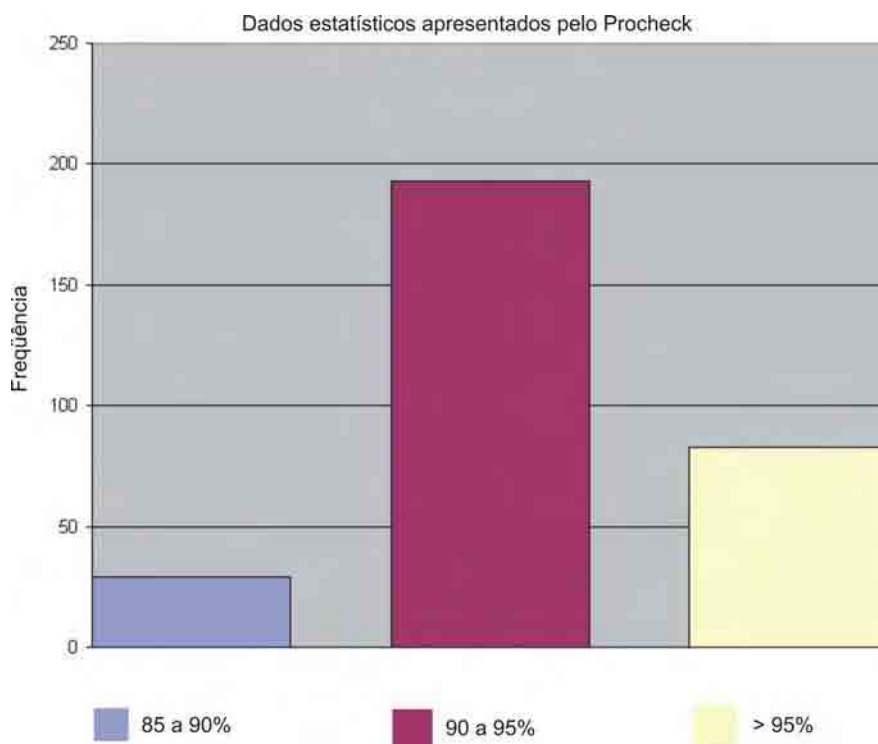


Figura 22. Histograma representando as regiões mais favoráveis do gráfico de Ramachandran para todas as estruturas do banco de dados geradas com o programa Procheck.

A tabela 5 nos dá uma visão estatística geral sobre a qualidade estrutural dos modelos presentes no banco de dados com relação à qualidade estereoquímica, apresentando os intervalos de qualidade, as frequências absoluta e relativa, a porcentagem de ocorrência e cada um dos intervalos e a média para todas as estruturas.

Tabela 5. Dados estatísticos para a região mais favorável do gráfico de Ramachandran.

Intervalos (%)	Frequência Absoluta	Frequência Relativa	Porcentagem (%)	Porcentagem Média dos Valores (%)
Excelente (> 95)	83	0,27	27,0	96,6
Bom (90-95)	193	0,63	63,0	92,9
Regular (85-90)	29	0,10	10,0	88,6
Total	305	1,00	100,0	93,5

Dados estatísticos gerados a partir do RMSD $C_{\alpha} - C_{\alpha}$ de sobreposição das estruturas modeladas e que posteriormente foram resolvidas experimentalmente por cristalografia de raios X ou RMN, mostraram uma alta concordância entre dados teóricos e experimentais. Tal concordância é expressa pelo gráfico da figura 23, mostrando a frequência de medidas de RMSD para 29 estruturas presentes no DBMODELING que foram posteriormente resolvidas, cuja média é de 0,88 Å.

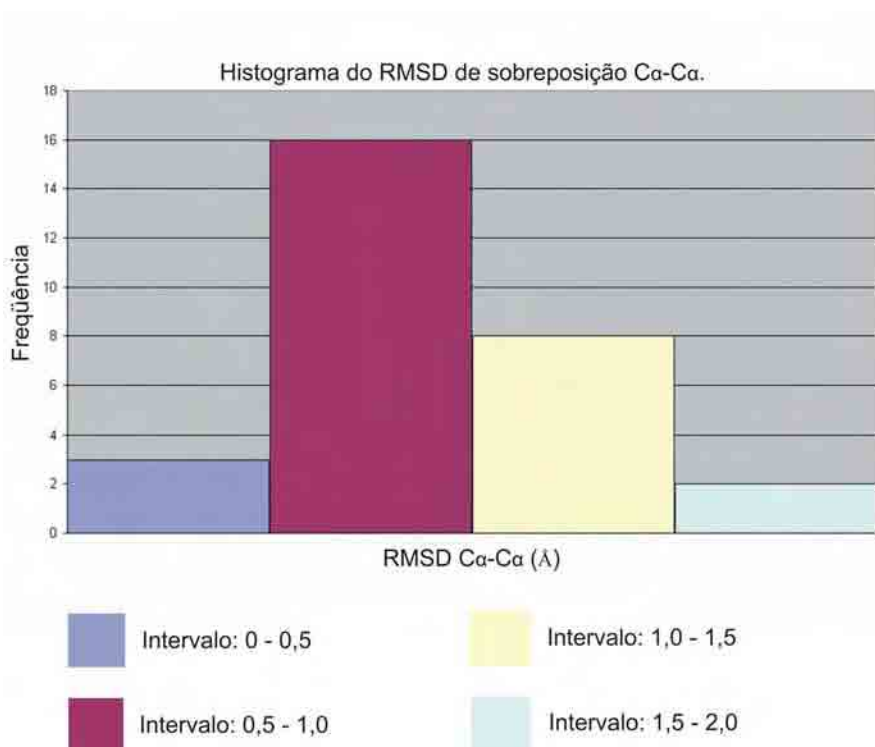


Figura 23. Histograma mostrando a frequência de proteínas com relação aos intervalos dos valores de RMSD de sobreposição $C_{\alpha} - C_{\alpha}$.

Outro dado de extrema importância que corrobora a precisão dos modelos presentes no banco de dados é a dispersão dos dados estimados pelo programa Procheck (porcentagem total de resíduos na região mais favorável no gráfico de Ramachandran) para cada estrutura com relação aos dados de RMSD da geometria

ideal (Ângulos de ligação) obtidos pelo programa X-PLOR. Os dados apresentados pelo gráfico de dispersão da figura 24 mostram uma evidência que relaciona a alta qualidade estereoquímica aos baixos valores dos ângulos de ligação com relação à geometria ideal. Isto é observado pela inclinação da reta de tendência de dispersão dos dados, ratificando os métodos de análise e o protocolo utilizado na modelagem das proteínas no DBMODELING.

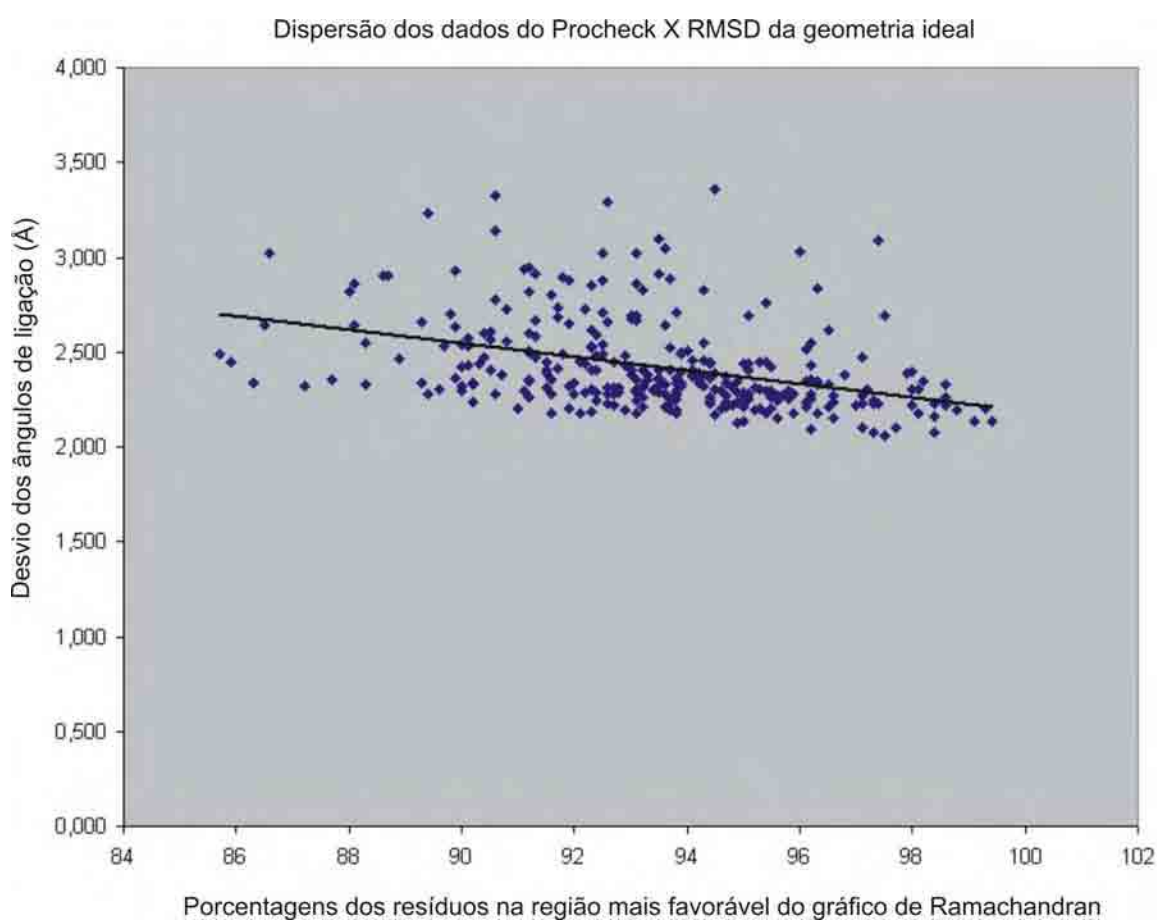


Figura 24. Gráfico de dispersão dos dados do Procheck e RMSD da geometria ideal com relação a todas as estruturas de proteínas contidas no DBMODELING. A inclinação da reta nos remete a crer em uma tendência importante, que é o decréscimo dos desvios dos ângulos de ligação da cadeia principal com o aumento do número de resíduos na região mais favorável do gráfico de Ramachandran, mostrando a importância e precisão dos métodos utilizados.

4.5 Análises realizadas para uma estrutura contida no DBMODELING

Para ilustrar a aplicação do DBMODELING no estudo estrutural de proteínas alvo para desenho de drogas antituberculose, discutiremos o modelo da glucose-1-fosfato timidilil-transferase de *M. tuberculosis* (*MtRmlA*), a qual é a primeira enzima na via biossintética da dTDP-L-rhamnose.

Após a seleção da enzima *MtRmlA*, é apresentado na página do banco de dados a imagem da proteína e uma relação de dados de análises mostrados nas tabelas 6 e 7. O primeiro passo é verificar a precisão do modelo selecionado no DBMODELING, observando a identidade, a qualidade estereoquímica, o perfil 3D da enzima, o RMSD $C_{\alpha} - C_{\alpha}$ e o RMSD da geometria ideal.

A síntese da desoxi-timidina di-fosfato (dTDP)-L-rhamnose, um importante componente da parede celular de muitos microorganismos, é um alvo para intervenção terapêutica. A *RmlA* é inibida pela dTDP-L-rhamnose, regulando a produção de L-rhamnose em bactérias (BLANKENFELDT *et al.*, 2000). Devido a sua importância, a *RmlA* é um alvo potencial para drogas principalmente por ser uma proteína envolvida na síntese da parede celular de micobactérias, e por seu produto enzimático, dTDP-Glc, não ser encontrado em humanos (MA *et al.*, 1997).

A L-rhamnose é derivada de uma base de glucose em quatro passos, iniciando com a glucose-1-fosfato (G-1-P) e desoxi-timidina tri-fosfato (dTTP), resultando na desoxi-timidina di-fosfato (dTDP)-L-rhamnose. As enzimas que catalisam a conversão são glucose-1-fosfato timidilil-transferase (*RmlA*, E.C. 2.7.7.24), dTDP-

D-glucose 4,6-desidratase (RmlB), dTDP-6-desoxi-D-xylo-4-hexulose 3,5-epimerase (RmlC) e dTDP-6-desoxi-L-lyxo-4-hexulose redutase (RmlD).

