Universidade Federal de Santa Catarina
Centro Tecnológico
Departamento de Engenharia Química e Engenharia de Alimentos
Programa de Pós-Graduação em Engenharia Química

Criação da Base de Dados Via/Genoma da *Chromobacterium violaceum* - CvioCyc e Análise das Informações Geradas pelo Software Pathway Tools

ARTIVA MARIA GOUDEL

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Química da Universidade Federal de Santa Catarina como requisito parcial para obtenção do grau de Mestre em Engenharia Química

Orientador: Prof. Luismar Marques Porto, *PhD*.

Florianópolis, SC

Setembro 2005

Artiva Maria Goudel

Criação da Base de Dados Via/Genoma da *Chromobacterium violaceum* - CvioCyc e Análise das Informações Geradas pelo Software Pathway Tools

Esta dissertação foi julgada e aprovada para a obtenção de grau de Mestre em Engenharia Química no Programa de Pós-Graduação em Engenharia Química da Universidade Federal de Santa Catarina

Área de Concentração

Desenvolvimento de Processos Químicos e Biotecnológicos

_____

Prof. Dr. Luismar Marques Porto Orientador

_____

Prof. Dr. Agenor Furigo Junior

Coordenador do Programa de Pós-Graduação em Engenharia Química

Banca examinadora:

_____ _____

Prof. Dr. Luismar Marques Porto Dra. Cristiana Gomes de Oliveira

_____

Prof. Dra. Selene Maria de A. Guelli Ulson de Souza

Florianópolis – SC, setembro de 2005.

Goudel, Artiva Maria

Criação da Base de Dados Via/Genoma da *Chromobacterium violaceum* - CvioCyc e Análise das Informações Geradas pelo Software Pathway Tools. 183p.

Tese (Mestrado) – Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Engenharia Química.

1. Via metabólica – 2. Base de dados – 3. Operon *VioABCD* – 4. *Chromobacterium violaceum* - 5. Erros de anotação

Este trabalho é parte integrante das pesquisas realizadas pelo Grupo de Engenharia Genômica e foi desenvolvido no Laboratório de Tecnologias Integradas (Intelab) do Departamento de Engenharia Química e Engenharia de Alimentos, Universidade Federal de Santa Catarina.

DEDICATÓRIA

À memória de dois grandes mestres Francis Crick e James Watson, que em 1953 descobriram que a molécula do DNA tem a estrutura de uma dupla hélice, pois deste marco em diante a biologia galgou um novo patamar dando origem à ciência genômica, que nos abre possibilidades infinitas.

À memória de Heloísa Vieira Goudel, minha mãe, que sempre me inspira a superar dificuldades e me motiva a ir além.

Ao meu filho, Gregório Goudel Azevedo, que desde o início acreditou no nosso novo projeto e incentivou-me em todos os momentos.

A todos os que estudam, ensinam ou admiram as ciências em geral.

AGRADECIMENTOS

Agradeço a Deus por ter mantido minha família e a mim saudáveis, dando-me a chance de estudar e aprender em harmonia.

Agradeço à amiga Derce de Oliveira Recouvreux, pelo convite a iniciar um novo projeto; agradeço pela confiança e amizade de sempre.

Agradeço ao Professor Luismar Marques Porto pelo exemplo de motivação, pela orientação e disponibilidade em aceitar-me no grupo de estudos genômicos.

Agradeço a todos os mestres, que transmitiram seus conhecimentos, com muita paciência, competência e amizade. Em especial a Profa. Dra Regina Vasconcellos Antônio que introduziu-me no mundo da ciência bioquímica.

Agradeço aos colegas Cristiana, Luciana, Claudimir, Diogo, Simão e Itamar, por terem me recebido no laboratório InteLAB; seus exemplos, deram-me a certeza de que poderia desenvolver um relevante trabalho neste grupo.

À minha família pelo carinho e atenção, em especial a Elaine Vieira Cortiano, que me ajudou na execução do Abstract. Aos meus amigos da Celesc, aos colegas dos outros laboratórios da Engenharia Química, e a tantas outras pessoas, que de diferentes formas contribuíram para que este trabalho fosse realizado.

À Coordenadoria de Pós-Graduação em Engenharia Química, nas pessoas do Professor Agenor e Edivilson, sempre muito solícitos e atenciosos.

Ao grupo de pesquisa em bioinformática do SRI International (SRI's Bioinformatics Research Group), desenvolvedores do software Pathway Tools, que sempre prestou total apoio à utilização do mesmo.

RESUMO

Artiva Maria Goudel. **Criação da Base de Dados Via/Genoma da *Chromobacterium violaceum* - CvioCyc e Análise das Informações Geradas pelo Software Pathway Tools.** 2005 183p. Dissertação (Mestrado em Engenharia Qmica) – Programa de Pós-Graduação em Engenharia Química, UFSC, Florianópolis, SC.

Microrganismos, cujo genoma já foram completamente seqüenciados, como é o caso da *Chromobacterium violaceum*, possuem dados de anotação genômica em geral produzidos por uma equipe que analisa a fisiologia do organismo e os correlaciona com os dados produzidos pelo projeto de seqüenciamento. Esses dados são geralmente armazenados em bases de dados públicas, e precisam muitas vezes ser verificados, ou seja, validados por via experimental. Todavia, os recentes avanços na área de bioinformática e biologia computacional permitem que certas condições fisiológicas sejam verificadas computacionalmente como, por exemplo, a existência ou não de uma dada via metabólica. Vários grupos têm desenvolvido técnicas para predição de vias metabólicas de organismos a partir da anotação do genoma, produzindo bases de dados integradas via/genoma (*pathway/genome database*). Este trabalho teve como objetivo construir uma base de dados para a *C. violaceum* (CvioCyc), uma bactéria Gram-negativa, seqüenciada pelo *Brazilian National Genome Project Consortium*, de grande potencial biotecnológico e biomédico. A *C. violaceum* produz um pigmento violeta conhecido como violaceína, ao qual são atribuídas propriedades anti-tumorais, anti-chagazíticas, entre outras; além disso, essa bactéria é capaz de produzir biopolímeros de grande interesse comercial. A base de dados via/genoma — CvioCyc (cviocy.intelab.ufsc.br), criada a partir do conjunto de softwares Pathway Tools, mostrou-se uma importante ferramenta de análise do genoma da *C. violaceum*. Através da análise de 61 vias metabólicas, de um total de 233 geradas automaticamente pelo software, 17 vias foram removidas (27,86%) e 44 mantidas (72,13%). Além da inclusão da via de biossíntese da violaceína a partir do aminoácido triptofano, essas 61 vias foram diretamente curadas a partir dos resultados do Pathway Tools, da análise de dados da literatura, e pelo uso de várias outras ferramentas de bioinformática disponíveis na web, tais como ferramentas BLAST e bases de dados de enzimas (KEGG, ENZYME e BRENDA). Trinta e nove ORFs (quadros abertos de leitura), relacionadas às vias analisadas, foram alteradas na base CvioCyc. Isto representa aproximadamente 24,3% de erro numa amostra de 160 ORFs analisadas. Este resultado está dentro da faixa de erros comumente encontrada na literatura, que varia de 8% a 25%. Vários erros se encaixam nos erros de anotação mais comumente encontrados na literatura, como ORFs falso-positivas, falso-negativas, erros de digitação, e falta de padronização nos nomes de enzimas e de genes. A análise dos genes envolvidos na biossíntese da violaceína nos permite sugerir que a ORF CV3270 pode fazer parte do operon *vio*, de acordo a predição do Pathway Tools. É, portanto, provável a existência de mais um gene no operon *vioABCD*. Entre essa ORF e o gene *vioD*, há uma distância de apenas 12 pares de bases, e observa-se uma estrutura em grampo, à jusante desta ORF, o que indica o término da transcrição após a ORF CV3270.

Palavras Chave: Via metabólica, base de dados, operon vioABCD, Chromobacterium violaceum, erros de anotação

ABSTRACT

Artiva Maria Goudel. **Pathway/Genome Database Generation for Chromobacterium violaceum – CvioCyc, and Analysis of Pathway Tools Results**. 2005 183p. Master's Thesis - Chemical Engineering Graduate Program, UFSC, Florianópolis, SC, Brazil.

Microorganisms that have already had their genomes completely sequenced, as in the case of Chromobacterium violaceum, have their genome annotation generally produced by a team that analyze the organism's physiology and its correlation to data produced by the sequencing project. These data are generally stored in public domain databases and most of the time they need to be verified, i.e., experimentally validated. However, recent advances in bioinformatics and computational biology allow us to computationally verify some physiological conditions, such as, for instance, the presence or lack of a particular metabolic pathway. Several research groups have developed techniques to predict metabolic pathways from genomic annotation, thus producing integrated pathway/genome database. The main objective of this work was to build a database for C. violaceum (CvioCyc), a Gram-negative bacterium sequenced by the Brazilian National Genome Project Consortium. C. violaceum is a microorganism of great biotechnological and biomedical potential. C. violaceum produces a violet pigment known as violacein, to which it is attributed anti-tumoral and anti-Trypanosoma cruzi properties, among others; moreover, this bacterium is able to synthesize biopolymer of great commercial interest. The pathway/genome database — CvioCyc (cviocyc.intelab.ufsc.br), was built from a suite of programs in the Pathway Tools package, and was shown to be an important tool to analyze the C. violaceum genome. It allowed us re-annotating 61 out of 233 automatically generated metabolic pathways, 17 were removed (27.86%) and 44 were remained (72.13%). Besides the violacein biosynthesis pathway included in the database, from its precursors (tryptophan), those 61 metabolic pathways were directly curated from the results generated by Pathway Tools, and from literature research, and the use of bioinformatics tools available on the web, such as BLAST and enzyme databases (KEGG, ENZYME and BRENDA). Thirty-nine ORFs (open reading frames) were modified in CvioCyc. That represents approximately 24.3% error in the 160 examined ORFs. This result is in agreement with what we have found for other genome annotations (8 to 25%). Most of those errors are among the frequently found in the literature, such as false-positive and false-negative ORFs, typo errors, and lack of standardization in enzyme and gene names. The analysis done on violacein biosynthesis suggests that ORF CV3270 may be part of the vio operon, according to Pathway Tools predictions. Thus, it is likely that it might exist another gene in the vioABCD operon. We have found that there is only a 12 bp distance between the ORF3270 and the vioD gene; besides that, there is a stem-loop (hairpin) structure downstream that ORF, what suggests a transcription termination moiety after CV3270.

Keywords: Metabolic pathway, database, vioABCD operon, Chromobacterium violaceum, annotation errors.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

CAPÍTULO I

Introdução Motivação e Objetivos

1.1 Introdução e Motivação

O conhecimento das vias metabólicas de um organismo é um constante desafio para os cientistas devido à complexidade nos processos celulares. Uma via metabólica é composta de uma cadeia de reações bioquímicas catalisadas por enzimas que vão transformando substratos numa série de produtos intermediários até que o último destes é convertido no(s) produto(s) final(is). Este produto pode ser utilizado tanto para responder a uma necessidade momentânea como, por exemplo, para a produção de um antibiótico para a defesa da célula frente a um ataque, ou mesmo para atender a uma necessidade constante da célula, como a manutenção celular.

