UNIVERSIDADE
DE SÃO PAULO
Instituto de Física de São Carlos

APLICAÇÃO DA BIOINFORMÁTICA NOS ESTUDOS DOS GENES E ENZIMAS ENVOLVIDOS NA SÍNTESE DA GOMA FASTIDIANA PRODUZIDA PELA *Xylella fastidiosa*

João Renato Carvalho Muniz

Dissertação apresentada ao Instituto de Física de São Carlos, da Universidade de São Paulo, para obtenção do título de Mestre em Ciências - Física Aplicada com opção: Física Biomolecular.

Orientadora: Dra. Dulce Helena F. de Souza

São Carlos - SP
2003

DEDICATÓRIA

Ao meu pai, Wilson Renato Muniz. À minha mãe, Maria Lúcia P. C. Muniz. À minha tia, Maria Eugênia P. Carvalho. Às minhas irmãs, Marília e Luciana.

AGRADECIMENTOS

À Dra Dulce Helena Ferreira de Souza, pela orientação, compreensão, amizade, ricos ensinamentos, apoio, ajuda e paciência.

Aos professores Richard C. Garratt, Otavio Henrique Thiemann, Glaucius Oliva e Igor Polikarpov pelas discussões e dicas.

Ao colega Marcelo Castilho, pela importante contribuição no estudo de *docking* e discussões subseqüentes.

A TODOS os colegas e amigos (todos mesmo!) do grupo de cristalografia e de biofísica do IFSC, pelo companheirismo, atenção e sempre prestativos para boas conversas e ensinamentos, garantindo assim um convívio alegre e hospitaleiro em todos os congressos, cursos, festas e churrascos.

A todas as pessoas que de maneira direta ou indireta contribuíram para o andamento desse trabalho.

Aos sempre companheiros notívagos AndRe e Sandrinha.

A TODOS os meus colegas e amigos “EXTRA” cristalografia.

Aos funcionários do Instituto de Física de São Carlos.

E a você que está lendo essa dissertação (ou pelo menos os agradecimentos!).

SUMÁRIO

LISTA DE FIGURAS

• Todas as figuras de representações moleculares e estruturais em geral foram feitas utilizando-se o programa PYMOL v. 0.86 (Warren L. DeLano Scientific, San Carlos, CA, USA). http://www.pymol.org

LISTA DE TABELAS

RESUMO

Xylella fastidiosa é uma bactéria Gram-negativa, limitada ao xilema das plantas e o agente causador de diversas doenças em importantes plantações como citros, videiras, mirta, amêndoa, arbustos e café. Em citros, *X. fastidiosa* causa a Clorose Variegada dos Citros (CVC) ou "amarelinho". Nove enzimas (GumB, C, D, E, F, H, J, K e M) estão envolvidas nas etapas biossintéticas de um polissacarídeo extracelular (EPS), chamado de goma fastidiana, um dos mecanismos envolvidos na patogênese da bactéria. Essas enzimas catalisam reações de adição de açúcares, polimerização e exportação do EPS através da membrana da bactéria. No presente trabalho, ferramentas de bioinformática foram utilizadas para o estudo e entendimento da biossíntese da goma fastidiana. As nove enzimas foram estudadas quanto ao seu conteúdo de estrutura secundária, análise de hidrofobicidade e das regiões transmembrânicas, classificação quanto as suas funções. A construção de modelos estruturais para as enzimas Gums através de comparação por homologia seqüencial mostrou ser um processo impossível, devido a falta de moléculas homólogas com estruturas tridimensionais conhecidas. Por outro lado, métodos de reconhecimento de enovelamento mostraram bons resultados e comparações entre as estruturas secundárias das enzimas Gums foram calculadas com a utilização dos programas GenThreader e THREADER 3.3. Modelos tridimensionais para as enzimas GumB, GumK, GumM, GumJ e GumC foram construídos com o programa MODELLER 6.0a e validados com o programa Procheck e VERIFY 3D. Para construção do modelo da GumH (enzima que catalisa a adição da GDP-manose em um lipídio carreador poliprenol), o GenThreader encontrou similaridades quando comparada a MurG (*E.coli*), 2-epimerase (*E. coli*), GtfB (*Amycolatopsis orientalis*) e β-GT de fago T4. Todos os modelos são bastante semelhantes e compostos por dois domínios (α/β), ambos similares ao motivo de ligação de nucleotídeos *Rossmann fold* e separados por uma fenda profunda, que, provavelmente, forma o sítio de ligação da GDP-manose. Estudos da interação entre proteína e substrato foram obtidos com a utilização do programa FLO. O alinhamento seqüencial da GumH com outras onze glicosiltransferases mostrou regiões bastante conservadas, incluindo o motivo EX_7E presente no sítio de ligação do substrato na proteína. Considerações a respeito das interações do substrato GDP-manose com a enzima GumH e do mecanismo da reação foram feitas. Essas análises enfatizam o modelo obtido para a GumH, que representa a primeira estrutura proposta para as enzimas envolvidas na síntese da goma fastidiana.

ABSTRACT

Xylella fastidiosa is a xylem-dwelling, insect-transmitted gamma-protobacterium that causes pathogenicity in citrus plants and many others important crops such as grapevine, periwinkle, almond, oleander and coffee. In citrus plants, *X. fastidiosa* causes citrus variegated chlorosis (CVC) or "amarelinho". Nine enzymes (GumB, C, D, E, F, H, J, K and M) are involved in the biosynthetic pathway of an exopolysaccharide (EPS) called fastidian gum which could be involved in the pathogenicity of the bacterium. These enzymes catalyses sugars addition reactions, polymerization and discharge of the EPS through the bacteria's membrane. We have used bioinformatic tools to study these enzymes and to understand the gum biosynthesis. The nine enzymes were studied regarding to its secondary structure content, analysis of hidrophobicity and transmembrane regions, and yet function classification. The construction of structural models using sequential homology was shown to be impossible, due to the necessity of homologues molecules whose three-dimensional structures are known. On the other hand, pairwises comparisons of secondary structures showed good results and were realized with GenThreader and THREADER 3.3 programs. Three-dimensional structures to GumB, GumK, GumM, GumJ and GumC enzymes were constructed using MODELLER 6.0a and validated with Procheck and VERIFY 3D programs. To construct the model of GumH (enzyme that catalyse the addiction of a GDP-mannose on a polyprenol phosphate carrier), GenThreader found folding similarities when compared to MurG and UDP-Acetylglucosamine 2-Epimerase (from *E. coli*), GtfB (from *Amycolatopsis orientalis*) and β-GT (from T4 phage). The models are very similar consisting of two α/β open sheet domains, both alike in topology to the Rossmann nucleotide-binding folds, and separated by a deep cleft which probably forms the GDP-mannose binding site. Studies of the interaction between enzyme and docked substrate were carried out using the FLO program. The sequence alignment between GumH and another eleven glycosiltransferases showed several preserved regions including the EX_7E motif present on the substrate binding site. The interactions between enzyme-GDP-mannose substrate and the mechanism of the reaction were studied. These analyses emphasize the three-dimensional model constructed for GumH that represents the first structural information for enzymes involved in fastidian gum synthesis.

CAPÍTULO 1

INTRODUÇÃO

Devido às próprias características do território brasileiro, como grande área e terras novas e férteis, a agricultura ainda tem alta representatividade na economia. Alguns tipos de plantações, como soja, café e laranja, além de abastecer o mercado interno, contribuem de forma substancial nas exportações.

A cultura da laranja pode ser considerada uma grande fonte econômica quando comparada às demais produções agrícolas. No estado de São Paulo (Brasil) e no estado da Flórida (EUA), estão localizadas as principais áreas produtoras de laranjas do mundo. Em uma área de aproximadamente 630.200 ha no estado de São Paulo há 164 milhões de árvores, que resultam em uma produção anual de 374 milhões de caixas de laranjas, correspondendo a 87% da produção nacional e 30% da produção mundial. Mais de 70% das laranjas produzidas em São Paulo abastecem as companhias produtoras de suco de laranja representando 28% da produção de suco *in natura* (não concentrado) no mercado interno. Para o Brasil, esse fato é muito relevante visto que a exportação de suco de laranja representa aproximadamente 10% das divisas geradas pelas exportações (US$ 1 bilhão dos US$10 bilhões anuais) (www.laranjabrasil.com.br).

Apenas no Estado de São Paulo, existem 22 indústrias beneficiadoras de laranja que empregam cerca de 400.000 trabalhadores distribuídos em 204 cidades (www.fundecitros.com.br). É uma indústria grande, mas que tem passado por grandes dificuldades, pois os produtores de laranja têm convivido com doenças e pragas que podem comprometer uma safra toda ou até mesmo toda a plantação. No Brasil, as doenças que mais atacam a plantação de cítricos são o Cancro Cítrico e a Clorose Variegada dos Citros (CVC), esta última também conhecida como 'amarelinho'. No início de 2003 uma nova doença chamada "morte súbita" começou a aparecer em algumas plantações de laranjas. Ainda não se sabe quem é o agente causador dessa doença. Países como Estados Unidos, França e Espanha enfrentam situações semelhantes e buscam saídas empregando recursos humanos e financeiros para eliminá-los.

O agente patogênico da CVC é a bactéria *Xylella fastidiosa*, cujas diferentes linhagens causam doenças em diversos tipos de plantações como o cafezal, a videira, o pessegueiro e a pereira (Beretta *et al.*, 1996; Hendson *et al.*, 2000; Metha e Rosato, 2001).

1.1 Clorose Variegada dos Citros (CVC) ou Amarelinho

A Clorose Variegada dos Citros foi primeiramente detectada em 1987 no estado de São Paulo, Macaubal, na região de São José do Rio Preto e Colina, região de Barretos (regiões norte e noroeste do estado, respectivamente) (Rossetti *et al.*, 1990; Chang *et al.*, 1993; Purcel *et al.*, 1996 e Souza Pinto *et al.*, 2001).

A doença então se espalhou rapidamente pela técnica do enxerto com ramos contaminados e por insetos vetores, tornando-se largamente distribuída em todas as regiões citrícolas paulistas. A CVC é agora a maior preocupação das indústrias nacionais beneficiadoras de citros e considerada a mais devastadora. Segundo recente levantamento da Fundecitros nos próximos três anos o estado de São Paulo deverá perder 42 milhões de pés de laranja – o equivalente a 24% do total de árvores no Estado – devido ao amarelinho. Atualmente, cerca de 10 milhões de árvores estão sendo erradicadas, com prejuízo aproximado de R$ 650 milhões.

Os principais sintomas da doença são o surgimento de manchas amarelas nas folhas das plantas que progridem em toda sua extensão levando-as à necrose (figuras 1.1 e 1.2). Os frutos amadurecem mais cedo, produzindo, portanto, menor quantidade de suco por fruto. Ocorre também um endurecimento da casca das laranjas doentes podendo danificar as máquinas durante o processamento do suco (Rossetti *et al.*, 1990).

Figura 1.1: Estágios da CVC no fruto da laranjeira. À esquerda, fruto saudável e à direita, fruto já seco, endurecido e atrofiado (modificado de www.fundecitros.com.br).

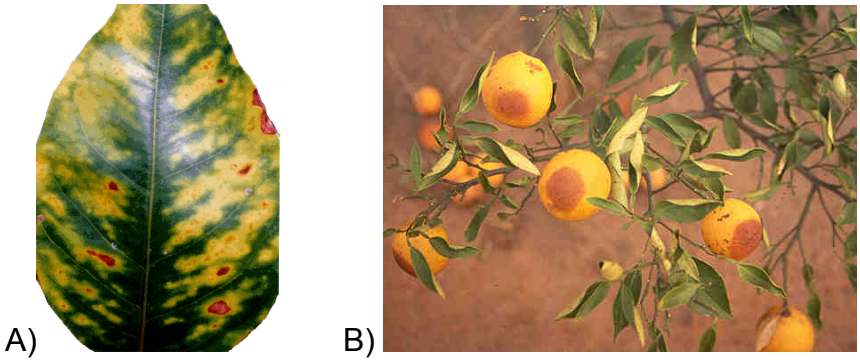


Figura 1.2: Necrose de folhas e frutos devido a CVC. A figura **A**, mostra as manchas amareladas e marrons que marcam o estágio de necrose da folha. A figura **B** mostra um ramo carregado de frutos em estágio já avançado da doença (modificado de www.fundecitros.com.br).

As plantas em sua fase inicial de contaminação mostram sintomas da doença apenas em uma região específica da árvore, aquela onde houve contato com o vetor, mas após um período de infecção de cerca de três anos, a CVC já se espalhou por toda a árvore. A CVC ocorre em plantas de qualquer idade, mas os sintomas mais severos ocorrem entre os dois e seis anos de idade.

1.2 A *Xylella fastidiosa*

X. fastidiosa é uma bactéria Gram–negativa que se aloja exclusivamente no xilema das plantas infectadas. As figuras 1.3 **A** e **B** ilustram o xilema da planta repleto de colônias de *X. fastidiosa*.

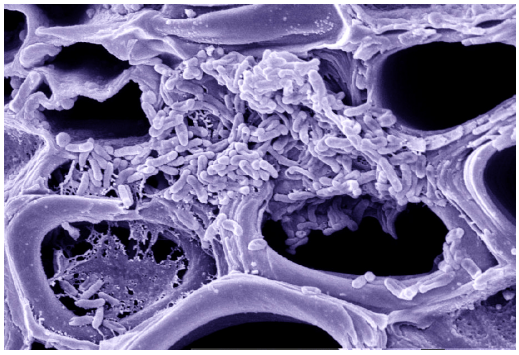
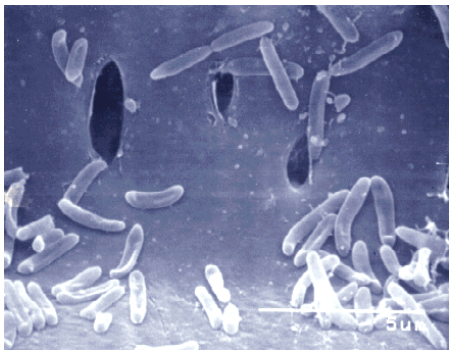
A)  B) 

Figura 1.3: Microscopia eletrônica de um vaso do xilema. Figura **A**, colônias de *X. fastidiosa* aderidas às paredes (modificado de http://aeg.lbi.ic.unicamp.br/xf/). Figura **B**, detalhe de *X. fastidiosa* aderidas à parede do xilema (modificado de www.fundecitros.com.br).

O termo "bactéria limitada ao xilema" tem sido usado para descrever patógenos procarióticos de plantas difíceis de serem isolados por procedimentos padrões de bacteriologia (Purcell e Hopkins, 1996). Outra característica importante de *X. fastidiosa*, e que deu origem ao nome 'fastidiosa', é a lenta velocidade de crescimento e a necessidade de um meio complexo para seu desenvolvimento. A disseminação da bactéria nas plantas se dá através de homópteros (cigarrinha) que, com seu aparelho bucal sugador alimentam-se do xilema das plantas contaminadas e adquirem a bactéria, que passa a alojar-se no intestino deste hospedeiro (Brlansky *et al.*, 1983 e Roberto *et al.*, 1996). Uma das espécies de cigarrinhas ou gafanhoto, transmissor da CVC, é mostrada na figura 1.4.

Figura 1.4: Exemplo de agente transmissor de CVC. A *Parathona gratiosa*, é uma das onze espécies de cigarrinhas consideradas vetores da CVC (www.fundecitros.com.br).

Embora todas as bactérias fastidiosas, Gram-negativas e limitadas ao xilema tenham sido incluídas na espécie *X. fastidiosa*, existe variabilidade suficiente de linhagens que justifica a separação em subespécies. Essas diferentes linhagens atacam plantações causando doenças importantes do ponto de vista econômico como, por exemplo, videiras (*doença de Pierce* - PD), mirta, amêndoa, café (Hendson *et al.*, 2000 e Van Sluys *et al.*, 2003), alfafa anã, olmo, plátano, amora (Monteiro *et al.*, 2001) e crucíferas, como repolho, rabanete e nabo (Williams, 1980).

A linhagem de *X. fastidiosa* que causa a CVC ou "amarelinho" teve seu genoma seqüenciado em estudos realizados por diversos laboratórios de pesquisa do Estado de São Paulo e financiado pela Fapesp (Fundação de Amparo à Pesquisa do Estado de São Paulo) e pela Fundecitros (Fundo de Defesa da Citricultura) (Simpson *et al.*, 2000). Este estudo foi um marco no meio científico por ter seqüenciado, pela primeira vez, o genoma de uma bactéria fitopatogênica e por ser o projeto que iniciou o Programa Genoma da Fapesp (www.fapesp.br). Mais recentemente, outras duas linhagens de *X. fastidiosa* causadoras de doenças em amendoeiras e loureiros também tiveram seus genomas seqüenciados (Bhattacharyya *et al.*, 2002).

O estudo do genoma e mais precisamente dos fatores que contribuem para a patogenicidade da *X. fastidiosa* são de interesse para a busca de mecanismos ou drogas no combate a CVC nos laranjais. Atualmente, o controle da CVC inclui a poda de galhos infectados, controles químicos dos vetores e melhorias no sistema de irrigação em pomares de citros com sintomas iniciais da CVC (Souza Pinto *et al.*, 2001).

O mecanismo pelo qual a bactéria *Xylella fastidiosa* causa a CVC ainda não é totalmente conhecido, mas muitos genes relacionados a patogenicidade foram encontrados no genoma da *X. fastidiosa* quando comparados aos de outras bactérias (Simpson *et al.*, 2000; Lambais *et al.*, 2000 e Van Sluys, 2003). Entre os fatores relacionados a patogenicidade podem-se destacar os processos de interação célula-célula, degradação da parede celular da planta, homeostase de íons, resposta antioxidante e síntese de toxinas, que devido a sua importância serão brevemente descritos abaixo.

Degradação da parede celular:

A mobilidade da bactéria entre os vasos do xilema é dependente da síntese de celulases e proteases que promovem a degradação da parede celular das plantas infectadas (Rajua e Well *et al.*, 1986; Purcell e Hopkins, 1996). O processo de degradação das paredes celulares promove também a liberação de carboidratos que serão utilizados no crescimento bacteriano no xilema (Lambais *et al.*, 2000). Genes responsáveis pela síntese de celulases e proteases assim como todo o mecanismo necessário para exportar essas moléculas através da membrana, foram encontrados no genoma da *X. fastidiosa*.

Homeostase de íons:

A homeostase de íons está presente no processo regulatório da expressão de genes envolvidos na patogenicidade em bactérias (Vasil e Ochsner, 1999). Diversos genes responsáveis pela codificação de proteínas seqüestradoras de íons de ferro que posteriormente serão transportados

através da membrana, foram encontrados no genoma da *X. fastidiosa* sugerindo que essas proteínas possam ter um papel importante na sobrevivência da bactéria no xilema, pois a concentração intracelular de íons de ferro é rigorosamente controlada na bactéria (Simpson *et al.*, 2000). Em *Xanthomonas campestris*, sob elevadas concentrações de ferro, a produção de enzimas serino proteases e polissacarídeos extracelulares (EPS) é reduzida (Wilson *et al.*, 1998) formando assim um mecanismo de controle da produção do polissacarídeo extracelular.

Reações antioxidantes:

Durante a contaminação pela *X. fastidiosa*, a planta dá início a um processo de defesa por meio da degradação parcial da parede celular, resultando na formação de oxigênios reativos (ROS), tais como superóxidos e peróxidos heterogêneos (Wojtaszek, 1997). A explosão oxidativa induzida sob a infecção pode inibir o crescimento bacteriano e o desenvolvimento da doença. Diversos genes envolvidos na desintoxicação que codificam para catalases, superóxido dismutases, glutationa peroxidases e glutationa S-transferases (Simpson *et al.*, 2000), assim como genes regulatórios como *Ohr* (***O**rganic **h**ydroperoxide **r**esistance gene*) sugerem que a *X. fastidiosa* pode, mesmo que em parte, responder a problemas oxidativos e contra-atacar ao primeiro sinal de resposta da planta devido à infecção causada pelo elemento patogênico (Cussiol *et al.*, 2003).

Síntese de toxinas:

Entre genes que codificam para diversas toxinas, foram identificados na *X. fastidiosa* genes que levarão à síntese de proteínas responsáveis pela

apoptose celular do hospedeiro, causando a liberação dos componentes intracelulares (Weinrauch e Zychlinsky, 1999).

O local em que a *X. fastidiosa* induz a apoptose nas células dos citros ainda é desconhecido, mas é certo que esse processo ocorre como prevenção ao mecanismo de defesa da célula hospedeira (Lambais *et al.*, 2000).

Regulação da patogenicidade:

O longo período necessário para a observação dos sintomas provocados pela CVC na planta contaminada, sugere que a síntese dos fatores patogênicos seja regulada por um número mínimo de sensores, equivalentes aos encontrados em um grande número de bactérias Gram-negativas como *Agrobacterium tumefaciens*, *Erwinia carotovora*, *Pseudomonas aeruginosa*, *P. aureofasciens*, *Escherichia coli* e *Rhizobium leguminosarum* (Cubo *et al.*, 1992; Fucqua e Winans, 1994; Fucqua *et al.*, 1996 e Pierson *et al.*, 1998).

Interação célula-célula:

O contato célula-célula é um processo essencialmente importante para o estabelecimento da interação planta-elemento patogênico e desenvolvimento da doença. Durante o processo de infestação e desenvolvimento da CVC, a bactéria interage com as células do intestino do inseto vetor e as células da planta hospedeira assim como as de outras bactérias, formando microcolônias.

Gene que codifica uma proteína conhecida como *fimbria IV* (proteína filamentosa e polar) foram encontrados no genoma da *X. fastidiosa*

(Simpson *et al.*, 2000). Essa proteína tem como papel principal a fixação e mobilidade da bactéria na parede do intestino do inseto hospedeiro através de interações eletrostáticas (polares), além de contribuir na formação de colônias e sobrevivência das células bacterianas aderindo-se à parede do xilema (Lambais *et al.*, 2000).

Outras moléculas envolvidas na interação célula-célula são os polissacarídeos extracelulares (EPS), que promovem a adesão das bactérias à superfície dos vasos vasculares da planta e também podem estar envolvidos no reconhecimento e determinação da natureza da associação planta-bactéria (Dharmapuri e Sonti, 1999).

Polissacarídeos extracelulares: envolvimento na patogenicidade

Diversos estudos têm mostrado que a produção de polissacarídeos extracelulares pode estar relacionada à patogenicidade de certas bactérias (Chan e Goodwin, 1999; Dow e Daniels, 2000; Lambais *et al.*, 2000; Garcia-Ochoa *et al.*, 2000 e Vojnov *et al.*, 2002).

A bactéria *Xanthomonas campestris* pv. campestris, que causa apodrecimento de crucíferas (como couve, nabo e mostarda), tem sido estudada em detalhes como um organismo modelo para investigação da genética de fitopatogenicidade bacteriana. Esta bactéria produz um polissacarídeo extracelular, chamado goma xantana, que tem sido extensivamente estudado (um *review* sobre a goma xantana foi publicado recentemente por Garcia-Ochoa *et al.*, 2000). A goma xantana é atóxica e não provoca nenhum tipo de disfunção no desenvolvimento normal dos humanos, tão pouco irritações na pele ou nos olhos. Sua aplicação atinge as

mais diversificadas áreas, desde a indústria alimentícia até a farmacêutica. Dentre suas aplicações pode-se mencionar a utilização da goma como emulsificante, espessante, agente dispersante e estabilizante, em formulações farmacêuticas, cosméticos e produtos agrícolas. É também utilizada nas indústrias têxteis (cola para tecidos), cerâmicas (como revestimento), em explosivos plásticos, como removedores de ferrugem e no enriquecimento do óleo nas indústrias petrolíferas (Baird *et al.*, 1983 e Garcia-Ochoa *et al.*, 2000).

Polissacarídeos extracelulares do tipo goma xantana apresentam três importantes características: são altamente hidratados, podendo assim garantir proteção contra o ressecamento e moléculas hidrofóbicas; são altamente aniônicos, permitindo a concentração de nutrientes e imobilização de substâncias tóxicas e finalmente, muito aderentes, proporcionando ao organismo a adsorção de superfícies biológicas facilitando a formação de colônias (Chan e Goodwin, 1999).

Diversos estudos têm mostrado o efeito da goma xantana na patogenicidade da bactéria *X. campestris*. Mutações no gene gumD, a primeira enzima da via sintética da goma xantana, e alterações nos últimos estágios da biossíntese da goma reduziram a virulência em brócolis e diminuíram a agressividade da bactéria na planta (Chou *et al.*, 1997 e Katzen *et al.*, 1998).

Genes homólogos aos encontrados em *X. campestris* envolvidos na síntese e exportação da goma xantana também foram encontrados no genoma da *X. fastidiosa* (Simpson *et al.*, 2000). Foram observados dois grupos de genes envolvidos na biossíntese da goma: os genes xpsIII, xpsIV

e xpsVI, que codificam para a síntese da glicose, ácido glucorônico e manose, respectivamente (Harding *et al.*, 1987; Hötte *et al.*, 1990 e Köplin *et al.*, 1992) e um *operon* (Operon gum) que agrupa os genes (*gum* ou xpsI), responsáveis pela expressão das enzimas que catalisam as reações de síntese e exportação da goma (que tem sido chamada de goma fastidiana) (Silva *et al.*, 2001).

Na *X. fastidiosa* os genes *gum* são agrupados em um Operon constituído por um conjunto de 12 kb contendo 9 genes designados gumB, gumC, gumD, gumE, gumF, gumH, gumJ, gumK e gumM. Estudos com a goma xantana têm levado a resultados sobre a função das enzimas Gums. Com base em comparações seqüenciais das Gums (de *Xanthomonas* e de *Xylella*) pode-se deduzir as funções de cada enzima. O gene gumD codifica a enzima glicosiltransferase I, responsável pela transferência da glicose-1-fosfato ao fosfato poliprenol para formar o monossacarídeo-lipídio; gumM codifica a enzima glicosiltransferase II que catalisa a adição da segunda glicose para formar o dissacarídeo-lipídio; gumH codifica a enzima glicosiltransferase III que catalisa a adição da manose formando um trissacarídeo-lipídio; gumK codifica a enzima glicosiltransferase IV que catalisa a adição do ácido glucorônico para formar o tetrassacarídeo-lipídio; gumF codifica a enzima acetiltransferase que catalisa a acetilação da manose; gumB, C, E e J estão envolvidos na polimerização e secreção do polissacarídeo pela membrana da bactéria. A figura 1.5 mostra o esquema proposto de todas as etapas do mecanismo de produção da goma fastidiana.

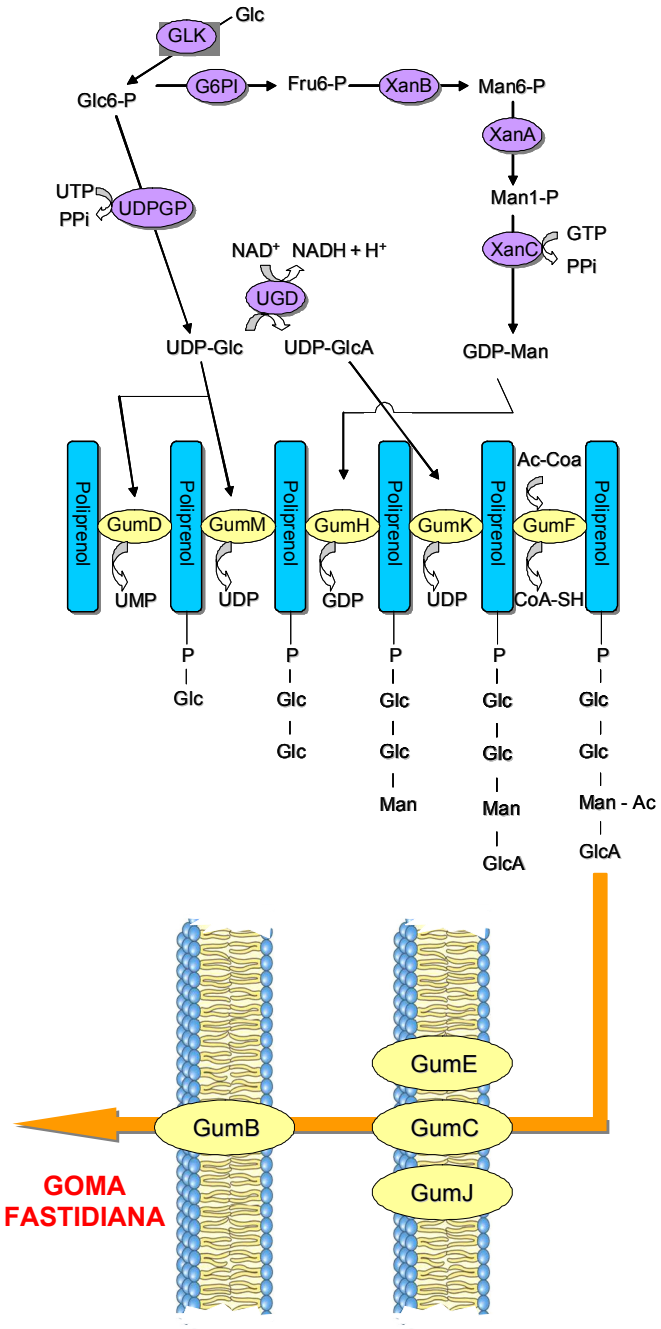


Figura 1.5: Mecanismo proposto para a biossíntese da goma fastidiana. As enzimas Gums estão mostradas em ovais amarelos: glicosiltransferases I, II, III e IV (GumD, M, H e K respectivamente); acetiltransferase I (GumF), as proteínas GumE, C, J e B relatadas como prováveis enzimas envolvidas no processo de polimerização e transporte da goma. As setas para as reações seqüenciais que adicionam açúcares ao poliprenol (lipídio carreador em azul) não são mostradas. Glc, D-glicose; GlcA, D-ácido glucorônico; Man, D-manose; Ac, éster acetil; PPi, pirofosfato; GLK, hexoquinase; G6PI, glicose fosfato isomerase; XanB, fosfomanoseisomerase; XanC, GDP-manose pirofosforilase; XanA, fosfoglicomutase; UDPGP, UDP-glicose pirofosfato; UGD, glucorônico-1-fosfato uridil transferase (adaptado de Silva *et al.*, 2001).

1.3 Estruturas das gomas xantana e fastidiana

A estrutura da goma xantana consiste de um β-1,4-D-glicose com uma cadeia lateral trissacarídica composta de uma manose, um molécula de ácido glucorônico e uma segunda manose. Os resíduos de manose são acetilados e piruvilados em sítios específicos, mas em diferentes graus (Jansson *et al.*, 1975; Cadmus *et al.*, 1976 e Stankowski *et al.*, 1993). Estudos mostram que a subunidade pentassacarídica é primeiro sintetizada, isto é, ligada a uma molécula de lipídio carreador (pirofosfato-poliprenol), de maneira seqüencial a partir de precursores D-glicose, D-manose, ácido glucorônico, acetil coenzima A e fosfoenolpiruvato (Ielpi *et al.*, 1981a, 1981b e 1983 e Harding *et al.*, 1987). As unidades pentassacarídicas são subseqüentemente polimerizadas e transportadas para fora da bactéria. Essa biossíntese é um processo complexo e controlado por diversas enzimas (Sutherland, 1977).

A estrutura da goma fastidiana deve ser composta por repetições de β-1,4-D-glicose e uma cadeia lateral formada por um dissacarídeo, uma molécula de manose acetilada e outra de ácido glucorônico. O esquema representado na figura 1.5 mostra a enzima GumD como sendo a glicosiltransferase I, responsável pela adição da glicose-1-fosfato ao fosfato poliprenol, formando o monossacarídeo-lipídio (P-Glc). A próxima enzima é a GumM, glicosiltransferase II, que adiciona a segunda glicose, formando um dissacarídeo-lipídio (P-Glc-Glc). Uma vez formado o dissacarídeo, a cadeia lateral começa a ser formada com a adição de uma manose pela manosiltransferase I, a GumH (P-Glc-Glc_Man).

A próxima etapa é a adição de uma molécula de ácido glucorônico à manose, catalisada pela GumK (P-Glc-Glc_Man_GlcAc). Finalmente, a GumF, uma acetiltransferase, catalisa a acetilação da manose. As Gums B, E, C e J, possivelmente, estão envolvidas na polimerização e exportação do polissacarídeo pela membrana bacteriana.

O tetrassacarídeo é formado por duas moléculas de glicose catalisadas pela GumD e GumM, uma molécula de manose catalisada pela enzima GumH e uma molécula de ácido glucorônico catalisada pela enzima GumK. A molécula de manose já acetilada pela enzima GumF está separada das demais apenas para facilitar a visualização de todo o processo (figura 1.6).

Todas essas etapas estão representadas nas figuras 1.6, 1.7 e 1.8.