Na reação catalisada pela RmlA, a enzima combina dTTP com G-1-P para produzir dTDP-D-glucose e pirofosfato (Figura 25). A reação é efetivamente transferir desoxi-timidina mono-fosfato (dTMP) para G-1-P. Pelo fato de não estar presente no organismo humano, a RmlA torna-se um candidato altamente atrativo na busca de inibidores contra a biossíntese da L-rhamnose.

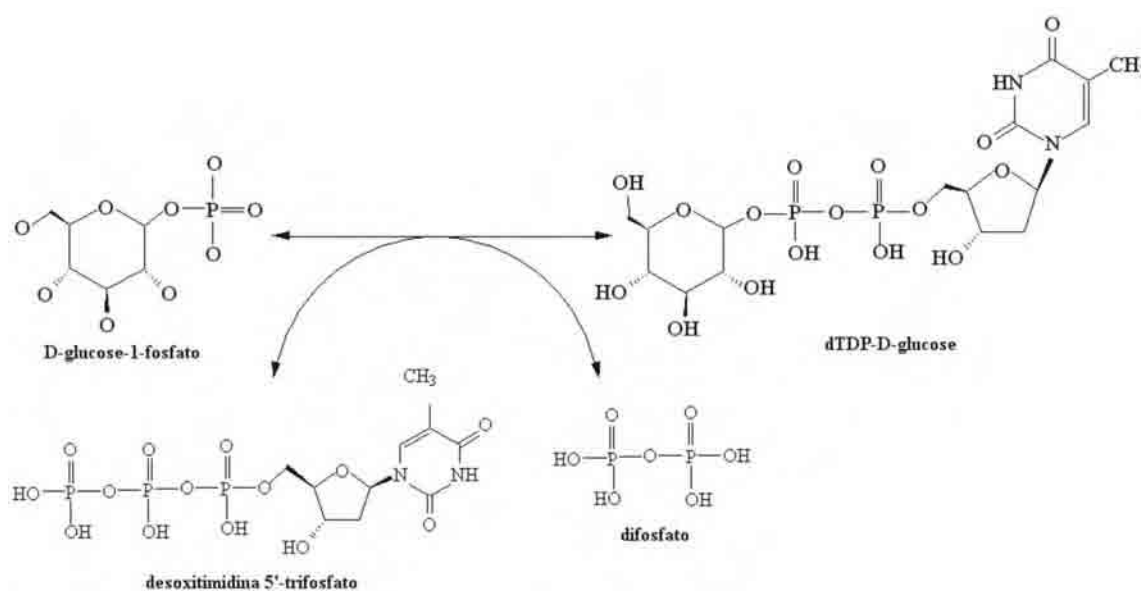


Figura 25. Reação catalisada pela Glucose-1-fosfato timidilil-transferase (RmlA)

4.5.1 Alinhamento das seqüências primárias e qualidade dos modelos

O alinhamento da *MtRmlA* foi realizado contra a cadeia A da enzima RmlA isolada de *Pseudomonas aeruginosa* (*PaRmlA*) selecionada como *template* (Código de acesso no PDB: 1FXO) (BLANKENFELDT *et al.*, 2000) e resolvida a 1,66 Å de resolução. O alinhamento entre as seqüências primárias da *MtRmlA* e da *PaRmlA*

(Figura 26) apresenta 60,1% de identidade e 74,7% de similaridade, indicando que a enzima *PaRmlA* é um *template* que irá gerar modelos de alta precisão. A qualidade estereoquímica do modelo da *MtRmlA* foi analisada pelo programa PROCHECK (Tabela 6) e apresentados na página de informações do DBMODELING mostrado na figura 27, juntamente com o gráfico de Ramachandran, dados como RMSD da geometria ideal e a média do G-factor. A identidade de 60,1% pertence a uma faixa de alta precisão para modelos comparativos, e em conjunto com a excelente qualidade estereoquímica e o baixo RMSD $C_{\alpha} - C_{\alpha}$, com o valor de 0,151 Å, torna o modelo ideal para utilização em simulações de *docking*.

```

           10      20      30      40      50      60
1fxo      KRKGIILAGGSGTRLHPATLAISKQLLPVYDKPMIYYPLSTLMLAGIREILIIISTPQDTPRFQQLLGD
Rv0334    -MRGIILAGGSGTRLYPITMGISKQLLPVYDKPMIYYPLTTLMAGIRDIQLITTPHDAPGFHRLLLGD
           ***** * * ***** * * * * * * * * * * * * * * * * * * * * * * * *
           70      80      90      100     110     120     130
1fxo      GSNWGLDLQYAVQPSPDGLAQAFDIGESFIGNDLSALVLDNLVYGHDFHELLGSASQRQTGASVFAY
Rv0334    GAHLGVNISYATQDQPDGLAQAFVIGANHIGADSVLVLDNIFYGPGGLGTSLKRFSI-SGGAIFAY
           * *  * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
           140     150     160     170     180     190     200
1fxo      HVLDPERYGVVEFDQGGKAISLEEKPLEPKSNYAVTGLYFYDQQVVDIARDLKPSPRGELEITDVMRA
Rv0334    WVANPSAYGVVEFGAEGMALSLEEKPVTPKSNYAVPGLYFYDNDVIEIARGLKKSARGEYEITEVNQV
           * *  * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
           210     220     230     240     250     260     270
1fxo      YLERGQLSVEIMGRGYAWLDTGTHDSLLEAGQFIATLENRQGLKVACPEEIAVRQKWIDAAQLEKLA
Rv0334    YLNQGRlavevLARGTAWLDTGTFDSLDDAADFVRTLERRQGLKVSIPPEEVAWRNGWIDDEQLVQRAR
           ** * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
           280     290
1fxo      PLAKNGYGQYLKRLLTETVY
Rv0334    ALVKSGYGNYLLELLERN--
           * *  * * * * * * * *

```

Figura 26. Alinhamento das seqüências de aminoácidos da *MtRmlA* e da *PaRmlA*. Marcados com asterisco estão os resíduos idênticos, apresentando apenas quatro gaps em toda a extensão da seqüência e uma identidade de 60,1% utilizando o algoritmo de programação dinâmica para alinhamento de seqüências proposto por Needleman e Wunsch em 1970 (Needleman & Wunsch, 1970).

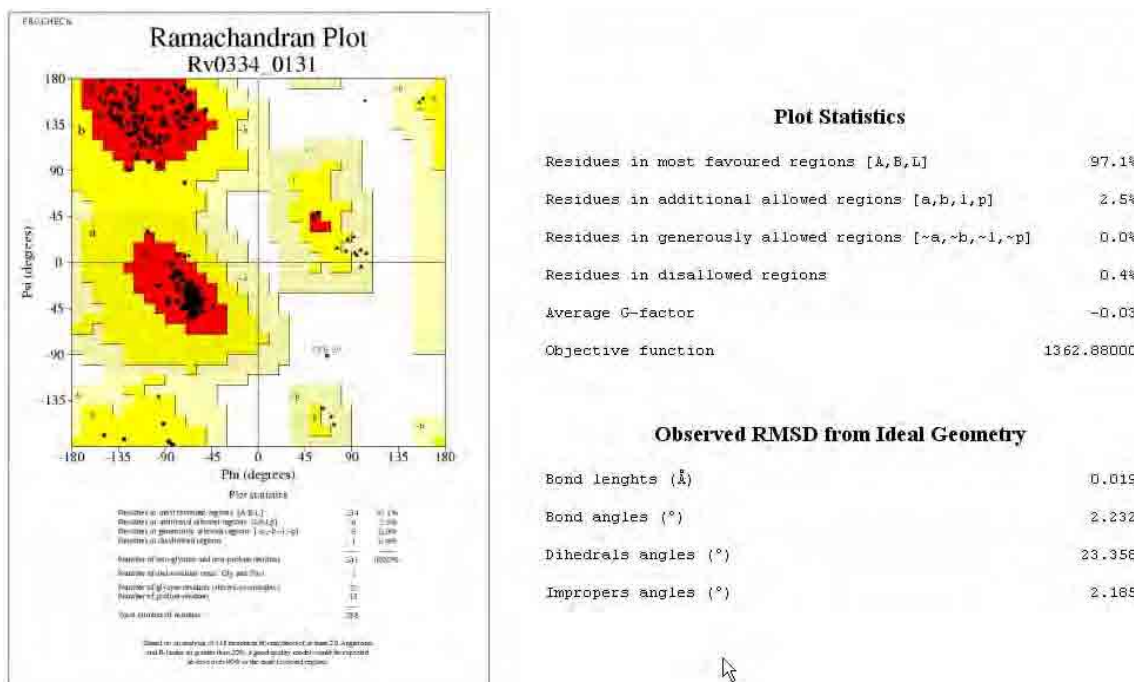


Figura 27. Gráfico de Ramachandran da modelagem da enzima *MtRmlA* da via metabólica da biosíntese do corismato. Na região mais favorável em vermelho concentram-se 97,3% dos resíduos e na região favorável apenas 2,7%, não apresentando resíduos nas regiões desfavoráveis.

O modelo da estrutura de *MtRmlA* (Figura 28A) foi avaliado ainda pelo VERIFY 3D para verificar a confiabilidade na compatibilidade seqüência/estrutura e estes valores indicam que a estrutura do modelo final tem compatibilidade entre a seqüência primária do modelo e a estrutura 3D construída, pois os valores gerados para o modelo final ficaram acima do limite de 0,45 S_{Ideal} (Tabela 6). A figura 29 descreve os resultados do VERIFY 3D dinamicamente na interface *web* do banco de dados para cada enzima solicitada.

As estruturas do modelo e do *template* foram sobrepostas, considerando somente a sobreposição $C_{\alpha} - C_{\alpha}$ com o auxílio do programa LSQKAB do pacote CCP4 (COLLABORATIVE COMPUTATIONAL PROJECT N° 4, 1994). Com a

sobreposição é possível verificar se há possíveis alterações no posicionamento das cadeias laterais dos resíduos ou se há alguma alteração na conformação da estrutura da proteína, além de determinar o RMSD da sobreposição (Tabela 7).

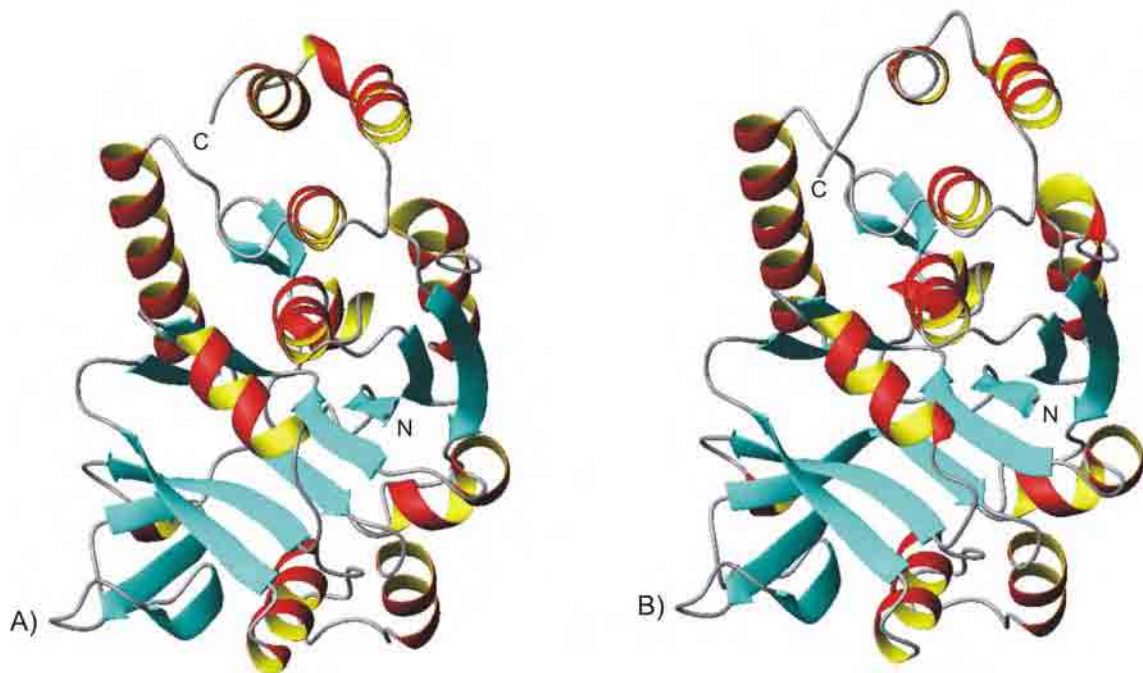


Figura 28. A) estrutura 3D da enzima Glucose-1-fosfato timidilil-transferase de *M. tuberculosis* (Rv0334) e da estrutura 3D da enzima Glucose-1-fosfato timidilil-transferase de *P. aeruginosa* em B), pertencentes à via biossintética do dTDP-rhamnose.