Uma das principais vantagens do estudo das vias metabólicas é que elas permitem visualizar a forma que os componentes moleculares codificados em um genoma interagem uns com os outros e também com outros elementos moleculares para formar a base bioquímica das funções celulares (PANGEASYSTEMS, 2002).

Durante a atribuição das vias metabólicas a um organismo, são encontradas muitas evidências de anotações falso negativas (ausências de anotações de genes existentes), e de falso positivas (anotações de genes que não ocorrem no organismo) (KARP et al., 1999). A atribuição de genes em vias metabólicas também ajuda na validação das anotações genômicas.

Há diversos estudos em diferentes vias metabólicas procurando elucidar os passos dessas vias, muitos com o objetivo de inferir maior conhecimento para serem aplicados em processos de interesse biotecnológico e médicos, tais como mutação, mudanças na regulação de genes, e intervenções nestas vias através de drogas. Neste caso, o objetivo maior é o de se chegar a drogas mais eficientes para a cura de doenças. Sabe-se que muitas vias metabólicas se mantêm através do tempo e que, durante a evolução, estas vias têm sido propagadas inclusive para

organismos filogeneticamente distantes, vias estas que são de vital importância para as células em geral como, por exemplo, as vias que produzem proteínas ribossomais (ROGOZIN *et al.*, 2002).

Devido a importância da violaceína, o estudo da sua via metabólica de biossíntese tem sido estudada em raros organismos produtores deste composto. A violaceína é um pigmento de cor metálica violeta escura, quimicamente bem caracterizado e que possui um grande potencial farmacêutico e biotecnológico. Com característica antibiótica de largo espectro, a violaceína é um componente de grande interesse que algumas bactérias sintetizam, dentre as quais a bactéria Gram-negativa *Chromobacterium violaceum* (AUGUST *et al.*, 2000; BROMBERG e DURAN, 2001). Estudos identificam outras características de grande potencial biotecnológico neste composto, como agente anti-*Trypanosoma cruzi* (MOMEN e HOSHINO, 2000; DURAN e MENCK, 2001), anti-tumoral (MELO *et al.*, 2000), e anti-leishemania (LEON *et al.*, 2001), dentre outros. Pesquisas recentes indicam que a violaceína pode ser eficaz contra células humanas de melanoma da íris (SARAIVA *et al.*, 2004). Muitos estudos têm sido feitos no sentido de usar esta substância como uma alternativa para o tratamento quimioterápico contra o câncer, pois há indícios de que a violaceína em conjunto com outras moléculas destroem as células cancerosas e minimizam a deterioração das células saudáveis (SARAIVA *et al.*, 2004).

A maquinaria celular executa, de forma coordenada, a complexa série de interações bioquímicas que dão à célula condições de se adaptar a flutuações nutricionais e ambientais, e nisto consiste o metabolismo celular. Para o entendimento destes mecanismos é desejável que se dê ênfase ao estudo das vias metabólicas. É através das vias metabólicas do organismo que ele responde às mudanças sofridas por agentes externos. Um agente externo, ou mesmo a manipulação humana sobre as condições normais de vida deste organismo, pode forçar a ativação e/ou desativação de certas vias.

Vários grupos têm desenvolvido técnicas para predição de vias metabólicas de organismos a partir da anotação do genoma, produzindo bases de dados integradas via/genoma (*pathway/genome database* –

PGDB). Tais grupos incluem os projetos KEGG (KANEHISA e GOTO, 2005), o projeto EMP (SELKOV *et al.*, 1998) e o projeto do *SRI International*, o BioCyc *database collection* (KRIEGER *et al.*, 2004).

O KEGG (KANEHISA,1997; KANEHISA e GOTO,2005) é uma base de dados de conhecimento que integra o conhecimento corrente sobre as redes de interações moleculares, genes e proteínas e informações sobre compostos e reações bioquímicas. A base de dados PATHWAY, que faz parte do conjunto de softwares do KEGG, é a base que suporta a rede de proteínas, ou o mapa metabólico, representando as várias funções celulares. Há 235 vias metabólicas de referência (até a data deste trabalho) na base de dados PATHWAY. No KEGG é possível se fazer pesquisas sobre as vias metabólicas de vários organismos. Na consulta ao mapa metabólico de um organismo de interesse é mostrado o mapa de referência que serve para qualquer espécie, porém os genes anotados no genoma são destacados por uma coloração diferente. Os mapas metabólicos propostos pelo KEGG são manualmente compostos e continuamente atualizados.

EMP (*Enzyme and Metabolic Pathways*) é uma base de dados que contém informações bioquímicas, e cobre vários aspectos da enzimologia e do metabolismo; ela continha 3000 diagramas metabólicos até 14 de agosto de 2005.

Outra base de dados via/genoma é o BioCyc, que é uma coleção de bases vindas de uma grande variedade de organismos, principalmente microrganismos e plantas. O objetivo destas bases de dados é conter uma amostra representativa das vias que foram elucidadas experimentalmente. Além disso, muitas destas vias possuem informações complementares, com citações da literatura. Estas bases foram criadas e são atualizadas por um conjunto (ou suíte) de softwares chamado Pahtway Tools (KARP *et al.*, 2002). Outras bases de dados desenvolvidas com este conjunto também estão disponíveis na Internet, como EcoCyc (KARP *et al.*, 1999), base de dados do microrganismo *Eschericchia coli*, AraCyc (MUELLER *et al.*, 2003), base de dados da planta *Arabidopsis thaliana*, primeiro genoma de planta completamente seqüenciado, e PlasmoCyc (YEH *et al.*, 2004), base de dados do microrganismo *Plasmodium falciparum*.

Um grande número de "model-organism databases" ou MODs, bases de dados de organismos-modelos, e software para gerenciá-los, tem sido criado por vários grupos de pesquisa. Por causa do grande número de seqüenciamento genômico de organismos, um enorme esforço de desenvolvimento deve ser feito para cada um deles; em face disto, a reutilização de softwares existentes deve ser levada em consideração. Os softwares para manipular dados biológicos contêm algoritmos tão complexos que inviabiliza a sua reprodução em alguns grupos de pesquisa, e poderia tomar anos de esforços para desenvolvê-los; além disso, a proliferação de softwares incompatíveis dificulta os estudos comparativos entre múltiplos MODs.

Este trabalho trata da estruturação dos dados genômicos da *C. violaceum* e das suas vias metabólicas; um passo inicial nesta direção é a criação de uma base de conhecimento a partir de seus dados genômicos disponíveis após o seu seqüenciamento. É desejável que esta base de conhecimento contenha além das informações genômicas também as informações integradas de vias metabólicas inferidas a partir dos genes anotados.

Com vistas à produção da violaceína em maior escala, em *C. violaceum* ou, heterologamente, em organismos recombinantes, o estudo genômico deste organismo e suas vias metabólicas é de grande interesse científico e tecnológico.

Entre os softwares avaliados para a criação e desenvolvimento de uma base de dados Via/Genoma para a *C. violaceum*, optou-se pelo software Pathway Tools, em virtude das suas ferramentas de análise e manutenção, e da sua abordagem na criação de vias metabólicas específicas. O KEGG, embora utilize uma conceituação genérica de vias multi-espécies, não cria um objeto distinto na base de dados para as variações relativas à espécie do organismo. Além disso, nem o projeto EMP nem o KEGG, ao contrário do Pathway Tools, possuem uma ferramenta de análise de expressão de genes, nem conseguem mostrar o organismo com todas as suas vias metabólicas integradas numa mesma visão, possibilitando inúmeras simulações.

1.2 Objetivos

Este trabalho tem como objetivo a construção e análise de uma base de dados que incorpore as vias metabólicas da *Chromobacterium. violaceum*, em especial a via de biossíntese da violaceína. E foi a partir da construção da base de dados da *C violaceum*, nomeada de CvioCyc, que uma gama de informações pôde ser vista de forma organizada e clara: genes, proteínas, vias metabólicas, e comentários da literatura sobre estas entidades, metabólitos, reações, a rede metabólica integrada, os genes transcritos em conjunto (operons), dentre outras.

Através da análise de algumas vias metabólicas da *C. violaceum*, algumas questões foram levantadas a respeito das anotações das funções dos genes e dos "*EC_numbers*" (*Enzyme Comission numbers*).

Neste trabalho são feitas algumas sugestões para a alteração na anotação da *C. violaceum*, que serão discutidas em Resultados e Discussão, e estão sendo implementadas no Banco de Dados Via/Genoma CvioCyc.

CAPÍTULO II

Fundamentação Teórica e Revisão Bibliográfica

As informações com que se tratam neste trabalho têm como origem o código genético (DNA) de um microrganismo. Para se saber como estas informações chegam ao projeto em questão, e também para ajudar num melhor entendimento do trabalho, algumas definições são descritas a seguir, tendo em vista apenas os organismos procariotos.

2.1 DNA

O DNA, ácido desoxirribonucléico (Figura 1), é o material genético de todos os organismos vivos exceto em alguns vírus (vírus de RNA). Estas moléculas são todas iguais para um mesmo organismo, não importando em que tipo de célula ela está. É no DNA que estão as informações necessárias para que o organismo cresça e se mantenha sob as diversas flutuações ambientais (LEWIN, 2001).

O DNA é uma molécula que define as características hereditárias entre as gerações (LEWIN, 2001). A molécula do DNA consiste de duas fitas compostas de desoxirribose-fosfato e unidas por pontes de hidrogênio entre bases nitrogenadas. Estas estruturas se enrolam e se torcem em si mesmas, lembrando uma escada em caracol, os corrimãos representando as fitas e os degraus as pontes. O modelo da estrutura do DNA veio de um modelo obtido através de difração de Raios-X, por Rosalind Franklin e Maurice Wilkins, juntamente com a observação experimental que em qualquer DNA a composição de adeninas e timinas era igual e que a composição de guaninas e citosinas também eram iguais (MATHEWS *et al.*, 1999; LEWIN, 2001).

Inspirados nestas descobertas dois cientistas, James Watson e Francis Crick, propuseram, em 1953, a estrutura em dupla-hélice do DNA (WATSON e CRICK, 1953).

As estruturas de composição do DNA consistem em uma longa cadeia de monômeros, os desoxirribonucleotídeos, comumente chamados de

nucleotídeos, podendo ser compostas de milhares a milhões deles, dependendo do organismo. As diferentes combinações destas seqüências é que definem os genes de cada indivíduo (MATHEWS *et al.*, 1999).

Em humanos o DNA fica enrolado em proteínas, formando as histonas.