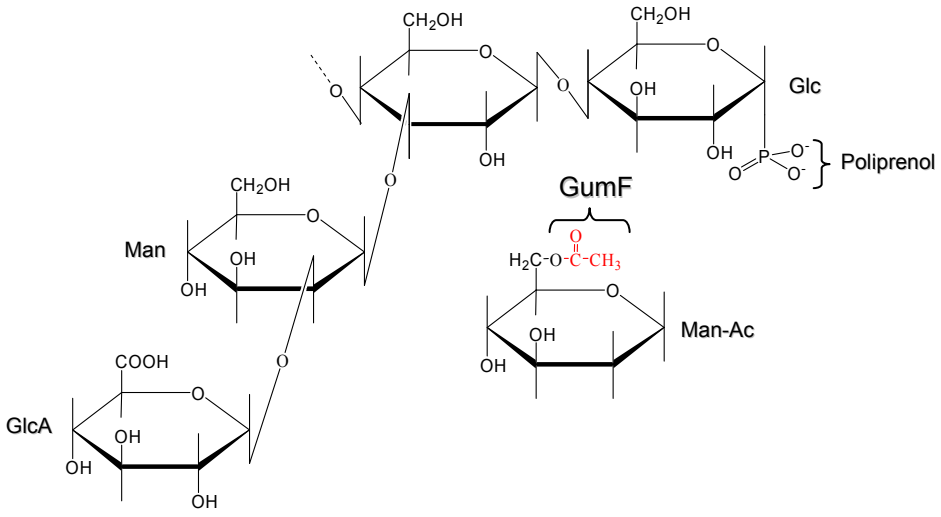


Figura 1.6: Estrutura tetrassacarídica da goma fastidiana. A figura ilustra a estrutura do tetrassacarídeo em formação, já que a molécula de manose ainda não foi acetilada pela GumF, mostrado à parte em destaque.

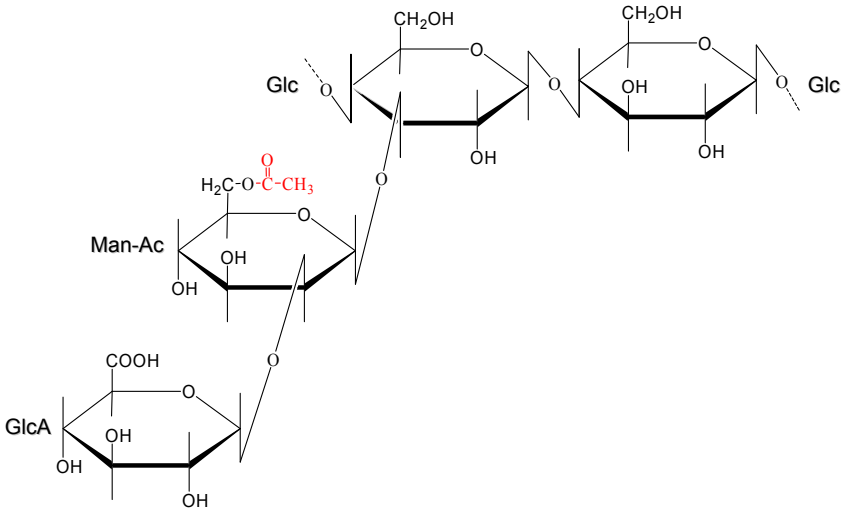


Figura 1.7: Tetrassacarídeo acetilado. O tetrassacarídeo está acetilado pela GumF (radical destacado em vermelho), formando uma unidade repetidora durante o processo de polimerização (figura 1.8).

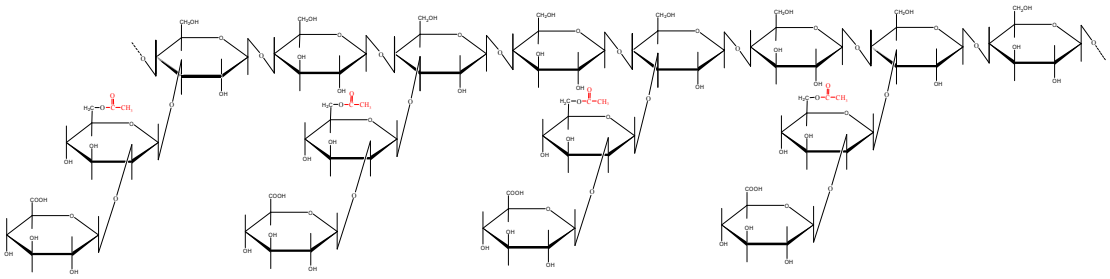


Figura 1.8: Polimerização da goma fastidiana. Uma vez formadas as unidades tetrassacarídicas fundamentais, dá-se início a polimerização do tetrassacarídeo, resultando na formação da goma fastidiana.

Temos visto que, frente a necessidade de desenvolvimento de mecanismos de controle da CVC, o estudo dos fatores envolvidos na patogenicidade da *X. fastidiosa* é de importância fundamental. A análise da goma fastidiana e sua relação com a CVC podem trazer esclarecimentos e avanços neste sentido.

Neste contexto, no presente trabalho, nos propomos a realizar o estudo das enzimas envolvidas na síntese da goma fastidiana utilizando ferramentas de bioinformática.

1.4 Referências bibliográficas

Baird, J. K., Sandfordm P. A. and Cottrell, L. W. (1983) *Biol. Technology*, **1**, 778-783.

Bhattacharyya, A., Stilwagen, S., Ivanova, N., D'Souza, M., Bernal, A., Lykidis, A., Kapatral, V. *et al.* (2002) *Microbiology*, **99**, 12403-12408.

Beretta, M. J. G., Harakava, R. and Chagas, C. M. (1996) *Plant Dis.*, **80**, 821.

Brlansky, R. H., Timmer, I. W., French, W. J. and McCoy, R. E. (1983) *Phytopathology*, **73**, 530-535.

Cadmus, M. C., Rogovin, S. P., Burton, K. A., Pittsley, J. E., Knutson, C. A. and Jeanes, A. (1976) *Can. J. Microbiol.*, **22**, 942-948.

Chan, J. W. Y. F. and Goodwin, P. H. (1999) *Biotechnology Advances*, **17**, 489-508.

Chang, C. J., Garnier, M., Zreik, L. Rossetti, V. and Bove, J. M. (1993) *Curr Microbiol*, **27**, 137-142.

Chou, F., Chou, H., Lin, Y., Yang, B., Lin, N., Weng, S. and Tseng, Y. (1997) *Biochem. and Biophys. Res. Commun.*, **233**, 265-269.

Cubo, M. T., Economou, A., Murphy, G., Johnston, A. W. B. and Downie, J. A. (1992) *J Bacteriol*, **174**, 4026-4035.

Cussiol, J. R., Alves, S. V., Oliveira, M. A. and Netto, L. E. (2003) *J Biol Chem*, ***in Press***.

Dharmapuri, S. and Sonti, R. V. (1999) *FEMS Microbiol Lett*, **179**, 53-59.

Dow, J. M. and Daniels, M. J. (2000) *Yeast*, **17**, 263-271.

Fucqua, C., Winams, S. C. and Greenberg, E. P. (1996) *Annu Rev Microbiol*, **50**, 727-751.

Fucqua, C. and Winams, S. C. (1994) *J Bacteriol*, **176**, 2796-2806.

Garcia-Ochoa, F., Santos, V. E., Casas, J. A. and Gómez, E. (2000) *Biotechnology Advances*, **18**, 549-579.

Harding, N. E., Cleary, J. M., Cabanas, D. K., Rosen, I. G. and Kang, K. S. (1987) *J. Bacteriol*, **169**, 2854-2861.

Hendson, M., Purcell, A. H., Chen, D., Smart, C., Guilhabert, M. and Kirkpatrick, B. (2000) *Applied and Enviromental Microbiology*, **67**, 895-903.

Hötte, B., Rath-Arnold, I., Puhler, A. and Simon, R. (1990) *J. Bacteriol.*, **172**, 2804-2807.

Ielpi, L., Couso, R. And Dansert, M. (1981a) *Biochem. Biophys. Res. Commun.*, **102**, 1400-1408.

Ielpi, L., Couso, R. And Dansert, M. (1981b) *FEBS Lett.*, **130**, 253-256.

Ielpi, L., Couso, R. And Dansert, M. (1983) *Biochem. Int.*, **6**, 323-333.

Jansson, P. E., Kenne, L. and Lindberg, B. (1975) *Carbohydr. Res.*, **45**, 275-282.

Katzen, F., Ferreiro, D. U., Oddo, C. G., Ielmini, V., Becker, A., Puhler, A. and Ielpi, L. (1998) *J. Bacteriol.*, **180**, 1607-1617.

Köplin, R., Arnold, W., Hötte, B., Simon, R., Wang, G. and Puhler, A. (1992) *J. Bacteriol.*, **174**, 191-199.

Lambais, M. R., Goldman, M. H. S., Camargo, L. E. A. and Goldman, G. H. (2000) *Current Opinion in Microbiology*, **3**, 459-462.

Metha, A. and Rosato, Y. B. (2001) *Inst. J. Syst. Evol. Microbiol.*, **51**, 311-318.

Monteiro, P. B., Teixeira, D. C., Palma, R. R., Garnier, M., Bové, J-M and Renaudin, J. (2001) *Applied and Envaronmental Microbiology*, **67**, 2263-2269.

Pierson, L. S., Wood, D. W. and Pierson, E. A. (1998) *Annu Rev Phytopathol*, **36**, 207-225.

Purcel, A. H. and Hopkins, D. L. (1996) *Annu. Rev. Phytopathol.*, **34**, 131-151.

Rajua, B. C. and Well, J. M. (1986) *Plant Disease*, **70**, 182-186.

Roberto, S. R., Coutinho, A., De Lima, J. E. O., Miranda, V. S. and Carlos, E. F. (1996) *Fitopatol Brás*, **21**, 517.

Rossetti, V. M., Garnier, J. M., Bové, M. J. G., Beretta, A. R. R., Teixeira, J. A., Quaggio and Dagoberto de Negri, J. (1990) *C. R. Acad. Sci.*, Paris, **310**, 345-349.

Silva, F. R., Vettore, A. L., Kemper, E. L., Leite, A. And Arruda, P. (2001) *FEMS Microbiol. Lett.*, **203**, 165-171.

Simpson, A. J., Reinach, F. C., Arruda, P. *et al.,* (2000) *Nature*, **406**, 151-157.

Souza Pinto, W. B., Basile, G. B. e Gonzales, M. A., (2001) *CATI - Coordenadoria de Assistência Técnica Integral* - Secretaria de agricultura e Abastecimento do Estado de São Paulo.

Stankowski, J. D., Mueller, B. E. and Zeller, S. G. (1993) *Carbohydr. Res.*, **17**, 241-321.

Sutherland, I. W. (1977) Academic Press, Inc., New York.

Van Sluys, M. A., De Oliveira, M. C., Monteiro-Vitorello, C. B., Miyaki, C. Y., *et al.* (2003) *J Bacteriol*, **3**, 1018-1026.

Vasil, M. L. and Ochsner, U. A. (1999) *Mol. Microbiol.*, **3**, 399-413.

Vojnov, A. A., Bassi, D. E., Daniels, M. J. And Dankert, M. A. (2002) *Carbohydrate Research*, **337**, 315-326.

Weinrauch, Y. and Zychlinsky, A. (1999) *Annu Rev Microbiol*, **53**, 155-187.

Williams, P. H. (1980) *Plant Dis.*, **64**, 736-774.

Wilson, T. J. G., Bertrand, N., Tang, J-L, Feng, J-X, Pan, M-Q, Barber, C. E., Dow, J. M. and Daniels, M. J. (1998) *Mol Microbiol*, **28**, 961-970.

Wojtaszek, P. (1997) *Biochem J*, **322**, 681-692.

CAPÍTULO 2

OBJETIVOS

A proposta deste trabalho é a utilização de ferramentas computacionais no estudo dos genes e proteínas envolvidos na síntese da goma fastidiana produzida pela bactéria *Xylella fastidiosa*, agente patogênico da Clorose Variegada dos Citros (CVC ou amarelinho). Espera-se que as informações obtidas neste estudo auxiliem na caracterização da via biossintética da goma, responsável pela obstrução do xilema das plantas, fato que as leva à morte. O total conhecimento desta via tem como objetivo principal a inibição da síntese da goma, pois, uma vez bem estabelecido o sítio ativo e identificados os principais aminoácidos nele contido, torna-se possível o estudo da modelagem de inibidores para o sítio ativo da enzima, o que poderá extinguir ou diminuir a patogenicidade da bactéria.

Objetivos específicos

A proposta deste projeto é obter o máximo de informação sobre os genes e as enzimas envolvidas na biossíntese da goma fastidiana, por meio de ferramentas computacionais.

- Analisar as seqüências do Operon gum e das nove enzimas envolvidas na via biossintética da goma fastidiana utilizando bancos de dados de seqüências de DNA e proteínas.
- Comparar e correlacionar as seqüências com as de outras espécies através de programas específicos.
- Classificar as nove enzimas em famílias de enzimas envolvidas na biossíntese de carboidratos.
- Procurar por estruturas conhecidas de proteínas homólogas às das enzimas da via biossintética da goma.
- Construção de modelos tridimensionais das moléculas.
- Análise do sítio ativo das enzimas.

CAPÍTULO 3

BIOINFORMÁTICA & FERRAMENTAS

A bioinformática é uma área da ciência que utiliza a tecnologia da computação para organizar, analisar e distribuir informações biológicas com a finalidade de responder perguntas mais complexas neste campo da ciência. É uma área de investigação multidisciplinar que pode ser amplamente definida como a interface entre duas ciências, a biologia e a computação, e está impulsionada pelas descobertas de novos genomas e a promessa de uma nova era na qual a investigação genômica poderá ajudar a melhorar a qualidade de vida humana. Avanços nos diagnósticos, tratamentos de doenças, na procura e desenvolvimento de inibidores específicos através de estudos de *docking* e na produção de alimentos geneticamente modificados são exemplos de benefícios mais freqüentes da utilização da bioinformática.

A bioinformática também está envolvida no agrupamento, organização, armazenamento e recuperação das informações biológicas que se encontram em bancos de dados, tais como seqüências de nucleotídeos, de aminoácidos, estruturas e domínios específicos das mais variadas proteínas de diferentes famílias. O processo de análise e interpretação dos

dados é conhecido como biocomputação. Dentre os diversos objetivos da bioinformática e da biocomputação existem dois principais:

- Desenvolvimento e implementação de ferramentas que permitem o acesso, uso e manejo de vários tipos de informações, alinhamentos seqüenciais, estruturais, visualizações de enovelamentos protéicos, etc.
- Desenvolvimento de novos algoritmos e estatísticas, com os quais é possível relacionar partes de um conjunto enorme de dados, como, por exemplo, métodos para localização de um gene dentro de uma seqüência, predizer a estrutura ou função de proteínas e poder agrupar seqüências em famílias relacionadas na busca por homologia ou uma eventual classificação, etc.

Para se ter uma idéia da abrangência da bioinformática, bancos de dados de bases nucleotídicas e protéicas como o NCBI (www.ncbi.nlm.nih.gov/), o PDB (www.rcsb.org/pdb) e o EXPASY (www.expasy.ch), recebem mais de 3.000.000 de acessos diários provenientes de pesquisadores situados ao redor de todo o mundo, seja na busca de alinhamentos seqüenciais ou na utilização de algum tipo de ferramenta de bioinformática que esses servidores provem.

No presente trabalho, técnicas de bioinformática foram utilizadas com o objetivo de se estudar as características das enzimas Gums e na construção de modelos tridimensionais das mesmas. Essas metodologias serão brevemente discutidas neste capítulo.

3.1 Alinhamentos, matrizes e algoritmos

Dentre as ferramentas mais utilizadas na bioinformática encontram-se aquelas relacionadas à comparação seqüencial de DNA ou proteínas.

O estudo de comparação através dos alinhamentos das seqüências pode representar o primeiro passo na busca de informações sobre um gene ou seu produto. Uma vez que em proteínas de uma mesma família a região catalítica ou do sítio ativo tende ser mais bem conservada, um alinhamento feito em um banco de dados onde as seqüências estão agrupadas em diferentes famílias terá como resultado a inclusão da seqüência alvo em uma dessas famílias, como acontece no banco de seqüências Pfam (Bateman *et al.*, 2002).

Alinhamentos de seqüências também podem auxiliar na busca por proteínas homólogas. Nesses casos, o grau de identidade seqüencial ou a predição das estruturas secundárias que os aminoácidos assumirão na seqüência alvo faz parte do processo inicial na construção de estruturas tridimensionais através da modelagem molecular por comparação.

Métodos para alinhamentos de seqüências determinam tipicamente a similaridade entre as seqüências de aminoácidos ou bases nitrogenadas utilizando uma matriz de valores para cada substituição possível durante o alinhamento, ou seja, a permutação entre as bases/resíduos. As matrizes mais utilizadas são baseadas no modelo de Dayhoff (Dayhoff e Eck, 1968) de taxas evolucionárias e no modelo de Needleman e Wunsch (Needleman e Wunsch, 1970) de probabilidades.

O alinhamento pode acontecer em pares seqüenciais ou entre três ou mais seqüências. A comparação entre duas seqüências recebe o nome de "alinhamento de pares", já o alinhamento entre três ou mais seqüências recebe o nome de "alinhamento múltiplo". Os alinhamentos podem ser **globais** ou **locais**. Em cada tipo, os alinhamentos serão validados de acordo com o esquema de valores determinados pela matriz de substituição utilizada a fim de se estimar o grau de similaridade entre as seqüências, discernindo os pares realmente significativos dos não significativos.

Embora diversos métodos tenham sido propostos (Dayhoff e Eck, 1968; Needleman e Wunsh, 1970; McLachlan, 1971; Feng *et al.*, 1985; Rao, 1987, Risler *et al.*, 1988; Smith e Smith, 1990 e George *et al.*, 1990), as matrizes de mutação (ou substituição) de Dayhoff (Dayhoff, 1978; George *et al.*, 1990 e Altschul, 1991) são geralmente consideradas como padrão em programas de alinhamentos (bases ou proteínas) múltiplos ou entre pares. Neste modelo, as taxas de substituições são derivadas dos alinhamentos de seqüências ao menos 85% idênticas. Entretanto, a tarefa mais comum está em envolver matrizes de valores na detecção de muitas outras relações, tais como acúmulo de mutações durante o processo evolutivo ou a semelhança nas propriedades químicas e estruturais, que são apenas inferidas a partir das matrizes de Dayhoff.

Durante a evolução, algumas posições tanto de bases nitrogenadas como de aminoácidos, sofrem mutações, ou seja, bases nitrogenadas ou aminoácidos podem ser inseridos ou deletados de uma seqüência, resultando na formação de lacunas (chamadas *gaps*) entre resíduos de uma determinada seqüência que será então alongada por essas lacunas a fim de

se preservar o maior número de bases/aminoácidos alinhados. Qualquer medição de similaridade deve ser feita em relação a melhor possibilidade de alinhamento entre duas ou mais seqüências.

Uma vez que eventos de inserção e/ou deleção são mais raros quando comparados às substituições de resíduos (mutações), torna-se necessária a utilização de diferentes pesos durante o processo de alinhamento, de modo que o mesmo convirja para o caminho de menor penalidade. As penalidades podem ser divididas em dois tipos: penalidade ao primeiro *gap* a ser inserido na seqüência ou *gap* de abertura (*open gap*) e penalidade a cada novo *gap* inserido na seqüência (*extend gap*). A fim de preservar a seqüência de resíduos com um número menor de inserções e/ou deleções, a penalidade para o *gap* de abertura é maior que o *gap* de extensão, principalmente no alinhamento do tipo global.

Por definição, o alinhamento possuidor do maior valor de similaridade (medida pela qual os alinhamentos são quantificados) será o melhor alinhamento. Basicamente, em um alinhamento, a matriz de substituição *mxn* criada é proporcional ao numero de bases/resíduos. Needleman e Wunsch conceituaram o problema de alinhamento como sendo um problema de Programação Dinâmica (Needleman e Wunsh, 1970).

3.1.1 Tipos de matrizes

Na construção de um alinhamento biologicamente significativo, devem ser levadas em consideração a influência das propriedades químicas dos aminoácidos e a degenerescência do código genético durante o processo evolutivo. Trocas ou substituições quimicamente conservativas tendem a

ocorrer mais freqüentemente que substituições entre aminoácidos quimicamente diferentes. Por exemplo, é muito mais comum encontrar substituições entre uma leucina e uma isoleucina, ambas apolares, do que substituições entre o ácido aspártico, que é negativamente carregado, por uma leucina.

Matrizes de alinhamentos foram construídas com a finalidade de equacionar o peso que cada par de aminoácidos poderá receber quando alinhados. Para cada par de aminoácidos alinhados, um valor correspondente a esse alinhamento é adicionado ao valor final da similaridade entre as seqüências. Sendo assim, o melhor alinhamento obterá um valor final maior, ou seja, acumulou mais pontos durante o alinhamento.

As matrizes mais comumente utilizadas são a PAM 250 (**P**oint **A**ccepted **M**utation) (Dayhoff *et al.*, 1978) e a BLOSUM 62 (**Blo**cks **Su**bstitution **M**atrix) (Henikoff e Henikoff, 1992), ilustradas nas figuras 3.1 e 3.2 respectivamente.

	Cys	Gly	Pro	Ser	Ala	Thr	Asp	Glu	Asn	Gln	His	Lys	Arg	Val	Met	Ile	Leu	Phe	Tyr	Trp
Cys	12																			
Gly	-3	5																		
Pro	-3	-1	6																	
Ser	0	1	1	1																
Ala	-2	1	1	1	2															
Thr	-2	0	0	1	1	3														
Asp	-5	1	-1	0	0	0	4													
Glu	-5	0	-1	0	0	0	3	4												
Asn	-4	0	-1	1	0	0	2	1	2											
Gln	-5	-1	0	-1	0	-1	2	2	1	4										
His	-3	-2	0	-1	-1	-1	1	1	2	3	6									
Lys	-5	-2	-1	0	-1	0	0	0	1	1	0	5								
Arg	-4	-3	0	0	-2	-1	-1	-1	0	1	2	3	6							
Val	-2	-1	-1	-1	0	0	-2	-2	-2	-2	-2	-2	-2	4						
Met	-5	-3	-2	-2	-1	-1	-3	-2	0	-1	-2	0	0	2	6					
Ile	-2	-3	-2	-1	-1	0	-2	-2	-2	-2	-2	-2	-2	4	2	5				
Leu	-6	-4	-3	-3	-2	-2	-4	-3	-3	-2	-2	-3	-3	2	4	2	6			
Phe	-4	-5	-5	-3	-4	-3	-6	-5	-4	-5	-2	-5	-4	-1	0	1	2	9		
Tyr	0	-5	-5	-3	-3	-3	-4	-4	-2	-4	0	-4	-5	-2	-2	-1	-1	7	10	
Trp	-8	-7	-6	-2	-6	-5	-7	-7	-4	-5	-3	-3	2	-6	-4	-5	-2	0	0	17

Figura 3.1: Matriz de valores de alinhamentos PAM 250. A substituição do ácido aspártico por um ácido glutâmico adiciona um valor igual a 3 ao *score* (função de custo) do alinhamento. Substituição da positivamente carregada Lys por uma apolar Pro adiciona o valor -1 a função de custo do alinhamento. Geralmente, quanto mais conservativo for o alinhamento, maior será a contribuição de cada par de aminoácidos na função de custo do alinhamento (Nicholas Jr *et al.*, 1998).

	Gly	Pro	Asp	Glu	Asn	His	Gln	Lys	Arg	Ser	Thr	Ala	Met	Val	Ile	Leu	Phe	Tyr	Trp	Cys
Gly	7																			
Pro	-2	9																		
Asp	-1	-1	7																	
Glu	-2	0	2	6																
Asn	0	-2	2	0	6															
His	-2	-2	0	0	1	10														
Gln	-2	-1	0	2	0	1	6													
Lys	-2	-1	0	1	0	-1	1	5												
Arg	-2	-2	-1	0	0	0	1	3	7											
Ser	0	-1	0	0	1	-1	0	-1	-1	4										
Thr	-2	-1	-1	-1	0	-2	-1	-1	-1	2	5									
Ala	0	-1	-2	-1	-1	-2	-1	-1	-2	1	0	5								
Met	-2	-2	-3	-2	-2	0	0	-1	-1	-2	-1	-1	6							
Val	-3	-3	-3	-3	-3	-3	-3	-2	-2	-1	0	0	1	5						
Ile	-4	-2	-4	-3	-2	-3	-2	-3	-3	-2	-1	-1	2	3	5					
Leu	-3	-3	-3	-2	-3	-2	-2	-3	-2	-3	-1	-1	2	1	2	5				
Phe	-3	-3	-4	-3	-2	-2	-4	-3	-2	-2	-1	-2	0	0	0	1	8			
Tyr	-3	-3	-2	-2	-2	2	-1	-1	-1	-2	-1	-2	0	-1	0	0	3	8		
Trp	-2	-3	-4	-3	-4	-3	-2	-2	-2	-4	-3	-2	-2	-3	-2	-2	1	3	15	
Cys	-3	-4	-3	-3	-2	-3	-3	-3	-3	-1	-1	-1	-2	-1	-3	-2	-2	-3	-5	12

Figura 3.2: Matriz de valores de alinhamentos BLOSUM 62. Assim como a matriz PAM 250, os valores de substituição ou alinhamento preservam a idéia de quanto mais conservativo for o alinhamento, maior será a contribuição de cada par de aminoácidos na função de custo do alinhamento (Nicholas Jr *et al.*, 1998).

Matrizes de substituição, como a PAM 250, foram construídas através da observação da freqüência de troca ou substituição em um grande conjunto de seqüências protéicas baseadas em uma árvore filogenética de

alinhamento. Para uma dada substituição, o valor PAM é proporcional ao logaritmo natural da freqüência com que a troca foi observada. A matriz PAM 1 é calculada a partir da comparação entre seqüências com menos de 1% de divergência. Outras matrizes, como a PAM 100 são extrapoladas a partir da matriz PAM 1. Em muitos programas que fazem busca em bancos de dados, a matriz PAM 250 é a mais utilizada, uma vez que grandes bancos tenderão a possuir conjuntos de proteínas muito distantes em termos evolutivos. Inicialmente, as unidades PAM foram calibradas com a utilização de pares de resíduos de proteínas intimamente relacionadas, pois podiam ser alinhadas à mão. Entretanto, as unidades PAM podem ser extrapoladas para altos valores, como citado anteriormente, às custas de um resultado de menor confiabilidade.

As matrizes do tipo BLOSUM foram calculadas com a utilização de um banco de dados de blocos de proteínas pouco relacionadas (blocks.fhcrc.org). Em princípio, a matriz BLOSUM deveria ser mais realista, quando comparada a PAM, em se tratando de proteínas com baixa similaridade justamente por ter sido construída para tal finalidade. Por outro lado, pode-se pensar que quanto menor for a relação entre as seqüências, menor será a confiança no alinhamento, o que poderia contribuir com erros no cálculo da matriz BLOSUM.

A matriz BLOSUM 62 é calculada a partir de comparações entre seqüências com identidade máxima de 62% podendo ser extrapolada também para uma matriz BLOSUM 80, o que significaria uma identidade máxima de 80% entre as seqüências, logo, a escolha de matrizes diferentes implicará em resultados ligeiramente distintos. Já que esse tipo de matriz

não utiliza uma árvore filogenética no processo de alinhamento, torna-se impossível gerar um modelo evolutivo.

3.1.1.1 Outras matrizes utilizadas em alinhamentos

Matriz identidade

É o tipo mais simples de matriz, alinhando apenas os resíduos idênticos; pares idênticos recebem um valor igual a 1 e pares não idênticos, valor igual a 0 (Schwartz e Dayhoff, 1978; Feng *et al.*, 1985).

Códigos genéticos

Utilizam um mesmo peso para a transição entre todos os aminoácidos. Considera o mínimo de mudanças (0, 1, 2 ou 3) necessárias nas bases de DNA/RNA durante a conversão dos códons genéticos em aminoácidos (Fitch, 1966).

Similaridades químicas

Levam em consideração durante o alinhamento de aminoácidos, propriedades físico-químicas similares (polaridade, tamanho, forma e carga) premiando alinhamentos onde essas características físico-químicas são preservadas e penalizando pares onde os aminoácidos possuem características muito distintas (McLachlan, 1972; Feng *et al.*, 1985).

Estrutura terciária

Matriz obtida por comparações das estruturas terciárias de proteínas determinadas experimentalmente (Rao, 1987; Overington *et al.*, 1990 e Bowie *et al.*, 1991).

3.1.2 Alinhamento global

O alinhamento global considera a seqüência completa de bases/aminoácidos. Nesse tipo de alinhamento, as penalidades tanto para os *gaps* de abertura como para os *gaps* 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 *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 a sua extensão (Fitch, 1966; Gibbs e McIntyre, 1970 e Collins e Coulson, 1987).

3.1.3 Alinhamento local

Os alinhamentos locais podem ser representados como blocos desprovidos de *gaps*. A formação de blocos é facilitada pela baixa penalidade imposta aos *gaps* de abertura e para os *gaps* de extensão. Logo, uma seqüência de resíduos poderá ter uma maior “mobilidade” e deslocar um grande número de resíduos através da inserção ou deleção de *gaps*. Esse tipo de alinhamento é apropriado quando as seqüências mostram

regiões isoladas de similaridade, por exemplo, múltiplos domínios ou repetições.

A figura 3.3 mostra uma comparação entre alinhamentos global e local.

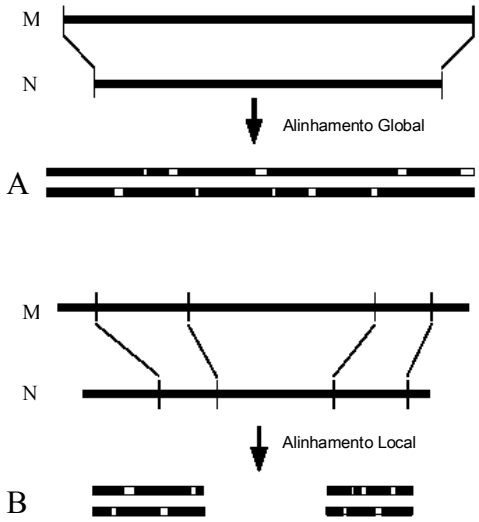


Figura 3.3: Comparação entre alinhamentos global e local. O alinhamento global é otimizado ao longo de todo o comprimento das seqüências M e N enquanto que o local encontra o melhor alinhamento entre frações de M e N, podendo haver diversas maneiras de se alinhar localmente as duas seqüências (adaptado de Sternberg, M. J. (1996) *Protein Structure Prediction*, Oxford University Press, p. 38).

3.1.4 Alinhamento múltiplo

Quando se dispõe de um banco de dados protéico, um alinhamento múltiplo sempre será a melhor opção, já que um grande grupo de proteínas será alinhado e as regiões semelhantes serão ainda mais destacadas.

O alinhamento de três ou mais seqüências de resíduos pode ser dividido em quatro categorias (Barton, 1996):

- Extensão de um alinhamento *pairwise* utilizando programação dinâmica.

- Extensões hierárquicas de métodos *pairwise*
- Métodos de segmentação
- Métodos consensuais

Desses, o segundo método é o mais prático e utilizado e será discutido a seguir.

3.1.4.1 Extensões hierárquicas de métodos *pairwise*

Métodos práticos para alinhamentos múltiplos baseados em três ou mais seqüências foram desenvolvidos em diferentes laboratórios (Dayhoff e Eck, 1968; Needleman e Wunsch, 1970; McLachlan, 1971; Feng *et al.*, 1985; Rao, 1987; Risler *et al.*, 1988; Smith e Smith, 1990; George *et al.*, 1990).

Os passos envolvidos neste tipo de alinhamento podem ser resumidos da seguinte maneira (a figura 3.4 também ilustra as etapas deste alinhamento):

I. Formar os pares possíveis entre as seqüências e compará-las entre N seqüências, tem-se um número total de pares dado por $\frac{N \times (N-1)}{2}$ pares.

II. Construção de um alinhamento hierárquico. Isso pode ser feito na forma de uma árvore binária ou simplesmente reordenando as seqüências.

III. Realizar o alinhamento múltiplo, respeitando a ordem dos pares anteriormente formados de acordo com a similaridade.

Essa seqüência de etapas fornecerá um alinhamento com *gaps* que pode ser aplicado a um grande número de seqüências, ou seja, um alinhamento múltiplo.

A figura 3.4 ilustra um alinhamento entre quatro seqüências (A, B, C e D). A seqüência A ficou inicialmente alinhada com C e a seqüência B alinhada com D, formando dois grupos separados, como pode ser observado na árvore de alinhamento gerada (os outros alinhamentos possíveis, porém descartados, seriam AB, AD, BD e CD). O alinhamento entre A, B, C e D é então efetuado pela comparação dos dois grupos BD e AC.

Figura 3.4: Estágios envolvidos em alinhamento múltiplo de seqüências utilizando um método hierárquico. (adaptado de Sternberg, M. J. (1996) *Protein Structure Prediction*, Oxford University Press, p. 46.).