Tabela 6. Análises da estrutura do *template* e do modelo.

Enzima	3D Profile ^a		PROCHECK				RMSD da geometria ideal		
	Score Total	Score Ideal	Score S _{ideal}	Região mais favorável (%)	Região permitida (%)	Região generosamente permitida (%)	Região não permitida (%)	Comprimento de Ligação (Å)	Ângulos de Ligação (°)
IFXO_A	123,34	133,24	0,93 IS	92,3	7,3	0,0	0,4	0,021	1,843
RV0334	133,08	131,40	1,01 IS	97,1	2,5	0,0	0,4	0,019	2,232

^aScore Total: é a soma dos scores 3D-ID (preferências estatísticas) de cada resíduo presente na proteína.

Score Ideal: $S_{ideal} = \exp(-0.83 + 1.008 \times \ln(L))$; onde L é o número de aminoácidos.

Score S_{ideal}: é a compatibilidade da sequência com sua estrutura 3D. Este score é obtido pela divisão do

Score Total pelo Score Ideal (Score Total / Score Ideal). Score S_{ideal} deve estar acima de 0.45S_{ideal}.

Tabela 7. Dados gerais apresentados no banco de dados sobre a o modelo e o *template* selecionado para análises.

Enzima	Template (código de acesso no PDB)	Identidade (%)	Similaridade (%)	Nº de aminoácidos	Gene	G-Factor ^a		
						Ângulos de Torção	Geometria Covalente	Total
RV0334	1FXO_A	60,1	74,7	288 (292)	rfbA	0,06 (-0,01)	-0,19 (0,17)	-0,03 (0,06)

^aIdealmente, os scores devem estar acima de - 0.5. Os valores abaixo de -1.0 devem ser investigados.

* Entre parêntesis estão os valores obtidos para a estrutura 3D do *template*.

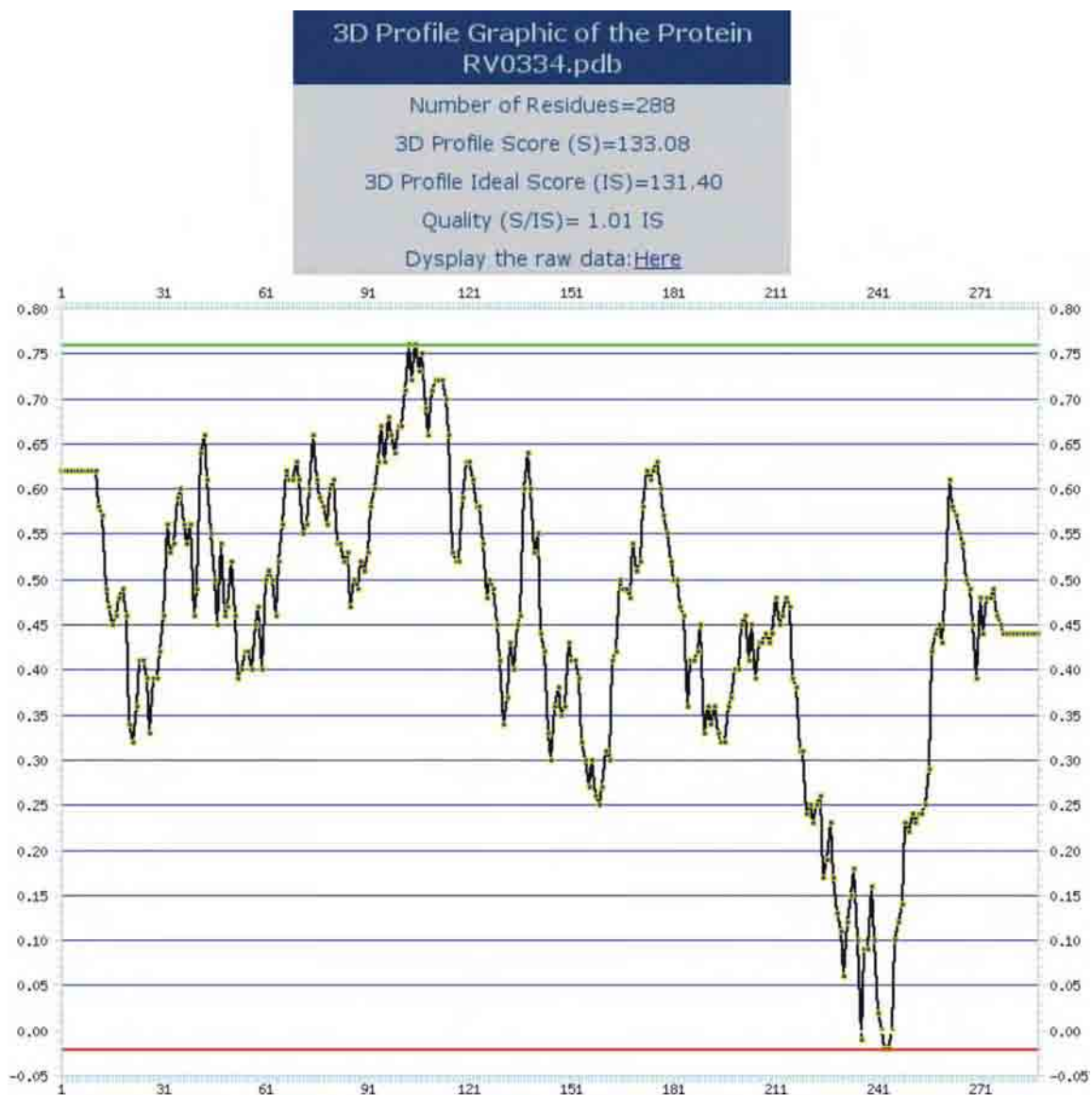


Figura 29. Gráfico gerado pelo programa VERIFY 3D da enzima *MtRmlA*, onde o eixo horizontal corresponde ao número de aminoácidos da seqüência e o eixo vertical ao escore total 3D-1D para cada aminoácido.

Após a sobreposição C_{α} - C_{α} , o programa também gera um arquivo de coordenadas atômicas da sobreposição. Este arquivo de coordenadas atômicas pode ser utilizado como entrada no programa MolMol (KORADI *et al.*, 1996) em conjunto com o modelo obtido para gerar imagens, visualizar e analisar estas estruturas com base em sua sobreposição (Figura 30).

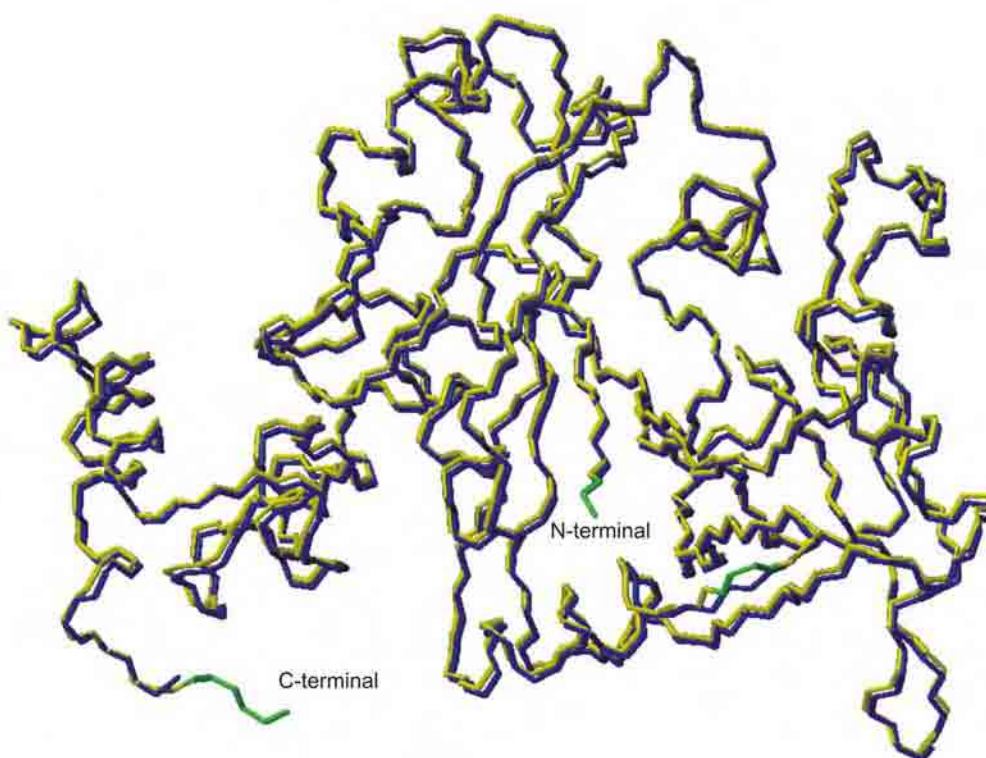


Figura 30. Sobreposição do modelo da *MtRmlA* (Rv0334) em azul com o *template PaRmlA* (código de acesso do PDB: 1FXO_A) em amarelo, mostrando as diferenças conformacionais entre as estruturas e a localização do N-terminal e C-terminal. As faixas de *gaps* apresentadas no alinhamento estão representadas em verde.

Contudo, a utilização de modelos estruturais de proteínas em simulações de *docking* contra possíveis inibidores, torna-se mais confiável com análises do ambiente químico das estruturas e conservação das características estruturais observadas com as sobreposições C_{α} - C_{α} e da geometria ideal. Um banco de dados de estruturas de proteínas obtidas por modelagem molecular comparativa de alta qualidade, utilizando vias metabólicas como objeto de seleção de proteínas alvo para desenho de drogas, é sem dúvida um grande passo no desenvolvimento de novos fármacos e terapias com drogas menos agressivas à pacientes acometidos com

patógenos como o *M. tuberculosis*, o qual estabelece seletividade a determinadas drogas.

5. Conclusões

A modelagem molecular comparativa de proteínas em larga escala, na qual bancos de dados inteiros de seqüências ou genomas completos são usados como entrada em algoritmos de modelagem automatizada, tem sido utilizado por diversos grupos de pesquisa em bioinformática estrutural. Pelo uso de poderosos sistemas de computadores com múltiplos processadores, estes esforços têm permitido a criação de grandes bancos de dados de modelos de proteínas determinadas por modelagem molecular comparativa. O DBMODELING tem acesso público aos dados estruturais e às publicações. Este estudo de proteínas alvo do genoma do *M. tuberculosis* enfatiza que a modelagem molecular é uma ferramenta de uso intensivo em biologia estrutural e que ela pode ser muito valiosa na anotação de seqüências de genomas e contribuir para a genômica estrutural e funcional. Além do mais, a sobreposição de modelos estruturais presentes no DBMODELING mostrou estar de acordo com as estruturas cristalográficas usadas como *templates*, validando o protocolo de modelagem e as análises realizadas.

A disponibilidade de modelos precisos tem aplicação em problemas biologicamente significantes tais como especificidade de substratos e ligantes e áreas como desenho de drogas. O desenho de drogas baseado em estrutura tornou-se uma tecnologia altamente desenvolvida que está em uso ativo nas maiores empresas farmacêuticas do mundo. Modelos moleculares de proteínas que são alvos preferenciais para desenho de drogas têm sido utilizados contra bases de dados de ligantes, com o objetivo de estimar sua afinidade usando *screening* virtual e

simulações de *docking*, além de avaliar a energia livre de ligação absoluta da interação proteína-ligante tão precisamente quanto possível.

A utilização destes modelos implica em grau de precisão, no qual a identidade, o r.m.s.d., as análises de ambiente químico da proteína são de extrema importância, tendo que obedecer a alguns parâmetros de precisão (Figura 4 e Tabela 3). Tais critérios foram estabelecidos na seleção dos modelos e inseridos no DBMODELING para fornecer ao usuário modelos precisos para uso em simulações de *docking* e identificação de alvos terapêuticos.

O DBMODELING está acessível no *site* do Laboratório de Sistemas Biomoleculares (BMSys) em <http://www.biocristalografia.df.ibilce.unesp.br/tools>.