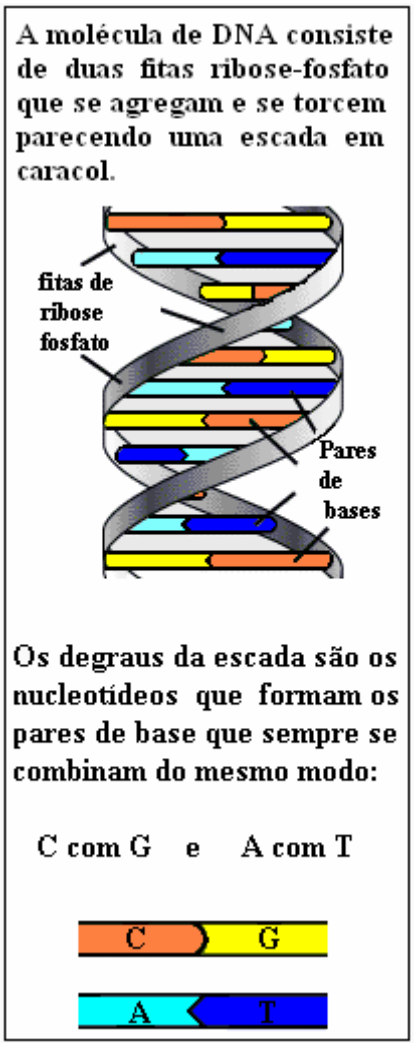


Figura 1: Esquema simbólico da molécula do DNA em sua formação em dupla hélice. Adaptado de BRUCE *e colaboradores* (2002).

2.2 Nucleotídeo

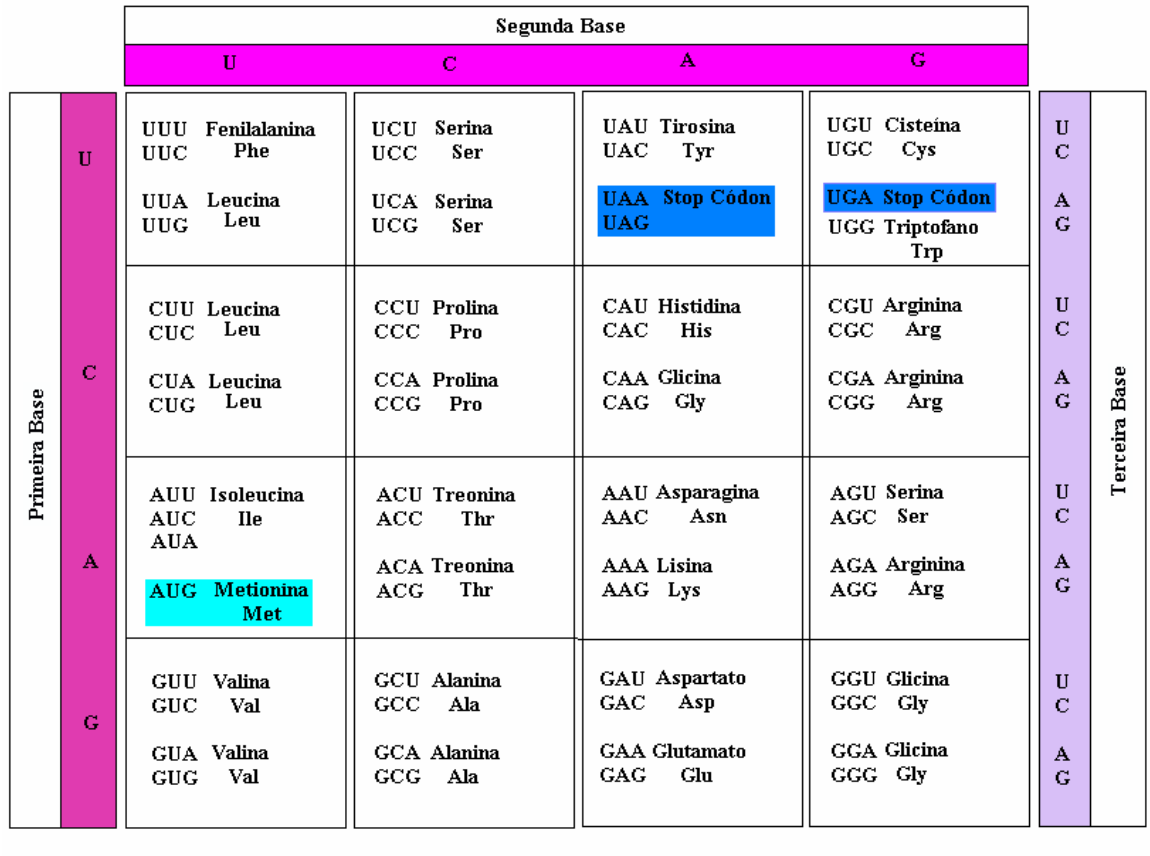
Nucleotídeo é composto por uma molécula de açúcar-fosfato e uma base nitrogenada. Estas bases é que se ligam umas com as outras (pares de base), promovendo a união das duas fitas.

Quimicamente, a ligação entre as bases nitrogenadas, chamada de ponte de hidrogênio, é uma ligação não-covalente. Isto facilita a abertura da fita quando necessário. As bases podem ser adenina (A), guanina (G), citosina (C) ou timina (T) na molécula de DNA e ser adenina (A), guanina (G), citosina (C) ou uracila (U) na molécula de RNA.

Estas bases sempre se ligam duas a duas (pareiam-se) da seguinte forma: Guanina com Citosina e Adenina com Timina ou Uracila (MATHEWS *et al.*, 1999; LEWIN, 2001).

2.3 Genes

Genes são fragmentos de DNA cromossômico capaz de determinar a síntese de uma cadeia polipeptídica, podendo ser uma proteína ou enzima ou parte das mesmas, ou seqüências do DNA que correspondem a rRNA, tRNA ou a outros tipos de RNA. Estes fragmentos são seqüências de nucleotídeos que formarão um composto do qual o organismo necessita, por exemplo, uma determinada proteína ou um ribossomo. Este fragmento de DNA, quando codificar um polipeptídeo, será copiado em uma nova molécula, chamada mRNA e será traduzida em aminoácidos; de modo que cada três bases (códon) codificam um aminoácido. Na Tabela 1 está mostrada esta codificação; ela contém o chamado código genético e diz-se que este código é degenerado porque um mesmo aminoácido pode ser determinado por diferentes códons. A elucidação do código genético foi completada em 1966 (ROSKOSKI e BRAZDA,1996). Na Tabela 1 está sendo usada a base U ao invés da T porque está implícito que as bases já foram copiadas para formar o mRNA. É a partir do mRNA que serão traduzidos os códons para os aminoácidos correspondentes (MATHEWS *et al.*, 1999). Um gene inclui seqüências que participam do início e do fim da transcrição e da tradução, porém não são nem transcritas nem traduzidas. A expressão de vários genes também depende de seqüências regulatórias que podem não estar diretamente adjacentes às regiões codificantes, mas sim estar em regiões bastante distantes destas, do ponto de vista da seqüência, pelo menos (VOET *et al.*, 2000).



Primeira Base	Segunda Base: U	Segunda Base: C	Segunda Base: A	Segunda Base: G	Terceira Base
U	UUU Fenilalanina UUC Phe UUA Leucina UUG Leu	UCU Serina UCC Ser UCA Serina UCG Ser	UAU Tirosina UAC Tyr UAA Stop Códon UAG	UGU Cisteína UGC Cys UGA Stop Códon UGG Triptofano Trp	U C A G
C	CUU Leucina CUC Leu CUA Leucina CUG Leu	CCU Prolina CCC Pro CCA Prolina CCG Pro	CAU Histidina CAC His CAA Glicina CAG Gly	CGU Arginina CGC Arg CGA Arginina CGG Arg	U C A G
A	AUU Isoleucina AUC Ile AUA AUG Metionina Met	ACU Treonina ACC Thr ACA Treonina ACG Thr	AAU Asparagina AAC Asn AAA Lisina AAG Lys	AGU Serina AGC Ser AGA Arginina AGG Arg	U C A G
G	GUU Valina GUC Val GUA Valina GUG Val	GCU Alanina GCC Ala GCA Alanina GCG Ala	GAU Aspartato GAC Asp GAA Glutamato GAG Glu	GGU Glicina GGC Gly GGA Glicina GGG Gly	U C A G

Tabela 1: Código genético. Os códons grifados em azul escuro são os Stop códons, os quais irão parar a tradução. O códon grifado com azul claro codifica a metionina que é o aminoácido que irá iniciar a tradução. Adaptado de Voet e colaboradores (2000).

2.3.1 Genes parálogos

Genes parálogos são genes de um conjunto de genes homólogos que divergiram uns dos outros como conseqüência da duplicação do gene. Por exemplo, os genes beta-globulina da galinha e alfa-globulina do rato (NCBI Hand Book, 2003).

2.3.2 Genes ortólogos

Genes ortólogos são genes de diferentes espécies que derivam de um único gene ancestral do último descendente comum das respectivas

espécies. Por exemplo o gene *WRN* em humanos e o gene *wrn-1* em *Caenorhabditis.elegans*, responsável pela síndrome de Werner (LEE *et al.*, 2004).

2.4 Quadro aberto de leitura (ORF)

Toda a região codificadora do genoma onde se encontram potenciais genes pertence ao contexto das ORFs "open reading frames", em português, quadros abertos de leitura. As ORFs começam com um códon de iniciação, usualmente ATG, mas não sempre, e terminam com um dos códons de terminação: TGA, TAG ou TAA (BROWN, 2002).

2.5 Transcrição

Em 1960, François Jacob e Jacques Monod, do Instituto Pasteur, na França, foram os primeiros a questionar a idéia de que o que representava o conjunto de moldes para a síntese de proteínas eram os RNAs ribossomais (rRNAs). Primeiro porque os rRNAs eram homogêneos em tamanho (5S, 16S e 23S em bactérias), e por outro lado as proteínas têm pesos moleculares variáveis na ordem de no mínimo duas vezes maiores que os rRNAs.

Baseados nesta direção, e em paralelo com outro cientista André Lwoff que estudava o bacteriófago lambda, Jacob e Monod em 1961 propuseram a hipótese da regulação do gene, no qual o gene seria regulado em nível de iniciação. Hipoteticamente, elementos regulatórios chamados repressores e operadores controlariam a síntese de outras entidades chamadas RNAs mensageiros (mRNAs). O mRNA foi postulado ser uma cópia complementar do DNA que contém um conjunto de genes estruturais, os quais codificam proteínas.

O RNA mensageiro é quimicamente um ácido ribonucléico. Ele é composto por moléculas de ribose-fosfato mais base nitrogenada. É diferenciado da molécula do DNA pelo açúcar, pela substituição da timina pela uracila e principalmente, ele é formado de apenas uma fita. Três etapas são necessárias para a formação do mRNA: início, alongamento e

término. A etapa em que o DNA é copiado para o mRNA é chamada de transcrição e está esquematizada na Figura 2 (MATHEWS *et al.*, 1999).