3.1.5 BLAST (*Basic Local Alignment Search Tool*)

BLAST (Altschul *et al.*, 1990) é um método heurístico para encontrar o melhor alinhamento local entre uma dada seqüência (de DNA ou protéica) e um banco de dados. Uma importante simplificação que o BLAST faz é a de não permitir *gaps*, e sim múltiplos resultados de alinhamentos para uma mesma seqüência. O algoritmo do BLAST faz uso de estatísticas de alinhamentos seqüenciais sem *gaps* idealizadas por Karlin e Altschul (Karlin e Altschul, 1990). Esse algoritmo procura eliminar estatisticamente homologias casuais e pode ser configurado com parâmetros tais como: penalidade para a introdução de inserções e deleções (*gaps*), matriz de substituição, etc. As estatísticas mostram a probabilidade de se obter um alinhamento com o menor número de *gaps* possível (MSP – ***M**aximal **S**egment **P**air*) com um valor mínimo *T* pré-fixado pelo usuário, dentro de uma margem de corte *S* ou um valor de E (*E-value*) menor que o máximo especificado.

Basicamente, o algoritmo opera em três passos:

- Para uma dada seqüência a ser estudada de N resíduos, a mesma será fragmentada em partes de *w* resíduos e este valor *w* será o número de resíduos a ser utilizado durante a busca em um banco de dados (usualmente *w* = 3 no caso de proteínas). Ou seja, é utilizada uma trinca de aminoácidos e um valor máximo *T* em uma matriz de alinhamento para cada comparação realizada pela trinca de resíduos.
- A busca em um banco de dados é feita utilizando-se *w* resíduos, na tentativa de se encontrar esses resíduos correspondentes nas outras

seqüências do banco de dados. Esta busca é feita com a utilização de uma matriz de substituição.

- Se durante os alinhamentos *T* for alcançado, *w* é estendida em ambas as direções para gerar um alinhamento ótimo e sem *gaps* ou MSP com valor de no mínimo *S* ou valor E (*E value*).

As três etapas envolvidas no algoritmo do programa BLAST estão ilustradas na figura 3.5.

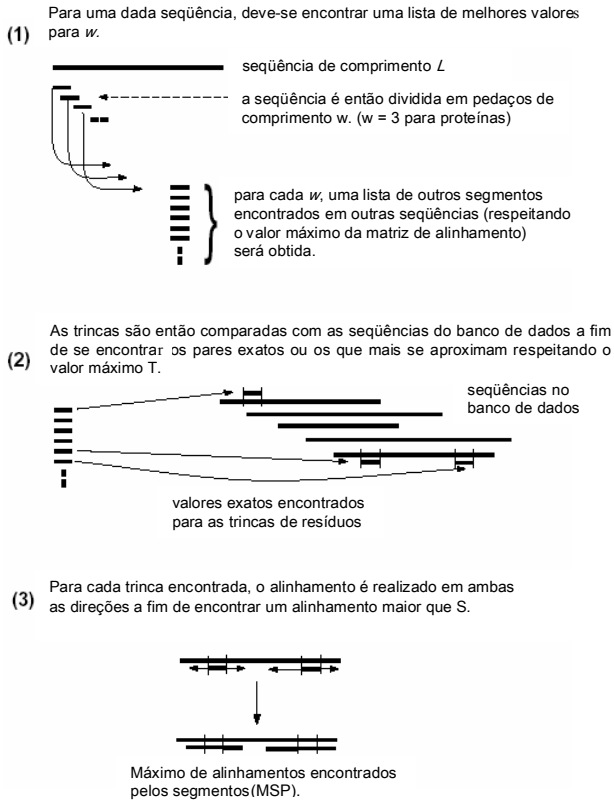


Figura 3.5: Algoritmo do programa BLAST. Três etapas envolvidas durante a busca em um banco de dados (adaptado de Sternberg, M. J. (1996) *Protein Structure Prediction*, Oxford University Press, p. 58.).

O programa BLAST não é tão sensitivo na busca por proteínas em banco de dados, porém, estatísticas mostram que os resultados podem ser considerados muito bons (Sternberg, 1996).

A quantificação do alinhamento (*score*), é o valor associado a um alinhamento baseado em punições relacionadas a *gaps* e a substituições em premiações relacionadas a identidades. Os valores das punições e premiações são obtidos através das matrizes de valores (Barton e Sternberg, 1990).

3.2 Predição de estrutura secundária

O estudo de estruturas secundárias de proteínas traz informações importantes sobre seu enovelamento. A técnica de Dicroísmo Circular (CD) tem sido utilizada para esse tipo de estudo. No presente trabalho, a predição da estrutura secundária foi fundamental para a construção do modelo tridimensional da enzima GumH.

Existe uma enorme quantidade de métodos que fazem a predição da estrutura secundária a partir da seqüência primária e que podem ser divididos, de maneira abreviada, em estatísticos, redes neurais e sistemas híbridos, que mesclam diferentes métodos (Barton, 1996).

Métodos estatísticos baseiam-se no estudo de bancos de dados protéicos de estruturas primárias e secundárias conhecidas. Esses estudos têm como finalidade procurar relações empíricas entre esses dois tipos de estruturas.

A metodologia empregada pelas redes neurais pode ser dividida em diferentes tipos: gradiente conjugado, correlação em cascata, máquinas de Boltzmann, etc (Simpson, 1990). Entretanto, quase todas as redes neurais que são utilizadas na predição de estrutura secundária foram baseadas no principio do "aprendizado em várias camadas" composto por várias unidades de processamento, cujo funcionamento é bastante simplificado. Essas unidades, geralmente, são conectadas por canais de comunicação que estão associados a determinado peso. Energia de formação de pares, energia de solvatação, valor do alinhamento *E-value*, comprimento do alinhamento, comprimento do alinhamento, comprimento da estrutura e comprimento da seqüência são alguns exemplos de pesos utilizados pelo programa GenThreader (Jones, 1999 e 2002). As unidades fazem operações apenas sobre seus dados locais, que são entradas recebidas pelas suas conexões. O comportamento inteligente de uma Rede Neural Artificial vem das interações entre as unidades de processamento da rede (King e Sternberg, 1990).

Nesse trabalho, o programa PSIPRED (***P**osition **S**pecific **I**terated **Pred**iction* de Jones, 1999) foi o utilizado nas predições das estruturas secundárias de todas as enzimas estudadas. O PSIPRED tem a característica de ser um programa híbrido, pois envolve métodos estatísticos e de redes neurais (figura 3.6).

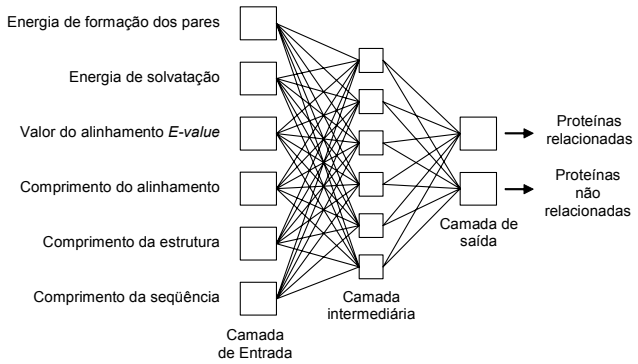


Figura 3.6: Diagrama representativo da arquitetura de uma rede neural (modificado de Jones, 1999).

3.2.4 Análise de hidrofobicidade e predição de regiões transmembrânicas

Proteínas de membrana estão envolvidas em diversas funções celulares e fazem parte de uma grande fração entre todos os tipos proteínas. É estimado que aproximadamente 40% dos genes podem codificar para proteínas integradas à membrana (Goffeau *et al.*, 1993a; Goffeau *et al.*, 1993b).

Duas classes básicas de proteínas de membrana são conhecidas de acordo com a estrutura do segmento que atravessa a membrana. A primeira classe é aquela em que todos os segmentos transmembrânicos formam uma estrutura de α-hélice com cerca de 17 – 25 aminoácidos (von Heijne, 1994). Membros da segunda classe são aquelas estruturas que formam poros constituídos por barris β de 16 fitas (Weis e Schulz, 1992).

Na maioria dos métodos de predição de regiões transmembrânicas, um gráfico de hidrofobicidade de toda a seqüência é construído mostrando a hidrofobicidade local de cada resíduo protéico. Para se fazer um gráfico de

hidrofobicidade, cada aminoácido deve receber inicialmente uma escala de hidrofobicidade, que varia de programa para programa. A melhor escala é aquela que, quando comparada à estrutura resolvida, consegue predizer o maior número de regiões que correspondem à da estrutura real.

Os diversos programas que fazem predições de regiões transmembrânicas usam diferentes tamanhos de peptídeos nas análises, e em geral, utilizam uma seqüência de 17 - 25 aminoácidos. A figura 3.7 ilustra um gráfico típico de hidrofobicidade (Claros e von Heijne, 1994). As regiões acima dos limites são as candidatas a apresentar motivos transmembrânicos.

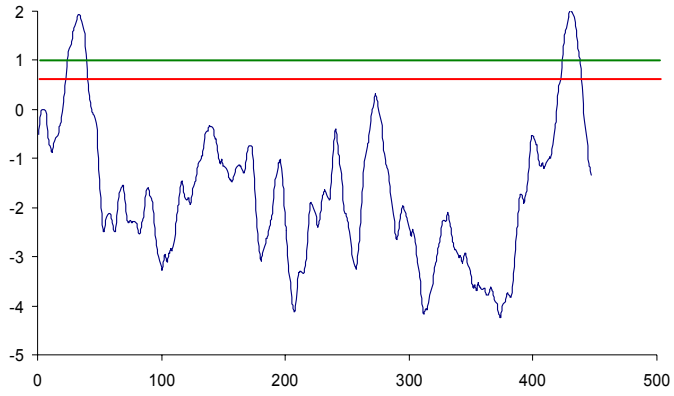


Figura 3.7: Exemplo de gráfico de hidrofobicidade. A linha verde indica o limite superior e a linha vermelha o limite inferior. Figura gerada pelo programa TOPPRED (***TOP**ology **PRED**iction* de Claros e von Heijne, 1994).

Há programas que utilizam um algoritmo de programação dinâmica baseado em probabilidades e estatísticas, como é o caso do programa MEMSAT 2 (***MEM**brane protein **S**tructure **A**nd **T**opology* de Jones *et al.*, 1994) e o HMMTOP (***H**idden **M**arkov **M**odel **TOP**ology* de Tusnady e Simon,

1998 e 2001) levam em consideração não só a hidrofobicidade de cada aminoácido, mas também a composição e estereoquímica de cada resíduo.

Esses programas foram utilizados na predição das regiões transmembrânicas das nove enzimas envolvidas na biossíntese da goma fastidiana (os resultados e as discussões serão mostrados nos capítulos 4 e 5).

3.3 Modelagem molecular

Um dos principais objetivos dos projetos de Genomas Estruturais é o estudo das estruturas tridimensionais com o enfoque no desenvolvimento de novos fármacos e estudos relacionados à interação proteína-proteína.

A modelagem comparativa representa um ganho de tempo na obtenção dessas estruturas 3D relacionadas, já que independe de técnicas experimentais como difração de raios X e RMN (**R**essonância **M**agnética **N**uclear). Muitas aproximações diferentes têm sido utilizadas para predizer estruturas de proteínas a partir de sua seqüência de aminoácidos, com vários níveis de sucesso. Métodos *ab initio* têm como característica o cálculo das posições (ou coordenadas) atômicas de uma dada seqüência protéica pelos "primeiros princípios", isto é, sem a referência de uma estrutura tridimensional de proteína, mas o sucesso decorrente dessa técnica tem sido relativamente baixo (Simons *et al.*, 2001). Métodos comparativos, como modelagem por homologia, predizem estruturas de proteínas enfatizando a semelhança seqüencial de aminoácidos em relação a uma outra proteína de estrutura tridimensional conhecida baseado na premissa de que essa similaridade seqüencial implica em uma similaridade estrutural. Essa é uma

aproximação confiável para a construção de estruturas terciárias, mas existe a limitação da dependência em existir uma proteína homóloga de estrutura conhecida à proteína em estudo.

Nos últimos anos, novas metodologias tem sido desenvolvidas para modelagem molecular em casos onde a proteína em estudo (molécula alvo) não apresenta elevado grau de identidade com a proteína de estrutura tridimensional conhecida (molécula molde). Neste caso, o que é importante para a construção do modelo é a predição da estrutura secundária da proteína em estudo e comparação da mesma com as estruturas secundárias depositadas nos bancos de proteínas. A base desta metodologia está na observação de que duas proteínas podem apresentar enovelamentos semelhantes, e, portanto, estruturas tridimensionais parecidas, sem terem elevado grau de identidade seqüencial (McGuffin e Jones, 2002). A este tipo de abordagem para modelagem molecular tem sido dado o nome de "*threading* protéico" e pode ser usado mesmo que o tamanho das bibliotecas ou acervos de estruturas atualmente disponíveis sejam limitados.

3.3.1 THREADER 3.3

THREADER 3.3 (Jones *et al.*, 2002) é um programa que realiza a predição da estrutura secundária de uma proteína por meio de comparações de enovelamentos corretos e conhecidos presentes em uma biblioteca de estruturas dos mais diferentes motivos. Se o enovelamento da proteína alvo (proteína em estudo) não for semelhante à de proteínas disponíveis na biblioteca, a busca não terá êxito. Felizmente, certos enovelamentos se

repetem muitas vezes e os métodos de reconhecimento e predição da estrutura podem ser efetivados.

Dos diversos programas (Xu *et al.*, 1999; Thiele, 1999 e Albrecht *et al.*, 2002) que aplicam a técnica do *threading* como característica, o THREADER 3.3 foi o que apresentou os melhores resultados na reunião realizada pela CASP (***C**ritical **A**ssessment of Techniques for Protein **S**tructure **P**rediction*) de 2000, onde diversos programas foram testados e seus resultados comparados e classificados de acordo com seu desempenho. O programa THREADER 3.3 apresentou uma margem de 77% de acerto nas predições realizadas quando comparadas às proteínas com enovelamentos bem determinados (http://predictioncenter.llnl.gov/).

3.3.2 Etapas envolvidas na modelagem por *threading*

Basicamente, existem quatro etapas envolvidas no processo de modelagem por *threading*: *seleção do molde* feita através da predição da estrutura secundária da molécula alvo e comparação em banco de dados de estruturas, *alinhamento alvo-molde*, a *construção do modelo* e finalmente, a *validação* do modelo obtido.

3.3.2.1 Selecionando o molde e alinhando as estruturas

A seleção dos moldes é feita através da comparação das estruturas secundárias da molécula alvo (estrutura predita) e das moléculas cujas estruturas estão depositadas nos bancos de dados. Programas fazem buscas nos mais diversos bancos seqüenciais e protéicos disponíveis na

rede como, por exemplo, o PDB (***P**rotein **D**ata **B**ank*) (Westbrook *et al.*, 2002), SCOP (Lo Conte *et al.*, 2002), DALI (Holm *et al.*, 1999) e CATH (Orengo *et al.*, 2002).

3.3.2.2 construção do modelo

Uma vez realizado o alinhamento alvo-molde, uma variedade de métodos podem ser usados na construção de modelos tridimensionais para a proteína alvo, como, por exemplo, programas computacionais e servidores via rede (*Internet*) que realizam um processo automático de modelagem comparativa, como o *Swiss-Model* (http://www.expasy.ch/swissmod/), CPHModels (http://www.cbs.dtu.dk/services/CHPmodels/), SDSC1 (http://cl.sdsc.edu/hm.html), FAMS (http://physchem.pharm.kitasato-u.ac.jp/FAMS/fams.html), MODWEB (http://guitar.rockefeller.edu/modweb/) e EsyPred3D (http://www.fundp.ac.be/urbm/bioinfo/esypred/). Muito desses servidores estão sendo avaliados pela EVA-CM (Eyrich *et al.*, 2001), um servidor que realiza predições de estruturas de proteínas por diversos métodos de maneira automática, contínua e em larga escala (Marti-Renom *et al.*, 2002).

A técnica de modelagem, em geral, é cercada por diversas dificuldades, tais como alinhamentos problemáticos, modelagem de *loops*, existências de múltiplos estados conformacionais e a modelagem de sítios de ligação. A metodologia original e ainda largamente utilizada é a da modelagem por corpo rígido (Taylor, 1994 e Barton, 1998). Outra metodologia é a modelagem por segmentos, que utiliza as posições conservadas dos átomos do molde (Jones *et al.*, 1986; Urger *et al.*, 1989;

Claessens *et al.*, 1989 e Levitt, 1992). O outro grupo de método utiliza técnicas de otimização de distâncias geométricas a fim de satisfazer as restrições espaciais abstraídas de uma estrutura homóloga conhecida e de seu alinhamento seqüencial ou estrutural, levando em consideração a estrutura primária ou secundária da molécula alvo. Essa abordagem automatizada para modelagens moleculares comparativas, baseadas em restrições espaciais, foi implementada pelo programa MODELLER 6.0a (Sali *et al.*, 1993), um dos programas mais utilizados em modelagem.

O alinhamento entre as moléculas alvo e molde serve de arquivo de entrada para o programa MODELLER 6a, que, juntamente com as informações contidas no arquivo das coordenadas tridimensionais (.pdb ou .atm), passa a calcular as restrições para a seqüência alvo.

Entre as restrições obtidas das entradas encontram-se:

- Distâncias entre pares de carbono α (Cα).
- Distâncias entre nitrogênios e oxigênios na cadeia principal visando manter as pontes de hidrogênio.
- Ângulos ϕ, ψ e ω, ângulos diédricos e pares de ângulos diédricos.
- Ângulos definidos por átomos ou pseudo-átomos.
- Restrições impostas pelo usuário.
- Restrições da estereoquímica padrão de proteínas obtidas empiricamente.

Estas restrições são combinadas com termos de energia produzidos pelo programa CHARMM (Brooks *et al.*, 1983) gerando uma função objetiva completa. Finalmente esta função é otimizada por gradientes conjugados e dinâmicas moleculares simuladas. As referidas dinâmicas são baseadas no

aumento de energia do sistema seguida de uma lenta diminuição (*simulated annealing*) resultando em um modelo que satisfaça todas as restrições espaciais (Sali *et al.*, 1993).

3.3.2.3 validação do modelo

Avaliar a qualidade do modelo reconstruído é fundamental e com este objetivo, diversos pacotes de programas têm sido desenvolvidos nos últimos anos. O modelo pode ser avaliado como um todo ou por regiões. A qualidade do modelo construído será proporcional à qualidade da estrutura 3D utilizada como molde (Sanchez *et al.*, 1998).

Alguns programas para validação da estereoquímica como o PROCHECK (Laskowski *et al.*, 1998) e o WHATIF (Vriend, 1990; Hooft *et al.*, 1996) são bastante utilizados. Os parâmetros testados por esses programas são os comprimentos das ligações, ângulos das ligações, planaridade das ligações peptídicas e dos anéis das cadeias laterais, quiralidade, ângulos torcionais da cadeia principal e das cadeias laterais e impedimentos estéricos entre pares de átomos não ligados.

Existem também programas que realizam comparações entre os modelos obtidos e proteínas resolvidas com alta resolução, como o VERIFY 3D (Luthy *et al.*, 1992) e o pacote de programas WHATIF.

Os programas de validação de estruturas tridimensionais utilizados neste trabalho foram o VERIFY 3D, WHATIF e PROCHECK, que serão brevemente descritos a seguir.

WHATIF

Apesar de possuir uma amplitude de aplicações, esse programa foi utilizado unicamente para a avaliação dos modelos obtidos. O método utilizado (QUALTY/OLDQUA), checa a normalidade da estrutura baseando-se em parâmetros de empacotamentos que envolvem proximidades atômicas.

O algoritmo do programa utiliza um conjunto de átomos representantes de um resíduo (um total de 80 grupos rígidos) para a superposição em todos os resíduos do mesmo tipo no arquivo PDB. Assim, a vizinhança de átomos de cada um desses resíduos será detectada.

Com as coordenadas dos átomos vizinhos, é calculada uma função que possui maiores valores nos locais onde o maior número de átomos são encontrados, levando-se em consideração o número de resíduos e suas exposições ao solvente a fim de normalizá-los. Posteriormente os valores gerados são comparados com valores padrões.

Os arquivos gerados pelo programa apresentam os desvios em relação à média para cada resíduo e um resultado global para os resíduos envolvidos, seguindo a classificação: desvio menor que -0,5 a estrutura da proteína é considerada perfeita; desvio entre -0,5 e -1,0 o modelo é considerado de muito boa qualidade; -1,5 o modelo é considerado bom, mas possivelmente contém alguns pequenos erros; -2,0 o modelo é considerado muito pobre e finalmente -3,0 o modelo é considerado ruim.

Os resíduos com valores individuais abaixo de -5,0 devem ser checados, pois provavelmente pertencem a um dos seguintes casos:

- Está envolvido em simetrias de contato ou na extremidade da cadeia.

- Está ligado a um cofator, ligante ou íon.
- Está localizado no sítio ativo.
- Está com conformação incorreta.

PROCHECK

Este programa analisa a geometria global da estrutura ou cada resíduo individualmente, utilizando para isso parâmetros estereoquímicos derivados de estruturas de alta resolução.

As entradas do programa são um arquivo de coordenadas 3D (formato .pdb) e a resolução desta estrutura (em casos de estruturas modeladas, utiliza-se a resolução de 2,0 Å).

Os parâmetros utilizados pelo PROCHECK podem ser sumarizados:

- Geometria covalente: inclui distâncias e ângulos de ligação da cadeia principal.
- Planaridade: checa a planaridade de anéis aromáticos (Phe, Tyr, Trp e His) e de grupos terminais (Arg, Asn, Asp, Gln e Glu) através de RMS (***R**oot **M**ean **S**quare*).
- Ângulos diédricos: distribuição phi-psi dos resíduos no diagrama de Ramachandran, onde os resíduos podem ser localizados em uma das seguintes regiões: central (*core*), permitida (*allowed*), generosamente permitida (*generously allowed*) e não permitida (*disallowed*). Também avalia a distribuição dos resíduos quanto à chi1-chi2, ângulos torcionais chi3 e chi4 (para os que possuírem) e ângulos torcionais ômega.

- Quiralidade: checa o ângulo torcional zeta definido pelos planos dos átomos Cα-N-C e N-C-Cβ.
- Interações não covalentes: investiga maus contatos, definido por átomos não ligados com distâncias menores que 2,6 Å, sendo previamente descartadas possíveis ligações de pontes de hidrogênio.
- Pontes de hidrogênio da cadeia principal: a checagem é realizada baseada nas energias de ligações de hidrogênio na cadeia principal.
- Pontes de dissulfeto: As distâncias de ligações S-S são comparadas ao valor ideal (2,0 Å).

VERIFY 3D

A metodologia do programa consiste em medir a compatibilidade entre uma determinada seqüência de aminoácidos e a estrutura tridimensional de uma proteína.

O VERIFY 3D utiliza três operações básicas em sua metodologia:

- Redução da estrutura tridimensional em uma seqüência unidimensional dentro de um ambiente. Esses ambientes estão categorizados de acordo com a área da cadeia lateral enterrada na proteína, a fração de área da cadeia lateral que está exposta a átomos polares e a estrutura secundária local.
- Geração de uma matriz de comparação dependente da posição, conhecida como perfil 3D. Esse cálculo é feito de acordo com o ambiente de cada resíduo da seqüência, isto é, a probabilidade de se encontrar cada um dos 20 aminoácidos em cada uma das classes de

ambientes, como observado em um banco de dados protéico e suas respectivas seqüências resultando na formação de uma matriz 18x20 (18 ambientes possíveis x número total de aminoácidos).

- Um alinhamento da seqüência primária com o perfil tridimensional. O resultado da qualidade do alinhamento é a medida da compatibilidade da seqüência com sua estrutura 3D descrita por seu perfil tridimensional.

Um resumo das etapas envolvidas na modelagem por *threading* é mostrado na figura 3.8.

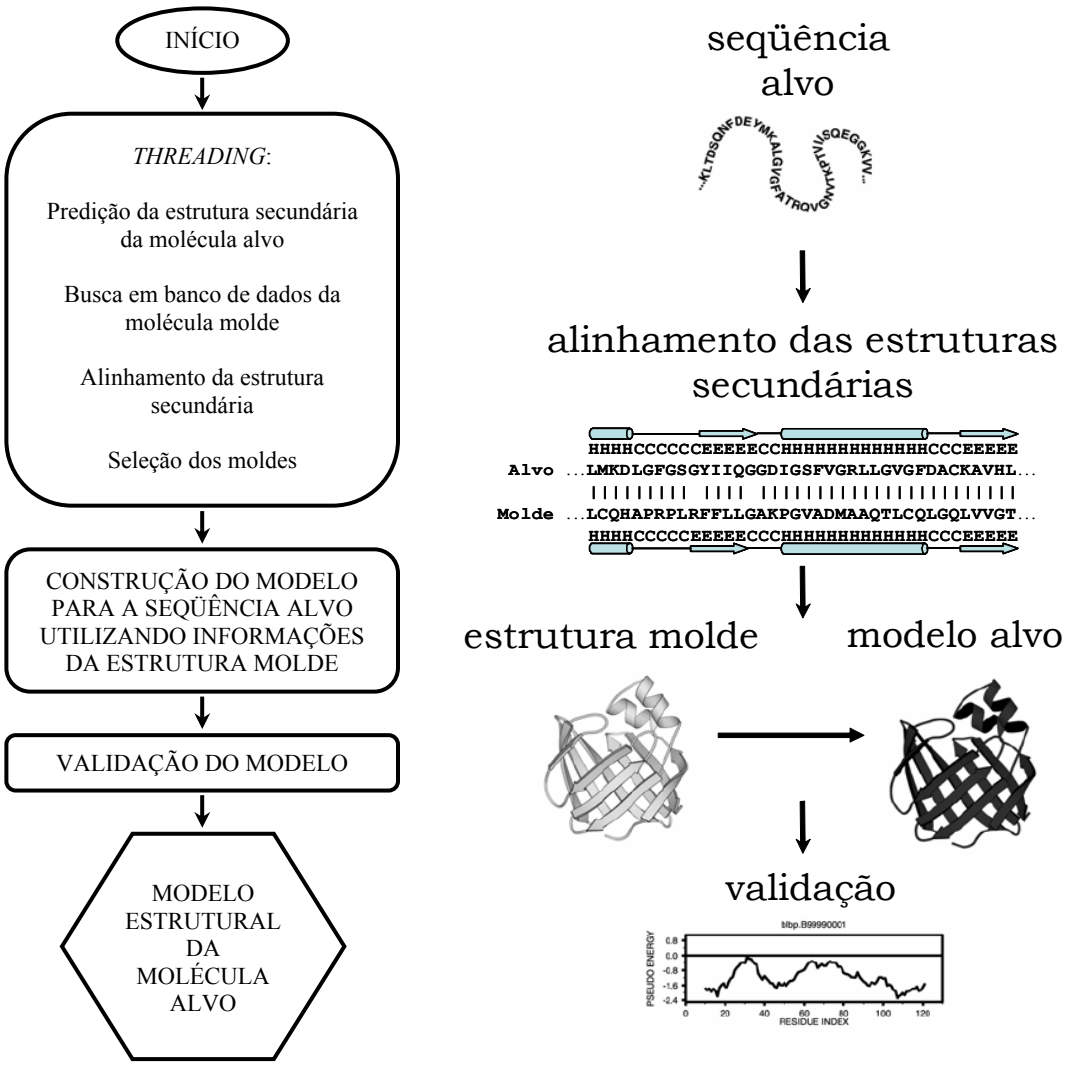


Figura 3.8: Etapas envolvidas na modelagem por comparação.

3.4 Referências bibliográficas

Albrecht, M., Hanisch, D., Zimmer, R and Lengauer, T. (2002) *In Silico Biology*, **2**, 30.

Altschul, S. F. (1991) *J. Mol. Biol.*, **219**, 555-565.

Altschul, S. F., Boguski, M. S., Gish, W. and Wootton, J. C. (1994) *Nat. Genet.*, **6**, 119-129.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) *J. Mol. Biol.*, **215**, 403.

Barton, G. J. (1998) *Acta Crystallogr. D Biol. Crystallogr.*, **54**, 1139-1146.

Barton, G. J. (1996) *Protein Structure Prediction – a practical approach*, Oxford University Press.

Barton, G. J. and Sternberg, M. J. E. (1990) *J. Mol. Biol.*, **212**, 389.

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S. R., Griffiths-Jones, S., Howe K. L., Marshall, M. and Sonnhammer, E. L. (2002) *Nucleic Acids Research*, **1**, 276-280.

Bonneau, R., Baker, D. (2001) *Annu. Rev. Biophys. Biomol. Struct.*, **30**, 173-189.

Bowie, J. U., Luthy, R., and Eisenberg, D. (1991) *Science*, **253**, 164.

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. and Karplus, M (1983) *J. Comp. Chem*, **4**, 187-217.

Claessens, M., Van Cutsem, E., Lasters, I. and Wodak, S. (1989) *Protein Eng.*, **2**, 335-345.

Claros, M.G., and von Heijne, G. (1994) *CABIOS*, **10**, 685-686.

Collins, J. F. and Coulson, A. F. W. (1987) In *Nucleic acid and protein sequence analysis: a practical approach* (ed. M. J. Bishop and C. J. Rawlings), pp. 323-358. IRL Press, Oxford.

Dayhoff, M. (1978) *Atlas of Protein Sequence and Structure* (Natl. Biomed. Res. Found., Washington), Vol. **5**, Suppl. 3, pp. 345-358.

Dayhoff, M. O. and Eck, R. V., eds. (1968) *Atlas of Protein Sequence and Structure* (Natl. Biomed. Res. Found., Silver Spring, MD), Vol. **3**, p. 33.

Eyrich, V. A., Marti-Renom, M. A., Przybylski, D., Madhusudham, M. S., Fiser, A., Pazos, F., Valencia, A., Sali, A. and Rost, B. (2001) *Bioinformatics*, **17**, 1242-1243.

Feng, D. F., Johnson, M. S. and Doolittle, R. F. (1985) *J. Mol. Evol.*, **21**, 122-125.

Fitch, W. M. (1966) *J. Mol. Biol.*, **16**, 9.

George, D, G., Barker, W. C. and Hunt, L. T. (1990) *Methods Enzymol.*, **183**, 333-351.

Gibbs, A. J. and McIntyre, G. A. (1970) *Eur. J. Biochem.*, **16**, 1.

Goffeau, A., Nakai, K., Slonimski, P. and Risler, J. L. (1993a) *FEBS Lett.*, **325**, 112.

Goffeau, A., Slonimski, P., Nakai, K. and Risler, J. L. (1993b) *Yeast*, **9**, 691.

Henikoff, S. and Henikoff, J. G. (1992) *Proc. Natl. Acad. Sci. USA*, **89**, 10915-10919.

Holm, L. and Sander, C. (1996) *Science*, **273**, 595-603.

Holm, L. and Sander, C. (1999) *Nucleic Acids Res.*, **27**, 244-247.

Hooft, R. W., Sander, C. and Vriend, G. (1996) *Proteins*, **26**, 363-376.

Jones, D. T. (2002) University College London, Dept. of Comp. Sci., Bionf. Unit, Gower Street, London, UK.

Jones, D.T. (1999) *J. Mol. Biol.*, **292**, 195-202.

Jones, D. T., Taylor, W. R. and Thornton, J. M. (1994) *Biochem.*, **33**, 3038-3049.

Jones, T. A., Thirup, S. (1986) *EMBO J.*, **5**, 819-822.

Jorja Henikoff, Shmuel Pietrokovski and Steven Henikoff (http://blocks.fhcrc.org).

Karlin, S. and Altschul, S. F. (1990) *Proc. Natl. Acad. Sci. USA*, **87**, 2264

King, R. D. and Sternberg, M. J. E. (1990) *J. Mol. Biol.*, **216**, 441.

Laskowski, R. A., MacArthur, M. W. and Thornton, J. M. (1998) *Curr. Opin. Struct. Biol.*, **8**, 631-639.

Levitt, M. (1992) *J. Mol. Biol.*, **226**, 507-533.

Lo Conte, L., Brenner, S. E., Hubbard, T. J., Chotia, C. and Murzin, A. G. (2002) *Nucleic Acids Res.*, **30**, 264-267.

Luthy, R., Bowie, J., U. and Eisenberg, D. (1992) *Nature*, **356**, 83-85.

Marti-Renom, M. A., Madhusudham, M. S., Fiser, A., Rost, B. and Sali, A. (2002) *Structure*, **10**, 435-440.

McGuffin, L. J. and Jones, D. T. (2002) *Proteins*, **1**, 44-52.

McLachlan, A. D. (1971) *J. Mol. Biol.*, **61,** 409-424.

McLachlan, A. D. (1972) *J. Mol. Biol.*, **64**, 417.

Needleman, S. B. and Wunsch, C. D. (1970) *J. Mol. Biol.*, **48**, 443.

Nicholas Jr, H. B., Deerfield II, D. W. and Ropelewski, A. (1998) *A tutorial on searching sequence databases an scoring methods*. http://www.psc.edu/biomed/training/tutorials/sequence/db

Orengo, C. A., Bray, J. E., Buchan, D. W., Harrison, A., Lee, D., Pearl, F. M. Sillitoe, I., Todd, A. E. and Thornton, J. M. (2002) *Proteomics*, **2**, 11-21.