6. Desenvolvimentos Futuros

Posteriormente, o DBMODELING aumentará a quantidade de organismos e ferramentas de análise da qualidade estrutural para refletir a importância do estudo de proteínas pertencentes a vias metabólicas específicas como alvos potenciais para desenho e seleção de drogas tais como as enzimas da via metabólica do ácido chiquímico e outros alvos potenciais.

O desenvolvimento de ferramentas é constante e dinâmico, podendo no futuro ser agregadas ao banco de dados ferramentas de *docking* e de análise proteína-ligante para proteínas selecionadas no DBMODELING, além de dados químicos das reações catalisadas, integração de novos bancos de dados para referências cruzadas e *plugins* para visualização 3D animada da proteína em tempo real. O potencial futuro deste banco de dados tem importância relevante para a pesquisa estrutural de proteínas e vias metabólicas de diversos genomas de interesse biotecnológico e farmacêutico.

7. Bibliografia

ABOLA, B.B., BERNSTEIN, F.C., BRYANT, S.H., KOETZLE, T. & WENG, J. (1987) Protein data bank. In: Allen, F.H., Bergerhoff, G., Sievers, R. (eds) *Crystallographic Databases-Information, Content, Software Systems, Scientific Applications*. Data Commission of the International Union of Crystallography, Bonn Cambridge Chester, pp. 107-132.

ADOBE SYSTEMS INC. (1985) *PostScript language Reference Manual*. Reading, MA: ADDISON-WESLEY.

ALLEN, F.H., BELLARD, S., BRICE, M.D., *et al.* (1979) The Cambridge Crystallographic Data Centre: computer-based search, retrieval, analysis and display of information. *Acta Cryst.* **B35**: 2331-2339.

ANDERSON, A.C. (2003) The Process of Structure-Based Drug Design. *Chemistry & Biology* **10**: 787-797.

BAIROCH, A. & APWEILER, R. (1999) The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. *Nucleic Acids Res.* **27**: 49-54.

BASSO, L.A. & BLANCHARD, J.S. (1998) Resistance to antitubercular drugs. *Adv. Exp. Med. Biol.* **456**:115-144.

BASSO, L.A., ZHENG, R., MUSSER, J.M., JACOBS, W.R.JR. & BLANCHARD, J.S. (1998) Mechanisms of isoniazid resistance in *Mycobacterium tuberculosis*: enzymatic characterization of enoyl reductase mutants identified in isoniazid-resistant clinical isolates. *J. Infect. Dis.* **178**: 769-775.

BAYAT, A. (2002) Science, medicine, and the future: Bioinformatics. *BMJ* **324**:1018-1022.

BERMAN, H.M., WESTBROOK, J., FENG, G., GILLILAND, G., BHAT, T.N., WEISSIG, H., SHINDYALOV, I.N. & BOURNE, P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.* **28**: 235-242.

BETTS, J.C., DODSON, P., QUAN, S., LEWIS, A.P., THOMAS, P.J., DUNCAN, K. & MCADAM, R.A. (2000) Comparison of the proteome of the *Mycobacterium tuberculosis* strain H37Rv with clinical isolate CDC1551. *Microbiology* **146**:3205-3216.

BLANKENFELDT, W., ASUNCION, M., LAM, J.S. & NAISMITH, J.H. (2000) The structural basis of the catalytic mechanism and regulation of glucose-1-phosphate thymidyltransferase (RmlA). *The EMBO Journal* **19**:6652-6663.

BOISSEL, J.P., LEE, W.R., PRESNELL, S.R., COHEN, F.E. & BUNN, H.F. (1993) Erythropoietin structure-function relationships. Mutant proteins that test a model of tertiary structure. *J Biol Chem* **268**:15983-15993.

BOWIE, J.U., LUTHY, R. & EISENBERG, D. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **256**: 164-170.

BRENNAN, P.J. & DRAPER, P. in *Tuberculosis: Pathogenesis, Protection, and Control* (ed. Bloom, B. R.) 271-284 (Am. Soc. Microbiol., Washington DC, 1994).

BRENNER, S.E. (2001) A tour of structural genomics. *Nat Rev Genet* **2**(10):801-809.

BRÜNGER, A.T. (1992). X-PLOR, A System for Crystallography and NMR (Yale Univ. Press, New Haven, CT), Version 3.1.

BUJNICKI, J.M., ELOFSON, A., FISCHER, D. & RYCHLEWSKI, L. (2001) LiveBench-1: continuous benchmarking of proteins structure prediction servers. *Protein Sci.* **10**:352-361.

BURLEY, S.K., ALMO, S.C., BONANNO, J.B., CARPEL, M., CHANCE, M.R.L., GAASTERLAND, T., LIN, D., SALI, A., STUDIER, F.W. & SWAMINATHAN, S. (1999) Structural genomics: beyond the human genome project. *Nat. Genet.* **23**:151-157.

CAMPOS, H.S. (1999) *Mycobacterium tuberculosis* resistente: de onde vem a resistência?. *Boletim de Pneumologia Sanitária* **7**(1): 51-64.

CAMUS, J-C., PRYOR, M.J., MÉDIGUE, C. & COLE, S.T. (2002) Re-annotation of the genome sequence of *Mycobacterium tuberculosis* H37Rv. *Microbiology* **148**:2967-2973.

COLLABORATIVE COMPUTATIONAL PROJECT N° 4., The CCP4 suite: programs for proteins crystallography. *Acta Crystallogr.* **D50**:760-763, 1994.

COLE, S.T., *et al.* (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **11**(393):537-44.

DA SILVEIRA, N.J.F., UCHÔA, H.B., PEREIRA, J.H., CANDURI, F., BASSO, L.A., PALMA, M.S., SANTOS, D.S. & DE AZEVEDO JR., W.F. (2005) Molecular models of proteins targets from *Mycobacterium tuberculosis*. *J. Mol. Model.* **11**:160-166.

DUJON, B. (1996) The yeast genome project: what did we learn? *Trends Genet.* **12**:263-270.

EU 3-D VALIDATION NETWORK (1998) Who checks the checkers? Four validation tools applied to eight atomic resolution structures. *J. Mol. Biol.* **276**(2): 417-436.

ENGH, R.A. & HUBER, R. (1991) Accurate bond and angle parameters for X-ray protein structure refinement. *Acta Cryst.* **A47**: 392-400.

EYRICH, V.A., MARTI-RENOM, M.A., PRZYBYLSKI, D., MADHUSUDHAN, M.S., FISER, A., PAZOS, F., VALENCIA, A., SALI, A. & ROST, B. (2001) EVA: continuous automatic evaluation of protein structure prediction servers. *Bioinformatics* **17**:1242-1243.

FISER, A., DO, R.K. & ŠALI, A. (2000) Modeling of loops in protein structures. *Protein Sci.* **9**:1753-1773.

FISER, A., FEIG, M., BROOKS III, C.L. & SALI, A. (2002) Evolution and Physics in Comparative Protein Structure Modeling. *Acc. Chem. Res.* **35**: 413-421.

FOSTER, M.J. (2002) Molecular modelling in structural biology. *Micron* **33**:365-384.

FOX, H.H. (1951) *Chem. Eng. News* **29**: 3963-3964.

GIBAS, C. & JAMBECK, P. (2001) *Desenvolvendo Bioinformática*, Rio de Janeiro, Ed. Campus, p.179-180.

GUENTHER, B., ONRUST, R., ŠALI, A., O'DONNELL, M. & KURIYAN, J. (1997) Crystal structure of the delta' subunit of the clamp-loader complex of E. coli DNA polymerase III. *Cell* **91**:335-345.

GOULDING, C.W., PERRY, L.J., ANDERSON, D., SAWAYA, M.R., CASCIO, D., APOSTOL, M.I., CHAN, S., PARSEGHIAN, A., WANG, S.S., WU, Y., CASSANO, V., GILL, H.S. & EISENBERG, D. (2003) Structural genomics of *Mycobacterium tuberculosis*: a preliminary report of progress at UCLA. *Biophys. Chem.* **105**:361-370.

HEYM, B., HONORE, N., TRUFFOT-PERNOT, C., BANERJEE, A., SCHURRA, C., JACOBS, W.R.JR., VAN EMBDEN, J.D., GROSSET, J.H. & COLE, S.T. (1994) Implications of multidrug resistance for future of short course chemotherapy of tuberculosis: a molecular study. *Lancet*, **344**: 293-298.

HEYM, B., ALZARI, P.M., HONORE, N. & COLE, S.T. (1995) Missense mutations in the catalase-peroxidase gene, katG, are associated with isoniazid resistance in *Mycobacterium tuberculosis*. *Mol. Microbiol.* **15**:235-245.

HOOFT, R.W., VRIEND, G., SANDER, C. & ABOLA, E.E. (1996) Errors in protein structures. *Nature* 381: 272.

HOWELL, P.L., ALMO, S.C., PARSONS, M.R., HAJDU, J. & PETSKO G.A. (1992) Structure determination of turkey egg-white lysozyme using Laue diffraction data. *Acta Crystallogr* **B48**:200-207.

ISEMAN, M.D. (1994) Evolution of drug-resistant tuberculosis: a tale of two species. *Proc. Natl. Acad. Sci USA* **91**:2428-2429.

JACOBSON, M. & ŠALI, A. (2004) Comparative Protein Structure Modeling and its Applications to Drug Discovery. *Annual Reports in Medicinal Chemistry* **39**:259-276.

JOHNSON, M.S., SRINIVASAN, N., SOWDHAMINI, R. & BLUNDELL, T.L. (1994) Knowledge-based protein modeling. *Crit. Rev. Biochem. Mol. Biol.* **29**:1-68.

JUNGBLUT, P.R., MULLER, E.C. MATTOW, J. & KAUFMANN, S.H. (1999) Proteomics reveals open reading frames in *Mycobacterium tuberculosis* H37Rv not predicted by genomics. *Infect. Immun.* **69**:5905-5907.

KABSCH, W. & SANDER, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**(12): 2577-2637.

KANEHISA, M. (2000) Post-genomic Informatics. Oxford University Press Inc., New York, pp.147.

KARP, P.D., RILEY, M., PALEY, S.M. & PELLEGRINI-TOOLE, A. (2002) The MetaCyc Database. *Nucleic Acids Res.* **30**:59-61.

KEARSLEY, S.K. (1989) On the orthogonal transformation used for structural comparisons. *Acta Crystallogr.* **45A**: 208-210.

-
- KOCHI, A. (1991) The global tuberculosis situation and the new control strategy of the World Health Organization. *Tubercle* **72**:1-6.
- KOEHL, P. & LEVITT, M. (1999) A brighter future for protein structure prediction. *Nature Struct. Biol.* **6**:108-111.
- KOONIN, E.V. & MUSHEGIAN, A.R. (1996) Complete genome sequences of cellular life forms: glimpses of theoretical evolutionary genomics. *Curr. Opin. Gen. Dev.* **6**:757-762.
- KORADI, R., BILLETER, M. & WUTHRICH, K (1996) MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graph.* **14**: 51-55.
- KWON, H.H., TOMIOKA, H. & SAITO, H. (1995) Distribution and characterization of lactamases of mycobacteria and related organisms. *Tubercle Lung Dis.* **76**:141-148.
- LASKOWASKI, R.A., MACARTHUR, M.W., MOSS, D.S. & THORNTON, J.M. (1993) PROCHECK – A program to check the stereochemical quality of protein structures. *J. Appl. Cryst.* **26**:283–291.
- LASKOWASKI, R.A., MACARTHUR, M.W. & THORNTON, J.M. (1998) Validation of protein models derived from experiment. *Curr Opin Struct Biol.* **8**: 631-639.
- LESK, A.M. (2001) “Introduction to Protein Architecture”. Oxford University Press, Oxford, Great Britain.
- LIU, J. & ROST, B. (2002) Target space for structural genomics revisited. *Bioinformatics* **18**(7):922-33.
- LUTHY, R., BOWIE, J.U. & EISENBERG, D. (1992) Assessment of protein models with three-dimensional profiles. *Nature* **356**:83-85.
- MA, Y., MILLS, J.A., BELISLE, J.T., VISSA, V., HOWELL, M., BOWLIN, K., SCHERMAN, M.S. & MCNEIL, M. (1997) Determination of the pathway for rhamnose biosynthesis in mycobacteria: cloning, sequencing and expression of the Mycobacterium tuberculosis gene encoding α -D-glucose-1-phosphate thymidyltransferase. *Microbiology* **143**:937-945.