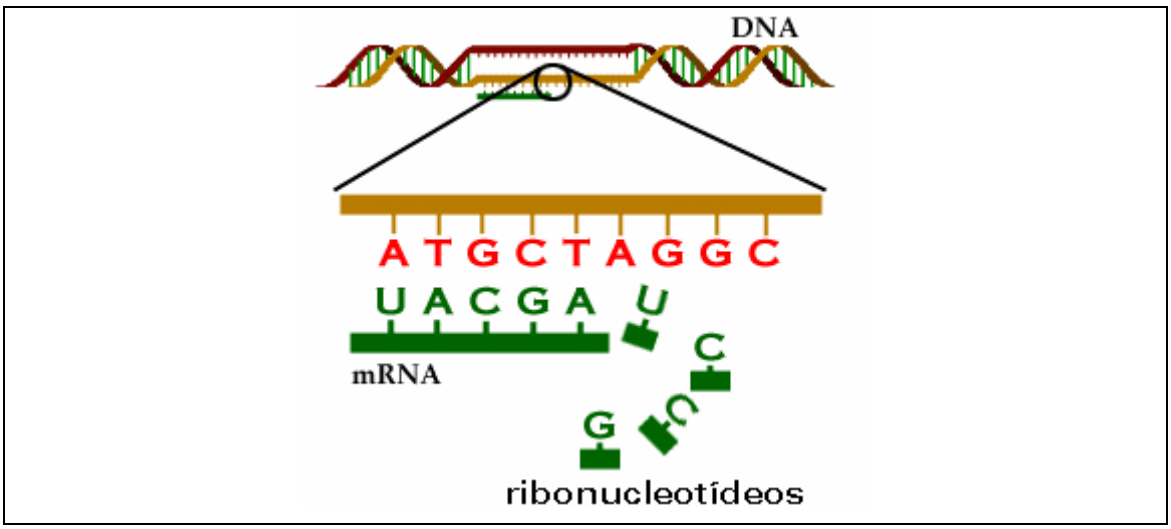


Figura 2: Esquema do processo de transcrição. O RNA mensageiro (mRNA) é sintetizado a partir da fita molde do DNA. Adaptado de BLAMIRE (2000).

2.6 Tradução

É um processo que decodifica o mRNA em proteínas. São 20 os aminoácidos que normalmente participam da síntese, como visto na Tabela 1. Trincas de nucleotídeos, chamados códons, são usados para que cada aminoácido seja traduzido; já que há 4 diferentes bases formando um código de 3 bases, há uma variação de 64 combinações (4^3). Este número é mais que suficiente para codificar os 20 aminoácidos, como pode ser observado na Tabela 1; assim a maioria dos aminoácidos tem múltiplos códons. Um esquema de transcrição em procariotos é mostrado na Figura 3.

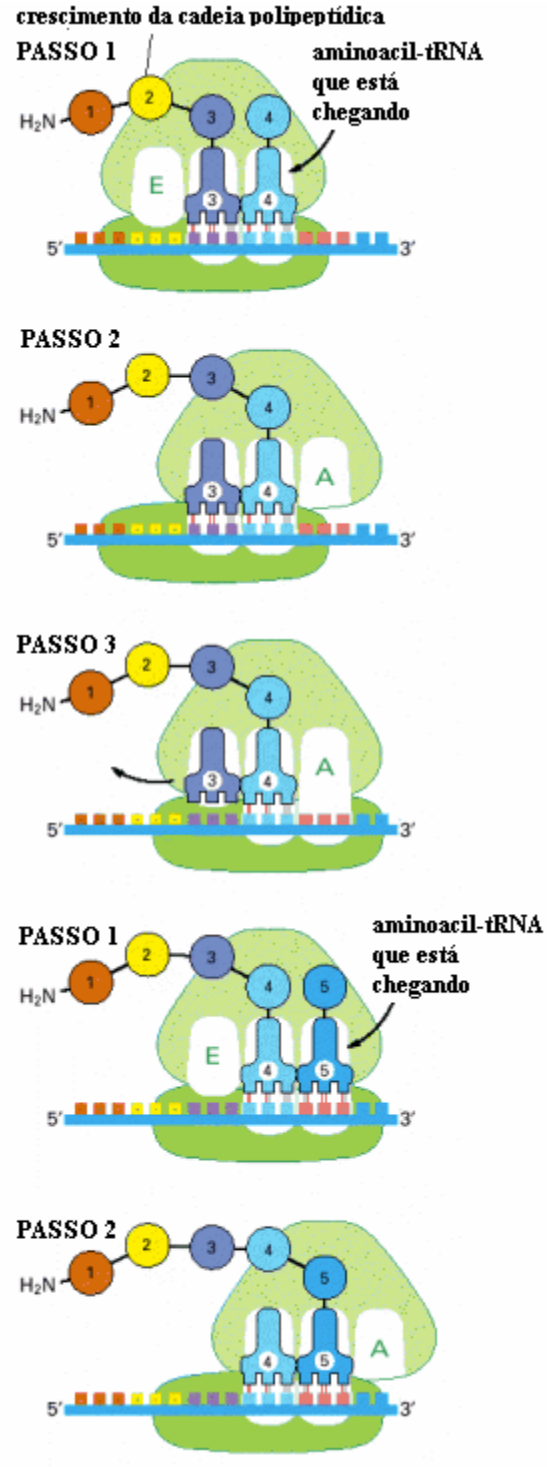


Figura 3: Tradução de uma molécula de mRNA. Este ciclo de 3 passos é repetido várias vezes durante a síntese de proteína. Um tRNA se liga ao local A do ribossomo no passo 1, uma nova ponte peptídica é formada no passo2, e o mRNA move uma distância de 3 nucleotídeos sobre o ribossomo no passo 3, liberando a molécula de tRNA enviada e resetando o ribossoma; assim a próxima molécula de aminoacil-tRNA pode ligar-se ao ribossomo e continuar o processo até que o stop-códon seja encontrado finalizando a cadeia. Adaptado de BRUCE e colaboradores.(2002).

O processo da tradução se dá quando o ribossomo se liga ao mRNA, o aminoacil-tRNA também se liga neste local, e um por um, casando os seus anti-códons com os códons do mRNA, o aminoácido é carregado por cada tRNA que está entrando e é transferido para a cadeia de peptídeo que está se formando. O primeiro tRNA é então liberado, e o ribossomo move-se um códon de comprimento sobre o mRNA, permitindo que o próximo tRNA venha tomar o seu lugar, carregado com o seu aminoácido correspondente. O gasto de energia vindo da hidrólise de um fosfato de alta energia é necessário a cada códon traduzido. Como o ribossomo move-se sobre o mRNA, ele eventualmente encontra um códon de parada (*stop codon*). Neste ponto a cadeia polipeptídica é liberada (MATHEWS *et al.*, 1999).

2.7 Enzyme Commission number

"Enzyme Commission number", conhecido como EC_number, foi estabelecido por uma comissão internacional chamada "International Commission on Enzymes" em 1956 através do professor M. Florkin, do "International Union of Biochemistry", sob anuência do "International Union of Pure and Applied Chemistry" (IUPAC). Esta comissão ficou responsável por estabelecer uma classificação e nomenclatura de enzimas e coenzimas, suas unidades de atividade e métodos padrões para ensaios, juntamente com os símbolos usados para a descrição de cinéticas enzimáticas. Hoje o EC_number significa o número, único para cada enzima, que traduz a classificação padronizada da uma enzima de acordo com este grupo. Em 1977 o NC-IUBMB (*Nomenclature Committee of the International Union of Biochemistry and Molecular Biology*) passou a ser responsável pela lista de classificação das enzimas de acordo com a reação que ela catalisa. Este comitê trabalha com a IUPAC-IUBMB – Joint Comission on Biochemical Nomenclature (JCBN) e de tempos em tempos inclui, atualiza e exclui enzimas da lista estabelecida e a publica para a comunidade científica (BARRETT, 1997). Os quatro dígitos do EC_number descrevem a atividade enzimática da função geral até a específica. O primeiro dígito define a classe da enzima: 1, oxidorredutases; 2, transferases; 3, hidrolases; 4, liases; 5, isomerases; e 6, sintetases. O significado dos dígitos subseqüente depende da classe da enzima e disponibiliza informações sobre cofatores ou

substrato aceptor. O último dígito representa a especificidade pelo substrato, mecanismo molecular ou o tipo de ligação química. Por exemplo, o último dígito é que diferencia as enzimas beta-glicosidase (EC 3.2.1.21) e a beta-galactosidase (EC 3.2.1.23), que são enzimas hidrolases, glicosilases agindo sobre compostos O-glicosil mas que se ligam a diferentes substratos açúcares (DEVOS e VALENCIA, 2001). O sítio do Comitê está disponível no endereço URL http://www.chem.qmul.ac.uk/iubmb/.

2.8 Plasmídeo

É um pequeno pedaço de DNA circular que freqüentemente é encontrado em bactérias. Esta molécula pode ajudar por exemplo, a bactéria a sobreviver na presença de antibióticos, devido aos genes que ela carrega. Para os cientistas, contudo, plasmídeos são importantes por se poder isolá-los em várias cópias do DNA recombinado, porque pode haver centenas de cópias de plasmídeos por célula. Os genes resistentes aos antibióticos são úteis porque este fenótipo pode ser usado para selecionar as células transformadas pelos plasmídeos e também para selecionar os vetores com o DNA recombinante. Os plasmídeos replicam seus DNAs independentemente do cromossomo da bactéria. Eles têm tamanho de 2,5 a 20 k de pares de bases (GRIFFITHS *et al.*, 2000).

2.9 Iniciador (“Primer”)

Um pequeno oligonucleotídeo (de 6 a 50 nucleotídeos) é usado para iniciar a síntese de DNA. Ele serve de ponto de partida para as reações de polimerização adicionais, catalisadas pela enzima DNA-polimerase, para formar um polinucleotídeo. O iniciador pareia-se com as bases de um segmento no polinucleotídeo molde, de modo a formar um segmento de dupla fita curto, que pode então ser estendido pela polimerização direcionada pelo molde (VOET *et al.*, 2000).

2.10 Reação em cadeia da polimerase (PCR)

É uma técnica, criada em 1987, para a amplificação (multiplicação em larga escala de um segmento específico de DNA *in vitro*. O método PCR necessita de conhecimento prévio das seqüências que delimitam a região a ser amplificada para que iniciadores sejam sintetizados juntamente com quantidades dos quatro nucleotídeos e enzimas Taq-polimerases. Pelas passagens de ciclos da temperatura, o DNA destino é repetidamente desnaturado , isto é, a dupla fita do DNA é separada em 2 fitas simples, e os iniciadores pareiam-se com estas fitas simples. Uma cópia única do DNA de interesse pode ser amplificada para obtenção de bilhões de replicações (VOET *et al.*, 2000). O esquema gráfico da reação em cadeia da polimerase é mostrada abaixo na Figura 4.