Overington, J., Johnson, M. S., Sali, A., and Blundell, T. L. (1990) *Proc. R. Lond. Ser. B*, **241**, 132.

Pieper, U., Eswar, N., Ilyin, V. A., Stuart, A. and Sali, A. (2002) *Nucleic Acids Res.*, **30**, 255-259.

Rao, J. K. M. (1987) *Int. J. Pept. Protein Res.*, **29**, 276-281.

Risler, J. K. M., Delorme, M. O., Delacroix, H, and Henaut, A. (1988) *J. Mol. Biol.*, **204**, 1019-1029.

Sali, A. and Blundell, T. L. (1993) *J. Mol. Biol.*, **234**, 779-815.

Sali, A., Fiser, A Sanchez, R., Marti-Renom, M. A., Jerkovic, B., Badretdinov, A., Melo, F., Overington, J. and Feyfant, E. (2001) *http://guitar.rockefeller.edu/modeller/*

Sanchez, R. and Sali, A. (1998) *Proc. Natl. Sci. USA*, **95**, 13597-13602.

Schwartz, R. M. and Dayhoff, M. O. (1978) *Atlas of Protein Sequence and Structure* (ed. M. O. Dayhoff), Vol. **5**, pp. 353-362. (Natl. Biomed. Res. Found., Washington).

Simons, K. T., Strauss, C. and Baker, D. (2001) *J. Mol. Biol.*, **306**, 1191-1199.

Simpson, P. F. (1990) *Artificial neural systems*. Pergamon Press.

Smith R. F. and Smith, T. F. (1990) *Proc. Natl. Acad. Sci. USA*, **87**, 118-122.

Sternberg, M. J. E (1996) *Protein Structure Prediction – A Practical Approach*, p.58 (Oxford University Press).

Taylor, W. R., Flores, T. P. (1994) *Protein Sci.*, **3**, 1858-1870.

Thiele, R., Zimmer, R. and Lengauer, T. (1999) *J. Mol. Biol.*, **290**, 757-779.

Tusnady, G.E. and Simon, I. (1998) *J Mol Biol*, **2**, 489-506.

Tusnady, G.E. and Simon, I. (2001) *Bioinformatics*, **9**, 849-50.

Urger, R., Harel, D., Wherland, S. and Sussman, J. L. (1989) *Proteins*, **5**, 355-373.

von Heijne, G. (1992) *J. Mol. Biol.*, **225**, 487-494.

Vriend, G. (1990) *J. Mol. Graph.*, **8**, 52-56.

Weiss, M. and Schultz, G. (1992) *J. Mol. Biol.*, **227**, 493-509.

Westbrook, J., Feng, Z., Jain, S., Bhat, T. N., Thanki, N., Ravichandran, V., Gilliland, G. L., Bluhm, W., Weissig, H., Greer, D. S., Bourne, P. E. and Berman, H. M. (2002) *Nucleic Acids Res.*, **30**, 245-248.

Xu, Y., Xu, D., Crawford, O. H., Einstein, J. R., Larimer, F., Uberbacher, E., Unseren, M. A. and Zhang, G. (1999) *Protein Engineering*, **12**, 899-907.

CAPÍTULO 4

RESULTADOS E DISCUSSÕES DOS ESTUDOS COM A ENZIMA GUMH

Neste capítulo serão apresentados e discutidos os resultados obtidos no estudo da enzima GumH. Análises de comparação da seqüência de nucleotídeos dos genes do Operon gum das bactérias *Xanthomonas campestris* pv campestris e *Xylella fastidiosa* mostraram que a enzima GumH de *X. fastidiosa* apresenta 61% de identidade com a GumH da *X. campestris*, uma GDP-manosiltransferase.

Foram realizados estudos de classificação da GumH, análise da estrutura secundária e construção do modelo tridimensional da molécula. Com base no modelo construído, foram feitas considerações sobre o mecanismo da reação catalisada pela enzima.

4.1 Glicosiltransferases

Nos últimos anos, um número crescente de seqüências de proteínas envolvidas na biossíntese de polissacarídeos foram depositadas em bancos de dados específicos como Pfam, CAZy, NBCI. Uma parte dessas proteínas são glicosiltransferases, enzimas que catalisam a transferência de açúcar de

moléculas doadoras para um aceptor específico, que pode ser um sacarídeo, proteína, lipídio, ou DNA (Radominska-Pandya *et al.*, 1999). Em bactérias, essas enzimas estão relacionadas à biossíntese de exopolissacarídeos e lipopolissacarídeos. O entendimento das bases funcionais da diversidade de estruturas relacionadas à síntese de açúcares observadas na natureza é uma das questões centrais na glicobiologia (Sinnott, 1990 e Campbell, 1997).

O papel realizado pelas glicosiltransferases na etiologia das doenças, assim como seus potenciais alvos terapêuticos, são agora bastante apreciados (Dennis *et al.*, 1999a e 1999b). Apesar dos avanços, o fluxo de novos dados na relação estrutura/função caminha a passos lentos e até meados de 2000, poucas estruturas de glicosiltransferases haviam sido reportadas. Os estudos dessas estruturas têm promovido não só uma rica informação em relação à ligação do substrato, especificidade e catálise, como também um discernimento entre suas classificações e prováveis origens evolucionárias.

As glicosiltransferases podem ser divididas em dois grupos dependendo do tipo do doador glicosil que utilizam, sejam os fosfoaçúcares nucleotídeos ou oligossacarídeos. Em todos os casos a reação catalisada é uma substituição no carbono anomérico da molécula de açúcar do substrato e ocorre com retenção ou inversão da configuração deste centro (Cid *et al.*, 2000). Em ambos os mecanismos é prevista a participação de duas carboxilas catalíticas, que estariam envolvidas nos ataques nucleofílicos que ocorrem nessas reações (Sinnott, 1990; McCarter e Withers, 1994). Somente duas estruturas de enzimas que apresentam o mecanismo de

retenção da conformação anomérica (Mulichak *et al.*, 2001; Gastinel *et al.*, 2001) e poucas estruturas de glicosiltransferases com mecanismo de inversão são conhecidas (Ünligil e Rini, 2000; Hu *et al.*, 2003).

O agrupamento das glicosiltransferases em famílias tem sido realizado considerando-se suas características seqüenciais e pela razão produto/substrato (Campbell *et al.*, 1997 e 1998). Essa classificação é um processo difícil devido à baixa identidade seqüencial normalmente encontrada entre as glicosiltransferases. Os bancos de dados mais comuns contendo a classificação dessas enzimas são o Pfam (***P**rotein **fam**ilies*) (Bateman *et al.*, 2002) e o CAZy (***C**arbohydrate **A**ctive en**Zy**mes*) (http://afmb.cnrs-mrs.fr/CAZY/index.html).

4.2 Classificação da enzima GumH

A fim de classificar a enzima GumH quanto a família a que pertence, sua seqüência (387 aminoácidos, deduzidos a partir da seqüência de DNA) foi submetida ao banco de seqüências de proteínas Pfam e CAZy. O programa Pfam fez um alinhamento múltiplo entre a seqüência da proteína alvo e as demais seqüências protéicas presentes em seu acervo já dividido em famílias e respectivas funções biológicas.

O resultado do alinhamento no banco de proteínas Pfam revelou que a GumH possui uma região compreendida entre os aminoácidos Asp184 (ácido aspártico) e Phe356 (fenilalanina) bastante conservada (como pode ser observado na tabela 4.1) quando comparada a outras glicosiltransferases, fato que possibilitou o enquadramento da mesma como

sendo pertencente ao **grupo 1** (EC: 2.4.1.-) das **glicosiltransferases** (número de acesso PF00534) e família 4 (GT4), segundo o banco de dados CAZy. A figura 4.1 ilustra a seqüência da proteína GumH, deduzida da seqüência de DNA, e destaca a região conservada e o motivo SX_2EX_7E presente na grande maioria das manosiltransferases (Geremia *et al.*, 1996 e Abdian *et al.*, 2000), sendo "X" um aminoácido qualquer.

Tabela 4.1: Resultado do alinhamento entre GumH e banco de dados Pfam/CAZy. O domínio glicosiltransferase está determinado entre os aminoácidos Asp184 e Phe356. A qualidade do alinhamento pode ser expressa pelo valor "E" ou *E-value*, que para esse alinhamento é bastante representativo.

Domínio	Início	Final	*E-value*
Glicosiltransferase	184	356	$4{,}7e^{-14}$

```
         10        20        30        40        50        60
....|....|....|....|....|....|....|....|....|....|....|....|
MEWEHCLMKVVHVVRQFHPSIGGMEDVVFNIAMQLHLHAGIDVDVVTLNRVFTQSDVLLP

         70        80        90       100       110       120
....|....|....|....|....|....|....|....|....|....|....|....|
CTDKYQGVSIQRIGYRGSSRYPLAPWVLRMLDKADVIHVHGIDFFYDFLALTRVLHGKPM

        130       140       150       160       170       180
....|....|....|....|....|....|....|....|....|....|....|....|
VVSTHGGFFHTDYASRLKLLWFNTLTRLSALAYARIIASSESDGALFSKIVAPSRLRVIE

        190       200       210       220       230       240
....|....|....|....|....|....|....|....|....|....|....|....|
NGVDVEKYARCGASEAGRTLLYFGRWSMNKGLLETLQLLAVLYVLDPRWRLIIAGREYDY

        250       260       270       280       290       300
....|....|....|....|....|....|....|....|....|....|....|....|
DQAALAYEVDRLGLSEQVHFHCSPSQSQLRFLMEQAQFFISLSRHEGFGIAAVEAMSAGL

        310       320       330       340       350       360
....|....|....|....|....|....|....|....|....|....|....|....|
IPVLSDIPPFARLHRESGLGVLVDPLQPQQAAVAVQGLAVQVDTHFIDWRSQAMAFSDRY

        370       380
....|....|....|....|....|..
HWRYVIGCYQDEYCRALGLGGEQEFLR
```

Figura 4.1: Localização da região de domínio conservado na seqüência de aminoácidos da GumH. A região conservada está compreendida entre os aminoácidos Asp184 e Phe356. Os aminoácidos em azul representam a região SX_2EX_7E.

Este resultado atesta que a GumH é uma glicosiltransferase, já que mostrou ser responsável pela catálise da adição de uma molécula do sacarídeo manose a uma base guanina difosfato (GDP), ou seja, uma manosiltransferase. A região altamente conservada está localizada no domínio C-terminal da proteína que deve englobar o sítio ativo da enzima.

Membros da família glicosiltransferase também transferem açúcares ativados para uma variedade de substratos, incluindo o glicogênio, frutose-6-fosfato, lipossacarídeos, e transferem ainda açúcares ligados a UDP, ADP, GDP ou CMP (CAZy).

Podemos observar na figura 4.2 um alinhamento da seqüência da GumH com as seqüências de 11 enzimas da família das glicosiltransferases. Embora tenham uma variação na identidade seqüencial na faixa de 12 a 61%, como mostrado na tabela 4.2, essas enzimas apresentam o motivo conservado SX_2EX_7E.

Tabela 4.2: Resultado de identidade/similaridade dos alinhamentos da GumH com 11 glicosiltransferases. Apesar do ótimo alinhamento local representado pela região SX_2EX_7E, existe uma considerável ausência de similaridade seqüencial quando todos os aminoácidos ao longo da seqüência são considerados. GumH (*Xylella fastidiosa*), RfbU (*Salmonela typhimurium*), Arcf (*Archaeoglobus fulgidus*), ML2443 (*Mycobacterium leprae*), Rv3032 (*Mycobacterium tuberculosis*), Metj (*Methanocaldococus jannaschii*), Salty (*Salmonela typhimurium*), HI1698 (*Haemophilus influenzae*), AceA (*Acetobacter xylinus*), GumHX (*Xanthomonas campestris*), MtfB (*Escherichia coli*), MtfA (*Escherichia coli*).

	GumH	AceA	GumHX	Arcf	HI1698	ML2443	Salty	Metj	Rv3032	MtfA	MtfB
AceA	37/63										
GumHX	61/77	47/71									
Arcf	20/38	19/36	19/45								
HI1698	18/40	17/45	17/45	24/52							
ML2443	21/40	21/45	22/44	21/46	19/47						
Salty	18/41	18/44	17/47	21/55	18/58	28/55					
Metj	20/44	19/47	21/47	33/53	22/55	33/61	35/65				
Rv3032	22/43	22/44	23/49	26/47	19/45	29/49	24/47	35/55			
MtfA	16/38	21/42	16/41	21/46	12/46	19/47	22/48	25/51	23/46		
MtfB	19/38	17/44	21/45	21/49	18/49	22/46	24/43	27/47	27/48	39/63	
RfbU	21/44	19/41	24/45	18/37	14/43	19/43	19/42	27/50	21/43	20/47	24/50

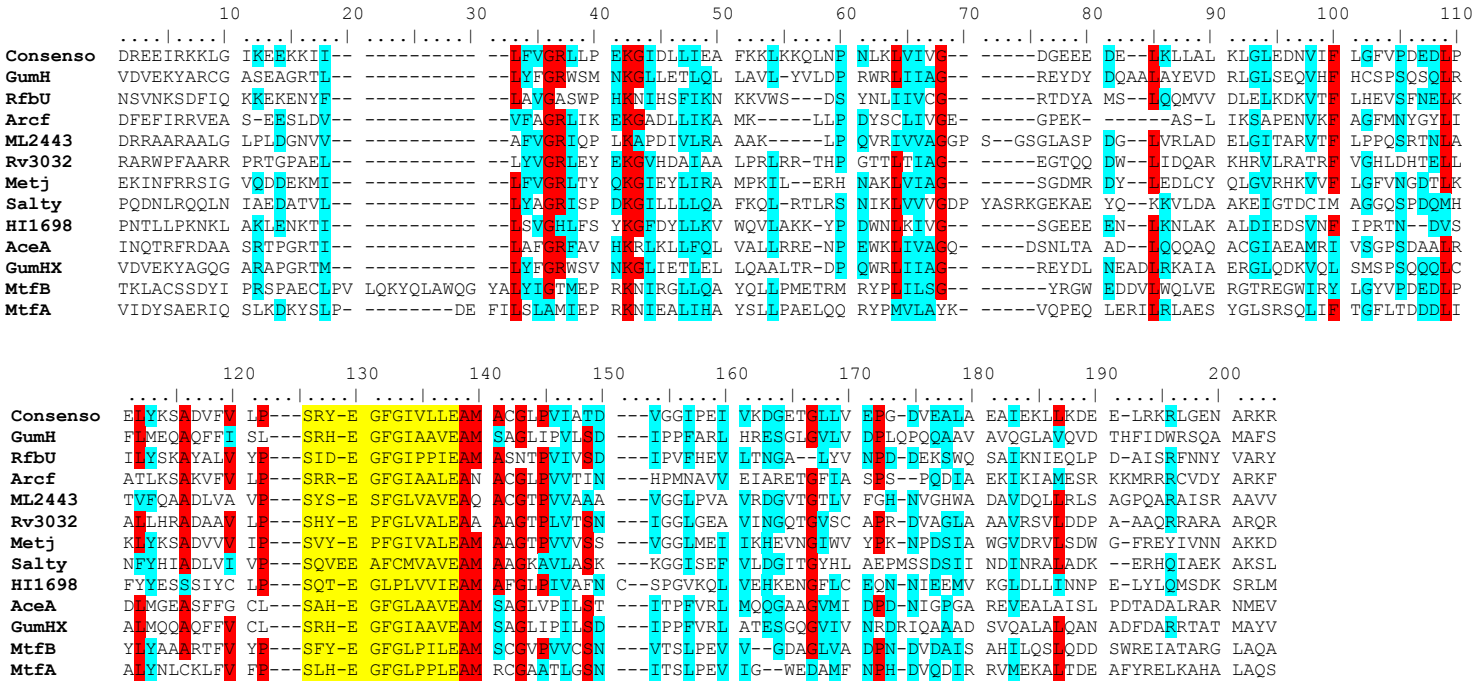


Figura 4.2: Alinhamento das doze glicosiltransferases: GumH (*Xylella fastidiosa*), RfbU (*Salmonela typhimurium*), Arcf (*Archaeoglobus fulgidus*), ML2443 (*Mycobacterium leprae*), Rv3032 (*Mycobacterium tuberculosis*), Metj (*Methanocaldococus jannaschii*), Salty (*Salmonela typhimurium*), HI1698 (*Haemophilus influenzae*), AceA (*Acetobacter xylinus*), GumHX (*Xanthomonas campestris*), MtfB (*Escherichia coli*), MtfA (*Escherichia coli*). A cor azul caracteriza aminoácidos conservados e a cor vermelha os aminoácidos altamente conservados quando comparados a outras proteínas da mesma família. A região destacada em amarelo mostra o motivo conservado SX_2EX_7E. As numerações dos aminoácidos estão dispostas acima das seqüências.

O peptídeo consensual (SX_2EX_7E) também é encontrado em duas proteínas de células eucarióticas, PigA e Gpi3 envolvidas na síntese de uma âncora de glicosilfosfatidilinositol (Miyata *et al.*, 1993; Schonbachler *et al.*, 1995), provavelmente relacionada com a transferência de *N*-acetilglucosamina formando *N*-acetilglucosaminilinositol. Essa enzima putativa catalisa a adição de um açúcar alternativo, sugerindo que a assinatura encontrada pode também ocorrer em glicosiltransferases que utilizam outros doadores de açúcar (Geremia *et al.*, 1996).

Nenhuma proteína da família 4 (GT4) teve, até o momento, sua estrutura tridimensional determinada.

4.3 Predição da estrutura secundária da GumH

O objetivo da predição da estrutura secundária é obter uma primeira informação estrutural que a seqüência de aminoácidos pode fornecer, além de visualizar a região onde os motivos conservados estão localizados. A predição da estrutura secundária da GumH foi realizada com o programa PSIPRED v2.3 (Jones, 1999), que fez uma comparação da seqüência de aminoácidos da GumH com bancos de dados de estruturas secundárias estabelecidas experimentalmente (figura 4.3). A predição das estruturas secundárias da GumH revelou a presença de 46,8% de α-hélices, 18,1% de folhas β e finalmente 35,1% de conexões do tipo *loops*, *turns* e *coils*.

Pela figura 4.3 pode-se verificar que as regiões conservadas englobam um *loop* e uma α-hélice.

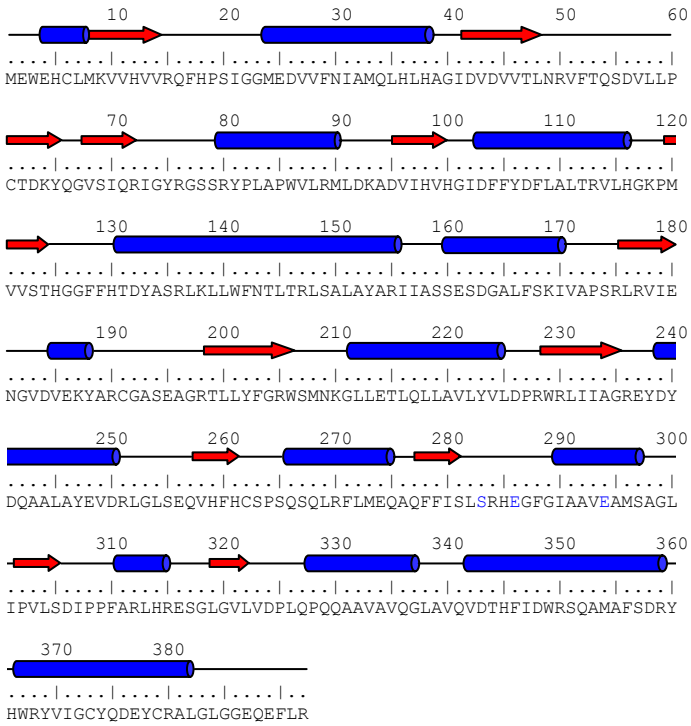


Figura 4.3: Predição da estrutura secundária da GumH. A predição da estrutura secundária (PSIPRED) está sobre a seqüência. Os cilindros azuis representam α-hélices e as setas vermelhas representam as folhas β enquanto que os *loops* estão representados por um traço preto. Os aminoácidos pertencentes ao motivo SX_2EX_7E estão destacados em azul.

4.5 Predição de regiões transmembrânicas

A seqüência primária da enzima foi submetida a diversos programas que fazem a predição das possíveis regiões transmembrânicas. Cada programa possui o seu próprio algoritmo, sendo assim, pequenas diferenças em relação ao número de regiões transmembrânicas podem ser esperadas. Como pode ser observado na tabela 4.3, a GumH possui no máximo uma região transmembrânica, podendo ainda ser considerada uma proteína solúvel dependendo do programa utilizado (capítulo 3 – 3.3.4).

Tabela 4.3: Predição da região transmembrânica para a GumH. O número de regiões não é o mesmo para todos os programas.

Programas Utilizados e Regiões Transmembrânicas Previstas							
HMMTOP	**DAS**	**MEMSAT2**	**SOSUI**	**SPLIT 4.0**	**TMHMM**	**Tmpred**	**TOPPRED 2**
----	1	----	----	1	----	1	1

A figura 4.4 ilustra um gráfico obtido pelo programa TOPPRED 2 (von Heijne, 1994) do momento hidrofóbico em relação a cada um dos resíduos da seqüência primária da enzima GumH. O programa TOPPRED 2 acusou apenas uma provável região transmembrânica e que está compreendida entre os resíduos 284 e 291. Devido ao tamanho da seqüência de aminoácidos da provável região transmembrânica (7 resíduos) a maioria dos programas foram contrários a predição do TOPPRED 2 e em uma análise estrutural no modelo da GumH foi verificado que essa região dificilmente poderia fazer qualquer tipo de contato com a membrana da bactéria.

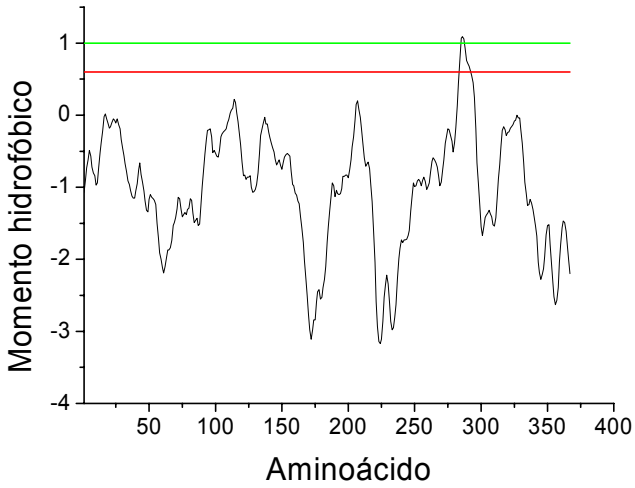


Figura 4.4: Gráfico do momento hidrofóbico em relação aos resíduos da enzima GumH.

4.6 A construção dos modelos

O modelo estrutural da GumH foi construído seguindo as 4 etapas descritas no item 3.3.2 do capítulo 3.

4.6.1 Busca por proteínas homólogas a GumH através de *threading*

A busca por proteínas cujas estruturas secundárias sejam homólogas a estrutura secundária predita para a GumH foi feita com o programa GenThreader (Jones, 1999). O arquivo gerado pelo programa mostrou níveis de confiança significativos estatisticamente para a GumH quando comparada a 4 estruturas depositadas nos bancos de dados:

1- 1F0K, código PDP da glicosiltransferase MurG envolvida na biossíntese do peptídeoglicano em *E.coli*. Esta enzima catalisa a transferência de N-acetil glucosamina da UDP para o carbono hidroxil do pentapeptídeo N-acetilmuramoil ligado a lipídio (Ha *et al.*, 2000).

2- 1F6D, código PDB da enzima UDP-N-Acetilglucosamina 2-Epimerase de *E.coli*. Esta enzima (chamada neste trabalho de Epimerase) catalisa a interconversão reversível de UDP-N-acetilglucosamina e UDP-N-acetilmanosamina (Campbell *et al.*, 2000).

3- 1IIR, código PDB da glicosiltransferase GtfB de *Amycolatopsis orientalis*, enzima que catalisa a transferência de glicose de UDP-glicose para o resíduo 4-OH-Phegly4 para formar o glicopeptídeo antibiótico da família da vancomicina (Mulichak *et al.*, 2001).

4- 1C3J, código PDB da β-glicosiltransferase , chamada neste trabalho de β-GT, do bacteriófago T4. Esta enzima catalisa a transferência de glicose de

UDP-glicose para 5-hidroximetilcitosina em folha dupla de DNA (Vrielink *et al.*, 1994; Moréra *et al.*, 1999).

Das quatro enzimas citadas, três são glicosiltransferases, porém nenhuma delas é GDP-manosiltransferase.

O resultado do *threading* protéico obtido pelo programa GenThreader pode ser visualizado na tabela 4.4. A primeira coluna da tabela (*Conf*), informa a estatística da qualidade do alinhamento que forma o nível de confiança, a segunda mostra o valor *E-value* do alinhamento que depende do número de aminoácidos estruturalmente alinhados, da energia de formação dos pares durante o alinhamento, da energia de solvatação (colunas não mostradas), a terceira coluna (*Alen*) informa o número total de resíduos das proteínas alvo-molde alinhados, a quarta coluna (*Dlen*) indica o comprimento total de aminoácidos da proteína molde, (*Tlen*) caracteriza a número total de resíduos da proteína alvo e finalmente (*COD PDB*) indica o código PDB padrão composto por quatro caracteres alfanuméricos seguidos do código indicador de cadeia do oligômero “A”, “B”, “C”, etc. e do número do domínio da proteína.

Tabela 4.4: Resultado da busca realizada pelo programa GenThreader. As proteínas encontradas foram a 1F6D (UDP-N-Acetilglucosamina 2-Epimerase, *E. coli*), 1F0K (MurG, *E. coli*), 1C3J (β-GT, fago T4) e 1IIR (GtfB, *Amycolatopsis orientalis*).

Conf	E-value	Alen	DLen	Tlen	COD PDB
CERT	0.875 e-1	347	376	387	1F6DA0
CERT	0.857 e-6	334	351	387	1F0KA0
CERT	0.846 e-1	317	333	387	1C3JA0
CERT	0.779 e-5	342	382	387	1IIRA0

Além da busca, o programa faz alinhamentos das seqüências, baseado nas estruturas secundárias das moléculas, que serão utilizados posteriormente na construção dos modelos. Um resultado típico de alinhamento seqüencial feito pelo programa GenThreader está mostrado na figura 4.5, que representa o alinhamento entre a GumH e a MurG (1F0KA0). Na figura 4.5, hélices alinhadas estão em azul, folhas β alinhadas estão em vermelho e as conexões (*loops* e *turns*) em negrito. As barras verticais verdes referem-se a aminoácidos idênticos, que foram alinhados independendo da estrutura secundária. Durante a análise do modelo, as duas extremidades que não obtiveram pareamento durante o alinhamento foram retiradas (aminoácidos destacados em cinza).

```
                                10        20        30        40
        -------CEEEE----ECCCCHHHHHHHHHHHH-HCCCCEEEEEECCCCCHHHHHHHH
1f0kA0  -------KRLMV----MAGGTGGHVFPGLAVAHHL-MAQGWQVRWLGTADRMEADLVPKH
                  ||          |||||||||||||| | ||||||||||||||     |
GumH    MEWEHCLMKVVHVVRQFHPSIGGMEDVVFNIAMQLHLHAGIDVDVVTLNRVFTQSDVLLP
        CCCHHHHHHHEEEECCCCCCCCCHHHHHHHHHHHHHHHHCCCCEEEEEECCCCCCCCCCCC
                10        20        30        40        50        60

                   60        70        80        90       100
        CCEEEECCCCCCCCCCHHHHHCCHHHHHHHHHHHHHHHHHHCCCEEEECCCCCHH-HHHH
1f0kA0  GIEIDFIRISGLRGKGIKALIAAPLRIFNAWRQARAIMKAYKPDVVLGMGGYVSG-PGGL
        |||||  |      |||| ||||  |||||||            ||||||||   || ||||
GumH    CTDKYQGVSIQRIGYRGSSRYPLAPWVLRML---------DKADVIHVHGIDFFYDFLAL
        CCEEECCEEEEECCCCCHHHHHHHHHHHHHHC---------CCCCEEEECCHHHHHHHHHH
                70        80        90                100       110

        110       120                     130       140       150
        HHHHCCCCCEEEEECCCCCC--------------HHHHHHCCCCCEEEECCCCCCCCEEC
1f0kA0  AAWSLGIPVVLHEQNGIAG--------------LTNKWLAKIATKVMQAFPGAFPNAEVV
        |||||||||||| ||||||               |||||||||| ||||||
GumH    TRVLHGKPMVVSTHGGFFHTDYASRLKLLWFNTLTRLSALAYA-RIIASSESDGALFSKI
        HHHHCCCCCEEEECCCCCCCHHHHHHHHHHHHHHHHHHHCCCC-EEEECCHHHHHHHHHH
              120       130       140       150        160       170

            160       170         180       190       200
        CCCCCHHHHCCCCHHHHHCC--CCCCEEEEEECCCCCCHHHHHHHHHHHH---HHHEE
1f0kA0  GNPVRTDVLALPLPQQRLAG--REGPVRVLVVGGSQGARILNQTMPQVAAKL---GDSVT
        |||||  | |||| ||||||  |||| |||||  |  ||   ||||||||||      ||
GumH    VAPSRLRVIENGVDVEKYARCGASEAGRTLLYFGRWSMNKGLLETLQLLAVLYVLDPRWR
        CCCCCEEEECCCCCHHHHCCCCCCCCCCEEEEEEEECCCCCHHHHHHHHHHHHCCCEE
               180       190       200       210       220       230

                 220         230         240       250
        EEEECCCCCHHH--HHHHHHHCCCCCCEEEC---CCCCHHHHHHHCCEEE----ECCCHH
1f0kA0  IWHQSGKGSQQS--VEQAYAEAGQPQHKVTE---FIDDMAAAYAWADVVV----CRSGAL
        |||||||| |||  |||||||||||| |||    |   ||||||||||||     ||| |
GumH    LIIAGREYDYDQAALAYEVDRLGLSEQVHFHCSPSQSQLRFLMEQAQFFISLSRHEGFGI
        EEEECCCCHHHHHHHHHHHHCCCCCEEEECCCCCHHHHHHHHCCEEEECCCCCCCCCH
               240       250       260       270       280       290

                270       280       290       300       310
        HHHHHHHHCCCEEECCCCCCCCCHHHHHHHHHHHHCCCEEECCCCCCCHHHHHHHHHCCCHH
1f0kA0  TVSEIAAAGLPALFVPFQHKDRQQYWNALPLEKAGAAKIIEQPQLSVDAVANTLAGWSRE
        |||||||||||||||||||        ||||  |||||| ||||  |||||||||  | |
GumH    AAVEAMSAGLIPVLSDIPP-------FARLHRESGLGVLVDPLQPQQAAVAVQGLAVQVD
        HHHHHHHCCCCEEECCCCC-------CHHHHCCCCCCEEEECCCCHHHHHHHHHHHCCCH
               300               310       320       330       340

                  330       340       350
        HHH--HHHHHHHHCCCCCHHHHHHHHHHHHHCCC----------
1f0kA0  TLL--TMAERARAASIPDATERVANEVSRVARAL----------
        |||  ||||||||   |||||||||||||||||
GumH    THFIDWRSQAMAFSDRYHWRYVIGCYQDEYCRALGLGGEQEFLR
        HHHHHHHHHHHHHHHHCCCHHHHHHHHHHHHHHHCCCCCCCCCCC
            350       360       370       380
```

Figura 4.5: Alinhamento gerado pelo programa GenThreader após a busca em banco de proteínas. A proteína mostrada refere-se a 1F0K ou MurG (*E. coli*). “H”, “E” e “C” significam α-hélice, folha β e conexões respectivamente.

Após uma análise da figura 4.5, é possível notar a grande diferença no alinhamento entre as duas proteínas ao levar-se em consideração a

seqüência de aminoácidos, representados por traços verdes (não passaria de 21% de identidade) e o alinhamento, considerando-se a predição da estrutura secundária (que atinge a marca de 78%). Estudos, cujos resultados validam a técnica de *threading*, têm mostrado que algumas proteínas de baixa identidade seqüencial apresentam enovelamentos bastante semelhantes (Mulichak *et al.*, 2001).

Os alinhamentos seqüenciais efetuados entre a GumH e as demais proteínas Epimerase, GtfB, e β-GT não são mostrados, mas foram tão bons quanto o realizado com a MurG.

4.6.2 Modelagem molecular

Uma vez gerados os alinhamentos pela técnica de *threading*, a próxima etapa foi a da construção dos modelos através da modelagem molecular com a utilização do programa MODELLER 6.0a (Sali *et al.*, 1993).

As coordenadas atômicas das quatro estruturas cristalográficas que obtiveram os melhores índices de confiança, ou seja, a MurG, Epimerase, GtfB, e β-GT que foram depositadas no banco de dados PDB com uma resolução maior que 2,5 Å foram utilizadas individualmente como molde na construção dos modelos para a GumH. As coordenadas das águas e ligantes foram removidas de suas respectivas estruturas durante o processo de modelagem.