MACKERELL JR., A.D., BASHFORD, D., BELLOTT, M., *et al.* (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B.* **102**: 3586-3616.

MARTI-RENOM, M.A., STUART, A.C., FISER, A., SANCHEZ, R., MELO, F. & SALI, A. (2000) Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **29**:291-325.

MARTI-RENOM, M.A., MADHUSUDHAN, M.S., FISER, A., ROST, B. & SALI, A. (2002) Reliability of assessment of protein structure prediction methods. *Structure* **10**: 435-440.

MARTI-RENOM, M.A., MADHUSUDHAN, M.S. & ŠALI, A. (2004) Alignment of protein sequences by their profiles. *Protein Sci.* **13**:1071-1087.

MEYER, H. & MALLY, J. (1912) *Monatsch Chem* **33**:393-414.

MELO, F. & FEYTMANS, E. (1998) Assessing proteins structures with a non-local atomic interaction energy. *J. Mol. Biol.* **277**: 1141-1152.

MIDDLEBROOK, G., COHN, M.L. & SCHAEFER, W.B. (1954) Studies on isoniazid and tubercle bacilli. III. The isolation, drug-susceptibility, and catalase-testing of tubercle bacilli from isoniazid-treated patients. *Am. Rev. Tub.* **70**:852-872.

MIKLOS, G.L.G. & RUBIN, G.M. (1996) The role of the genome project in determining gene function: insights from model organisms. *Cell* **86**:251-259.

MIWA, J.M., IBANEZ-TALLON, O., CRABETREE, G.W., SANCHEZ, R., ŠALI, A., ROLE, L.W. & HEINTZ, N. (1999) lynx1, an endogenous toxin-like modulator of nicotinic acetylcholine receptors in the mammalian CNS. *Neuron* **23**:105-114.

MODI, S., PAINE, M.F., SUTCLIFFE, M.J., LIAN, L.Y., PRIMROSE, W.U., WOLF, C.R. & ROBERTS, G.C. (1996) A model for human cytochrome P450 2D6 based on homology modeling and NMR studies of substrate binding. *Biochemistry* **35**:4540-4550.

MOLLENKOPF, H.J., JUNGBLUT, P.R., RAUPACH, B., MATTOW, J., LAMER, S., ZIMNY-ARNDT, U. SCHAIBLE, U.E. & KAUFMANN, S.H. (1999) A dynamic two-dimensional polyacrylamide gel electrophoresis database: the micobacterial proteome via Internet. *Electrophoresis* **20**:2172-2180.

MORRIS, A.L., MACARTHUR, M.W., HUTCHINSON, E.G., THORNTON, J.M. (1992) Stereochemical quality of protein structure coordinates. *Proteins* **12**: 345-364.

NEEDLEMAN, S.B. & WUNSCH, C.D. (1970) A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *J. Mol. Biol.* **48**:443-453.

NIKAIDO, H. (1994) Prevention of drugs access to bacterial targets: Permeability barriers and active efflux. *Science* **264**:328-388.

OGATA, H., GOTO, S., SATO, K., FUJIBUCHI, W., BONO, H. & KANEHISA, M. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **27**:29-34.

OLIVER, S.G. (1996) From DNA sequence to biological function. *Nature (London)* **379**:597-600.

ORENGO, C.A., JONES, D.T. & THORNTON, J.M. (1994) Protein superfamilies and domain superfolds. *Nature (London)* **372**:631-634.

PASCOPELLA, L., COLLINS, F.M., MARTIN, J.M., LEE, M.H., HATFULL, G.F., STOVER, C.K., BLOOM, B.R. & JACOBS JR., W.R. (1994) Use of in vivo complementation in *Mycobacterium tuberculosis* to identify a genomic fragment associated with virulence. *Infect. Immun.* **62**:1313-1319.

PIEPER, U., ESWAR, N., BRABERG, H., MADHUSUDHAN, M.S., DAVIS, F.P., STUART, A.C., MIRKOVIC, N., ROSSI, A., MARTI-RENOM, M.A., FISER, A., WEBB, B., GREENBLATT, D., HUANG, C.C., FERRIN, T.E. & SALI, A. (2004) MODBASE, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Res.* **32**:D217-D222.

QUE, X., BRINEN, L.S., PERKINS, P., HERDMAN, S., HIRATA, K., TORIAN, B.E., RUBIN, H., MCKERROW, J.H. & REED, S.L. (2002) Cysteine proteinases from distinct cellular compartments are recruited to phagocytic vesicles by *Entamoeba histolytica*. *Mol. Biochem. Parasitol.* **119**:23-32.

RAMACHANDRAN, G.N., RAMAKRISHNAN, C. & SASISEKHARAN, V. (1963) Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* **7**:95-99.

RAMAKRISHNAN, C. (2001) Ramachandran and his map. *Resonance* **1**: 48-56.

RING, C.S., SUN, E., MCKERROW, J.H., LEE, G.K., ROSENTHAL, P.J., KUNTZ, I.D. & COHEN, F.E. (1993) Structure-based inhibitor design by using protein models for the development of antiparasitic agents. *Proc. Natl. Acad. Sci. USA* **90**:3583-3587.

ROSENKRANDS, I., KING, A., WELDINGH, K., MONIATTE, M., MOERTZ, E. & ANDERSEN, P. (2000) Towards the proteome of *Mycobacterium tuberculosis*. *Electrophoresis* **21**:3740-3756.

ŠALI, A. & BLUNDELL, T. L. (1993) Comparative Protein Modelling by Satisfaction of Spatial Restraints. *J. Mol. Biol.* **234**:779-815.

ŠALI, A. & OVERINGTON, J.P. (1994) Derivation of Rules for Comparative Protein Modeling from a Database of Protein Structure Alignments. *Proteins Sci.* **3**: 1582-1596.

ŠALI, A. (1995) Comparative protein modeling by satisfaction of spatial restraints. *Mol. Med. Today* **1**: 270-277.

SÁNCHEZ, R. & ŠALI, A. (1997) Advances in comparative protein-structure modelling. *Curr Opin Struct Biol* **7**:206-214.

SÁNCHEZ, R. & ŠALI, A. (1998) Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. *Proc. Natl. Acad. Sci. USA* **95**:13597-13602.

SAQI, M.A., RUSSELL, R.B. & STERNBERG, M.J. (1998) Misleading local sequence alignments: implications for comparative protein modelling. *Protein Eng.* **11**: 627-630.

SCHWIETERS, C.D., KUSZEWSKI, J.J., TJANDRA, N. & CLORE, G.M. (2003) The Xplor-NIH NMR Molecular Structure Determination Package. *J. Magn. Res.* **160**: 65-73.

SIPPL, M.J. (1990) 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.

SIPPL, M.J. (1993) Recognition of errors in three-dimensional structures of proteins. *Proteins* **17**: 355-362.

SNIDER JR., D.E., RAVIGLIONE, M. & KOCHI, A. (1994) in *Tuberculosis: Pathogenesis, Protection, and Control* (ed. Bloom, B.R.) 2-11 (Am. Soc. Microbiol., Washington DC).

SRINIVASAN, N. & BLUNDELL, T.L. (1993) An evaluation of the performance of an automated procedure for comparative modelling of protein tertiary structure. *Protein Eng.* **6**:501-512.

TAYLOR, W.R. (2002) A 'periodic table' for protein structures. *Nature* **11**(416): 657-660.

TERWILLIGER, T.C., *et al.* (2003) The TB structural genomics consortium: a resource for *Mycobacterium tuberculosis* biology. *Tuberculosis* **83**: 223-249.

TOPHAM, C.M., SRINIVASAN, N., THORPE, C.J., OVERINGTON, J.P. & KALSHEKER, N.A. (1994) Comparative modelling of major house dust mite allergen Der p I: structure validation using an extended environmental amino acid propensity table. *Protein Eng.* **7**: 869-894.

UCHÔA, H.B., JORGE, G.E., DA SILVEIRA, N.J.F., CAMERA JR., J.C. & DE AZEVEDO JR., W.F. (2004) *Parmodel*: a web server for automated comparative modeling of proteins. *Biochem. Biophys. Res. Commun.* **325**:1481-1486.

VAKSER, I.A. (1997) Evaluation of GRAMM low-resolution docking methodology on the hemagglutinin-antibody complex. *Proteins* **1**:226-230.

VAN VLIJMEN, H.W. & KARPLUS, M. (1997) PDB-based protein loop prediction: parameters for selection and methods for optimization. *J. Mol. Biol.* **267**:975-1001.

VEERAPANDIAN, P. (1997) Structure-based drug design, (Dekker, M., ed.), INC. New York.

WALL, L., CHRISTIANSEN, T. & ORWANT, J. (2000) Programming Perl. Ed. O'Reilly, 3a.ed., pp.1070.

WELDINGH, K., ROSENKRANDS, I., JACOBSEN, S., RASMUSSEN, P.B., ELHAY, M.J. & ANDERSEN, P. (1998) Two-dimensional electrophoresis for analysis of *Mycobacterium tuberculosis* culture filtrate and purification and characterization of six novel proteins. *Infect. Immun.* **66**:3492-3500.

WESTBROOK, J., FENG, Z., CHEN, L., YANG, H. & BERMAN, H.M. (2003) The Protein Data Bank and structural genomics. *Nucleic Acids Res.* **31**:489-491.

WHEELER, P.R. & RATLEDGE, C. (1994) in *Tuberculosis: Pathogenesis, Protection, and Control* (ed. Bloom, B.R.) p.353-385 (Am. Soc. Microbiol., Washington DC)

WHO / IUATLD. (1998) Guidelines for surveillance of drug resistance in tuberculosis. *Int. J. Tuberc. Lung Dis.* **2**(1):72-89.

WINDER, F.G. (1982) The biology of the Mycobacteria (Ratledge C e Sanford J, eds), pp. 353-438. Academic Press.

WU, G., FISER, A., TER KUILE, B., ŠALI, A. & MULLER, M. (1999) Convergent evolution of *Trichomonas vaginalis* lactate dehydrogenase from malate dehydrogenase. *Proc Natl Acad Sci USA* **96**:6285-6290.

XU, L.Z., SÁNCHEZ, R., ŠALI, A. & HEINTZ, N. (1996) Ligand specificity of brain lipid-binding protein. *J Biol Chem* **271**:24711-24719.

YOUNG, D.B. (1994) Tuberculosis. Beating the bacillus. *Curr. Biol.* **4**:351-353.

ZHANG, Y., HEYM, B., ALLEN, B., YOUNG, D. & COLE, S.T. (1992) The catalase-peroxidase gene and isoniazid resistance of *Mycobacterium tuberculosis*. *Nature*. **358**:591-593.

APÊNDICE A – Descrição dos softwares utilizados

I. MODELAGEM MOLECULAR

A modelagem molecular comparativa ou por homologia constrói um modelo tridimensional para uma proteína desconhecida baseada em uma ou mais proteínas de estruturas conhecidas relacionadas (*templates*).

O método consiste de três etapas: 1) alinhamento da seqüência alvo com a seqüência do *template* e segmentos relacionados; 2) extração das restrições espaciais das seqüências usando o alinhamento; e 3) satisfação de restrições para obter um modelo tridimensional (figura 31).

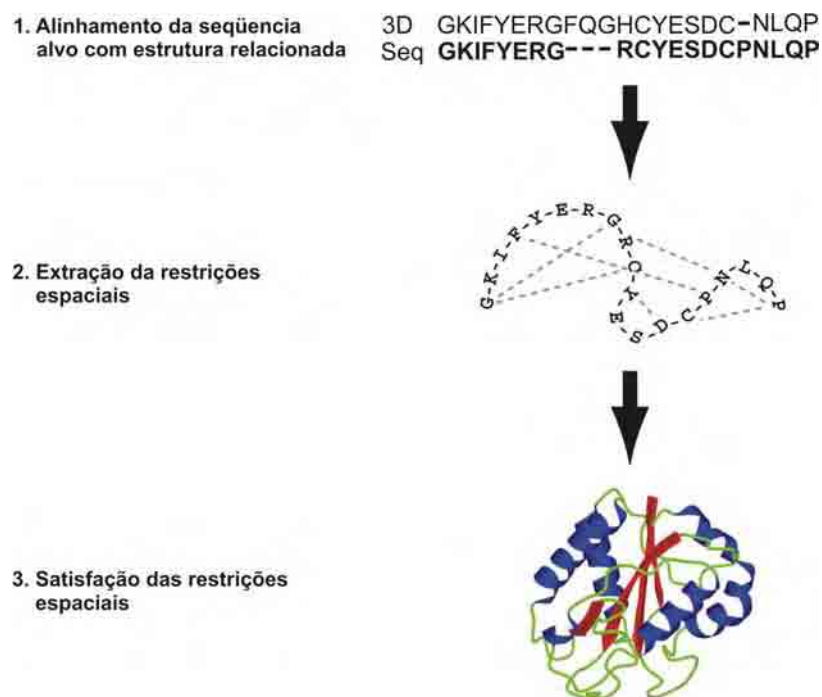


Figura 31. Passos para construção de um modelo utilizando modelagem comparativa por satisfação de restrições espaciais.