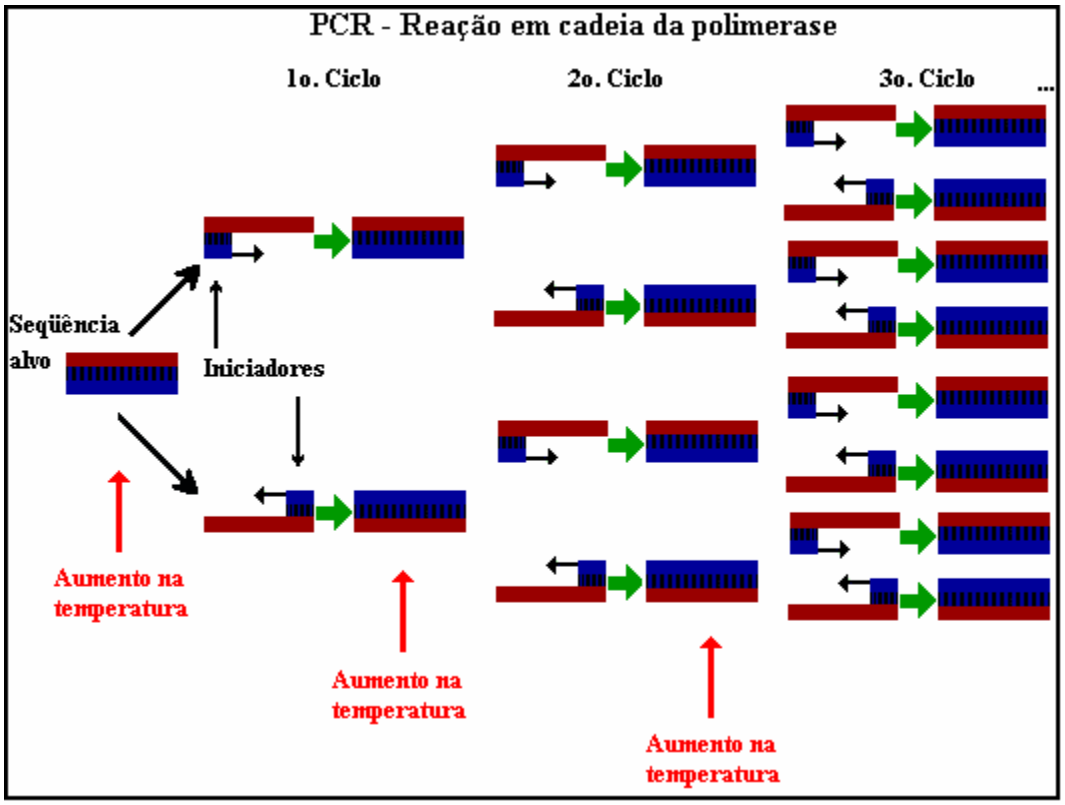


Figura 4: Os 3 primeiros ciclos da técnica para a amplificação de DNA, chamada de reação em cadeia da polimerase, ou mais popularmente de PCR ("polymerase chain reaction"). O aumento da temperatura provoca a abertura do DNA, desta maneira os iniciadores encontram os segmentos de DNA em fitas simples e iniciam a duplicação dos mesmos. Isto vai se repetindo e a seqüência alvo crescendo exponencialmente, em relação a fita origem.

2.11 Genômica

A Genômica é a área da ciência que estuda as informações provenientes do DNA. Um grande número de organismos tem sido seqüenciado, como plantas, bactérias, vírus, fungos e animais. Isto tem acontecido de uma maneira acelerada a partir de 2001, visto que as técnicas para um seqüenciamento mais rápido e confiável foram desenvolvidas através destes últimos anos. Neste novo campo da pesquisa genômica, espera-se que se dê um grande passo, em direção a:

- prevenção de doenças;
- medicamentos mais eficazes;
- soluções preventivas para diversos problemas genéticos;
- biociências.

2.11.1 Seqüenciamento de genoma

O seqüenciamento de genoma é o processo que faz o reconhecimento da molécula de DNA e a traduz em uma seqüência de nucleotídeos.

O método mais utilizado para fazer o seqüenciamento é chamado de "método didesoxi" conhecido também como "método terminadores de cadeia" ou "método de Sanger". Por causa deste método desenvolvido em 1976, Frederick Sanger recebeu Prêmio Nobel em 1980 (SANGER INSTITUTE, 2005).

Com base em VOET e colaboradores (2000), no método de Sanger, o primeiro passo a ser dado é a obtenção de fitas simples da cadeia de DNA a ser analisada, que pode ser feita através da clonagem do fragmento em um vetor que permita ser isolado em fita simples. Em seguida é usada a enzima DNA-polimerase I para sintetizar cópias complementares do DNA de fita simples de interesse. A DNA-polimerase I consegue adicionar nucleotídeos apenas na extremidade 3´ da seqüência, por isso a replicação é iniciada com uma pequena seqüência chamada de iniciador e deve ser complementar a extremidade 3´do DNA.

A próxima etapa é a incubação da mistura de DNA-polimerase I, um iniciador adequado, a fita molde, os quatro nucleosídeos-trifosfatos dATP,

dCTP, dGTP e dTTP (Figura 5). Além disto, mistura-se um composto marcado, que pode ser um dos dNTP ou o iniciador. Este composto marcado pode ser um isótopo radiotivo como o ^{32}P. Acrescenta-se ainda o componente chave da mistura, uma pequena quantidade de 2´,3´-didesoxinucleosídeo-trifosfato (ddNTP), visto na Figura 5.

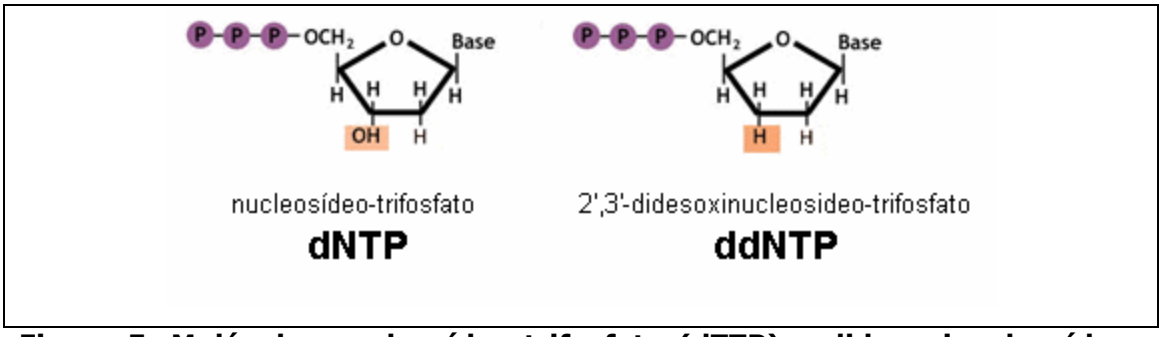


Figura 5: Moléculas nucleosídeo-trifosfato (dTTP) e didesoxinucleosídeo-trifosfato (ddNTP), no qual o grupo 3´-OH do desoxinucleosídeo está ausente, por isto o crescimento da cadeia termina pois a adição do próximo nucleotídeo requer uma posição 3´-OH livre. Adaptado de BROWN (2002).

Para gerar séries, por exemplo, de fragmentos terminados com Adenosina (A), processa-se a DNA polimerase na presença de iguais concentrações de dATP, dCTP, dGTP e dTTP, mais 1/10 desta concentração de ddATP. Quando T está na fita molde, a DNA polimerase ocasionalmente insere ddATP ao invés de dATP. Quando isto ocorre, a cadeia pára de crescer e é liberada da enzima. Portanto, séries de fragmentos com vários comprimentos diferentes acumulam-se, cada fragmento identificando a localização do nucleotídeo Timina (T) na seqüência de interesse. Similarmente, a identificação da localização dos nucleotídeos terminais G, C ou T se dá usando a mesma idéia. A inclusão de uma marcação radioativa na mistura de polimerização e a eletroforese em gel seguida por radioautografia gera quatro escalas de seqüenciamento, cada escala com uma base absolutamente específica. A seqüência de nucleotídeos é obtida pela leitura do menor para o maior fragmento (isto é, da porção inferior para a superior do gel), e é complementar a seqüência DNA de interesse.

Mais recentemente foi introduzida a tecnologia de automatização da determinação da seqüência usando o método de Sanger e ocorre da

seguinte maneira: é incluído no iniciador, no seu terminal 5', um corante fluorescente que pode ser vermelho, azul, verde ou amarelo. Cada uma das quatro reações é realizada com um iniciador colorido com uma cor própria, o qual resulta em uma característica diferente fluorescente para todos os fragmentos terminados por A, T, G ou C, respectivamente. Isto permite tanto a determinação da seqüência sem radioisótopos, evitando contaminação, como a leitura e processamento das seqüências no gel através de processamento computacional. Pode haver uma variação no composto que receberá o corante, tanto o iniciador quanto o didesoxinucleosídeo poderá recebê-lo (MATHEW *et al*., 1999). Na Figura 6 é mostrado o exemplo deste método com o corante específico para cada um dos didesoxinucleosídeos.

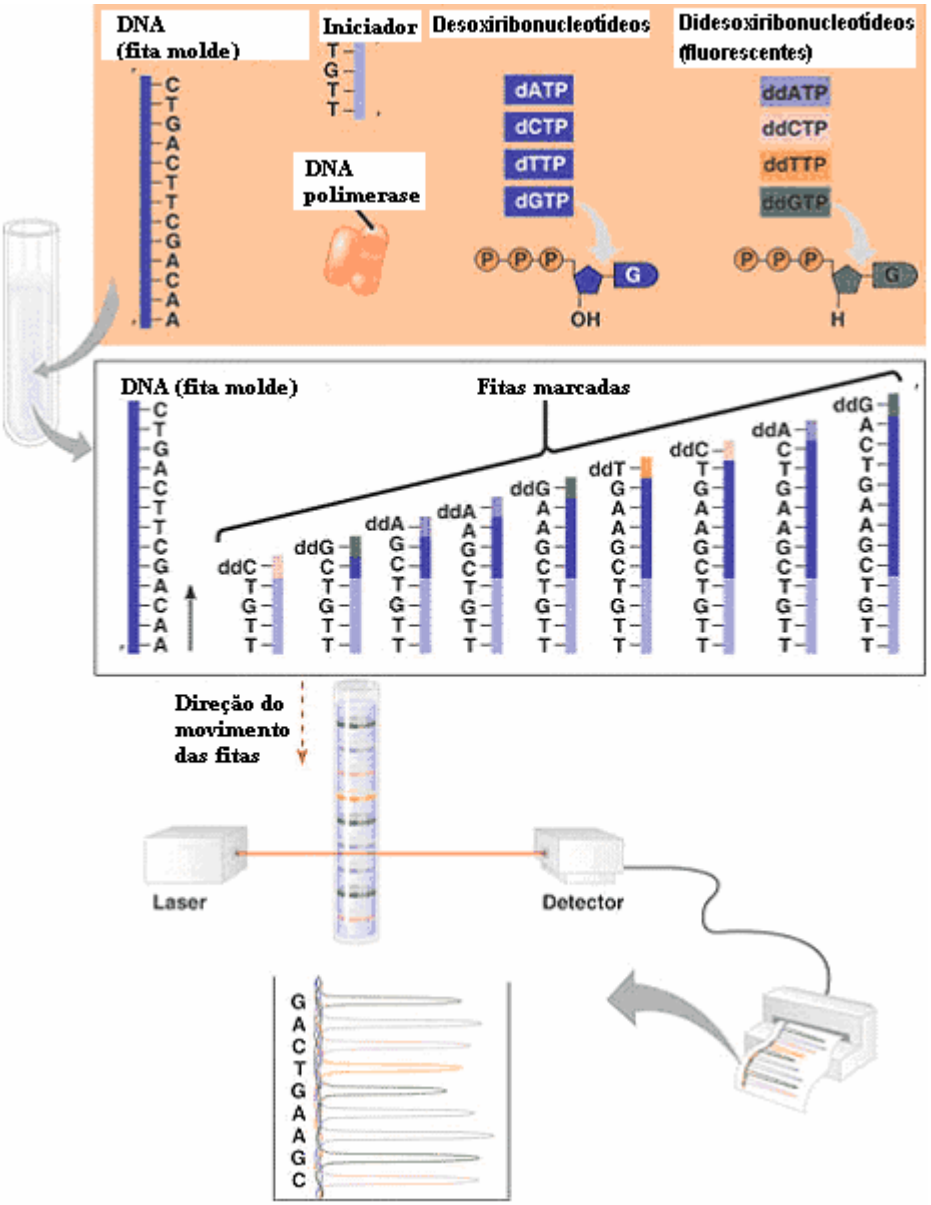


Figura 6: Variação no método de Sanger, os iniciadores utilizados nas quatro reações são ligados a grupos fluorescentes diferentes. Em seguida, através de processo computacional, a leitura do gel gera a seqüência de nucleotídeos. Adapatado de UMBiology (2005).