O programa MODELLER 6.0a gerou 50 modelos para cada um dos quatro moldes, e um trabalho de validação foi efetuado para que os modelos de melhor qualidade fossem filtrados. São considerados modelos ruins,

aqueles que, apesar de todas as restrições que o programa MODELLER 6.0a impõe, ainda possuem enovelamentos incorretos em diversas regiões de sua estrutura.

4.6.3 Escolha dos modelos

Dos 50 modelos gerados pelo programa MODELLER 6.0a, cinco representantes foram escolhidos de cada proteína molde empregando-se o critério inicial de menor energia fornecida pelo próprio programa. O gráfico de energia obtido para os 50 modelos gerados a partir da estrutura da MurG é mostrado na figura 4.6; os gráficos obtidos para os outros modelos estão no **apêndice A**.

Os cinco moldes escolhidos, de acordo com o critério de menor energia, foram os modelos de número 1, 7, 21, 34 e 44.

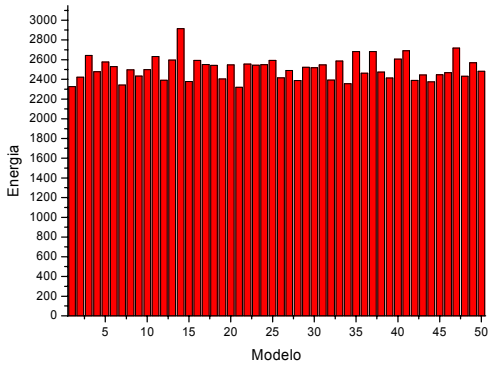


Figura 4.6: Gráfico da energia para os modelos da GumH tendo como molde a proteína 1F0K (MurG). A unidade de energia não possui nenhum significado físico, é um número atribuído a cada modelo gerado pelo programa MODELLER 6.0a com a finalidade de se obter um índice de classificação.

4.6.4 Validação dos modelos gerados

Escolhidos os cinco modelos, os programas PROCHECK (Laskowski *et al.*, 1998) e WHATIF (Vriend, 1990) (REFINE/REFI, significa que a ferramenta utilizada foi a REFI, que pertence ao programa REFINE que por sua vez está contido no pacote de programas WHATIF), foram utilizados para testes iniciais em relação à qualidade dos mesmos (dados não exibidos). Os quatro modelos que apresentaram os piores resultados foram descartados. Sendo assim, apenas um modelo foi levado em consideração e posteriormente otimizado por diversas técnicas, como a biblioteca de rotâmetros do programa "O" (Kleywegt e Jones, 1995) e *scripts* do MODELLER 6.0a para otimização de *loops*, voltas (*turns*) e conexões.

Os procedimentos de validação que serão descritos a seguir foram realizados para os quatro modelos finais de GumH, construídos baseados nas estruturas tridimensionais das proteínas MurG, Epimerase, MtfB e β-GT. Somente os resultados obtidos para o modelo da GumH construído através da estrutura da MurG serão discutidos.

Após a etapa inicial de escolha do melhor modelo, análises mais criteriosas e mais detalhadas foram efetuadas com a utilização do VERIFY 3D, o WHATIF (QUALTY/OLDQUA significa que a ferramenta utilizada foi a OLDQUA, que pertence ao programa QUALTY que por sua vez pertence ao pacote de programas WHATIF), além do PROCHECK.

Durante o processo de escolha dos melhores rotâmeros os aminoácidos rotacionados e/ou transladados podem sofrer algum tipo de distorção entre os ângulos da cadeia principal, lateral e até a planaridade

dos mesmos pode ficar comprometida. O programa REFI, citado anteriormente, faz a correção automática dessas distorções.

4.6.4.1 WHATIF

O modelo foi submetido à avaliação de suas vizinhanças atômicas com o programa WHATIF. De maneira geral os resíduos com valor menor que -5,0 devem ser checados quanto ao seu desvio excepcional da média.

A avaliação do modelo revelou que os ambientes químicos de resíduos individuais enquadram-se nos padrões freqüentemente encontrados em estruturas protéicas com um valor referente a todos os resíduos calculado em aproximadamente -1,5 (**Apêndice B**). Segundo os valores de referência do programa, o modelo pode ser considerado de boa qualidade.

4.6.4.2 PROCHECK

Os modelos tiveram suas estruturas avaliadas através de comparações de parâmetros freqüentemente encontrados nas proteínas depositadas em bancos de dados internacionais como, por exemplo, o PDB.

Quanto a posição dos resíduos analisados no diagrama de Ramachandran (356 resíduos no total), o modelo da GumH a partir da 1F0K (figura 4.7) exibiu 86,4% dos resíduos nas regiões mais favoráveis, 12,3% em regiões adicionais, 1,3% em regiões generosamente permitidas e finalmente 0% de resíduos em regiões não permitidas. Os resíduos de glicina (24 no total) possuem como cadeia lateral um átomo de hidrogênio,

logo, o Cα não possui quiralidade e são representados por triângulos no gráfico de Ramachandran.

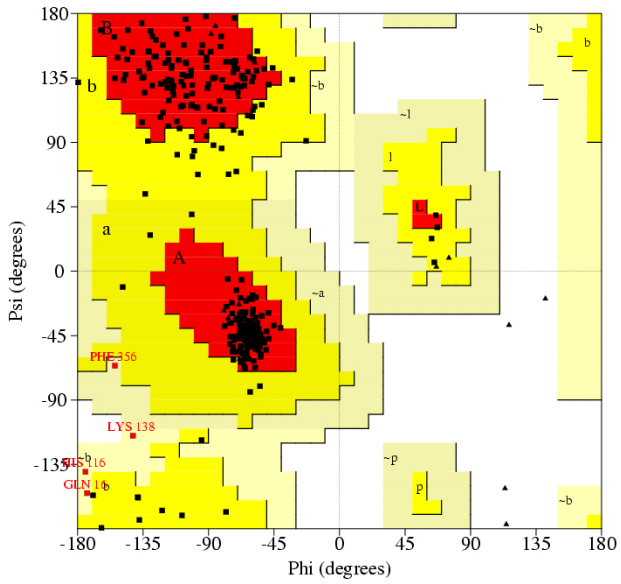


Figura 4.7: Gráfico Ramachandran do modelo final obtido para a GumH. Os ângulos Phi e Psi estão em graus.

Os aminoácidos que estão em destaque no gráfico Ramachandran (figura 4.7) Gln16, Lys138, His166 e Phe356 não fazem parte do sítio ativo da proteína. A Gln16, His166 e Phe356 estão localizadas em uma região periférica em contato com o ambiente (solvente), já a Lys138 está localizada em uma região intermediária entre o solvente e a sítio ativo da proteína, ou seja, “enterrada” na proteína. Uma localização mais precisa desses resíduos pode ser visualizada na figura 4.8.

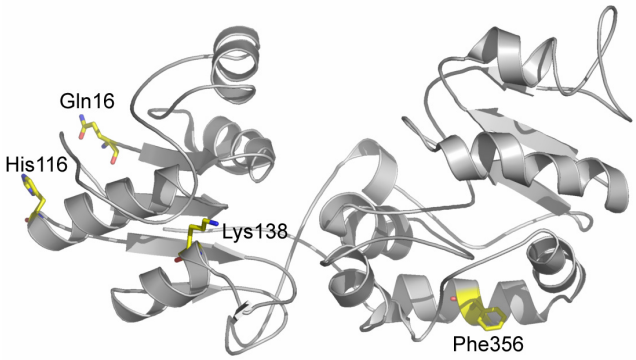


Figura 4.8: Modelo da GumH destacando os aminoácidos localizados nas regiões generosas no gráfico Ramachandran. A glutamina 16 (Gln16), a histidina 116 (His116), a fenilanlanina 356 (Phe356) e a lisina 138 (Lys138) representam o 1,3% dos resíduos em regiões generosamente permitidas no gráfico Ramachandran.

Na comparação estrutural da enzima com demais estruturas refinadas na resolução de 2,0 Å (no caso de modelagem molecular a resolução de 2,0 Å é comumente empregada), parâmetros da cadeia principal podem ser abordados (figura 4.9). As regiões destacadas na cor violeta nesses gráficos representam as faixas de resultados obtidos por essas estruturas, sendo a linha central o ajuste mínimo quadrático para a inclinação média em função da resolução enquanto que a largura da banda corresponde à variação de um desvio padrão sobre a média. A figura 4.9 **A** ilustra a qualidade do diagrama de Ramachandran a figura 4.9 **B** ilustra a planaridade da ligação peptídica (desvio padrão para os ângulos ω das diversas estruturas). Quanto menor o desvio, mais próximo de 180° (o que representa uma ligação peptídica perfeita). O G_{factor} total (figura 4.9 **C**), é uma medida da normalidade da estrutura como um todo. Representa uma média para os G_{factor} de cada resíduo na estrutura. A figura 4.9 **D** ilustra uma medida da distorção tetraédrica dos $C\alpha$. Esta propriedade é efetivada pelo cálculo do

desvio padrão do “ângulo torsional ζ”, um ângulo imaginário definido por quatro átomos em um dado resíduo: Cα, N, C e Cβ. Observando o gráfico verifica-se que as distorções do modelo estão abaixo do esperado para uma proteína com resolução de 2,0 Å. Sendo assim, o modelo pode ser considerado acima da média.

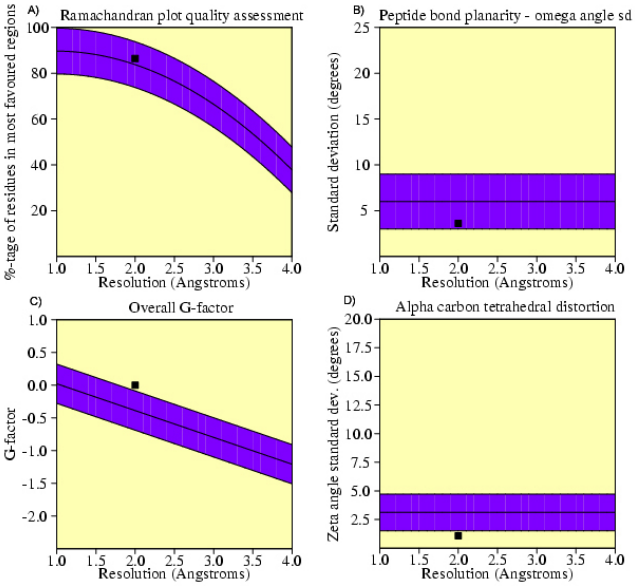


Figura 4.9: Estatísticas de alguns parâmetros estereoquímicos do modelo da GumH (representado pelo quadrado preto), quando comparado a estruturas refinadas a mesma resolução: **A** diagrama de Ramachandran, **B** planaridade da ligação peptídica, **C** G_{factor} e **D** distorção tetraédrica.

4.6.4.3 VERIFY 3D

Na verificação do modelo através do programa VERIFY 3D (http://www.doe-mbi.ucla.edu/Services/Verify_3D), foram encontradas algumas regiões nas quais os aminoácidos apresentaram uma conformação desfavorável, permanecendo abaixo da linha zero. Esses resultados são mostrados na figura 4.10. As regiões que estão abaixo do ponto zero no

gráfico possivelmente apresentam problemas quanto ao ambiente no qual os aminoácidos estão contidos.

Os aminoácidos acusados pelo VERIFY 3D são: Ala84, Pro85, Trp86, Phe260 e His261, que estão localizados em uma região de *loop*, Ala339, Val340, Gln341, Val342, Asp343, Thr344, His345 e Arg359 pertencentes a α-hélices. Nenhum desses aminoácidos localizados abaixo da linha zero (linha vermelha) pertence ao sítio ativo da proteína como pode ser observado na figura 4.11.

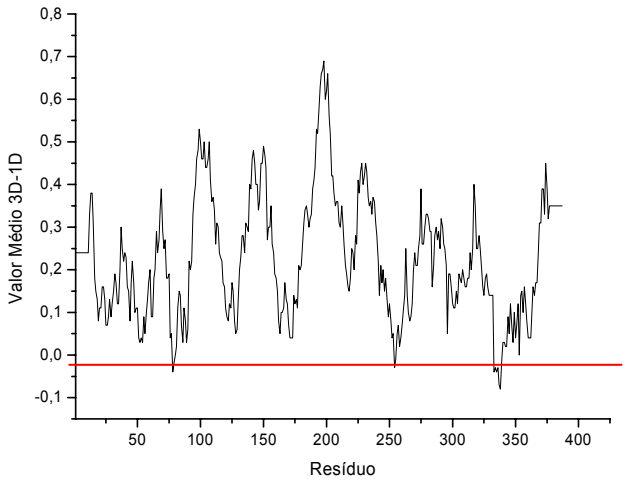


Figura 4.10: Gráfico do programa VERIFY 3D para o modelo da GumH. Os valores ficaram entre -0,08 (ALA 84) e 0,69 (GLY 204).

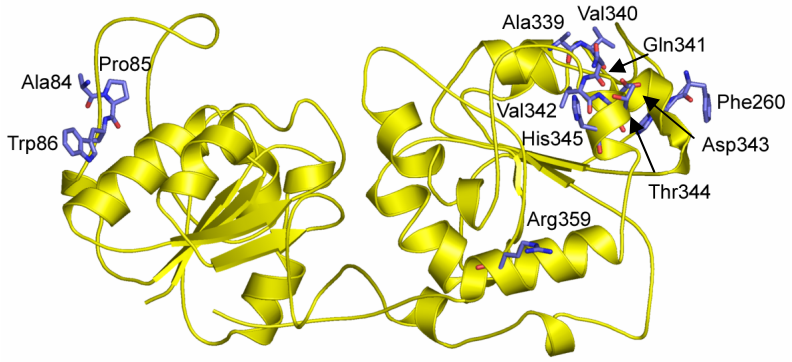


Figura 4.11: Modelo da GumH destacando os resíduos localizados abaixo da linha zero no VERIFY 3D.

4.7 Os modelos

Os quatro modelos finais construídos e validados para a proteína GumH são mostrados na figura 4.12 (**A, B, C** e **D**). Todos os modelos possuem certa homologia estrutural, com os domínios N- e C-terminais bastante similares em tamanho e topologia. Ambos domínios contém uma estrutura central similar de folhas β paralelas conectadas por α-hélices. Este motivo é chamado de *Rossmann fold* e é característico de domínios que ligam nucleotídeos (Branden e Tooze, 1998). No caso da GumH, o doador do açúcar é uma molécula de GDP-manose (Guanosina 5'-Difosfato-manose), sendo o nucleotídeo uma guanosina.

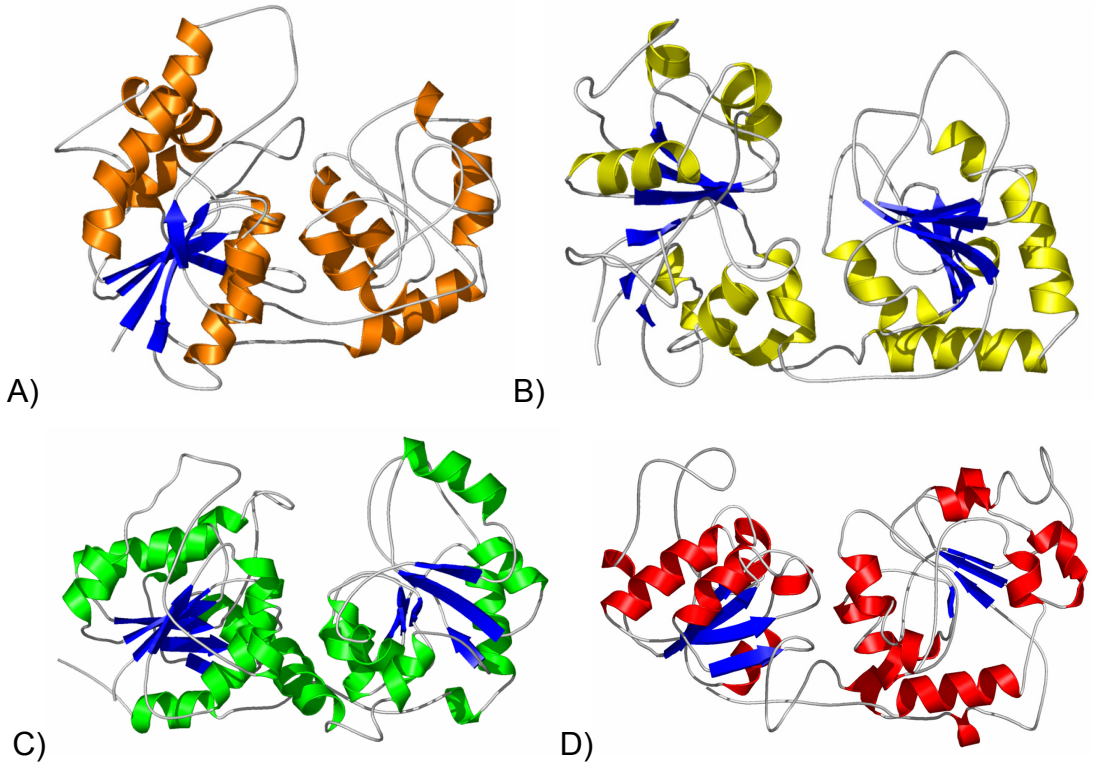


Figura 4.12: Modelos obtidos para a GumH a partir dos moldes sugeridos pelo programa GenThreader. As figuras A, B, C e D, ilustram modelos da GumH gerados tendo como molde a Epimerase, GtfB, β-GT e MurG, respectivamente.

A figura 4.13 mostra a comparação topológica do modelo gerado para a GumH e da estrutura cristalográfica da MurG (figura 4.13 **A**), onde é possível notar a semelhança entre as duas estruturas. A figura 4.13 **B** mostra que o domínio N-terminal do modelo é constituído por 4 folhas β paralelas e 6 α-hélices, já o domínio C-terminal é constituído por 4 folhas β paralelas e 8 α-hélices. Vale lembrar que o modelo da GumH apresenta nesta etapa 361 resíduos, pois 16 resíduos do N-terminal e 10 resíduos do C-terminal foram desconsiderados (já que não foram pareados durante o alinhamento), como pode ser visto na figura 4.4 na cor cinza.

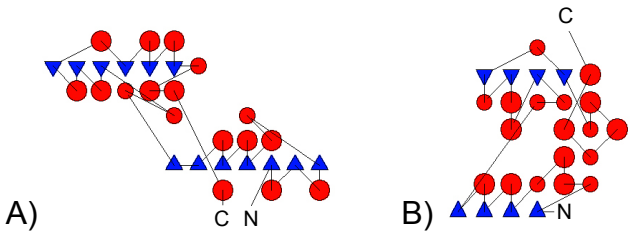


Figura 4.13: Comparação das topologias. Topologia obtida para a estrutura cristalográfica da MurG **A** e a topologia do modelo obtido para a enzima GumH **B**. Folhas β são mostradas em triângulos e α-hélices em círculos, enquanto que N e C indicam N-terminal e C-terminal respectivamente.

As figuras 4.14 **A, B, C** e **D** mostram a sobreposição dos modelos com as estruturas das moléculas moldes. Nota-se que, apesar de baixa identidade seqüencial, a estrutura secundária é bastante conservada devido ao tipo de alinhamento global/estrutural que a seqüência de aminoácidos da GumH foi submetida.

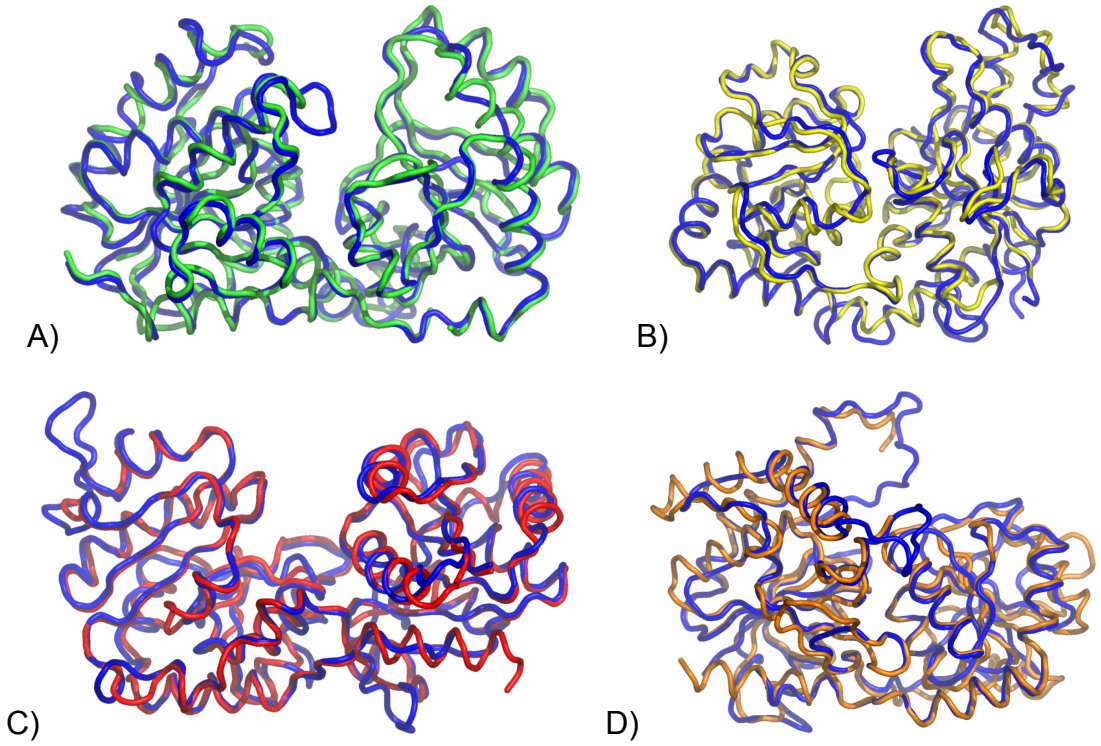


Figura 4.14: Superposição dos modelos gerados para a GumH a partir das estruturas das proteínas **A** β-GT, **B** GtfB, **C** MurG e **D** Epimerase. O modelo da enzima GumH está em azul.

Após a análise, com programas gráficos dos modelos gerados e da validação dos modelos obtidos, observou-se que o modelo estrutural da GumH construído a partir da estrutura cristalográfica da MurG é melhor em termos dos parâmetros analisados, anteriormente mencionados. Portanto, este modelo passou a ser mais bem estudado para verificação dos contatos entre resíduos e também foi o usado nos estudos de *docking*.

Das quatro estruturas utilizadas como moldes, Epimerase, GtfB, β-GT e MurG as duas últimas foram determinadas na presença do substrato UDP (Mulichak *et al.*, 2001; Hu *et al.*, 2003). A ligação do nucleotídeo no domínio C-terminal revelou a topologia do 'bolsão' de ligação e também mostra importantes contatos para o nucleotídeo. As figuras 4.15 **A** e **B** ilustram o

modelo da GumH ressaltando a provável região do sítio ativo, com a α-hélice que contém importantes resíduos que devem fazer contatos com o substrato GDP-manose.

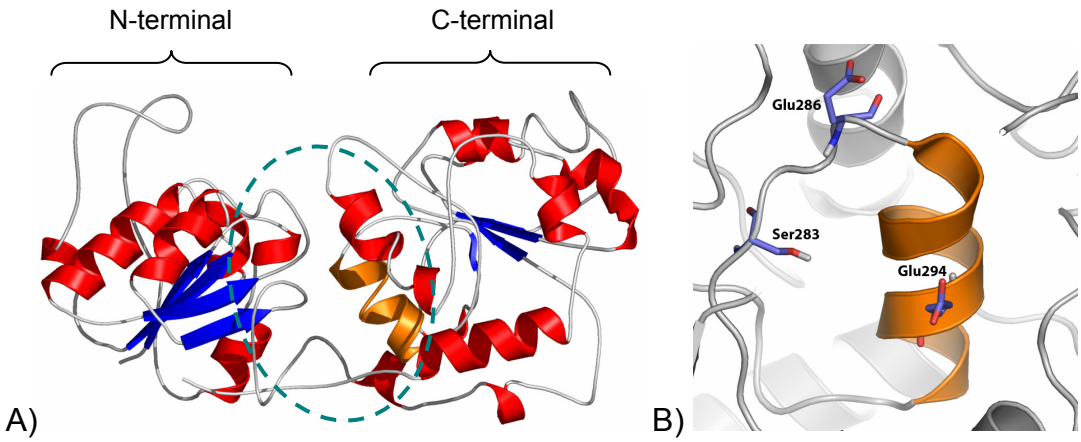


Figura 4.15: Região do provável sítio ativo e aminoácidos envolvidos na catálise da GDP-manose. Figura **A** ilustra a provável região do sítio ativo "pontilhados" e a α-hélice SX_2EX_7E (alaranjada) envolvida na catálise. A figura **B** ilustra a posição espacial dos aminoácidos presentes no motivo SX_2EX_7E conservado em diversas glicosiltransferases da família 4.

A figura 4.16 mostra em destaque a região conservada (cor "salmão") obtida durante o alinhamento global realizado pelo programa Pfam (região que compreende os aminoácidos Asp184 - Phe356) previamente discutidos na seção 4.2. Quando comparada às figuras 4.15 **A** e **B**, verifica-se que a região em destaque engloba praticamente todo o C-terminal da proteína, incluindo a α-hélice contendo o motivo conservado.

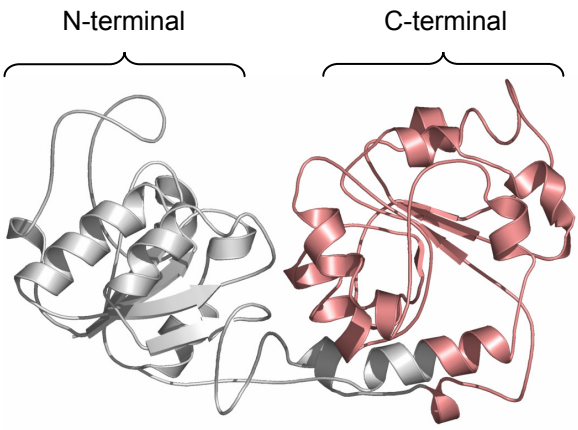


Figura 4.16: Região conservada da estrutura durante o alinhamento obtido pelo programa Pfam. A região conservada está representada em destaque na cor "salmão".

4.8 A região catalítica e estudos de *docking*

A reação catalisada pela GumH, a transferência de uma manose de GDP-manose (substrato doador) para uma molécula de celobiose ligada a um pirofosfato-poliprenol (substrato aceptor), pode ser ilustrada pela figura 4.17:

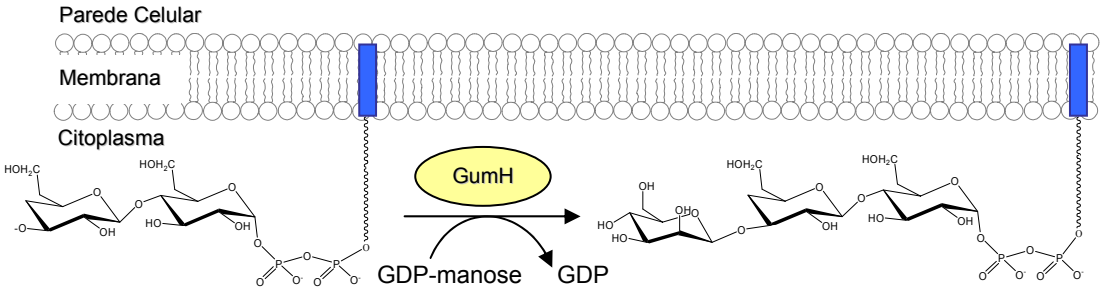


Figura 4.17: Reação catalisada pela GumH. Os retângulos azuis na membrana representam o lipídio carreador (poliprenol).

Com o objetivo de se determinar as interações existentes entre resíduos da enzima e as moléculas dos substratos doadores e aceptores do

açúcar, um estudo de *docking* entre o modelo da GumH e a GDP-manose (substrato doador) foi feito através do programa FLO (McMartin e Bohacek, 1997). Este programa tem como principal característica a minimização de um potencial energético AMBER (Weiner *et al.*, 1986) cuja finalidade é encontrar mínimos locais e globais dessa energia potencial dada uma região pré-definida do suposto sítio catalítico da proteína em estudo.

Durante o processo de *docking*, informações como provável região catalítica da proteína e possíveis aminoácidos envolvidos na catálise são pré-definidos pelo usuário. O programa faz buscas a partir de posições iniciais randômicas para o substrato que será ligado à enzima e através da minimização da energia potencial, o programa consegue ajustar o substrato na configuração de menor energia possível dentro do sítio ativo da enzima. No presente trabalho, 1200 diferentes posições para o substrato GDP-manose foram testadas pelo programa. O programa FLO conta ainda com um esquema de cores para interpretar e diferenciar os aminoácidos que fazem parte da estrutura protéica dos aminoácidos ou demais átomos (heteroátomos) que compõem o substrato ou ligante passível ao *dock*. Por exemplo, na enzima a cor vermelha e roxa caracterizam aminoácidos considerados rígidos e flexíveis, respectivamente. No ligante ou substrato a cor que desempenha igual característica é a cor alaranjada e verde, respectivamente.

A região da molécula de GumH que foi definida para estudos de *docking* é mostrada na figura 4.18. Em vermelho estão os aminoácidos definidos como parte do sítio ativo da enzima selecionados a partir de um raio de corte de 15 Å do centro de uma esfera imaginária centrada no motivo

conservado EX_7E. A parte cinza caracteriza a região de busca que o programa FLO irá considerar durante o processo de minimização energética.

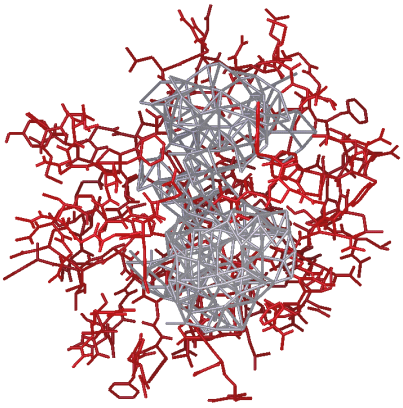


Figura 4.18: Região da molécula GumH definida para os estudos de *docking*.

Após as 1200 tentativas de *docking*, em um estudo mais refinado dos aminoácidos que fazem parte da região representada pela malha cinza na figura 4.18 que engloba os aminoácidos do motivo altamente conservado SX_2EX_7E, alguns aminoácidos tiveram suas cadeias laterais substituídas por cadeias laterais de rotâmeros (aminoácidos com diferentes conformações para suas cadeias laterais) presentes no banco de rotâmeros do programa "O". Este procedimento foi adotado para que o substrato GDP-manose ficasse em uma posição de menor energia possível e realizando interações entre os aminoácidos conservados presentes no motivo SX_2EX_7E.

Diversos refinamentos das posições das cadeias laterais dos aminoácidos e subseqüentes minimizações energéticas foram realizados com o objetivo de se evitar impedimentos estéricos e proporcionar a maximização do número de pontes de hidrogênio garantindo maior estabilidade na interação enzima/substrato. Alguns aminoácidos

apresentaram uma maior mudança em seu estado conformacional com relação aos átomos que compõem suas respectivas cadeias laterais para garantir maior acomodação do substrato GDP-manose. A figura 4.19 ilustra alguns desses aminoácidos. Um RMS (***R**oot **M**ean **S**quare*) de 0,15 Å sugere a conservação da cadeia principal (Cα) após o processo de minimização energética.

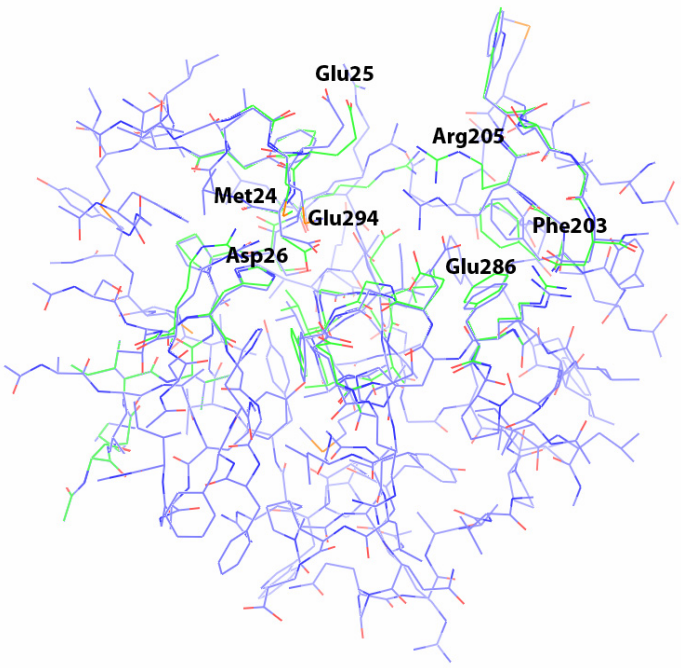


Figura 4.19: Sítio catalítico antes e após aplicações sucessivas de rotâmeros e minimizações energéticas. Os átomos de carbono do sítio catalítico antes e depois das minimizações energéticas estão representados em azul e verde, respectivamente.