As restrições espaciais da seqüência a ser modelada são obtidas de análises estatísticas de relações entre várias características de estrutura protéica. Estas relações foram descritas como uma função densidade de probabilidade condicional (pdf) $P(\mathbf{f}|\mathbf{I})$ para a característica da restrição espacial \mathbf{f} , dado diversas variáveis \mathbf{I} que foram encontradas para as características a serem preditas. Há três tipos de restrições, depende da origem e da natureza da informação \mathbf{I} (ŠALI & BLUNDELL, 1993; FISER *et al.*, 2002).

No primeiro tipo, as restrições são obtidas para aqueles resíduos do alvo que são alinhados com os resíduos do molde. Esta homologia é derivada das restrições limitada às distâncias entre os átomos da cadeia principal e cadeia lateral, bem como os ângulos diedros da cadeia principal (Φ , Ψ e Ω) e os ângulos diedros da cadeia lateral (χ_i). Estas restrições foram obtidas de uma análise estatística de 105 famílias alinhadas que incluem 416 proteínas definidas estruturalmente (ŠALI & OVERINGTON, 1994). Por exemplo, uma restrição numa certa distância entre dois $C\alpha$ - $C\alpha$ equivalentes em duas estruturas de proteína relacionadas é bem descrita por uma soma pesada de duas funções Gaussianas que correspondem às duas distâncias do molde, respectivamente (ŠALI & BLUNDELL, 1993; FISER *et al.*, 2002).

A segunda classe de restrições reflete as preferências estatísticas extraídas das estruturas de proteínas conhecidas em geral e são relacionadas aos potenciais estatísticos de força média (FISER *et al.*, 2002). As restrições dependem somente dos tipos restritos de átomos ou resíduos e não da estrutura do molde. Estas restrições são aplicadas à cadeia principal e aos ângulos diedros da cadeia lateral dos

resíduos alvos que não são alinhados com os resíduos do molde e para as distâncias entre todos os pares de átomos não ligados. São usados porque foram encontrados para resultar em modelos mais exatos do que os termos correspondentes do campo de força molecular (FISER *et al.*, 2000).

O terceiro tipo de restrição é obtido a partir do campo de força molecular do CHARMM-22 (MACKERELL *et al.*, 1998) e inclui restrições de ligações químicas. Estas restrições de forças moleculares reforçam estereoquimicamente o modelo correto. Depois que todas as pdfs são calculadas, seus logaritmos são somados para obter uma função objetiva que dependa do modelo e dê sua probabilidade. Finalmente, o modelo que contém todos os átomos não hidrogenados é calculado otimizando a função objetiva no espaço cartesiano (Figura 32). A otimização é realizada pelo método da função alvo variável que emprega gradientes conjugados e a mecânica molecular com *simulated annealing* (ŠALI & BLUNDELL, 1993; FISER *et al.*, 2002).

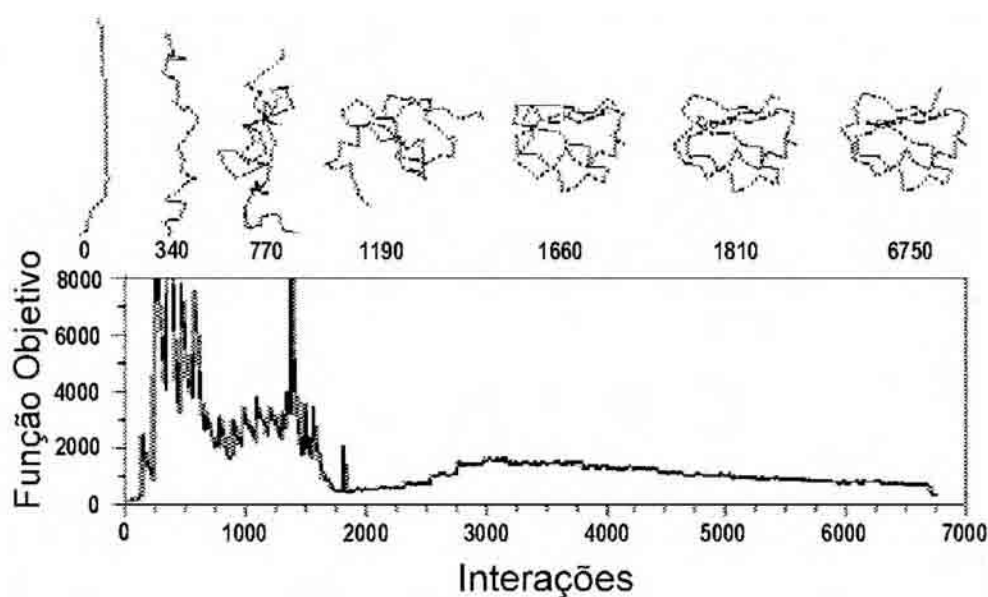


Figura 32. Mudança conformacional pela minimização da função objetiva.

A probabilidade finita atual de um evento $x_1 \leq x < x_2$ é obtida pela integração de p :

$$p(x_1 \leq x < x_2) = \int_{x_1}^{x_2} p(x) dx \quad (1)$$

Uma pdf útil para restringir certa característica x pode ser escrita como:

$$p(x/a, b, \dots, c) \quad (2)$$

Isto é uma pdf condicional que nos dá a densidade de probabilidade para x , quando a, b, \dots, c são conhecidos. Isso pode ser visto como uma pdf para x que também depende dos valores de outras variáveis.

Para que uma pdf seja útil na modelagem, todas as características associadas $(x/a, b, \dots, c)$ exceto x devem ser conhecidas no momento da predição. Além disso, x deve ser uma característica espacial no momento da predição. A característica conhecida mais importante é aquela do mesmo tipo de x e associada com posições equivalentes nas estruturas conhecidas relacionadas.

Após a determinação das pdfs individuais baseadas nas restrições espaciais, é formada a pdf molecular a partir do produto das pdfs características, onde f_i representa as características individuais. A pdf molecular deve fornecer uma probabilidade de ocorrência de alguma destas características simultaneamente. Então o modelo para a estrutura tridimensional desconhecido poderá corresponder ao máximo da pdf molecular.

$$P = \prod_i p^f(f_i)$$

Sendo assim, maximizando a função P pode-se encontrar o modelo mais provável para a estrutura 3D da seqüência, dado seu alinhamento com as estruturas conhecidas.

Finalmente, o modelo é obtido pela otimização da função objetiva, em um espaço cartesiano. Diversos modelos ligeiramente diferentes podem ser calculados, variando a estrutura inicial, e a variabilidade entre estes modelos pode ser usada para estimar a baixa ligação sobre os erros em regiões correspondentes do enovelamento.

Logo, a função que é realmente otimizada é uma transformação da pdf molecular de P :

$$F = -\ln(p)$$

onde todas as características são expressas em termos de coordenadas atômicas no espaço cartesiano. A função F é chamada de função objetivo. O mesmo conjunto de coordenadas cartesianas que maximiza P também minimiza F , pelo fato de que se substitui a multiplicação de termos em P pela adição de termos em F , o que reduz o problema de ponto flutuante.

A saída é um modelo 3D para a seqüência alvo contendo todas as cadeias principais e cadeias laterais dos átomos sem o hidrogênio. Além da construção do modelo o MODELLER (ŠALI & BLUNDELL, 1993) pode executar tarefas auxiliares adicionais, incluindo um alinhamento de duas seqüências de proteínas ou de seus perfis, alinhamento múltiplo de seqüências e/ou de estruturas de proteínas,

cálculo de árvores filogenéticas, e modelagem *de novo* de alças nas estruturas de proteínas (JACOBSON & ŠALI, 2004).

II. ALINHAMENTO (NEEDLEMAN-WUNSCH)

Needleman e Wunsch conceituaram o problema de alinhamento como sendo um problema de programação dinâmica (NEEDLEMAN & WUNSCH, 1970). Conhecido por algum tempo como algoritmo heurístico de homologias, a importância do método proposto por Needleman e Wunsch é que este método foi o primeiro a introduzir a noção de uma matriz de percurso – idéia central na versão moderna do algoritmo de alinhamento global de seqüências por programação dinâmica.

O algoritmo de programação dinâmica é um algoritmo geral para otimização de problemas. É também um algoritmo fundamental para o entendimento dos conceitos de alinhamento de seqüências. A figura 33 ilustra o princípio do algoritmo de programação dinâmica. As duas *strings* a serem comparadas são colocadas nos eixos horizontal e vertical da matriz, a qual foi chamada de matriz caminho. Sem mudanças na ordem das letras, o alinhamento é feito da esquerda para a direita em ambas as *strings*. Obviamente, há muitos caminhos alternativos e o problema de encontrar o alinhamento de seqüências ótimo torna-se equivalente a encontrar o caminho ótimo na matriz caminho. A estrutura em árvore mostrada na figura 33b é uma ilustração de todos os possíveis caminhos, iniciando no canto superior esquerdo da matriz caminho e resultando em um leque de três vias em cada nó. Este algoritmo

apresenta como problema importante o fato de o número de operações a realizar crescer como produto do comprimento das duas seqüências a serem alinhadas. Assim, o problema do alinhamento de seqüências é um típico problema de otimização combinatorial em computação.

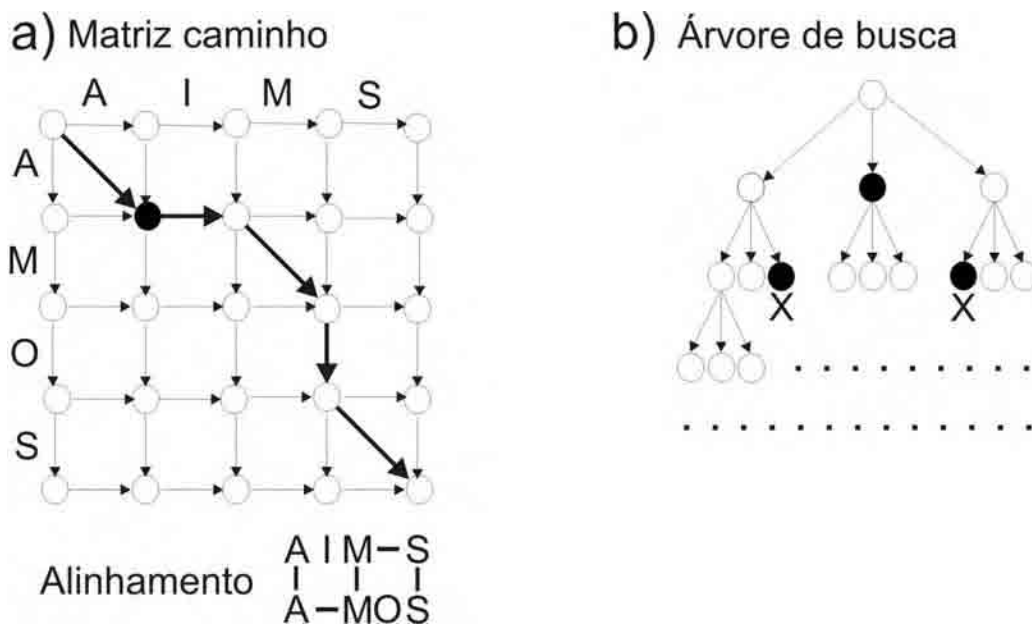


Figura 33. Alinhamento de seqüências por algoritmo de programação dinâmica. O algoritmo envolve a busca pelo caminho ótimo na matriz caminho a), o qual é equivalente a buscar uma solução ótima na árvore de busca b).