Os genomas completos são bastante longos e por mais simples que o organismo seja, no mínimo seu DNA vai se compor de milhares de nucleotídeos (os de vírus por exemplo); os DNAs maiores, como algumas plantas e animais, chegam a ter bilhões de pares de bases em seu DNA. Porém, por causa da limitação de processamento dos seqüenciadores, o número de bases que eles processam é normalmente de até 1.000 pares de

base. Por isto, estas quantidades muito grandes de informações genéticas não poderiam ser processadas ao mesmo tempo e dentro de um tempo razoável, sendo que esta questão tem sido um grande desafio. Desde os anos 70, várias técnicas começaram e ser desenvolvidas; porém o que tornou o seqüenciamento de um organismo inteiro viável foi o método "Shotgun".

O método "Shotgun", segundo BROWN (2002), consiste em cortar a molécula de DNA em vários pequenos fragmentos, e determinar a seqüência de cada um, e posteriormente, usando um computador para procurar as sobreposições entre as seqüências, determinar a seqüência principal ("master").

Este método é uma estratégia padrão para seqüenciamento de pequenos genomas de procariotos, mas para genomas maiores é muito mais difícil por dois problemas principais:

1 – Quanto maior o número de fragmentos mais sobreposições ("overlaps") têm que ser analisadas (para n fragmentos o número de possíveis sobreposições é dado por $2n^2 - 2n$).

2 – Podem ocorrer erros quando regiões repetitivas de um genoma são analisadas. Quando uma seqüência repetitiva é quebrada em fragmentos, muitos dos pedaços resultantes vão conter partes das seqüências iguais ou muito similares. Isso daria margem de haver seqüências sobrepostas incorretamente.

Como solução para este problema, um mapa genômico, antes da fragmentação, é feito. O mapa é um diagrama do genoma com as posições de marcadores genéticos e/ou físicos. Este mapa é como um guia mostrando as posições e outras características dos genes no genoma. Marcações são incluídas no genoma e os fragmentos vão estar com uma referência do local de onde vieram.

Uma vez disponível o mapa, há duas alternativas para o próximo passo do seqüenciamento:

- O Método Shotgun do genoma inteiro ("Whole-genome shotgun") usa a mesma abordagem do método shotgun padrão, porém usa as marcações do mapa genômico na hora de unir as milhares de seqüências para montagem da seqüência principal.

As referências do mapa também garantem que regiões repetitivas do DNA sejam unidas corretamente. Esta alternativa é um caminho mais rápido de obter-se a seqüência de um organismo eucarioto, mas ainda há dúvidas do grau de confiança de acerto que pode ser conseguido.

- Método do Clone do contig, onde o genoma é quebrado em segmentos, cada um com comprimento de poucas centenas ou milhares de pares de base, os quais são curtos o suficiente para serem corretamente seqüenciados pelo método *shotgun.* Uma vez que a seqüência de um segmento foi completada, ele é posicionado no seu exato local dentro do mapa. Esta abordagem passo a passo demora bem mais que a técnica anterior, mas é tido como um método que produz uma seqüência mais correta .

Em ambas as alternativas, o mapa genômico é a estrutura que suporta o seqüenciamento nesta fase. Se o mapa indica as posições dos genes, então ele pode também ser usado para direcionar a parte inicial da clonagem do contig para regiões de interesse de um genoma, assim as seqüências de genes importantes são obtidos mais rapidamente. O diagrama dos métodos de shotgun é mostrado na Figura 7.

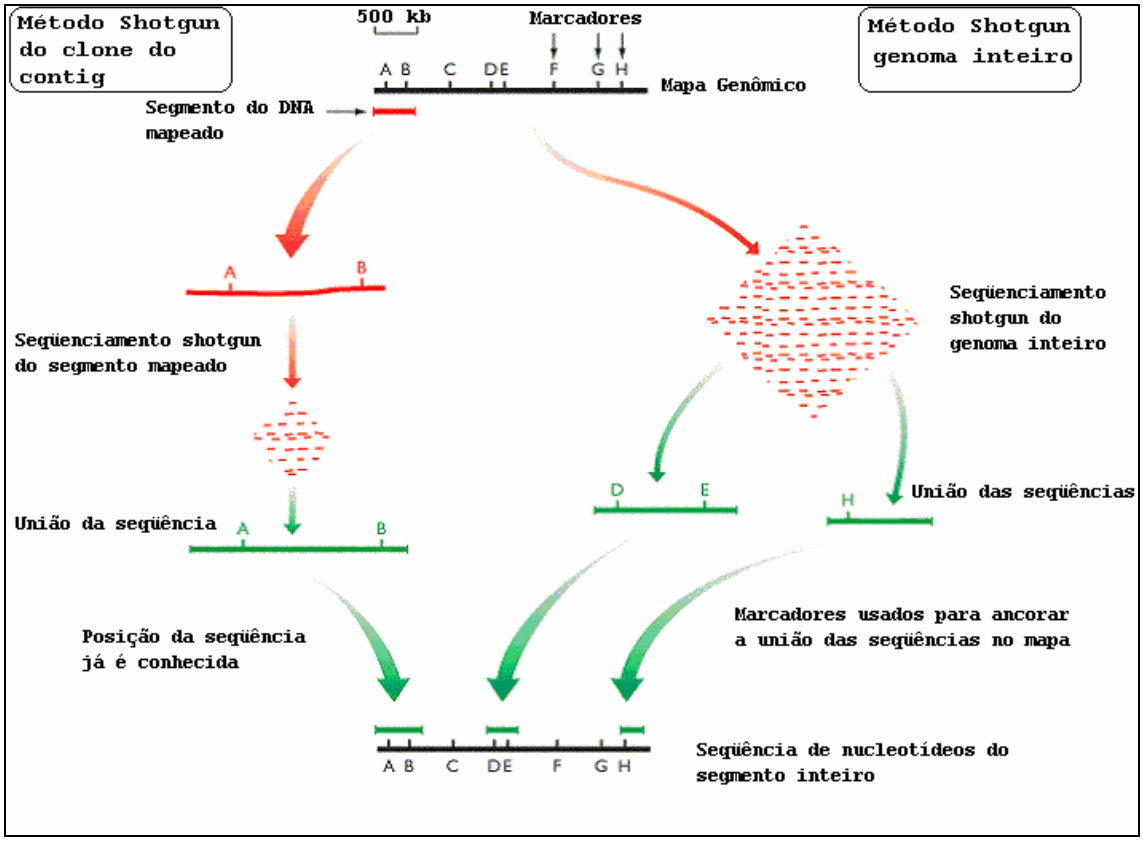


Figura 7: Método shotgun com todas as etapas. À esquerda o procedimento do uso do clone do contig e à direita a técnica shotgun do genoma inteiro. Adaptado de BROWN (2002).

Na fase de união de todas as seqüências vindas dos clones, a sobreposição das seqüências é feita por programas de computador, que comparam umas com as outras e alinham as seqüências em contigs.

Para cada base no contig é usual que se queira que ela seja independentemente confirmada por múltiplas sobreposições da seqüência em ambas as direções. Softwares para construir contigs têm sido desenvolvidos para que leve em conta a qualidade de cada base em uma seqüência (onde a qualidade é uma medida da confiança que o software tem que aquela base foi corretamente definida). Espaçamentos (*gaps*), discrepâncias ou ambigüidades nas seqüências podem ser marcadas para um re-seqüenciamento, possivelmente usando uma técnica quimicamente diferente.

Vários programas têm sido desenvolvidos para este fim; cada seqüenciador já vem com seus programas inclusos, porém os mais

utilizados atualmente nesta etapa de montagem de seqüências, são PHRAP/PHRED/CONSED, conjunto de softwares públicos desenvolvidos por EWING e colaboradores (1998). Estes softwares lêem os arquivos com as seqüências geradas pelos seqüenciadores, chamadas de "calls bases", e atribuem um valor de qualidade para cada "called bases". Foi esta a suíte de software usada no Projeto Genoma Humano do Consórcio Internacional. O "download" (gravação no PC local) deste pacote está disponível publicamente no endereço URL http://www.phrap.org (Bionet, 2005).

O próximo passo é a procura das regiões contaminantes, ou seja, as partes das seqüências obtidas que não fazem parte do DNA origem, por exemplo, pertenciam ao vetor de clonagem. Esta etapa é conhecida por mascaramento dos vetores e é uma função dos programas de análise das seqüências, como o PHRED (EWING *et al*., 1998).

Em seguida vem a etapa de juntar as "*calls bases*" que são contíguas no DNA, formando seqüências maiores chamados de contíguos ou "*contigs*". Esta etapa é executada por um software que se chama genericamente de "*assembler*", há inúmeros disponíveis além do PHRAP, por exemplo o TIGR Assembler, desenvolvido por SUTTON e colaboradores (1995), está disponível publicamente também..

A saída de um programa "*assembl*er" de genoma geralmente é uma coleção de seqüências de DNA que são contíguas (*contigs*), cuja localização dentro do genoma não está definida. Uma função chamada "*scaffolding*" é usada para ordenar e orientar estes *contigs*, nos programas *assemblers* mais recentes, esta função já está incorporada ao mesmo.

Na Figura 8, um exemplo da saída de um seqüenciador. Uma seqüência foi submetida em um seqüenciador onde cores são detectadas em uma lâmina de gel e lida do menor para o maior fragmento. O computador também interpreta as cores imprimindo os nucleotídeos da seqüência no topo do gráfico. Este é apenas um fragmento do arquivo inteiro, que vai até aproximadamente 800 nucleotídeos por seqüência por cada passagem. Este arquivo de saída gerado por um cromatograma é que vai servir de entrada ao programa de montagem do genoma descrito acima (PHRED).