O resultado final do *docking* apresentou energia total estimada para a ligação de E_{ass} = -88,8 kJ/mol, sendo geralmente observada entre - 30 a - 90 kJ/mol para compostos de ligações menores que micromolares. Logo, a energia obtida no estudo de *docking* foi bastante satisfatória. Outro parâmetro a ser analisado em um estudo de *docking* é a energia de ligação relativa ao mínimo global obtido durante a minimização de energia do

potencial, representado por E_{lig} e que foi igual a 8,4 kJ/mol. Este é um ótimo valor de energia, principalmente quando comparado ao intervalo geralmente obtido neste tipo de estudo, de 10 e 30 kJ/mol. Esses resultados indicam que o estudo de *docking* foi eficiente para determinar a região de ligação do substrato doador.

Como pode ser observado nas figuras 4.20 **A** e **B** o substrato doador GDP-manose fica localizado na região mais próxima ao C-terminal da proteína e na parte inferior da fenda que separa os dois domínios e onde se encontra o motivo SX_2EX_7E já mencionado (figura 4.15).

Um gráfico (figuras 4.21 **A** e **B**) de superfície da região catalítica da GumH e o ligante GDP-manose ilustra melhor o grau de interação e localização em termos da profundidade em que se encontra o substrato. A figura **A** ilustra toda a superfície da proteína, já a figura **B** mostra em detalhes, por meio de uma superfície ligeiramente transparente, a interação e o canal formado pelos átomos na vizinhança do grupo fosfato.

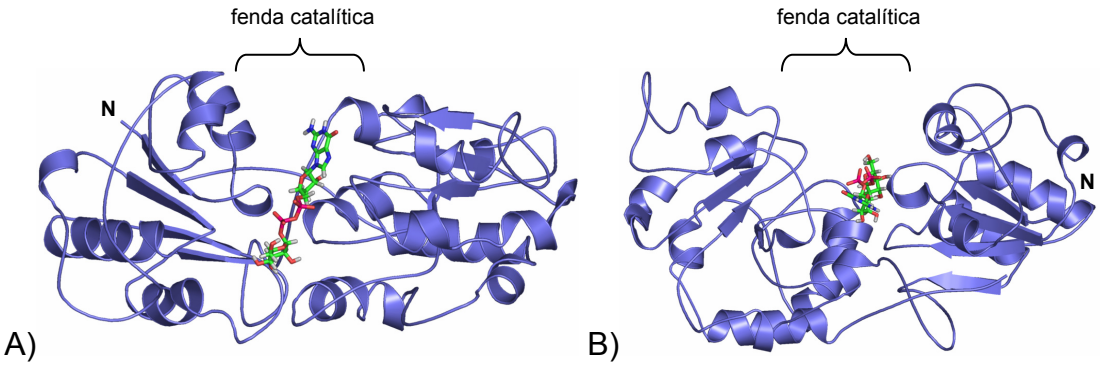


Figura 4.20: Complexo enzima/substrato. As figuras **A** e **B** ilustram o complexo enzima/substrato em duas perspectivas: superior e lateral, respectivamente. O N-terminal está indicado pela letra “N”.

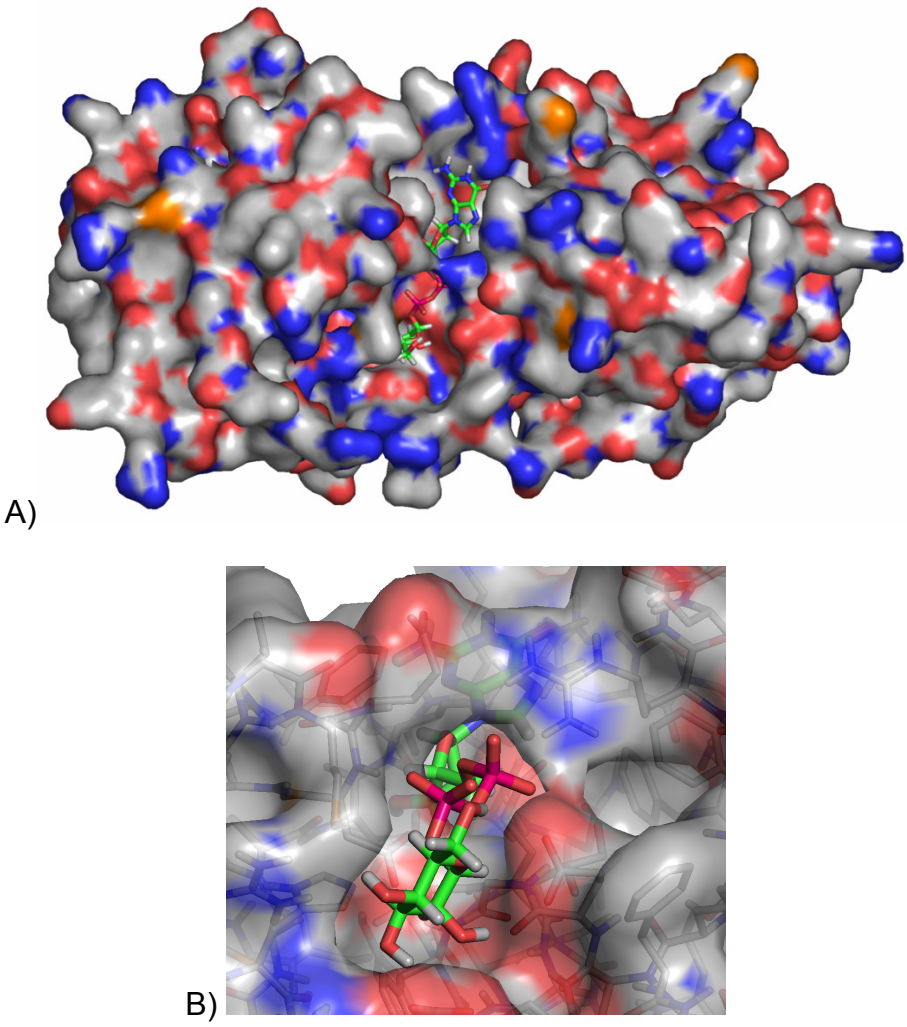


Figura 4.21: Gráfico de superfície da provável região catalítica da GumH e o ligante GDP-manose ao centro. As cores correspondem aos átomos de oxigênio (vermelho), nitrogênio (azul), enxofre (alaranjada) e cinza (carbono).

4.9 Interações enzima/substrato

Em termos de interações em níveis atômicos, algumas características como número de pontes de hidrogênio, interações hidrofóbicas, interações eletrostáticas, enfim, as complementaridades nas ligações ou interações entre enzima e substrato foram checadas. A figura 4.22 ilustra as principais pontes de hidrogênio existentes entre a GDP-manose e os aminoácidos ao seu redor. Observa-se que o sacarídeo manose faz pontes de hidrogênio

com a Met24 e com o Glu286, o grupo fosfato faz pontes de hidrogênio com o Asp26 e uma ponte hibridizada com dois hidrogênios da Arg205; já o anel ribose da base guanina faz uma ponte de hidrogênio com o Glu294. A tabela 4.5 traz descritos os átomos que realizam pontes de hidrogênio e os valores das mesmas.

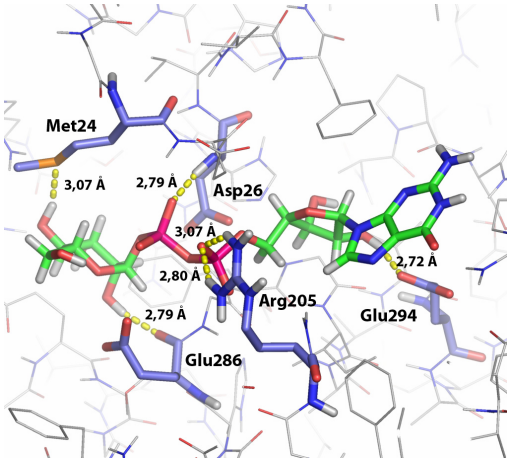


Figura 4.22: Interações entre enzima e substrato. Os átomos de carbono do ligante GDP-manose estão destacados pela cor verde e os átomos de carbono dos principais aminoácidos que interagem com a GDP-manose estão em azul e dos demais aminoácidos da proteína estão em cinza.

Tabela 4.5: Principais pontes de hidrogênio envolvidas na coordenação do substrato GDP-manose no proposto sítio ativo da enzima GumH, desprezando-se os hidrogênios.

Pontes de hidrogênio entre resíduos do sítio ativo e do substrato			**Distância (Å)**
GDP-manose	GumH		
H_{34}	Met24	S	3,07
O_{24}	Asp26	N	2,79
O_{20}	Arg205	N	3,07
O_{20}	Arg205	N	2,80
H_{32}	Glu286	O	2,79
O_{11}	Glu294	OE2	2,72

Recentes estudos comparativos entre estruturas de glicosiltransferases livres e complexadas com UDP-glicose (Ünligil e Rini,

2000) demonstraram que a arginina presente em uma região de *loop* é responsável pela interação direta com o grupo fosfato através de pontes de hidrogênio, caracterizando um mecanismo de "dobradiça" responsável por modular a ligação do substrato no sítio de ligação do doador e garantir a clivagem do fosfato durante a catálise. No presente modelo, a arginina 205, também presente em uma região de *loop*, faz interações através de pontes de hidrogênios diretamente com o grupo fosfato, de acordo com o modelo de dobradiça que pode ser observado na figura 4.23.

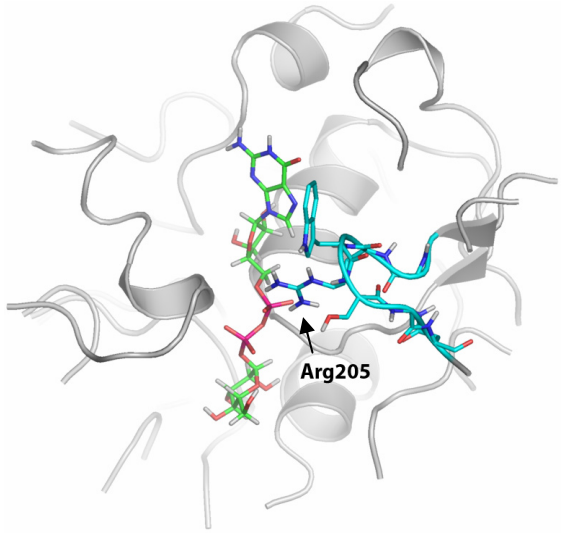


Figura 4.23: Mecanismo de "dobradiça". A figura ilustra o *loop* desordenado e flexível composto por seis aminoácidos (Gly204, Arg205, Trp206, Ser207, Met208 e Asn209) em azul claro. O mecanismo do tipo "dobradiça" garante maior liberdade no movimento e maior interação entre o substrato e a enzima.

Outro resultado importante para uma boa interpretação da qualidade do *dock* é um gráfico de complementaridade baseado na superfície acessível ao solvente (também obtido através do programa FLO), e o tipo de interação atômica, isto é, interações apolares (hidrofóbicas), polares (hidrofílicas) e o estado de protonação dos átomos. O gráfico de

complementaridade para o estudo de *docking* realizado é mostrado nas figuras 4.24 **A** e **B**. A superfície mostra as posições onde os átomos do substrato doador deveriam estar para satisfazer um bom contato. Átomos que fazem pontes de hidrogênio freqüentemente penetram à superfície (círculos vermelhos), enquanto que os outros átomos usualmente ficam próximos à superfície exposta ao solvente. Aceptores de hidrogênio permanecem próximos das regiões vermelhas e doadores de hidrogênio permanecem próximos às regiões azuis. As áreas amarelas representam interações apolares (hidrofóbicas). Quanto melhor o ligante, maior sua complementaridade.

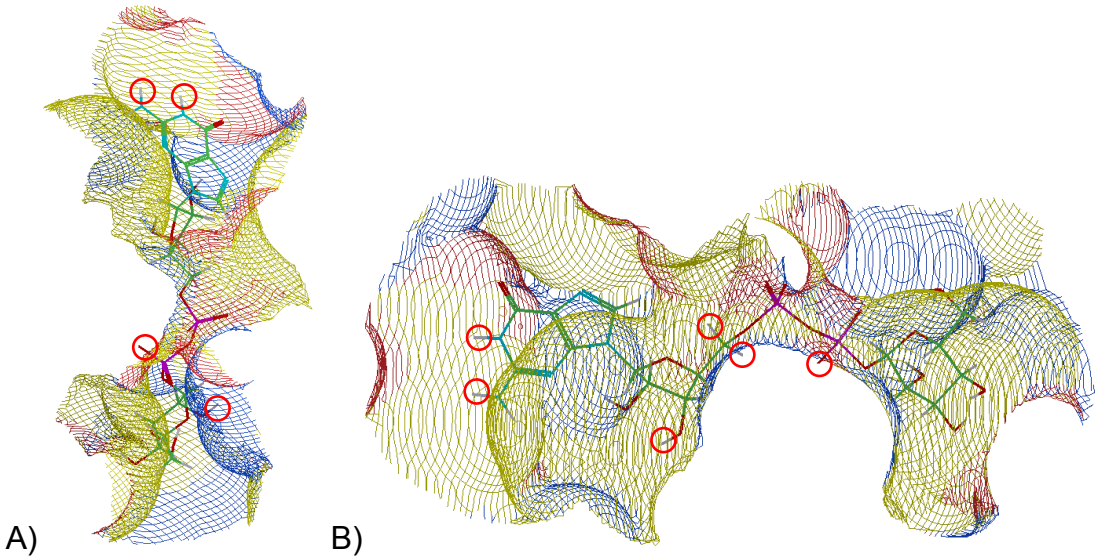
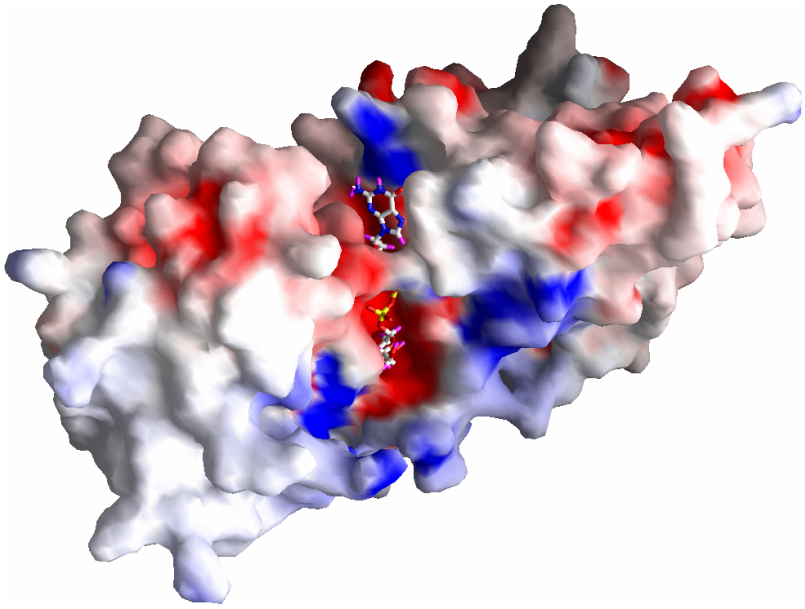


Figura 4.24: Gráfico de complementaridade das interações atômicas obtido com o programa FLO. A figura **A** ilustra um ângulo de visão perpendicular a GDP-manose e **B** é uma visão lateral a GDP-manose.

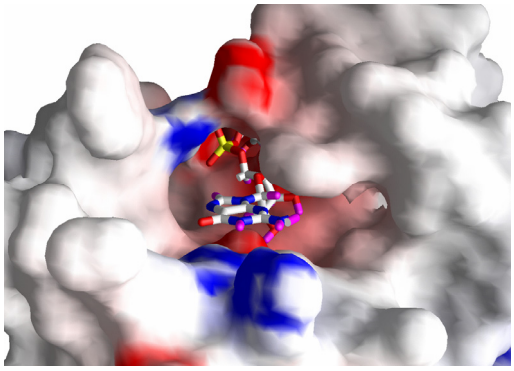
Da figura 4.24, verifica-se a boa complementaridade do complexo enzima/substrato, uma vez que todos os átomos que realizam pontes de hidrogênio podem ser facilmente localizados através dos pequenos círculos

vermelhos e os átomos de oxigênio do grupo fosfato estão próximos a superfícies vermelhas, enquanto que átomos como o nitrogênio estão próximos a superfícies azuis e átomos de carbono estão em regiões amarelas, respeitando o critério de complementaridade.

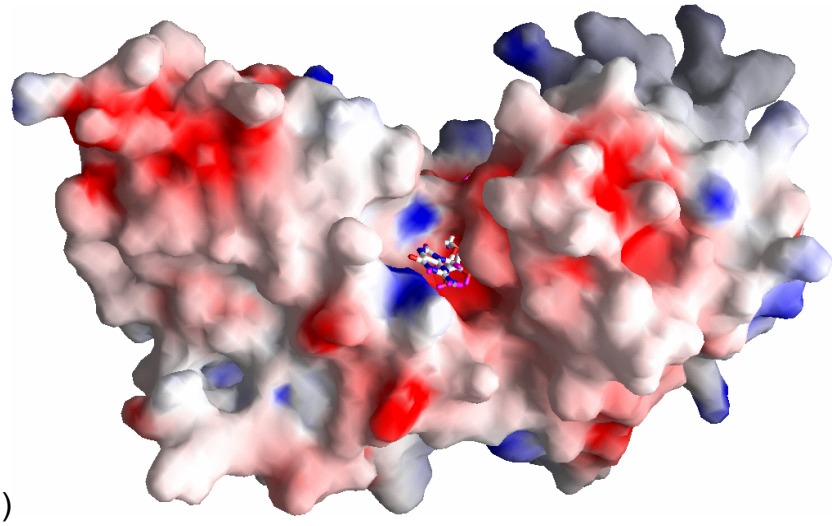
Com a finalidade de estudar as interações atômicas quanto à distribuição e complementaridade eletrostática das cargas na superfície da proteína, foram geradas figuras através do programa GRASP (Nicholls *et al.*, 1991). As figuras 4.25 **A, B, C** e **D** mostram a superfície da região do sítio ativo da GumH complexada com o substrato GDP-manose em diferentes ângulos. As figuras estão coloridas de acordo com o potencial eletrostático relacionado aos aminoácidos que possuem características polares básicas (em azul), aminoácidos polares ácidos (cor vermelha) e apolares (branco). O substrato GDP-manose está complexado em todas as figuras. A figura **A** mostra a proteína vista de 'cima' e ilustra as divisões das cargas eletrostáticas por toda estrutura. A parte superior da molécula é formada em grande parte por uma região hidrofóbica caracterizada pela cor branca, já as laterais são compostas, majoritariamente, por regiões carregadas negativamente (figura **C**) e positivamente (figura **D**). A figura **B** ilustra com maiores detalhes a região mais profunda do sítio ativo onde deve se ligar o substrato doador. Pode-se notar ainda a complementaridade entre os nitrogênios da base guanina e os oxigênios do sítio ativo, assim como os oxigênios do grupo fosfato e os nitrogênios da arginina. Essas complementaridades eletrostáticas são responsáveis por garantir maiores interações entre a enzima e o substrato e servem também como guia durante o processo catalítico.



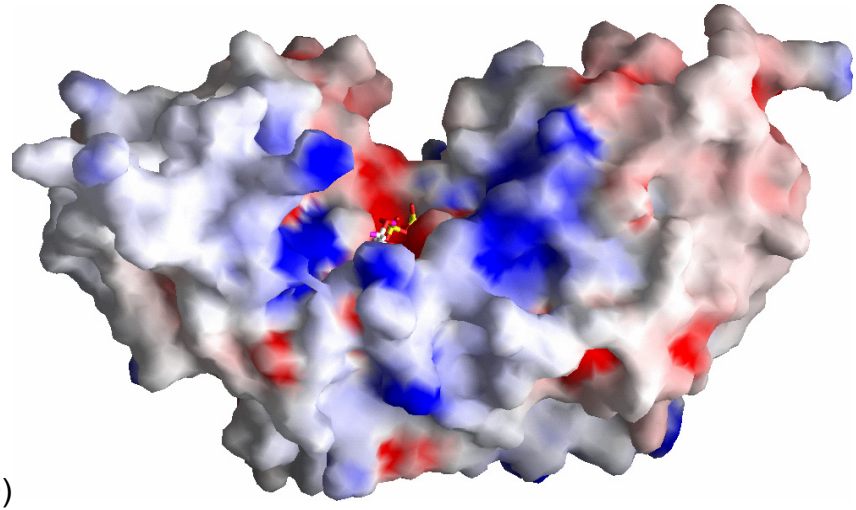
A)



B)



C)



D)

Figura 4.25: Potencial eletrostático e a complementaridade das cargas no modelo GumH complexado com GDP-manose. As figuras estão coloridas de acordo com o potencial eletrostático relacionado aos aminoácidos que possuem características polares básicas (em azul), aminoácidos polares ácidos (cor vermelha) e apolares (branco).

4.10 Implicações para o mecanismo catalítico

Como descrito no item 4.2 a enzima GumH foi classificada como sendo uma glicosiltransferase pertencente à família 4 (GT4), família a qual pertence GDP-manosiltransferases. Membros da família das GDP-manosiltransferases, cujas reações ocorrem com o mecanismo de retenção da conformação do carbono anomérico do substrato doador, devem apresentar dois substratos, o doador GDP-manose e a molécula aceptora da manose.

Embora nenhum representante da família GT4 tenha uma estrutura determinada até o momento, estudos bioquímicos com a AceA, têm permitido apontar quais são os resíduos essenciais para a catálise (Abdian *et al.*, 2000). A AceA é uma GDP-manosiltransferase de *Acetobacter xylinum*, que transfere manose de GDP-manose para a molécula de

celobiose ligada ao lipídio carreador pirofosfato-poliprenol (Geremia *et al.*, 1999), reação exatamente idêntica à catalisada pela GumH. Na *Acetobacter* a goma produzida, chamada de *acetan*, é uma molécula maior que a goma fastidiana, como pode ser observado na figura 4.26.

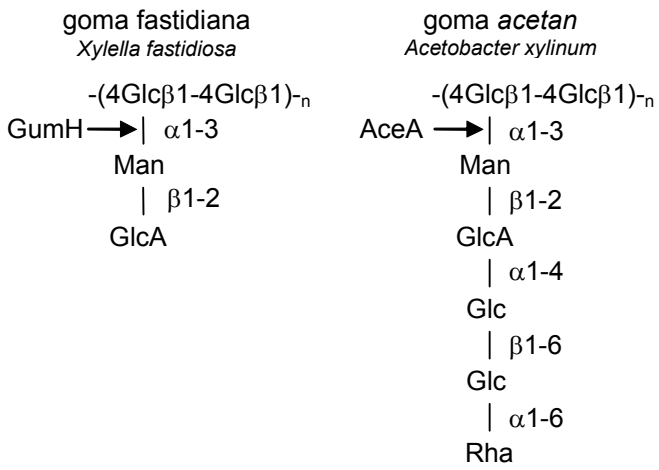


Figura 4.26: Estrutura da unidade repetidora das gomas fastidiana e *acetan*. (adaptado de Abdian *et al.*, 2000).

A seqüência de aminoácidos da AceA também apresenta o motivo conservado EX_7E, como pode ser observado no alinhamento total entre as seqüências das duas enzimas (figura 4.27). Os estudos realizados foram de mutação sítio dirigida nos resíduos glutâmicos (Glu) do motivo, além de mutações em outros resíduos que poderiam ser importantes para a catálise da AceA por serem conservados na família dessas enzimas. A importância funcional de cada mutante foi subseqüentemente determinada *in vivo*, por complementação de cepas mutantes gumH$^-$ de *Xanthomonas* campestris, e *in vitro* com a forma recombinante de AceA expressa em *E.coli.* Os resultados desses estudos mostraram que a mutação do segundo glutâmico

(Glu) do motivo resultou em uma enzima com atividade muito baixa, mas residual *in vivo*. Mutações no primeiro ácido glutâmico inativaram completamente a enzima tanto nos experimentos *in vitro* como *in vivo*. A conclusão do estudo é, portanto, que o primeiro glutâmico do motivo seja essencial para a catálise e seria a base catalítica que age no ataque nucleofílico. O segundo ácido glutâmico seria importante, porém não fundamental para a catálise.

Estudos estruturais, bastante recentes (Hu *et al.*, 2003), da enzima MurG complexada com o substrato doador UDP-GlcNAC mostram a interação entre o resíduo Glu269 (conservado em diversas proteínas homólogas a MurG) e o grupo hidroxila da ribose. Esses dados, somados aos experimentos de mutações neste resíduo levaram os autores à conclusão que o resíduo Glu269 tem uma importante função na habilidade da MurG em distinguir UDP e TDP, porém não está diretamente relacionado à catálise.

```
                10        20        30        40        50        60
        ....|....|....|....|....|....|....|....|....|....|....|....|
GumH    --MEWEHC-LMKVVHVVRQFHPSIGGMEDVVFNIAMQLHLHAGIDVDVVTLNRVFTQSDV
AceA    MNSKKRGDETLKVLHICRQFSPSVGGLEDSLLNLARSQRQRLGIDAEVLTLDTVFGRPG-

                70        80        90       100       110       120
        ....|....|....|....|....|....|....|....|....|....|....|....|
GumH    LLPCTDKYQGVSIQRIGYRGSSRYPLAPWVLRMLDKADVIHVHGIDFFYDFLALIRVLHG
AceA    KLPHRDVVDGIPVTRLAWRGSTKYPLAPQVLRHIGGFDLLHVHAIDFFFDFLAWIWPLHR

               130       140       150       160       170       180
        ....|....|....|....|....|....|....|....|....|....|....|....|
GumH    KPMVVSTHGGFFHTDYASRLKLLWFNTLTRLGALAYARIIASGESDGALFSKIVAPSRLR
AceA    KTMIASTHGGFFHTGALRRIKEIWFRTITPIGVRAYKKIVACSYSDADLF-RHVAAGRLI

               190       200       210       220       230       240
        ....|....|....|....|....|....|....|....|....|....|....|....|
GumH    VIENGVDVEKYARCGASEAGRTLLYFGRWSMNKGLLETLQLLAVLYVLDPRWRLIIAGRE
AceA    TIENGINQTRFRDAASRTPNRTILAFGRFAVHKRLKLLFQLVALLRAYNSGWNIIVAGQD

               250       260       270       280       290       300
        ....|....|....|....|....|....|....|....|....|....|....|....|
GumH    YDYDQAALAYEVDRLGLSEQVHFHCSPSQSQLRFLMEQAQFFISLSRHEGFGIAAVEAMS
AceA    SNLTADDLRAQARACGIEDSLRIVSGPSDAELRGLMGEASFFGCLSAHEGFGLAAVEAMS

               310       320       330       340       350       360
        ....|....|....|....|....|....|....|....|....|....|....|....|
GumH    AGLIPVLSDIPPFARLHRESGLGVLVDPLQPQQAAVAVQGLAVQVDTHFIDWRSQAMAFS
AceA    AGLVPILSNITPFARLMQQGAAGVMVNPDNLAPGAREAEDMAAALPETADALRARNMDVA

               370       380       390
        ....|....|....|....|....|....|....|
GumH    DRYHWRYVIGCYQDEYCRALG--LG-GEQEFL--R
AceA    SRYDWHSVAHEYARLYQQVLGRALPEANMAAAGAE
```

Figura 4.27: Alinhamento entre as enzimas GumH e AceA. As enzimas apresentaram 40% de identidade e 58% de similaridade entre as suas seqüências de aminoácidos. A cor vermelha representa o alinhamento entre aminoácidos idênticos, enquanto que a cor azul representa um alinhamento entre aminoácidos similares. O motivo EX_7E está desatacado na cor amarela.

O modelo da GumH revelou alguns aspectos importantes quanto a disposição dos aminoácidos fortemente conservados, principalmente o motivo SX_2EX_7E. Com base nos resultados obtidos nos estudos de *docking* aliados ao conhecimento dos estudos com as enzimas AceA e MurG, podemos propor o mecanismo da reação catalisada pela enzima GumH, que deve ocorrer com retenção da conformação do carbono anomérico da GDP-manose. Neste mecanismo deve ocorrer um duplo deslocamento da

configuração do carbono anomérico, ou seja, este centro deve sofrer inversão duas vezes de modo a preservar a configuração original.

A análise das interações do modelo da GumH complexada com a GDP-manose mostrou que os resíduos Asp26 e Glu286 (o primeiro "E" do motivo conservado $\underline{E}X_7E$) são potentes candidatos a bases carboxílicas, que promoveriam os ataques nucleofílicos. A figura 4.28 mostra as principais interações entre o substrato GDP-manose e a enzima GumH. O valor da distância entre os átomos de carbono das duas carboxilas teoricamente, deve ser por volta de 5,5 a 6,0 Å (CAZy). No presente modelo a distância entre os átomos (carbono) das duas carboxilas é de 7,42 Å respeitando as minimizações energéticas impostas pelo programa FLO. As distâncias entre os átomos de oxigênio e o carbono anomérico (destacado em amarelo na figura) são consideradas boas, já que propicia a catálise ácida através do ataque nucleofílico pelas bases carboxílicas (figura 4.28 **A**). O Glu286 da GumH é equivalente ao primeiro glutâmico da enzima AceA, que foi caracterizado como essencial para a catálise por meio de estudos bioquímicos.

O resíduo Glu294 (o segundo "E" do motivo conservado $EX_7\underline{E}$) da GumH faz interações com o anel ribose da base nitrogenada (o tracejado azul representa a ponte de hidrogênio presente entre o anel ribose e o Glu294). É importante ressaltar que este resíduo é estruturalmente equivalente ao Glu269 da enzima MurG e ao segundo glutâmico do motivo conservado da AceA. Observa-se que o anel da manose está na conformação do tipo cadeira, respeitando o ângulo ideal para o ataque das carboxilas como representado na figura 4.29, ou seja, perpendicular ao

plano das carboxilas. Na figura 4.28 **B** o anel da manose está sob a forma de um hexágono, indicando que o ângulo de visão está no plano das carboxilas. A preservação dessa conformação é fundamental para que o mecanismo ocorra.

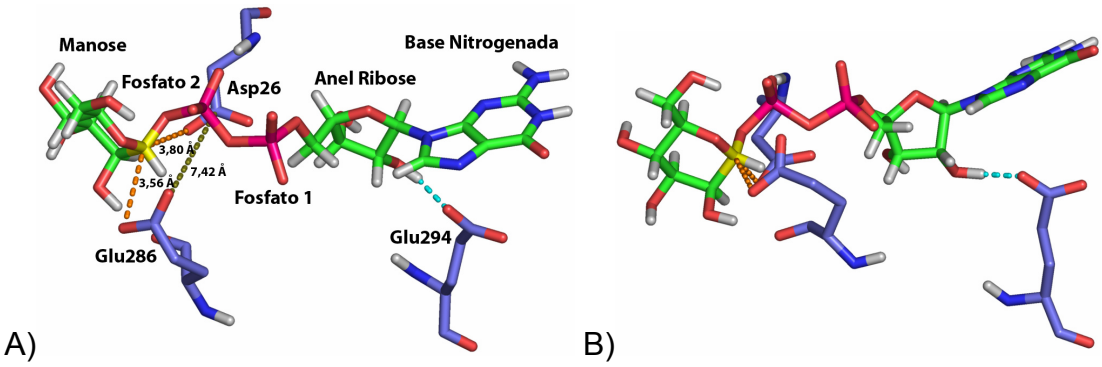


Figura 4.28: Interações entre a GumH e GDP-manose. Diferentes ângulos de visão: **A** perpendicular ao plano das carboxilas e **B** perpendicular ao plano das carboxilas.

Com base nessas considerações, propomos um esquema para o mecanismo de retenção da enzima GumH, que pode ser visualizado na figura 4.29. Neste esquema, as bases carboxílicas que participariam das duas etapas do mecanismo de duplo deslocamento na GumH seriam o ácido aspártico 26 (Asp26) e o ácido glutâmico 286 (Glu286). O Glu286 é responsável pelo primeiro ataque nucleofílico ao carbono anomérico da manose permitindo a formação do complexo enzima-manosil. Em uma segunda etapa, o Asp26 agiria como uma base para ativar o nucleófilo Glc-Glc-P-Poliprenol (ou celobiose-P-prenol) que hidrolisa ao complexo enzima-manosil. Na figura 4.29 a letra **R** simboliza a molécula de GDP.