O alinhamento de seqüências requer encontrar a melhor solução entre todas as possibilidades de acordo com um critério dado, ao invés de encontrar qualquer uma das muitas soluções como em resolver um quebra-cabeças. Um algoritmo de programação dinâmica encontra uma boa solução dividindo o problema original em menores problemas e solucionando-os depois, tornando-se um algoritmo ideal para se usar em paralelização computacional, pois não existe dependência seqüencial. O

objetivo é maximizar a pontuação geral para o alinhamento. Para isso, o número de pares de resíduos de alta pontuação deve ser maximizado e o número de espaços e pares de baixa pontuação deve ser minimizado. Contudo, a beleza do algoritmo de programação dinâmica é que mais ramos são sistematicamente cortados de acordo com a função *escore*. Devido o algoritmo avaliar eficientemente todas as possibilidades, o problema do alinhamento de seqüência por pares pode ser resolvido rigorosamente.

O alinhamento usado para identificar os *templates* e gerar os modelos foi o alinhamento global, o qual inclui todos os caracteres de cada uma das duas seqüências que estão sendo comparadas; o alinhamento ótimo é aquele com o mais alto *escore* possível. O alinhamento global é útil para comparar duas seqüências homólogas. Este alinhamento possui dependência quadrática, ou seja, para duas seqüências de comprimentos n e m , o tempo gasto na execução do alinhamento é proporcional a $n*m$.

O alinhamento global considera a seqüência completa de bases ou aminoácidos. Nesse tipo de alinhamento, as penalidades tanto para os espaços de abertura como para os espaços na extensão são bastante altos. Portanto, formações de blocos durante os alinhamentos são inexistentes e o que se observa são pequenas regiões ou alguns poucos espaços (*gaps*) espalhados ao longo da seqüência, preservando assim, o maior número possível de resíduos alinhados. Esse tipo de alinhamento é apropriado para seqüências que possuem grande similaridade em todo

seu comprimento, já que o alinhamento é otimizado em toda sua extensão, favorecendo sua utilização em modelagem molecular comparativa de proteínas.

A função escore a ser otimizada é a soma dos pesos de cada posição do alinhamento. Os pesos são definidos pela matriz de substituição entre os 20 aminoácidos e também pela penalidade do espaço (*gap*). Idealmente, eles devem refletir processos biológicos de mutações ou inserções/deleções.

Para a determinação do escore de alinhamento ótimo, devemos considerar $W_{s,t}$ como sendo o peso para a substituição da letra s pela letra t ou vice-versa, a qual é um elemento da matriz de substituição simétrica, e seja d o peso para um simples *gap*. De acordo com o algoritmo de programação dinâmica, o valor ótimo da função escore em cada nó é determinado por três possibilidades: diagonal, vertical e horizontal como mostra a figura 34a. A equação 1 representa esta maximização da função escore para uma matriz D :

$$D_{i,j} = \max(D_{i-1,j-1} + w_{s(i),t(j)}, D_{i-1,j} + d, D_{i,j-1} + d) \quad \text{Eq. 1}$$

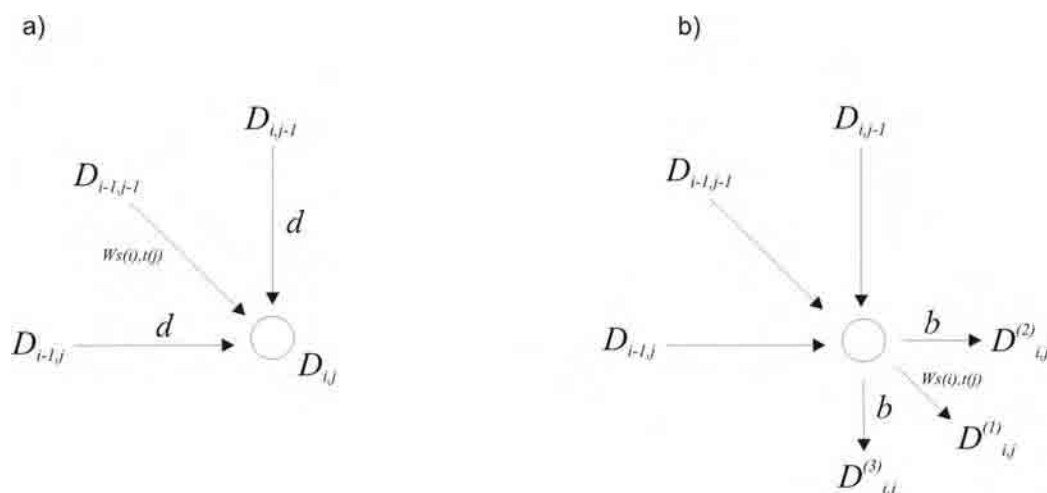


Figura 34. Método para calcular o escore ótimo no algoritmo de programação dinâmica. a) A penalidade do espaço (*gap*) é constante. b) A penalidade do espaço é uma função linear com relação ao comprimento do espaço.

Embora a programação dinâmica consuma tempo, ela encontra soluções mais rigorosas e é mais sensível para detectar similaridades sutis. É dito ser o método final para busca por similaridade de seqüências de aminoácidos, porque a função escore reflete similaridades entre aminoácidos (KANEHISA, 2000).

A matriz de alinhamento utilizada no programa foi a BLOSUM 62. A matriz BLOSUM é calculada a partir de comparações entre seqüências com identidade máxima de 62% podendo ser extrapolada também para uma matriz BLOSUM 80, o que significa uma identidade máxima de 80% entre as seqüências, logo, a escolha de matrizes diferentes implicará em resultados ligeiramente distintos.

As matrizes BLOSUM (Figura 35) são derivadas do banco de dados Blocks (<http://blocks.fhcrc.org>), um conjunto de alinhamentos contínuos de regiões de seqüência em famílias de proteínas relacionadas. Um método por agrupamento ordena as seqüências de cada bloco em grupos de relação à próxima e as freqüências

de substituições entre eles dentro de uma família determinam a probabilidade de uma substituição significativa.

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	4	0	-2	-1	-2	0	-2	-1	-1	-1	-1	-2	-1	-1	-1	1	0	0	-3	-2
C	0	9	-3	-4	-2	-3	-3	-1	-3	-1	-1	-3	-3	-3	-3	-1	-1	-1	-2	-2
D	-2	-3	6	2	-3	-1	-1	-3	-1	-4	-3	1	-1	0	-2	0	-1	-3	-4	-3
E	-1	-4	2	5	-3	-2	0	-3	1	-3	-2	0	-1	2	0	0	-1	-2	-3	-2
F	-2	-2	-3	-3	6	-3	-1	0	-3	0	0	-3	-4	-3	-3	-2	-2	-1	1	3
G	0	-3	-1	-2	-3	6	-2	-4	-2	-4	-3	0	-2	-2	-2	0	-2	-3	-2	-3
H	-2	-3	-1	0	-1	-2	8	-3	-1	-3	-2	1	-2	0	0	-1	-2	-3	-2	2
I	-1	-1	-3	-3	0	-4	-3	4	-3	2	1	-3	-3	-3	-3	-2	-1	3	-3	-1
K	-1	-3	-1	1	-3	-2	-1	-3	5	-2	-1	0	-1	1	2	0	-1	-2	-3	-2
L	-1	-1	-4	-3	0	-4	-3	2	-2	4	2	-3	-3	-2	-2	-2	-1	1	-2	-1
M	-1	-1	-3	-2	0	-3	-2	1	-1	2	5	-2	-2	0	-1	-1	-1	1	-1	-1
N	-2	-3	1	0	-3	0	1	-3	0	-3	-2	6	-2	0	0	1	0	-3	-4	-2
P	-1	-3	-1	-1	-4	-2	-2	-3	-1	-3	-2	-2	7	-1	-2	-1	-1	-2	-4	-3
Q	-1	-3	0	2	-3	-2	0	-3	1	-2	0	0	-1	5	1	0	-1	-2	-2	-1
R	-1	-3	-2	0	-3	-2	0	-3	2	-2	-1	0	-2	1	5	-1	-1	-3	-3	-2
S	1	-1	0	0	-2	0	-1	-2	0	-2	-1	1	-1	0	-1	4	1	-2	-3	-2
T	0	-1	-1	-1	-2	-2	-2	-1	-1	-1	0	-1	-1	-1	1	5	0	-2	-2	-2
V	0	-1	-3	-2	-1	-3	-3	3	-2	1	1	-3	-2	-2	-3	-2	0	4	-3	-1
W	-3	-2	-4	-3	1	-2	-2	-3	-3	-2	-1	-4	-4	-2	-3	-3	-2	-3	11	2
Y	-2	-2	-3	-2	3	-3	2	-1	-2	-1	-1	-2	-3	-1	-2	-2	-2	-1	2	7

Figura 35. Matriz de valores de alinhamento BLOSUM62.

III. PROCHECK

O PROCHECK (LASKOWSKI *et al.*, 1993) analisa a geometria global da estrutura ou cada resíduo individualmente, utilizando para isso parâmetros estereoquímicos derivados de estruturas de alta resolução ou bem refinadas (MORRIS *et al.*, 1992) que constituem sua base de dados. Os parâmetros estereoquímicos usados são aqueles descritos em detalhes em MORRIS *et al.* (1992). As verificações também fazem uso do comprimento e ângulo de ligação ideal, derivados de uma análise detalhada (ENGH & HUBER, 1991) de estruturas de pequenas moléculas em Cambridge Structural Database (CSD) (ALLEN *et al.*, 1979). Estes parâmetros utilizados como informações estereoquímicas checadas pelo

PROCHECK são: ligações covalentes, planaridade de grupos planares (aromáticos, ligações peptídicas, etc.) ângulos diedros, quiralidade, interações não covalentes, ligações de hidrogênio da cadeia principal e pontes de dissulfeto.

O PROCHECK requer como entrada um arquivo de coordenadas atômicas da estrutura da proteína a ser avaliada no formato “pdb” e produz representações coloridas em PostScript (ADOBE SYSTEM INC., 1985) facilmente, interpretadas, descrevendo a estrutura de uma proteína, e também comparando duas estruturas de proteínas relacionadas, juntamente com uma lista detalhada de resíduo por resíduo. Dando uma avaliação da qualidade total da estrutura, em comparação às estruturas bem refinadas da mesma definição, e destaca também as regiões que podem necessitar uma investigação adicional. Dentre as várias análises fornecidas pelo PROCHECK, foram utilizadas no DBMODELING o diagrama de Ramachandran e o G-factor.

G. N. Ramachandran descreveu as conformações disponíveis para os aminoácidos em uma cadeia polipeptídica. A conformação da cadeia peptídica é simplesmente descrita pelos valores dos ângulos diedros na estrutura principal da proteína (ângulo descrito pelo nitrogênio e carbono alfa (N - C_α) e o ângulo descrito pelo carbono alfa e carbono (C_α - C)) (Figura 36). Estes ângulos são denominados ângulos de torção ϕ e ψ , respectivamente (GIBAS & JAMBECK, 2001). Por convenção, ambos, ϕ e ψ , são definidos como 0° na conformação na qual as duas ligações peptídicas conectadas a um único carbono α estão em um mesmo plano. Em princípio, ϕ e ψ podem ter qualquer valor entre -180° e +180°, porém muitos valores

de ϕ e ψ são proibidos pelas interferências estéricas entre átomos pertencentes ao mesmo esqueleto polipeptídico e pelas cadeias laterais dos resíduos de aminoácidos. A conformação na qual ϕ e ψ são, ambos, iguais a 0° é proibida por esta razão. Esta convenção é meramente usada como ponto de referência para descrever os ângulos de torção.

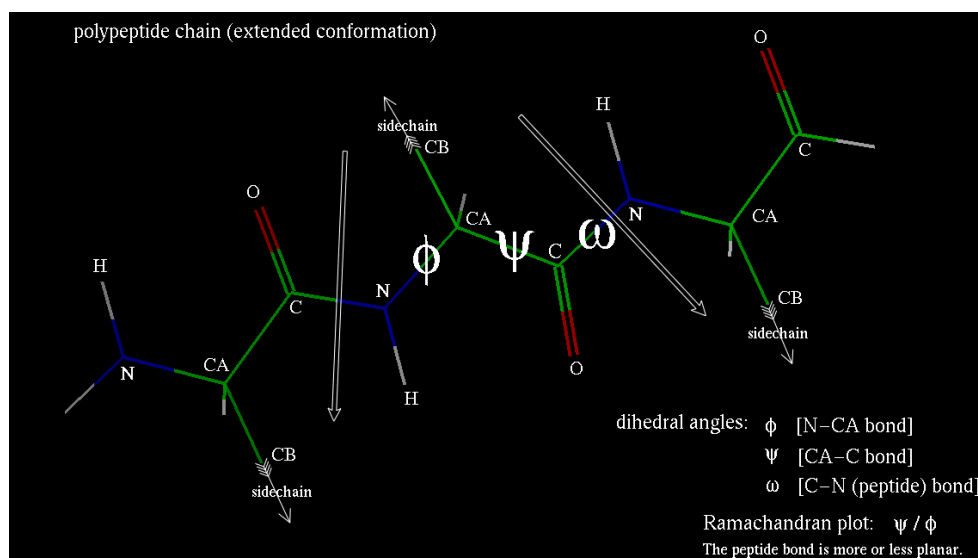


Figura 36. Representação dos ângulos de torção em uma cadeia polipeptídica.