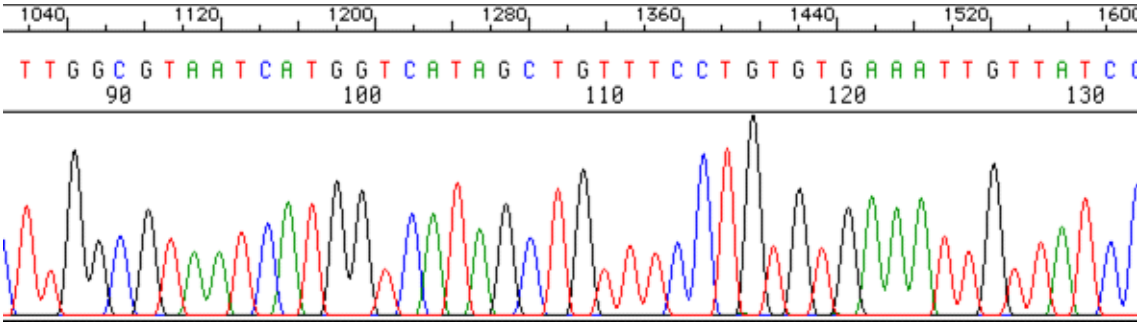


Figura 8: Parte de um relatório de saída do seqüenciador Amplicon Express. Fonte: Amplicon Express (2005).

2.11.2 Anotação genômica

O seqüenciamento de um genoma gera a informação da seqüência completa de bases nitrogenadas que o organismo codifica em seu DNA genômico.

A anotação consiste no processo de saber o que representa cada uma das seqüências nucleotídicas encontradas no seqüenciamento. Primeiramente é necessário identificar, para cada uma das seqüências, se a mesma está inserida em uma região gênica. Uma região gênica é aquela que contém um ou mais genes, que são seqüências do DNA que contêm informações codificantes, isto é codificam para algum produto, como proteína, rRNAs, etc. O principal objetivo desta etapa é construir o mapa genômico do organismo, encontrando cada um dos possíveis genes e as regiões não-gênicas. Alguns programas de predição gênica são usados, por exemplo o Glimmer (DELCHER *et al*., 1999), programa para encontrar genes especialmente em bactérias, arqueas e vírus. Este é um programa público.

Segundo Koonin e Galperin (2003), a "unidade" da anotação genômica é a descrição de um gene individual e seu produto. O ponto focal de cada registro é a função atribuída ao produto do gene. O registro pode também incluir uma breve descrição da evidência da atribuição da função, isto é, o percentual de identidade com a seqüência caracterizada funcionalmente homóloga ou as fronteiras dos domínios detectados na pesquisa a uma base de dados de domínio, mas não há lugar para informar qualquer detalhamento das análises.

A anotação do genoma vai demandar um grande volume de trabalho em termos de pesquisa para a anotação de genes, de agrupamentos de genes, de vias metabólicas, de regiões de regulação, dentre outros. Grande parte deste trabalho é realizada pelos grupos responsáveis pelo seqüenciamento, com base em suas experiências pessoais e, sobretudo, em dados da literatura e de bases de dados disponíveis na Web. Um recurso bastante usado pelos pesquisadores é a comparação das regiões gênicas do organismo em estudo com outras seqüências já existentes em bases de

dados públicas, estas comparações são feitas através de programas de computadores como por exemplo o BLAST (Altschul, 1990). Achando seqüências com similaridade viável, parte-se para a análise pessoal do curador da seqüência, que usará os recursos disponíveis para aceitar ou não aquela possível atribuição de função à sua seqüência (KOONIN e GALPERIN, 2003).

Muitas informações disponibilizadas na Internet, a exemplo do KEGG, utilizam as informações genômicas juntamente com outras informações sobre reações enzimáticas e vias metabólicas, para gerar informações que contribuiam com mais subsídios para novas descobertas científicas.

A anotação automática dos genomas pode gerar falsas expectativas quanto à previsão de determinadas vias metabólicas. Por exemplo, enzimas anotados corretamente podem, teoricamente, participar de mais de uma via metabólica, o que pode, em princípio, gerar previsão de vias não existentes no organismo em estudo.

No sentido de auxiliar no processo de anotação genômica, abordagens computacionais que automatizam parte do processo de organizar as vias metabólicas dos organismos têm sido desenvolvidas de maneira acelerada, cada uma tendo visões próprias das soluções para o mesmo problema e muitas vezes propõem a mesma solução, porém com interfaces diferentes.

Problemas em geral com a anotação equivocada de genomas têm ocorrido freqüentemente. Muitas anotações são baseadas em seqüências de baixas identidades. Em muitos casos, alinhamentos parciais são interpretados como se fossem globais (DEVOS e VALENCIA, 2001). Muitas vezes, apesar dos pares de bases se relacionarem, identificados por suas estruturas similares, eles não são funcionalmente correspondentes (DEVOS e VALENCIA, 2000). Enfim, muitas re-anotações estão sendo feitas ao longo do tempo, não só por causa da disponibilidade de técnicas mais recentes e confiáveis mas também por causa de erros de análise no processo de anotação.

2.11.3 Erros de anotação genômica

Vários pesquisadores têm estudado os erros nas anotações genômicas, principalmente para buscar as causas dos mesmos. BORK e BAIROCH (1996) estudaram alguns exemplos que mostram vários tipos de pistas falsas nas bases de dados usadas para as pesquisas para anotação de seqüências. Eles buscaram as causas destes equívocos e encontraram algumas possíveis origens:

Sinônimos: Em organismos que são alvos importantes nos estudos genéticos, é freqüente que vários grupos isolem os mesmos genes de interesse e cada grupo pode dar um nome diferente para o mesmo gene, como no caso do gene *bns* de *E. coli*, conhecido também como *bnsA, drdX, osnZ, blgY, msyA, cur, pilG* e *lopS*. Isto se dá com nomes de proteínas também, por exemplo, "annexin V" foi também chamada de "lipo cortin V", "endonexin II", "calphobindin I", "placenteral anticoagulant protein I", pp4, "thomboplastin inhibition", "vascular anticoagulant-alpha" e "anchorin CII";

Genes diferentes com mesmo nome: É frequente acontecer que o mesmo nome de gene é dado para genes diferentes. Geralmente um dos nomes é rapidamente alterado, mas em alguns casos os dois nomes fazem pressão e são simultaneamente promovidos, como, por exemplo, o gene de levedura *mrf1* que é tanto o gene para a cadeia de peptídeo mitocondrial fator 1 e também para a proteína 1 da função respiratória mitocondrial. Um exemplo famoso é o "cyclin", nome aceito para uma grande família de componentes do ciclo celular, no qual tornou-se tão proeminente que seu nome não é mais usado para uma proteína, agora conhecida como "proliferating cell nuclear antigen" (originariamente chamada de "cyclin");

Erros de escrita: Estes erros podem terminar como sinônimos. Por exemplo, o gene de levedura, *scd25* (antes *CDC25*), foi tão frequentemente anotado de forma errada, que se tornou um sinônimo aceito. Também as consultas às bases de dados podem ter seu acesso dificultado por: diferenças de sintaxe entre o inglês dos Estados Unidos e da Inglaterra (exemplo, "hemoglobin" e "haemoglobin"); representação de caracteres

especiais, tais como caracteres de acentuação (exemplo,Krüppel, Krueppel e Kruppel);

Erros de origem biológica e de contaminação: Há muitos erros nas anotações devido à origem biológica de uma seqüência. Por exemplo, para a localização do gene na célula, uma divisão específica na base de dados EMBL/GenBank é definida, porém há muitas anotações que são inadvertidamente colocadas em divisões não procedentes. Isto é um problema quando se troca o local do núcleo pelo local da mitocondria, pois o código para a tradução do mRNA é diferente para estes dois locais.

De acordo com GALPERIN e KOONIN (1998), quando atribuições dúbias de função são usadas como base em predições subsequentes, os erros tendem a se proliferar, levando a uma "explosão da base de dados" ("database explosion"), isto é, o erro vai se multiplicando nas bases de dados que usam como base de pesquisa estas informações dúbias. Os autores também identificaram algumas causas mais comuns de erros nas anotações funcionais:

1. A anotação do melhor resultado de busca na base de dados é reproduzida de forma não crítica, mesmo quando a função anotada para a proteína não seja conhecida ou não pode ser realisticamente esperada ocorrer no genoma que está sendo caracterizado.
2. Levar em conta somente a anotação da melhor resposta vinda de uma base de dados;
3. Mascaramento insuficiente de regiões de baixa complexidade nas seqüências de proteína (por ex., domínios não-globulares), implicando em resultados de pesquisa falsos e com isto resultados relevantes ficam relegados;
4. Ignorar organização de multi-domínios nas proteínas consultadas e/ou nas respostas das bases de dados;
5. Inferências funcionais não críticas baseadas nos produtos de genes vizinhos.

Brenner, em 1999, comparou três anotações independentes do orgasnismo *Mycoplasma genitalium*; ele examinou manualmente todas as

anotações conflitantes. Sua conclusão foi que há no mínimo uma taxa de 8% de erros entre os 340 genes anotados por no mínimo dois dos três grupos. Num estudo similar empregado por KOONIN e GALPERIN (2003), feito com base na base de dados COG - Cluster of Orthologous Groups (TATUSOV *et al.*, 2001), de 786 COGs, 194 tinham conflito na anotação no GenBank. Isto sugere, mais pessimistamente, uma taxa de erros de anotação de no mínimo 25% usando o mesmo critério aplicado por Brenner. Tanto na anotação manual como na automática são encontrados os mesmos problemas típicos; inevitavelmente a automatização no processo de anotação tende a aumentar a probabilidade destes problemas. BRENNER (1999) também sugere causas para as principais origens de erros nas anotações genômicas, listados mais ou menos na ordem decrescente de comprometimento:

(i) resultados espúrios das bases de dados, freqüentemente causado pelas regiões de baixa-complexidade da seqüência procurada ou na seqüência da base de dados;

(ii) transferência não crítica da predição funcional de um registro não confiável da base de dados;

(iii) interpretação incorreta (falta de reconhecimento) da arquitetura multidomínio ou na seqüência procurada ou na base de dados;

(iv) predição funcional demasiadamente específica;

(v) falta de predição de funções.

Segundo DEVOS e VALENCIA (2000), o grande vazio entre os milhares de seqüências de proteínas e suas funções tem levado a prática da atribuição de possíveis funções como se fossem funções consolidadas com base em seqüências similares. As dificuldades deste processo estão relacionadas, parte na definição teórica da função e parte em problemas práticos. Eles apontam alguns ítens que podem estar relacionados a estes problemas:

- As ferramentas de análise e as bases de dados, um campo em que se tem feito imensos progressos (PSI-BLAST, Hidden Markov Models);

- A persistência de erros sistemáticos na detecção de homologia devido a regiões equivocadamente homólogas, sendo de natureza diferentes;
- A anotação de mesmas seqüências feitas de maneiras diferentes em base de dados diversas, como tem sido amplamente descrito;
- E a propagação de erros simplesmente pela cópia repetida de uma anotação incorreta.