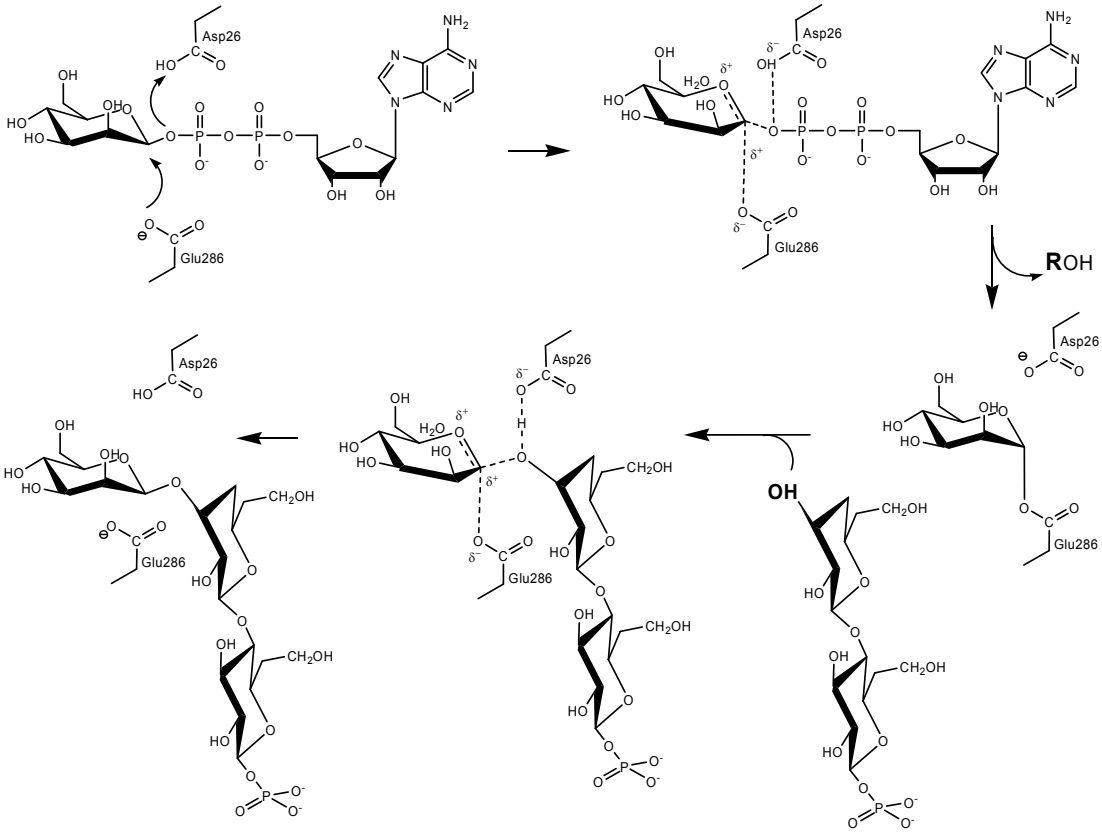


Figura 4.29: Esquema proposto para o mecanismo de retenção da GumH. A letra **R** representa a molécula de GDP.

Estudos recentes mostram a estrutura da enzima bovina α1,3-galactosiltransferase, uma proteína relacionada ao grupo sanguíneo ABO, glicoesfingolipídios e glicosiltransferases (Gastinel *et al.*, 2001), e uma das poucas enzimas que atuam por mecanismo de retenção que tiveram sua estrutura determinada. A proteína foi cristalizada na presença do substrato, fato que permitiu considerações interessantes a respeito de seu mecanismo. A proteína apresenta somente um resíduo polar ácido (o Glu317) que se encontra em posição ideal para o ataque nucleofílico ao carbono anomérico do substrato doador. Estudos bioquímicos mostraram que a enzima bovina é ativa somente na presença do substrato aceptor do açúcar. Com base

nesses dados os autores propõem que a segunda inversão do carbono anomérico seja promovida pelo próprio substrato aceptor. Neste caso, a hidroxila ligada ao átomo C3 do aceptor (lactose) sofreria uma desprotonação, em processo não totalmente esclarecido, e seria capaz de atacar o carbono anomérico do intermediário galactosil-enzima.

Essa nova proposta para o mecanismo catalítico abriu novas possibilidades na análise do mecanismo anteriormente proposto para a reação catalisada pela GumH. O resíduo Asp26, embora presente na seqüência da AceA, não é conservado nas demais GDP-manosiltransferases. Estes fatos nos levaram a indagar sua função na atividade catalítica da enzima GumH e a análise de todos os possíveis rotâmeros para o resíduo Asp26 foi minuciosamente estudada. Alguns dos rotâmeros levam o resíduo Asp26 para posições desfavoráveis para a interação com o substrato doador o que permite supor um mecanismo alternativo, parecido com o observado na enzima bovina.

Na tentativa de adaptar o que seria uma segunda proposta para o mecanismo catalítico da GumH, um estudo de comparação entre o modelo da GumH e a estrutura cristalográfica da proteína bovina α1,3-galactosiltransferase foi realizado.

A figura 4.30 ilustra as comparações entre os aminoácidos responsáveis pelo primeiro ataque nucleofílico na estrutura cristalográfica **A** e no modelo da GumH **B**. As figuras mostram as distâncias entre o carbono anomérico (destacado em amarelo) e o átomo OE2 do ácido glutâmico responsável pelo ataque nucleofílico e a primeira inversão da configuração do carbono anomérico em ambas moléculas.

Como pode ser observado na figura 4.30 **A**, o OE2 (segundo oxigênio) do ácido glutâmico (Glu317) da estrutura cristalográfica, responsável pelo primeiro ataque nucleofílico está a uma distância de 4,76 Å do primeiro carbono (C1) do anel da lactose (açúcar aceptor). Analogamente, no modelo obtido para a GumH a distância entre os átomos (OE2 do Glu286 e do C1 da manose) é de 3,56 Å (figura 4.30 **B**).

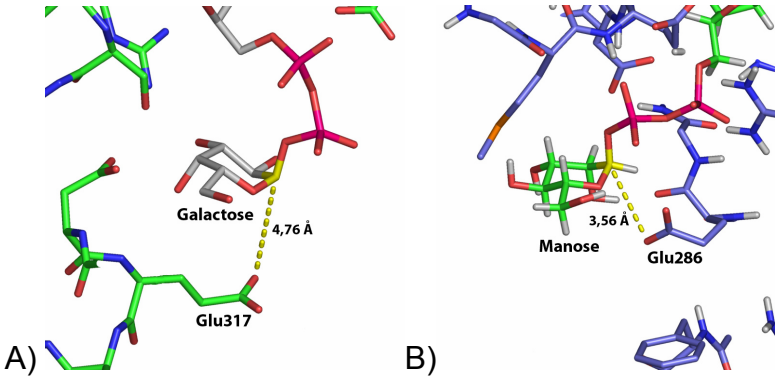


Figura 4.30: Detalhe da localização das bases catalíticas na estrutura cristalográfica da α1,3-galactosiltransferase bovina **A** e da GumH modelada **B**.

Glicosiltransferases inversoras da configuração do carbono anomérico possuem geralmente uma base (ácido aspártico ou glutâmico), que abstrai um próton do grupo hidroxila de uma molécula de açúcar do doador. O anel de açúcar do aceptor deve estar posicionado “acima” do plano do açúcar ou entre a base carboxílica e o carbono C1 do açúcar doador (Gastinel *et al.*, 2001). A figura 4.31 ilustra em **A** a disposição dos açúcares do aceptor e doador na estrutura cristalográfica bovina α1,3-galactosiltransferase e em **B** a disposição no modelo da GumH das moléculas de glicose do aceptor. Essa análise foi realizada através da comparação direta entre a estrutura

cristalográfica e o modelo, e também da posição do substrato doador ligado ao sítio da enzima GumH provido do estudo de *dock*.

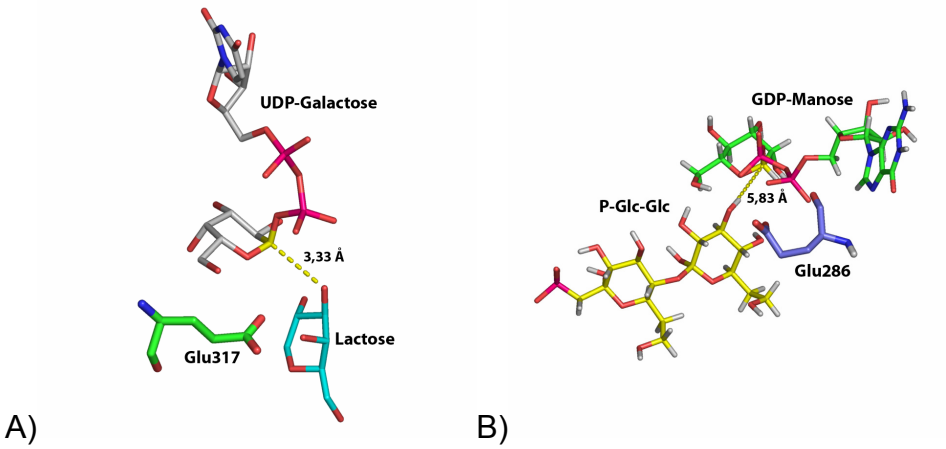


Figura 4.31: Posição do substrato aceptor embasado na estrutura cristalográfica da α1,3-galactosiltransferase bovina. As figuras **A** e **B** ilustram as posições dos substratos doadores e aceptores da estrutura cristalográfica e do modelo da GumH respectivamente. A linha pontilhada refere-se à distância entre o C1 (destacado em amarelo) e a hidroxila 3 (do carbono C3) em cada estrutura.

Dessas observações, um modelo esquemático do mecanismo de retenção pôde ser proposto, onde o ataque nucleofílico do Glu286 resulta na primeira inversão da configuração do átomo C1 da manose (figuras 4.32 **A** e **B**). O substrato aceptor (celobiose) deve estar posicionado no bolsão catalítico “abaixo” do plano do anel do açúcar doador, de modo que o átomo de oxigênio desse grupo desprotonado 3-hidroxil permaneça a uma distância ideal (3,33 Å e 5,83 Å nas estruturas cristalográfica e modelada respectivamente) para que o ataque sobre o carbono C1 da manose aconteça diretamente, invertendo a configuração do carbono anomérico uma segunda vez (figuras 4.32 **C** e **D**).

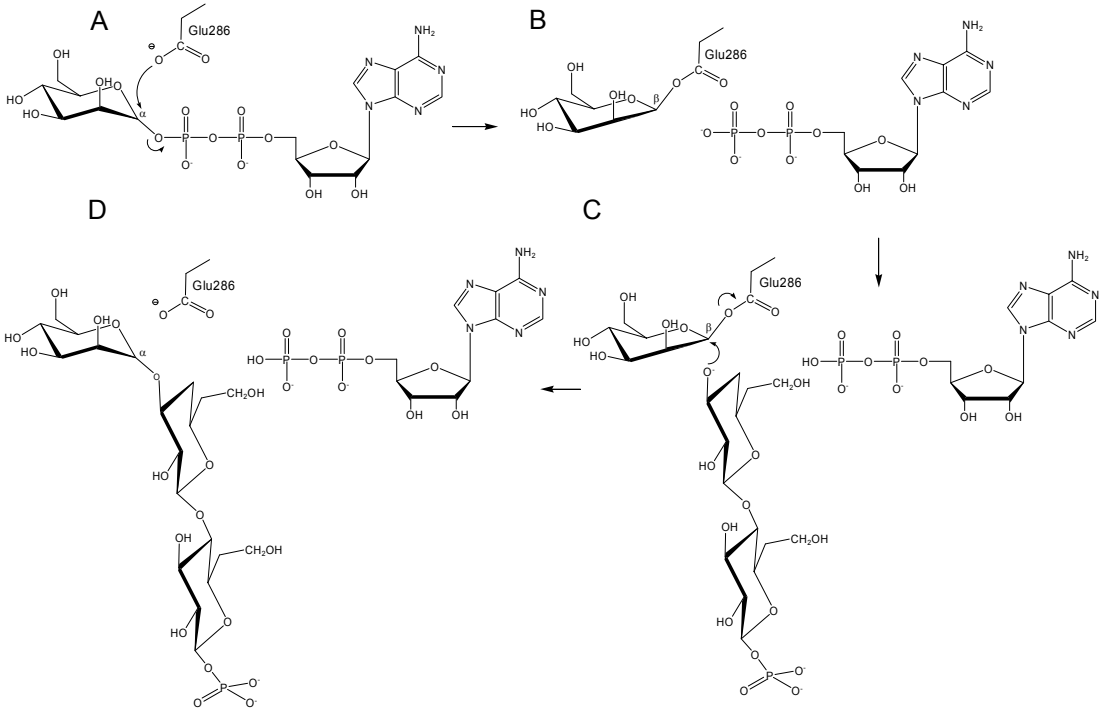
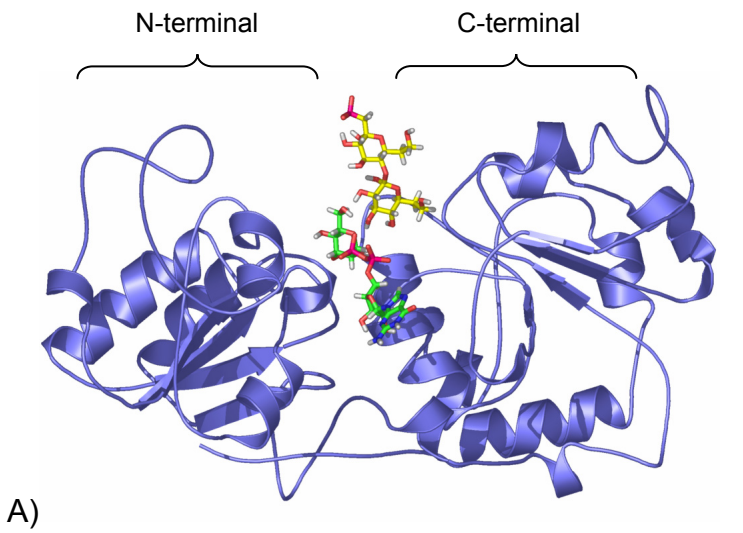
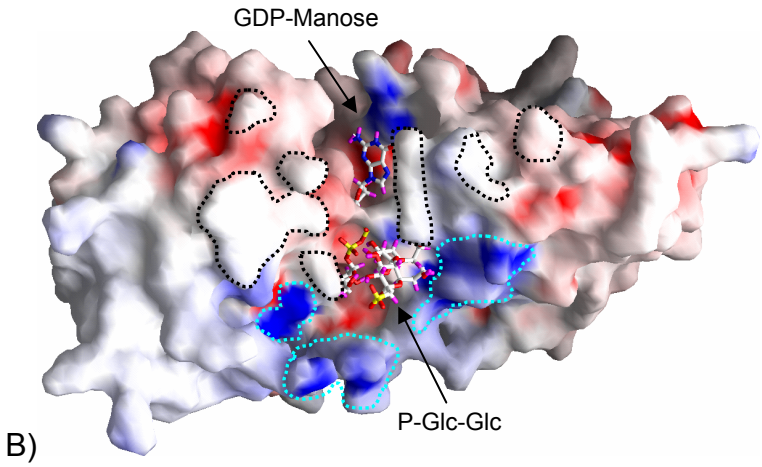


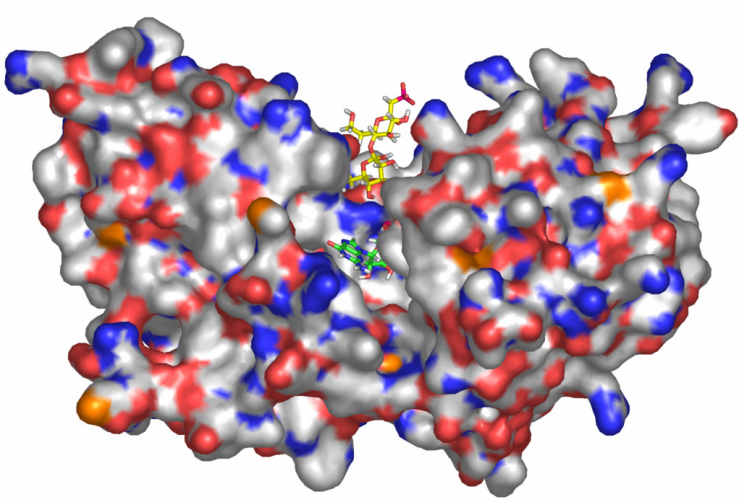
Figura 4.32: Representação esquemática do mecanismo de retenção da configuração do carbono anomérico do sacarídeo manose. As etapas **A** e **B** envolvidas na primeira inversão da configuração do carbono anomérico da manose e **C** e **D** estão envolvidas na segunda inversão da configuração do carbono anomérico preservando assim sua configuração inicial.

A figura 4.33 **A** ilustra a localização dos substratos na estrutura terciária do modelo gerado para a GumH. O substrato doador GDP-manose, como anteriormente mencionado, está localizado acima da α-hélice central e abaixo do *loop* flexível no domínio C-terminal da proteína e o substrato aceptor celobiose-P-prenol, embora ligeiramente afastado da α-hélice central, também se encontra próximo ao *loop* flexível. Na figura 4.33 **B** é mostrada a superfície de potencial eletrostático da molécula na qual a equivalência das cargas dos átomos vizinhos aos substratos podem ser verificadas. O substrato aceptor celobiose-fosfato está localizado entre uma região lateral positivamente carregada (indicada pelas linhas pontilhadas

azuis) e uma região predominantemente apolar na parte superior da proteína (indicada pelas linhas pontilhadas pretas). Essas regiões podem favorecer interações com outras proteínas presentes no complexo enzimático da goma fastidiana e na provável interação com a membrana. A figura **C** ilustra a superfície da estrutura da proteína levando-se em consideração o raio de van der Waals dos átomos dos resíduos expostos ao solvente, sendo que as cores indicam o tipo de átomo presente na superfície da proteína; oxigênio (vermelho), nitrogênio (azul), enxofre (alaranjada) e carbono (cinza).







C)

Figura 4.33: Diferentes representações da enzima GumH com o substrato doador e aceptor ligados. Nas figuras **A** e **C** a GDP-manose e a celobiose-fosfato estão representados nas cores verde e amarela para seus átomos de carbono respectivamente enquanto que em **B**, tanto a GDP-manose como a celobiose-fosfato têm seus átomos de carbono na cor branca, porém estão legendados e indicados por setas. A orientação C- e o N-terminal estão preservadas nas três figuras.

O fato de serem propostos dois possíveis mecanismos para a reação catalisada pela enzima GumH, demonstra as limitações de estudo estrutural por modelagem molecular, logo, informações obtidas de estudos bioquímicos da proteína ou de proteínas homólogas à de estudo são de grande importância na análise dos modelos construídos. Quando aliados esses estudos podem contribuir para o conhecimento estrutural de proteínas cujas estruturas não foram ainda determinadas experimentalmente.

Neste trabalho, foi possível construir o modelo da enzima GumH e especular a respeito do mecanismo da reação catalítica. Considerando-se que nenhuma GDP-manosiltransferase teve, até o momento, sua estrutura tridimensional determinada, os resultados obtidos neste trabalho resumem-se nas primeiras e únicas informações estruturais desta família de enzimas.

4.11 Referências bibliográficas

Abdian, P. L, Lellouch, C. A., Gautier, C., Ielp, L. and Geremia, R. A. (2000) *The Jornal of Biological Chemistry*, **275**, 40568-40575.

Branden, C. and Tooze, J. (1998) *Introduction to protein structure.* New York: Gerland Publishing, Inc.

CAZy - ***C**arbohydrate-**A**ctive En**zy**mes*, http://afmb.cnrs-mrs.fr/~pedro/CAZY/db.html

Campbell, J. A., Davies, G. J., Bulone, V. and Henrissat, B. (1997) *Biochem J.*, **326**, 929-939.

Campbell, J. A., Davies, G. J., Bulone, V. and Henrissat, B. (1998) *Biochem J.*, **329**, 719.

Campbell, R. E., Mosimann, S. C., Tanner, M. E. and Strynadka, N. C. J. (2000) *Biochem.*, **39**, 14993-15001.

Cid, E., Gomis, R. R., Geremia, R. A., Guinovart, J. J. and Ferrer, J. C. (2000) *The Jornal of Biological Chemistry*, **275**, 33614-33621.

Dennis, J. W., Granovsky, M. and Warren, C. E. (1999a) *Bioessays*, **21**, 412-421.

Dennis, J. W., Granovsky, M. and Warren, C. E. (1999b) *Biochim Biophys Acta*, **1473**, 21-34.

Gastinel, L. N., Bignon, C., Misra, A. K., Hindsgaul, O., Shaper, J. H. and Joziasse, D. H. (2001) *EMBO J.*, **20**, 638-649.

Geremia, R. A., Petroni, E. A., Ielpi, L. and Henrissat, B. (1996) *Biochem J*, **318**, 133-138.

Ha, S., Walker, D., Shi, Y. and Walker, S. (2000) *Protein Science*, **9**, 1045-1052.

Hu, Y., Chen, L., Ha, S., Gross, B., Falcone, B., Walker, D., Mokhtarzadeh, M. and Walker, S. (2003) *Biochemistry*, **100**, 845-849.

Jones, D. T. (1999) *J. Mol. Biol.*, **287**, 797-815.

Jones, D. T. (2002) Program and Documentation by David T. Jones. Copyright by University College London, Department of Computer Science, Bioinformatics Unit, London – UK.

Kleywegt, G. J. and Jones, T. A. (1995) *Structure*, **3**, 535-540.

Kido, N., Torgov, V. I., Sugiyama, T., Uchiya, K., Sugihara, H., Komatsu, T., Kato, N. and Jann, K. (1995) *J. Bacteriol*, **177**, 2178-2187.

Laskowski, R. A., MacArthur, M. W. and Thornton, J. M. (1998) *Curr. Opin. Struct. Biol.*, **8**, 631-639.

McCarter, J. D. and Withers, S. G. (1994) *Curr. Op. Struct. Biol*, **4**, 885-892.

McMartin, C. and Bohacek, R. S. (1997), *J. Comp. Aided Mol. Design*, **11**, 333-334.

Miyata, T., Takeda, J., Lida, Y., Yamada, N., Inoue, N. Takahashi, M., Maeda, K., Kitani, T. and Kinoshita, T. (1993) *Science*, **259**, 1318-1320.

Moréra, S., Imberty, A., Aschke-Sonnenbom, U., Rüger, W., Freemont, P. S. (1999) *J. Mol. Biol.*, **292**, 717-730.

Mulichak, A. M., Losey, H. C., Walsh, C. T. and Garavito, R. M. (2001) *Structure*, **9**, 547-557.

Murray, B. W., Takayama, S., Schultz, J. and Wong, C. H. (1996) *Biochemistry*, **36**, 823-831.

Nicholls, A., Sharp, K. and Honing, B. (1991) *Proteins: Struct., Funct., Genet.*, **11**, 281.

Radominska-Pandya, A., Czernik, P. J., Little, J., M., Battaglia, E. and Mackenzie, P. I. (1999) *Drug Metab Rev*, **31**, 817-899.

Sali, A. and Blundell, T. L. (1993) *J. Mol. Biol.*, **234**, 779-815.

Schonbachler, M., Horvath, A., Fassler, J. and Riezma, H. (1995) *EMBO J.*, **14**, 1637-1645.

Sinnott, M. L. (1990) Chem. Rev., **90**, 1171-1202

Ünligil, M. U. and Rini, J. M. (2000) *Current Opinion in Structural Biology*, **10**, 510-517.

Vrielink, A., Rüger, W., Driessen H.P.C. and Freemont, P.S. (1994) *EMBO J.*, **13**, 3413-3422.

Vriend, G. (1990) *J. Mol. Graph.*, **8**, 52-56.

Weiner, S. J., Kollman, P. A., Nguyen, D. T. and Case, D. A. (1986) *J. Comp. Chem.*, **7**, 230-252.

CAPÍTULO 5

AS OUTRAS ENZIMAS

Devido a falta de informações estruturais em se tratando de proteínas homólogas às Gums um estudo também foi desenvolvido para as demais proteínas (GumB, GumC, GumD, GumE, GumF, GumJ, GumK e GumM). Nesse estudo, foram utilizados programas computacionais que permitiram a comparação das seqüências de DNA e de aminoácidos com moléculas já estudadas presentes em bancos de dados com a finalidade de se obter informações quanto a classificação, estrutura secundária, de regiões transmembrânicas e modelagem molecular por comparação.

5.1 Classificação das enzimas quanto a suas funções

Com a utilização do programa Pfam foi possível obter a classificação das oito proteínas quanto às suas funções tendo como base as seqüências de proteínas presentes nos bancos de dados Pfam/CAZy/Swiss. O resultado da classificação pode ser visualizado na tabela 5.1. A GumB obteve melhores índices quando comparada às proteínas que fazem parte de uma família de proteínas periplasmáticas (*Enterobacteriaceae*) envolvidas na biossíntese/exportação de polissacarídeos. A GumC foi classificada como pertencente à família das proteínas responsáveis pela determinação do

tamanho das cadeias de polipeptídios conferindo forma aos mesmos. Nessa família estão incluídas proteínas envolvidas na biossíntese de lipopolissacarídeos como a proteína wzz de *E. coli*. A GumD foi agrupada à família de proteínas que fazem parte de diversas vias biossintéticas, são essencialmente transferases de moléculas de açúcar à um lipídio carreador e geralmente participam do início da síntese de um polissacarídeo. A GumF foi agrupada às acetiltransferases e hipoteticamente integrada a membrana ou com pelo menos diversos motivos transmembrânicos. A GumM foi comparada à glicosiltransferases e finalmente a GumJ foi relacionada à proteínas envolvidas na biossíntese de polissacarídeos e tida como proteína integrada a membrana. As enzimas GumK e GumE não obtiveram nenhum alinhamento quando comparadas às seqüências presentes nesses bancos de dados.

Tabela 5.1: Classificação quanto a função das demais Gums baseado no banco de seqüências protéicas Pfam/CAZy/Swiss.

Nome da Proteína	Aminoácido inicial	Aminoácido final	Número de resíduos alinhados	*E-value*	Provável função associada
GumB	73	194	121 (46%)	1,2e-56	Exportar polissacarídeos
GumC	18	187	169 (36%)	3e-28	Dar forma às cadeias dos polissacarídeos
GumD	293	484	191 (40%)	8,2e-35	Transferir açúcar ao lipídio carreador
GumF	25	356	331 (91%)	2e-61	Transferir um grupo acetil
GumM	75	245	170 (64%)	5,6e-93	Adicionar glicose
GumJ	20	287	267 (52%)	7,4e-59	Síntese de polissacarídeos

Com exceção da GumE e GumK, a classificação quanto às funções das enzimas foi bastante satisfatória, pois se enquadrou perfeitamente no esquema proposto da via biossintética da goma fastidiana (figura 1.5). É válido lembrar que, estudos em *X. campestris,* as quatro últimas enzimas (GumB, GumE, GumC e GumJ) têm sido associadas aos últimos processos da via biossintética da goma (polimerização e exportação), porém suas funções específicas não foram ainda esclarecidas. Com base nessas buscas e subseqüentes atribuições funcionais às enzimas, propomos que as enzimas GumB, GumC e GumJ tenham sua localização como ilustrado na figura 5.1. A GumC, próxima a membrana interna, seria responsável por garantir o tamanho da cadeia do polissacarídeo que posteriormente será exportado por ação da GumE e GumJ, que seriam proteínas integradas à membrana (transmembrânicas) e responsáveis pela exportação/síntese das unidades sacarídicas. Finalmente a GumB, localizada na parte externa da membrana bacteriana, seria responsável pelo processo de polimerização, formando assim a goma fastidiana.

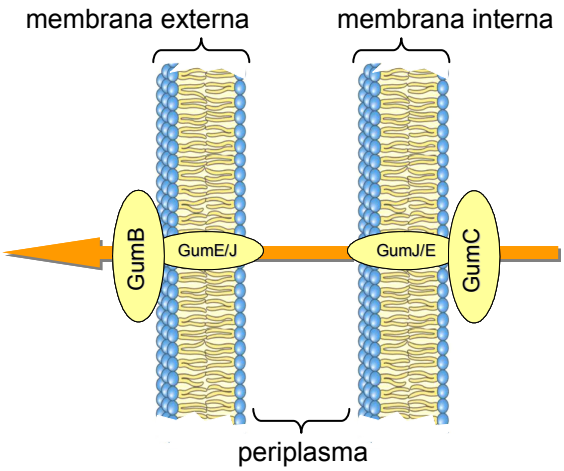


Figura 5.1: Distribuição das enzimas GumB, GumC, GumE e GumJ na membrana bacteriana. A seta alaranjada indica o sentido do processo de exportação do peptídeo.

5.2 Análise do padrão de hidrofobicidade

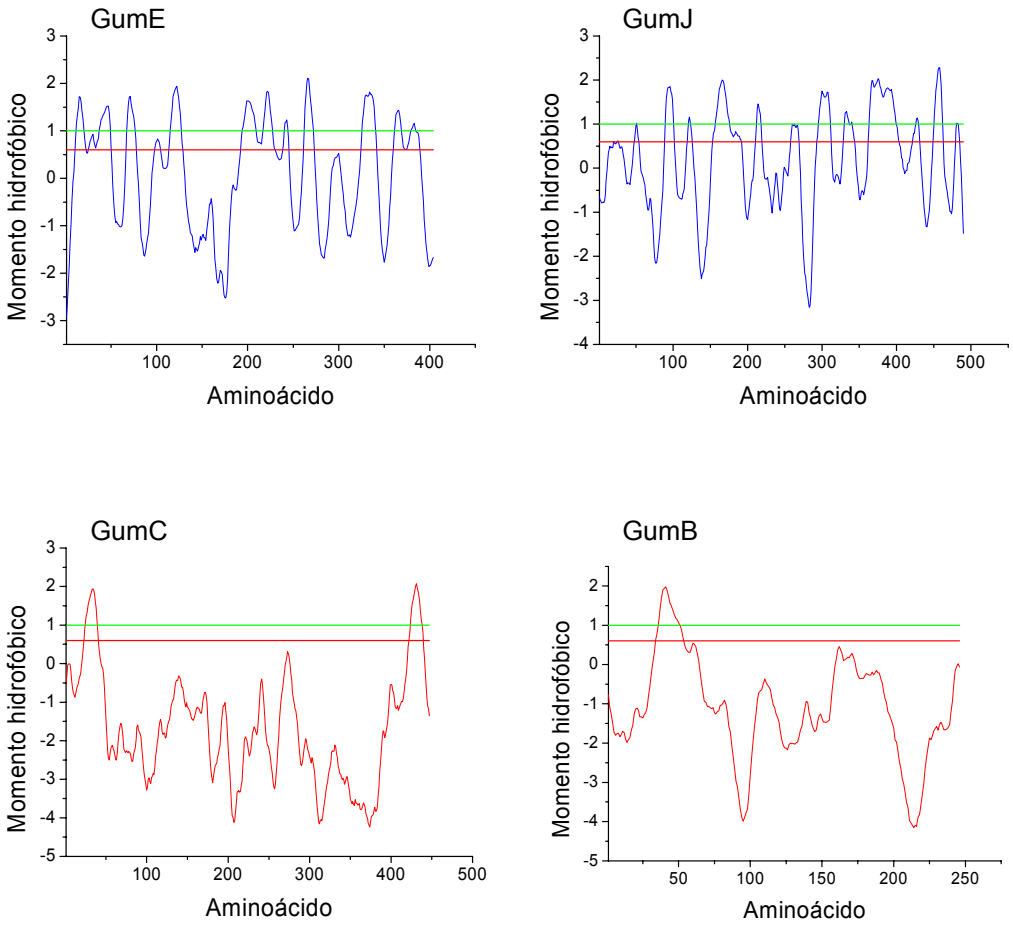
A análise do padrão de hidrofobicidade mostrou que das quatro enzimas envolvidas nas últimas etapas de síntese da goma (a polimerização e o transporte através da membrana celular) a GumB e a GumC apresentam uma e duas regiões transmembrânicas respectivamente, enquanto as outras duas, GumE e GumJ, apresentam onze e treze (em média), respectivamente. As outras enzimas são todas transferases (exceto a GumF), e apresentaram um variável conteúdo hidrofóbico. Os resultados obtidos nesta análise de hidrofobicidade são mostrados na tabela 5.2. Como já mencionado no item 4.5, diferenças nos resultados são explicadas pelo algoritmo utilizado pelos programas para o cálculo dos domínios transmembrânicos, pois os mesmos definem diferentes intervalos de aminoácidos hidrofóbicos na determinação da região em potencial.

Tabela 5.2: Predição de regiões transmembrânicas para as oito enzimas envolvidas na via biossintética da goma fastidiana e o nome dos programas utilizados.

	Regiões Transmembrânicas							
Proteína	**HMMTOP***	**DAS***	**MEMSAT2***	**SOSUI***	**SPLIT 4.0***	**TMHMM***	**Tmpred***	**TOPPRED 2***
GumB	1	1	1	----	1	----	1	1
GumC	2	2	2	2	2	2	2	2
GumD	7	6	5	5	7	5	5	6
GumF	10	10	10	9	10	9	9	8
GumK	----	----	2	----	1	----	1	2
GumM	3	----	----	----	1	----	2	2
GumJ	14	10	12	16	12	12	13	14
GumE	11	11	12	10	12	10	10	9

*referências na seção 5.4

A figura 5.2 ilustra as seqüências protéicas das oito enzimas em função do momento hidrofóbico de cada aminoácido. Dessa forma, torna-se possível visualizar a região tida como transmembrânica (entre as margens de corte inferior (linha vermelha) e superior (linha verde)).