A cadeia não é livre para girar ao redor do terceiro tipo de ligação na estrutura principal da proteína, a ligação peptídica, por ser uma ligação parcialmente dupla e, portanto quimicamente restrita a ser planar; sendo assim, os valores de ϕ e ψ para cada aminoácido fornecem uma descrição completa da estrutura principal da proteína.

G. N. Ramachandran usou modelos computacionais de pequenos peptídeos para variar sistematicamente os ângulos de torção ϕ e ψ com o objetivo de encontrar

conformações estáveis. Para cada conformação, a estrutura foi examinada para verificar a proximidade existente dos contatos entre os átomos. Os átomos foram tratados como esferas rígidas com as dimensões que correspondem a seus raios em van der Waals. Conseqüentemente, os ângulos de torção ϕ e ψ que fazem com que as esferas colidam correspondem a conformação estericamente não permitida (região não permitida) para a cadeia principal de um polipeptídeo (RAMAKRISHNAN, 2001).

O diagrama de Ramachandran (RAMACHANDRAN *et al.*, 1963) indica a distribuição dos ângulos ϕ e ψ dos resíduos pertencentes a uma determinada estrutura. Em 1993, Laskowski e colaboradores (LASKOWSKI *et al.*, 1993) usaram a estrutura cristalográfica de 118 proteínas resolvidas a resolução melhor que 2,0 Å e R-factor melhor que 20% para definição de regiões permitidas e não permitidas no gráfico ϕ x ψ . A análise das estruturas a alta resolução levou à definição das seguintes regiões: permitidas, adicionalmente permitidas, generosamente permitidas e proibidas. A análise que leva em consideração as novas regiões do gráfico ϕ x ψ , foram implementadas no programa PROCHECK (LASKOWSKI *et al.*, 1993) e encontra-se disponível para uso *on-line* no programa *Parmodel* (UCHÔA *et al.*, 2004). Nos seus critérios uma estrutura de boa qualidade deve ter 90% ou mais de seus resíduos nas regiões mais favoráveis (Figura 37). No diagrama abaixo, as áreas em branco correspondem a regiões onde existem os choques estereoquímicos na proteína. Estas regiões não permitidas estericamente são para todos os aminoácidos

exceto a glicina (Gly) que é o aminoácido que possui a cadeia lateral mais simples, possuindo apenas um hidrogênio, podendo assim, assumir qualquer ângulo de torção ϕ e ψ nos quadrantes do diagrama de Ramachandran e a prolina, a qual possui uma ligação cíclica, impedindo o ângulo ϕ de se movimentar livremente, deixando-a com pouca liberdade de torção (Figuras 38a e b). As regiões vermelhas correspondem à conformação onde nenhum choque estereoquímico é observado, ou seja, são as regiões permitidas e aonde se encontram as hélices- α e folhas- β . As áreas em amarelo mostram as regiões limite onde um raio ligeiramente mais curto de van der Waals é usado no cálculo. Isto faz com que tenha uma região adicional que corresponda a hélices esquerda e direita.

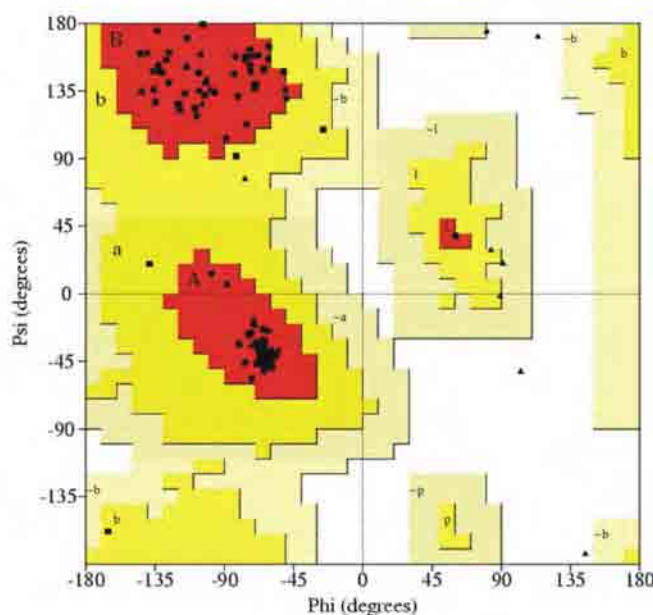


Figura 37. Diagrama de Ramachandran. A região mais favorável está expressa em vermelho, a região permitida está em amarelo, a região generosamente permitida em amarelo claro e a não permitida em branco. As regiões vermelhas no canto superior esquerdo, no centro esquerdo e no centro direito representam as formações de folhas- β paralelas e anti-paralelas, hélices- α à direita e hélices- α à esquerda respectivamente.

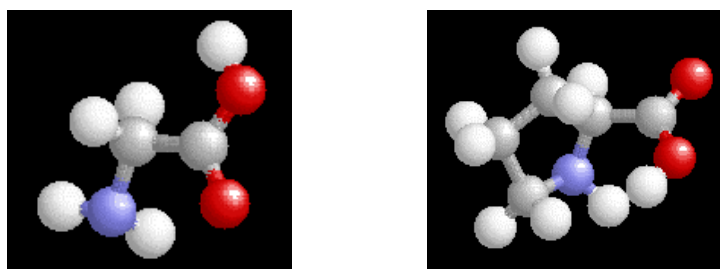


Figura 38. a) Representação estrutural da molécula de Glicina (Gly) e b) da molécula de Prolina.

IV. VERIFY 3D

O VERIFY-3D (BOWIE *et al.*, 1991; LUTHY *et al.*, 1992; KABSCH & SANDER, 1983) mede a compatibilidade entre a seqüência de aminoácidos de uma proteína e o modelo da sua estrutura tridimensional, usando um perfil 3D. Para tanto, o programa utiliza a seguinte metodologia: 1) reduz a estrutura 3D a uma seqüência de ambientes químicos (um para cada resíduo) categorizados entre 18 possibilidades de acordo com a área do resíduo exposta ao solvente, com a fração de contatos feitos com átomos polares e a estrutura secundária local (perfil 3D), 2) utilizando uma matriz (chamada matriz 3D-1D (BOWIE *et al.*, 1991)) que descreve a probabilidade de se encontrar cada um dos 20 aminoácidos em cada uma das 18 classes de ambientes químicos, o programa determina a probabilidade, por resíduo, de acordo com o tipo do aminoácido e a natureza do ambiente químico calculado no passo anterior. A matriz 3D-1D é previamente conhecida e derivada de uma análise de estruturas conhecidas de boa qualidade (BOWIE *et al.*, 1991; LUTHY *et al.*, 1992).

O resultado é uma medida da compatibilidade entre a seqüência e sua estrutura 3D descrita pelo seu perfil tridimensional. LUTHY *et al.* (1992) determinaram empiricamente o índice global esperado de $S_{\text{calc}} = \exp(-0.83 + 1.008 \times \ln(L))$ de compatibilidade entre a seqüência e a estrutura 3D onde L é o comprimento da seqüência e S_{calc} representando, então, a soma das probabilidades individuais dos resíduos. LUTHY *et al.* (1992) também sugere um limite de $0.45 \times S_{\text{calc}}$ para a confiabilidade da compatibilidade seqüência/estrutura. Valores menores que $0.45 \times S_{\text{calc}}$ indicam uma estrutura incorreta, valores em torno de $0.45 \times S_{\text{calc}}$ podem estar corretas, porém possuem qualidade questionável, e valores acima de $0.45 \times S_{\text{calc}}$ indicam uma estrutura correta. O VERIFY-3D requer como entrada um arquivo de coordenadas atômicas no formato “pdb”.

V. RMSD

O RMSD pode ser calculado de duas maneiras: (i) uma função de todos os átomos de uma proteína (ENGH & HUBER, 1991 – Equação 2) e (ii) uma função de algum subconjunto dos átomos, como a estrutura principal da proteína ou posições de C_{α} apenas (KABASH & SANDER, 1983 – Equação 3). O parâmetro mais comum que expressa a diferença entre duas estruturas protéicas é o RMSD, ou desvio médio quadrático, em posições atômicas entre as duas estruturas.

Quando o RMSD é calculado como uma função de todos os átomos de uma proteína, podemos adotar a definição de que o RMSD serve para analisar a medida da variação da posição de cada átomo de uma estrutura com relação a um vetor

aceito, e é denominado RMSD da geometria ideal (KEARSLEY, 1989). O RMSD da geometria ideal utiliza como parâmetros o comprimento e ângulo de ligação ideal, derivados de uma análise detalhada (ENGH & HUBER, 1991) de estruturas de pequenas moléculas em Cambridge Structural Database (CSD) (ALLEN *et al.*, 1979). O programa utilizado para executar os cálculos de comprimento e ângulos de ligação foi o programa X-PLOR (SCHWIETERS *et al.*, 2003; BRUNGER, 1992).

Quando o RMSD é calculado como uma função de algum subconjunto dos átomos é comum a utilização de um subconjunto de átomos da proteína, quando duas estruturas protéicas são comparadas, elas não serão idênticas entre si em seqüência e, por isso, os únicos átomos entre comparação de posição (um-a-um) que podem ser efetuados são os átomos da cadeia principal (C_{α} - C_{α}). Neste contexto, abordar a orientação de uma estrutura molecular se torna importante. Uma vez que as estruturas de proteínas são geralmente descritas em coordenadas atômicas cartesianas, e surgem como uma orientação incorporada em relação ao espaço. O RMSD é uma função da distância entre os átomos em uma estrutura e os mesmos átomos em outra estrutura. Portanto, se uma molécula começar em uma posição diferente do sistema de coordenadas de referência com relação à outra molécula, o RMSD entre as duas proteínas será calculado independentemente.

$$rmsd_{C_{\alpha}-C} = \sqrt{\sum_{j=1}^N (d_j - d_{C_{\alpha}-C})^2} / N$$

Eq. 2

$$rmsd = \sqrt{\sum_{j=1}^N d_j^2} / N$$

Eq. 3

Para computar RMSDs significativos, as duas estruturas em consideração devem primeiro ser superpostas, desde que possível. A superposição das estruturas de proteínas começa geralmente com uma comparação de seqüências. A comparação de seqüências define as relações um-a-um entre os pares de átomos onde o RMSD é computado. As relações átomo-a-átomo, para fins de comparação de estruturas, podem ocorrer entre resíduos que não estão na mesma posição relativa na seqüência de aminoácidos. As inserções e deleções da seqüência podem forçar duas seqüências a ficarem sem registro entre si, enquanto a arquitetura central das duas estruturas permanece similar.

Uma vez definida as relações átomo-a-átomo entre duas estruturas, com um programa de sobreposição como o LSQKAB do pacote CCP4 (COLLABORATIVE COMPUTATIONAL PROJECT N° 4., 1994) alcançamos uma superposição ótima entre as duas estruturas, isto é, a superposição com o menor RMSD possível. A sobreposição de um par de átomos pode perfeitamente deixar outro par de átomos à parte. Os algoritmos de superposição otimizam a orientação e a posição espacial de duas moléculas (translação e rotação) entre si. Uma vez efetuadas as superposições ótimas de todos os pares de estruturas, os valores de RMSD que são computados como resultado, podem ser comparados entre si, já que as estruturas foram removidas para a mesma estrutura de referência antes de fazer os cálculos de RMSD (GIBAS & JAMBECK, 2001).

APÊNDICE B – Produção bibliográfica

Citação:

Da Silveira, N.J.F., Uchôa, H.B., Pereira, J.H., Canduri, F., Basso, L.A., Palma, M.S., Santos, D.S. & De Azevedo Jr., W.F. (2005) Molecular models of protein targets from *Mycobacterium tuberculosis*. *J. Mol. Model.* **11**:160-166.