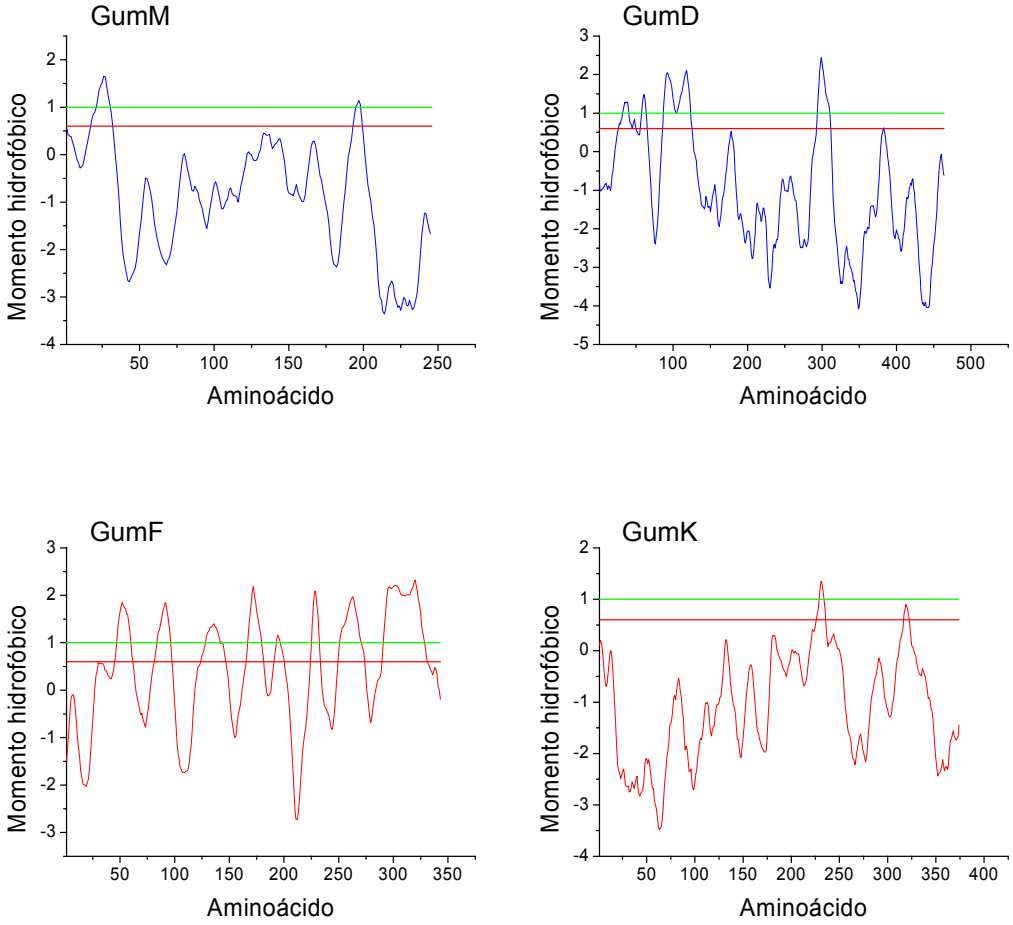


Figura 5.2: Predição de regiões transmembrânicas das enzimas GumE, GumJ, GumC, GumB, GumM, GumD, GumF e GumK. Predição realizada a partir de um gráfico do momento hidrofóbico para cada um dos resíduos.

5.3 Construção dos modelos tridimensionais

Assim como no caso da enzima GumH, os programas GenThreader e THREADER 3.3 foram utilizados nas buscas realizadas em bancos de dados protéicos com a finalidade de encontrar proteínas homólogas para cada uma das oito Gums. Os resultados obtidos estão mostrados na tabela 5.3, com os respectivos nomes das enzimas para as quais os programas encontraram estruturas de enovelamentos semelhantes apesar de baixa identidade

seqüencial. As enzimas GumE, GumF e GumD não estão presentes na tabela, uma vez que nenhum dos programas foi capaz de associar proteínas homólogas a elas.

A GumK obteve quatro moldes, sendo que o primeiro deles, a proteína 1BS0 (uma enzima envolvida na síntese das vitaminas H e B12 e dependente de piridoxal-5'-fosfato de *E. coli*) (Alexeev *et al.*, 1998) obteve índice máximo "*CERT*" aparentando ser um importante molde. A GumK também obteve bons índices ("*HIGH*") quando comparada a 1F0K (Ha *et al.*, 2000), 1C3J (Moréra *et al.*, 1999) e 1F6D (Campbell *et al.*, 2000) as mesmas encontradas para a GumH, o que sugere que a GumH e a GumK podem compartilhar enovelamentos semelhantes. Para a GumJ o programa encontrou quatro proteínas de bons níveis, representado pelo valor "HIGH": 1EHK (correspondente a citocromo oxidase de *Thermus thermophilus*, enzima da cadeia respiratória) e utilizada como molde para a construção do modelo (Soulimane *et al.*, 2000), 1GW5 (Collins *et al.*, 2002), 2OCC (Yoshikawa *et al.*, 1998) e 1KPL (Dutzler *et al.*, 2002). Para a GumC o programa encontrou a proteína 1C1G (tropomiosina que polimeriza para formar um filamento contínuo que associa com actina em músculo) (Whitby *et al.*, 2000). Outras duas proteínas, GumM e GumB, obtiveram valores máximos para moldes quando comparadas às proteínas 1L7V (proteína de membrana BtuCD, de *E.coli*, transportadora de vitamina B12 através da membrana) (Locher *et al.*, 2002) e 1FVI (uma DNA ligase do vírus *Chlorella*, envolvida na reparação de DNA) (Odell *et al.*, 2000) respectivamente.

Tabela 5.3: Resultados dos programas GenThreader e THREADER 3.3 para as enzimas: GumJ, GumC, GumB, GumM e GumK. Para cada enzima que o programa encontrou solução, são mostrados índices de confiança seguidos do código PDB da proteína encontrada pelo programa.

GumK	CERT	1BS0
	HIGH	1F0K
	HIGH	1C3J
	HIGH	1F6D
GumC	HIGH	1C1G
GumJ	HIGH	1EHK
	HIGH	1GW5
	HIGH	2OCC
	HIGH	1KPL
GumB	CERT	1FVI
GumM	CERT	1L7V

Modelos tridimensionais também foram construídos para as proteínas GumB, GumC, GumK, GumJ e GumM, baseados nas moléculas moldes que obtiveram os melhores índices sugeridos pelos programas. O procedimento adotado para a construção dos modelos foi semelhante ao adotado para a GumH, ou seja, inicialmente 50 modelos foram gerados pelo programa MODELLER 6.0a para cada uma das cinco enzimas e posteriormente o critério de menor energia foi empregado a fim de se filtrar cinco modelos de menor energia para cada enzima. Após uma etapa inicial de validação estrutural realizada pelos programas PROCHECK e VERIFY 3D, apenas um modelo de cada enzima foi qualificado para que estudos minuciosos de sua estrutura possam ser efetuados. Os modelos obtidos para as cinco enzimas podem ser visualizados na figura 5.3.

É interessante notar o predomínio de α-hélices nos modelos da GumJ e GumC em arranjo apropriado para interação com a membrana.

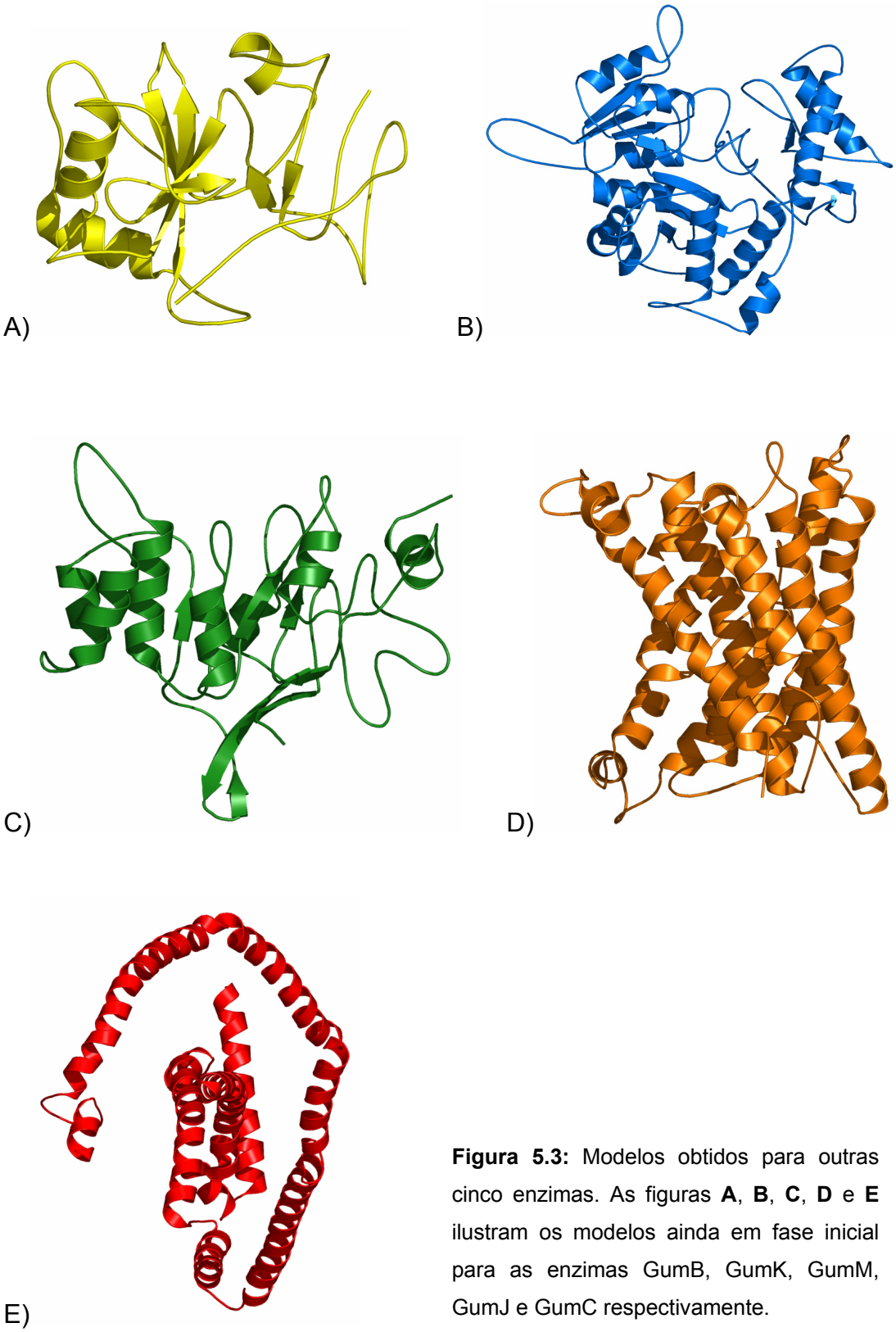


Figura 5.3: Modelos obtidos para outras cinco enzimas. As figuras **A**, **B**, **C**, **D** e **E** ilustram os modelos ainda em fase inicial para as enzimas GumB, GumK, GumM, GumJ e GumC respectivamente.

A predição da estrutura secundária também foi realizada paras as cinco enzimas modeladas com a utilização do programa PSIPRED v2.3 e o conteúdo de α-hélices, folhas β e *loops* estão na tabela 5.4. Observa-se que os conteúdos de estrutura secundária calculados são bastante coerentes com os observados nos modelos.

Tabela 5.4: Conteúdo de α-hélices, folhas β e *loops* para as enzimas GumB, GumK, GumM, GumJ e GumC. A predição das estruturas secundárias foram determinadas com a utilização do programa PSIPRED v2.3.

Enzima	α-hélices	folhas β	*loops*
GumB	36,0%	22,1%	42,9%
GumK	41,4%	16,2%	42,4%
GumM	42,3%	18,5%	39,2%
GumJ	79,4%	6,1%	14,5%
GumC	69,2%	7,7%	23,1%

Estudos mais aprofundados de validação dos modelos assim como a verificação de mecanismos de reação e as interações com os ligantes poderão ser realizados futuramente, a partir dos resultados obtidos neste trabalho.

5.4 Referências bibliográficas

Alexeev, D., Alexeeva, M., Baxter, R. L., Campopiano, D. J., Webster, S. P. and Sawyer, L. (1998) *J. Mol. Biol.*, **284**, 401.

Campbell, R.E., Mosimann, S.C., Tanner, M.E. and Strynadka, N.C.J. (2000) *Biochem.*, **39**, 14993-15001.

Collins, B. M., McCoy, A. J., Kent, H. M., Evans, P. R. and Owen D. J.(2002) *Cell*, **109**, 523-535.

DAS - Cserzo, M., Wallin, E., Simon, I., von Heijne, G. and Elofsson, A. (1997) *Prot. Eng.*, **10**, 673-676.

Dutzler, R., Campbell, E. B., Cadene, M., Chait, B. T., MacKinnon, R. (2002) *Nature*, **415**, 287-294.

Ha, S., Walker, D., Shi, Y. e and Walker, S. (2000) *Protein Science*, **9**, 1045-1052.

HMMTOP - Tusnády, G. E. and Simon, I. (1998) *J. Mol. Biol.*, **283**, 489-506.

Laskowski, R. A., McArthur, M. W., Moss, D. S. and Thornton, J. M. (1993) *J. App. Cryst.*, **26**, 283.

Locher, K. P., Lee, A. T. and Rees, D. C. (2002) *Science*, **296**, 1091.

MEMSAT2 - Jones, D. T. (1998) *FEBS letters*, **423**, 281-285.

Odell, M., Sriskanda, V., Shuman, S. and Nikolov, D. B. (2000) *Mol. Cell*, **6**, 1183.

SOSUI - Mitaku Group – Department of Biotechnology – Tokyo University of Agriculture and Technology – http://sosui.proteome.bio.tuat.ac.jp/sosuiframe0.html

Soulimane, T., Buse, G., Bourenkov, G. P., Bartunik, H. D., Huber, R. and Than, M. E. (2000) *Embo J.*, **19**, 1766.

SPLIT - Juretic, D., Zoranic, L. and Zucic, D. - http://indigo1.biop.ox.ac.uk/sundeep/sunny.html

TMH - Krogh, A., Larsson, B., von Heijne, G. and Sonnhammer, E. L. L. (2001) *J. Mol. Biol.*, **305**, 567-580.

TOPPRED 2 - von Heijne, G. (1992) *J. Mol. Biol.*, **225**, 487-494.

Whitby, F. G. and Phillips Jr., G. N. (2000) *Proteins: Struct., Funct., Genet.*, **38**, 49.

Yoshikawa, S., Shinzawa-Itoh, K., Nakashima, R., Yaono, R., Yamashita, E., Inoue, N., Yao, M., Fei, M. J., Libeu, C. P., Mizushima, T., Yamaguchi, H., Tomizaki, T. and Tsukihara, T. (1998) *Science*, **280**, 1723.

CAPÍTULO 6

CONCLUSÕES E PERSPECTIVAS

A aplicação da bioinformática na análise de seqüências de bases nitrogenadas (DNA) ou de seqüências protéicas (aminoácidos) teve um crescimento vertiginoso nos últimos anos. As técnicas computacionais têm sido fundamentais na atual corrida aos seqüenciamentos genômicos ou nos estudos de proteomas, onde a busca por informações sobre as seqüências genéticas recém determinadas exige um número rápido de respostas e associações quanto a função que esses códigos ainda "secretos" possam revelar.

A seqüência gênica da *Xylella fastidiosa* foi determinada em meados de 2000 em um projeto, onde o Brasil mostrou-se capaz e possuidor de tecnologia suficiente para que toda a seqüência gênica de uma bactéria pudesse ser desvendada. A presença de um operon, denominado "Operon gum" na *X. fastidiosa* revelou a existência de nove genes (gumE, gumJ, gumC, gumB, gumM, gumD, gumF, gumK e gumH) responsáveis pela síntese de um polissacarídeo extracelular denominado goma fastidiana.

No presente trabalho, análises dessas seqüências de DNA foram iniciadas por meio de comparações de alinhamentos das seqüências obtidas da *X. fastidiosa* com os parentes mais próximos, no caso a bactéria

Xanthomonas campestris. Os resultados destas comparações revelaram que as seqüências de aminoácidos são bastante semelhantes com índices de identidade que variaram de 55 a 70% de identidade, com exceção da GumF que apresentou 37% de identidade seqüencial.

A seqüência de aminoácidos das enzimas foram deduzidas a partir da seqüência de DNA, e o próximo passo foi a busca de informações quanto a classificação funcional dessas seqüências. Nesta etapa, os bancos de dados seqüenciais, gênicos e protéicos, revelaram-se importantes aliados da bioinformática. As nove seqüências de aminoácidos das enzimas foram submetidas a esses bancos de dados e através da técnica de alinhamento seqüencial, as seqüências foram comparadas e alinhadas a famílias pré-estabelecidas de seqüências conhecidas.

O fato de que a identidade seqüencial implica em atividades funcionais semelhantes, possibilitou que as enzimas fossem classificadas quanto as suas funções. A GumB mostrou-se uma enzima responsável pela exportação do polissacarídeo, a GumC como responsável no controle do tamanho da cadeia lateral do polissacarídeo, a GumD como um enzima transferidora de açúcares ao lipídio carreador, a GumF responsável pela transferência de um grupo acetil (acetilação do sacarídeo manose no caso da goma fastidiana), a GumM seria responsável pela adição de uma glicose, a GumJ presente na síntese de polissacarídeos e a GumH, também responsável pela transferência de açúcar durante a formação de um polissacarídeo. Entretanto, duas enzimas (GumE e GumK) não foram comparadas a nenhuma outra seqüência, porém, devido ao fato da GumK possuir uma certa identidade seqüencial (69%) quando comparada a GumK

da *X. campestris* a função de transferidora de ácido glucorônico foi a ela associada. No entanto, a GumE ainda permanece sem uma classificação funcional.

Na construção de modelos estruturais tridimensionais através da técnica de modelagem molecular comparativa, a homologia entre duas ou mais proteínas é condição necessária para que um modelo de boa qualidade possa ser obtido. Homologia estrutural não implica necessariamente em grande identidade seqüencial, pois duas proteínas seqüencialmente pouco semelhantes (9% por exemplo) podem possuir um enovelamento semelhante, revelando a presença de ancestrais comuns, mas com divergências evolutivas por meio de mutações seqüenciais. Essas premissas corroboram a utilização de programas que aproveitam a semelhança estrutural e não a semelhança seqüencial para a modelagem molecular. A busca por moléculas candidatas a moldes, estruturalmente homólogas às enzimas Gums, foi feita com os programas GenThreader e THREADER 3.3 utilizando-se bancos de dados locais de estruturas tridimensionais conhecidas,

Foram construídos modelos estruturais tridimensionais para as enzimas GumB, GumK, GumM, GumJ, GumC e GumH. Os modelos gerados para as GumB, GumK, GumM, GumJ e GumC foram validados pelos programas PROCHECK e VERIFY 3D e serão analisados quanto ao sítio catalítico em trabalhos futuros.

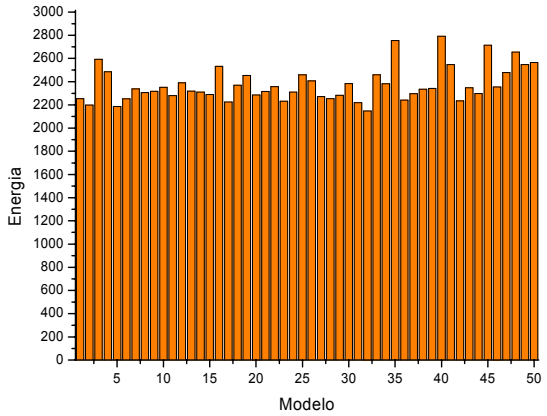
No caso da enzima GumH, quatro estruturas foram consideradas de boa homologia estrutural e, portanto, modelos foram construídos para a enzima, baseados nos quatro moldes. Dos quatro modelos construídos,

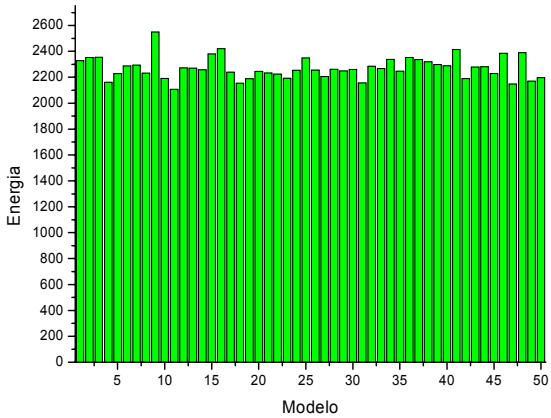
aquele baseado na estrutura da proteína MurG (de *Escherichia coli*) foi o que apresentou melhores resultados na validação de diversos parâmetros analisados. O modelo revelou informações interessantes quanto a disposição de um ácido glutâmico potencialmente envolvido no mecanismo de retenção efetuado pela GumH. A participação de um ou dois ácidos glutâmicos e o suposto mecanismo de retenção da configuração do carbono anomérico do substrato foi especulada após um alinhamento seqüencial em bancos seqüenciais (Pfam/CAZy/Swiss) que revelaram a presença de um motivo EX_7E fortemente conservado em manosiltransferases.

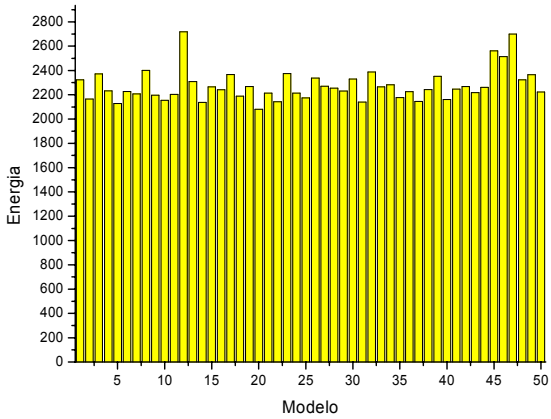
Com base nas informações seqüenciais e na presença desses motivos conservados, a região de ligação dos substratos doador (GDP-manose) e aceptor (celobiose monofosfato) foi proposta. Estudos de *docking* foram realizados e a partir dos resultados obtidos o mecanismo da reação catalisada pela enzima GumH foi proposto. A enzima GumH foi classificada como pertencente a família GT4, onde se enquadram GDP-manosiltransferases, que atuam em um mecanismo de retenção da conformação do carbono anomérico do substrato GDP-manose. A enzima GumH pode atuar por um mecanismo onde duas bases carboxílicas (Glu286 e Asp26) são diretamente responsáveis pelos ataques nucleofílicos e dupla inversão da configuração do carbono anomérico. A enzima pode ainda, de acordo com comparações com estruturas de enzimas que agem com retenção da configuração do carbono anomérico, proporcionar o primeiro ataque nucleofílico através do resíduo Glu286, porém a segunda inversão da conformação anomérica poderia ser feita pelo próprio substrato receptor (celobiose). Essas análises enfatizam o modelo obtido para a GumH

representando o primeiro modelo estrutural concreto proposto para as enzimas envolvidas na síntese da goma fastidiana. Considerando-se que nenhuma GDP-manosiltransferase teve, até o momento, sua estrutura tridimensional determinada, os resultados obtidos neste trabalho resumem-se nas primeiras e únicas informações estruturais desta família de enzimas.

Apêndice A – Gráficos para os modelos gerados para a enzima GumH a partir dos moldes GtfB (*Amycolatopsis orientalis*), 2-epimerase (*E. coli*), e β-GT de fago T4 respectivamente.







Apêndice B – Resultados da avaliação feita pelo programa WHATIF para o modelo GumH baseado no molde MurG (1F0K) de *E. coli*.

N°	AA	Valor	N°	AA	Valor	N°	AA	Valor	N°	AA	Valor
1	LEU	5.716	90	VAL	-0.302	179	VAL	-1.153	268	GLU	-3.238
2	MET	2.468	91	ILE	-1.246	180	GLU	-2.241	269	GLN	-2.540
3	LYS	0.165	92	HIS	-2.735	181	LYS	-5.242	270	ALA	-1.570
4	VAL	-1.167	93	VAL	-3.355	182	TYR	-4.891	271	GLN	-2.607
5	VAL	-2.031	94	HIS	0.599	183	ALA	-1.817	272	PHE	-4.041
6	HIS	-2.206	95	GLY	-2.566	184	ARG	-7.592	273	PHE	-1.930
7	VAL	-2.907	96	ILE	-2.060	185	CYS	-2.165	274	ILE	-3.139
8	VAL	-1.672	97	ASP	-0.427	186	GLY	-0.792	275	SER	3.037
9	ARG	-7.046	98	PHE	4.773	187	ALA	-3.378	276	LEU	-0.871
10	GLN	-2.042	99	PHE	2.396	188	SER	-2.572	277	SER	-0.462
11	PHE	-4.116	100	TYR	3.473	189	GLU	-5.530	278	ARG	-1.395
12	HIS	-2.386	101	ASP	0.411	190	ALA	-2.974	279	HIS	-3.373
13	PRO	-1.160	102	PHE	1.341	191	GLY	-1.621	280	GLU	-1.110
14	SER	-1.721	103	LEU	-2.111	192	ARG	-2.032	281	GLY	-2.861
15	ILE	-2.139	104	ALA	1.061	193	THR	0.014	282	PHE	-2.223
16	GLY	-0.365	105	LEU	-1.176	194	LEU	3.220	283	GLY	-3.997
17	GLY	-0.434	106	THR	-2.222	195	LEU	3.110	284	ILE	-2.286
18	MET	-0.266	107	ARG	-4.190	196	TYR	-0.733	285	ALA	-1.129
19	GLU	-1.643	108	VAL	-2.242	197	PHE	-3.060	286	ALA	0.789
20	ASP	-0.540	109	LEU	-5.008	198	GLY	-1.824	287	VAL	2.514
21	VAL	1.185	110	HIS	-4.331	199	ARG	-3.720	288	GLU	1.103
22	VAL	1.283	111	GLY	-1.306	200	TRP	-4.694	289	ALA	3.009
23	PHE	1.366	112	LYS	0.952	201	SER	-0.967	290	MET	4.826
24	ASN	1.310	113	PRO	1.609	202	MET	-5.870	291	SER	1.173
25	ILE	-0.470	114	MET	0.912	203	ASN	-3.686	292	ALA	1.169
26	ALA	-0.378	115	VAL	0.057	204	LYS	-3.959	293	GLY	-0.504
27	MET	-1.965	116	VAL	-1.846	205	GLY	-0.227	294	LEU	-1.928
28	GLN	-5.093	117	SER	-4.240	206	LEU	-2.286	295	ILE	-2.928
29	LEU	-0.336	118	THR	-2.887	207	LEU	-1.636	296	PRO	-0.271
30	HIS	-1.651	119	HIS	-6.544	208	GLU	-2.056	297	VAL	-3.320
31	LEU	-4.357	120	GLY	-3.208	209	THR	-2.691	298	LEU	-1.685
32	HIS	-5.565	121	GLY	-0.901	210	LEU	-4.152	299	SER	-0.992
33	ALA	-2.205	122	PHE	-6.735	211	GLN	0.250	300	ASP	-3.573
34	GLY	-1.321	123	PHE	-1.863	212	LEU	0.291	301	ILE	-3.931
35	ILE	-3.089	124	HIS	-3.773	213	LEU	-0.806	302	PRO	-1.882
36	ASP	-2.604	125	THR	-2.284	214	ALA	-2.378	303	PRO	0.890
37	VAL	-1.837	126	ASP	-3.423	215	VAL	-2.444	304	PHE	-2.765
38	ASP	-1.591	127	TYR	-0.843	216	LEU	-3.549	305	ALA	-0.891
39	VAL	3.951	128	ALA	2.866	217	TYR	0.610	306	ARG	-3.423
40	VAL	-0.002	129	SER	1.369	218	VAL	-2.478	307	LEU	-3.411
41	THR	0.724	130	ARG	-2.551	219	LEU	-2.630	308	HIS	-0.416
42	LEU	-4.791	131	LEU	-4.307	220	ASP	-3.702	309	ARG	-2.748
43	ASN	-4.089	132	LYS	0.001	221	PRO	-0.460	310	GLU	-4.208
44	ARG	-3.639	133	LEU	-0.847	222	ARG	-3.565	311	SER	3.094
45	VAL	-3.906	134	LEU	1.617	223	TRP	0.696	312	GLY	-3.082
46	PHE	-4.538	135	TRP	-1.912	224	ARG	1.467	313	LEU	-0.017
47	THR	-0.201	136	PHE	-1.127	225	LEU	4.500	314	GLY	1.085
48	GLN	0.313	137	ASN	-4.282	226	ILE	0.290	315	VAL	0.024
49	SER	5.017	138	THR	-0.664	227	ILE	-2.457	316	LEU	-0.360
50	ASP	1.654	139	LEU	2.259	228	ALA	-1.510	317	VAL	-1.277

51	VAL	-3.603	140	THR	1.869	229	GLY	-2.257	318	ASP	-0.658
52	LEU	-2.835	141	ARG	2.704	230	ARG	-0.507	319	PRO	-2.203
53	LEU	-2.972	142	LEU	4.269	231	GLU	0.429	320	LEU	-1.084
54	PRO	-1.888	143	SER	1.790	232	TYR	-5.368	321	GLN	1.257
55	CYS	-2.758	144	ALA	1.553	233	ASP	-1.201	322	PRO	-1.512
56	THR	-1.057	145	LEU	-0.461	234	TYR	-1.996	323	GLN	-0.997
57	ASP	-0.774	146	ALA	-0.354	235	ASP	-2.782	324	GLN	-3.536
58	LYS	-2.297	147	TYR	-6.437	236	GLN	-1.747	325	ALA	-3.527
59	TYR	-4.959	148	ALA	1.436	237	ALA	-1.005	326	ALA	0.686
60	GLN	-3.092	149	ARG	1.160	238	ALA	-0.524	327	VAL	0.280
61	GLY	-1.022	150	ILE	0.105	239	LEU	-2.288	328	ALA	0.145
62	VAL	-2.358	151	ILE	-0.493	240	ALA	-2.377	329	VAL	-0.328
63	SER	-1.712	152	ALA	-0.644	241	TYR	-1.459	330	GLN	1.381
64	ILE	0.695	153	SER	-2.115	242	GLU	-3.262	331	GLY	1.206
65	GLN	1.022	154	SER	1.746	243	VAL	-1.227	332	LEU	1.407
66	ARG	-4.213	155	GLU	-0.786	244	ASP	-2.148	333	ALA	1.571
67	ILE	-1.155	156	SER	-1.907	245	ARG	-3.340	334	VAL	1.020
68	GLY	-0.206	157	ASP	1.533	246	LEU	-2.725	335	GLN	1.948
69	TYR	-3.840	158	GLY	0.675	247	GLY	-0.782	336	VAL	0.442
70	ARG	-4.404	159	ALA	-0.793	248	LEU	-2.041	337	ASP	1.620
71	GLY	-1.051	160	LEU	-1.230	249	SER	-0.460	338	THR	1.233
72	SER	-3.555	161	PHE	-2.985	250	GLU	-1.673	339	HIS	-0.242
73	SER	-2.941	162	SER	-0.160	251	GLN	-3.304	340	PHE	-1.520
74	ARG	-6.679	163	LYS	1.725	252	VAL	-2.371	341	ILE	-2.335
75	TYR	-5.941	164	ILE	0.894	253	HIS	-3.550	342	ASP	-0.868
76	PRO	-1.038	165	VAL	-2.254	254	PHE	-4.420	343	TRP	-3.427
77	LEU	-6.469	166	ALA	0.592	255	HIS	-2.334	344	ARG	-6.690
78	ALA	-2.181	167	PRO	-2.144	256	CYS	-3.238	345	SER	-1.024
79	PRO	-1.358	168	SER	2.078	257	SER	0.242	346	GLN	-0.535
80	TRP	-3.570	169	ARG	-5.494	258	PRO	-1.789	347	ALA	0.810
81	VAL	-2.335	170	LEU	-1.357	259	SER	-2.181	348	MET	-2.346
82	LEU	0.956	171	ARG	-1.363	260	GLN	-5.366	349	ALA	-1.126
83	ARG	0.795	172	VAL	-1.809	261	SER	2.723	350	PHE	-0.147
84	MET	0.300	173	ILE	-2.601	262	GLN	-0.692	351	SER	2.312
85	LEU	1.507	174	GLU	-3.541	263	LEU	4.254	352	ASP	0.174
86	ASP	0.835	175	ASN	-1.319	264	ARG	1.400	353	ARG	-0.970
87	LYS	1.647	176	GLY	-0.928	265	PHE	0.356	354	TYR	-3.368
88	ALA	1.512	177	VAL	-1.716	266	LEU	-0.107	355	HIS	-6.392
89	ASP	1.544	178	ASP	-3.787	267	MET	-2.810	356	TRP	-1.775

N°: número seqüencial do resíduo na proteína; AA: Resíduo. O valor para toda a molécula foi -1.486. Aminoácidos em destaque são aqueles que apresentaram piores índices.