Muitas vezes, apesar dos pares de bases se relacionarem, identificados por suas estruturas similares, eles não são funcionalmente correspondentes. Além disso, tem sido observado que há pequena diferença de resultados derivados de análises automáticas para os derivados por especialistas. As diferenças entre as anotações feitas pelo software de anotação GeneQuiz (ANDRADE *et al.*1999) e aqueles feitos por especialistas, pode ser estimado em 10% aproximadamente, o que não é muito representativo (DEVOS e VALENCIA, 2000).

OUZOUNIS e KARP (2002) afirmam que as anotações são normalmente fundamentadas na hipótese de que funções e estruturas podem ser transferidas entre seqüências similares porque elas têm sido conservadas ao longo do tempo (Ortologia). Muitas anotações de genomas são baseadas em seqüências de baixas identidades, e em muitos casos em alinhamentos parciais das mesmas. Também, neste mesmo trabalho, os autores afirmam que, como as informações incorporadas às bases de dados estão crescendo e sendo refinadas o tempo todo, haverá mais oportunidades para gerar predições mais confiáveis e hipóteses mais complexas. Contudo, deveria ser bem enfatizado que o sucesso de tal cenário implica que as informações das bases de dados capturem o conhecimento sobre as funções moleculares de maneira específica e sensitiva, onde específica significa sem falso-positivas, que são anotações feitas de genes que na realidade não ocorrem no organismo, e sensitiva significa sem falso negativas, que são ausências de anotações de genes existentes.

A preocupação da transferência de função de uma molécula já caracterizada, tanto experimentalmente como computacionalmente, para

uma seqüência nova pesquisada, leva a necessidade de existir um balanço entre uma abordagem conservadora para anotação (risco de falso-negativa) e uma abordagem mais agressiva (risco de falso-positiva). Evidentemente, forçar este balanço para uma abordagem mais conservadora pode ser preferível a ocorrer o erro de propagação de resultados falso-positivos. Os autores concluem que a anotação genômica pode conter uma considerável quantidade de erros, desde o simples erro de digitação até erros complexos relacionados à análise da seqüência. O resultado mais surpreendente deste estudo é que, sistemas automáticos podem fazer a anotação de seqüências genômicas tão bem quanto um grupo de especialistas (OUZOUNIS e KARP, 2002).

Paralelamente às constatações dos erros de anotação, principalmente por causa das descobertas e caracterizações de novos genes e por surgirem ferramentas mais eficientes, há um movimento no sentido da re-anotação genômica por parte de grupos de pesquisas a sistemas metabólicos específicos, em diversos organismos já seqüenciados.

Um exemplo de re-anotação é o que fizeram DARASELIA e colaboradores (2003), eles realizaram a re-anotação do genoma da *Shewanella oneidensis*, onde chegou-se aos seguintes resultados: 51 novos genes foram identificados, e anotação funcional foi adicionada a 97 genes, incluindo 15 novos e 82 existentes com funções previamente não atribuídas. A identificação de novos genes foi aumentada pela predição de regiões codificantes de proteínas através do programa baseado na metodologia das cadeias ocultas de Markov (HMM Hidden Markov Model), o GeneMark.hmm. Foram usadas comparações subseqüentes de produtos de genes obtidos usando o BLAST contra bases de dados de proteínas não redundantes, e a base de dados COG (Cluster of Orthologous Groups), usando o programa COGNITOR para a anotação funcional.

No trabalho de BETTS e colaboradores (2004) há uma análise detalhada e uma re-anotação de genes referentes a dois conjuntos de genes (*cluster*) que estão relacionados ao sistema de secreção do tipo-III, da *C. violaceum*. A anotação original, revela a presença de genes associados com este sistema, porém incompleto, segundo os autores. Este trabalho também esclarece melhor, computacionalmente, a função de

vários genes codificados como "putative type-III effector proteins" e levam à predição computacional de que este organismo pode manipular o movimento vesicular, a actina do citoesqueleto e as vias apoptóticas dentro das células de mamíferos, para sua própria vantagem. Os dois clusters se compõem de 26 genes para o sistema Cpi-1 e 39 genes para o sistema Cpi-2, totalizando a re-anotação de 65 genes.

2.12 Vias metabólicas

Os caminhos pelos quais a célula converte compostos externos que entram pela membrana celular (proteínas, ácidos nucléicos e carboidratos) em compostos celulares que serão de seu interesse, são chamados vias metabólicas do organismo. Estes processos são usados inclusive para a obtenção da energia necessária para a sobrevivência e duplicação da célula.

Em um aspecto bastante geral, é possível dizer que as vias metabólicas são processos bioquímicos que os organismos utilizam com o objetivo bem definido de sobreviver e deixar descendentes para preservação da sua espécie.

Nestes processos bioquímicos, as enzimas executam papel primordial. Enzimas são moléculas que aceleram os processos bioquímicos, sem, no entanto, participar deles como reagente ou produto, cada interação bioquímica terá uma enzima específica para agir nesta reação.

O metabolismo pode ser dividido em dois grandes grupos:

ANABOLISMO: são os processos envolvidos principalmente com as sínteses de moléculas orgânicas complexas.

CATABOLISMO: são os processos relacionados à degradação e desassimilação da matéria. Envolvem reações de oxidação para liberação de energia (MATHEWS *et al.,* 1999). Na Figura 9 é mostrada o esquema gráfico do metabolismo celular em procariotos.

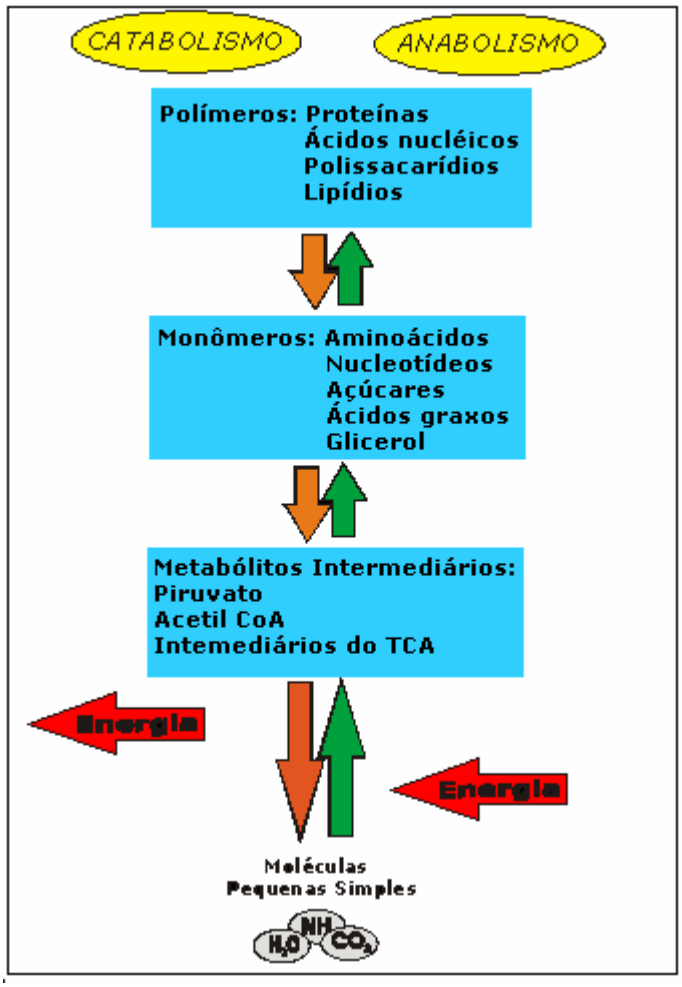


Figura 9: Esquema do metabolismo celular. Fonte: Adaptado de Mathews *et al.* (1999).

O mecanismo central nos processos metabólicos é controlado pela regulação da atividade enzimática, isto é, através da regulação da concentração e da atividade de enzimas presentes ou ausentes na célula. A maquinaria celular executa, de forma coordenada, a complexa série de interações bioquímicas, que darão à célula condições de se adaptarem a flutuações nutricionais e ambientais.

As vias metabólicas colocam os genes num contexto biológico amplo, e são primordiais para o entendimento dos processos complexos codificados pelo genoma.

Segundo ILIOPOULOS e colaboradores (2003) o desafio da era pós-genomica é caracterizar as múltiplas funções bioquímicas num contexto genômico, através de abordagens como a identificação de clusters de genes

participando nos mesmos processos celulares, assim como identificar o papel bioquímico para os genes de funções desconhecidas.

A análise de vias metabólicas vindas de um genoma pode ajudar a identificar anotações funcionais falso-negativas ou anotações funcionais falso-positivas. Através da análise de vias, também pode-se observar vias que, apesar de serem amplamente conhecidas e sabidamente fazer parte das funções básicas do organismo, possuem muitos passos ausentes; isto pode ser devido a erros de anotação ou mesmo refletir uma variação na via deste organismo. As anotações de genes pertencentes a vias com poucos passos existentes em comparação ao número total de passos devem ser alvos de uma cuidadosa análise (KARP *et al.*,1999).

Segundo LANGE e GHASSEMIAN (2005), a representação visual de vias ajuda os biólogos a entenderem a complexa relação entre os compostos das redes metabólicas e também, disponibilizam um recurso inestimável para a integração dos conjuntos de dados dos transcriptomas, proteomas e metabolomas.

2.13 Regulação da expressão gênica

Os organismos respondem rapidamente à mudança do ambiente em que vivem. A lógica metabólica envolvida é óbvia: Muitos genes são expressos somente quando os seus produtos são requeridos, por exemplo para utilizar um substrato disponível, para sintetizar um metabólito complexo que está ausente no meio, ou de alguma outra maneira de responder à mudanças nas condições do ambiente (MATHEWS *et al.* 1999).

Centenas de diferentes reações enzimáticas ocorrem simultaneamente durante um único ciclo do crescimento celular. A maioria dos microorganismos tem informação genética para codificar muito mais tipos de proteínas do que as proteínas presentes na célula sob determinadas condições. Portanto, a necessidade da regulação das reações bioquímicas em resposta a mudanças nas condições de crescimento, ou mesmo como parte dos processos de manutenção está clara (MADIGAN *et al*., 2000).

Entre um gene codificado no DNA, e a proteína pronta na célula, há diversas etapas, e todas elas, em princípio, podem ser reguladas. Portanto uma célula pode controlar a produção de proteínas, enzimas ou qualquer outro composto que ela produz. Em geral, mais de 10% dos genes do genoma são dedicados à expressão e regulação dos genes (SILVA *et al.*, 2004).

2.13.1 Controles na transcrição e tradução

Procariotos tem dois níveis de controle de expressão de genes. Um é o transcricional, se compõem dos mecanismos que controlam a síntese do mRNA, e o outro o traducional, mecanismos que controlam a síntese de proteínas depois que o mRNA já foi produzido.

Mecanismos Transcricionais: as múltiplas etapas na produção dos mRNAs possibilitam oportunidades para controles adicionais na regulação da transcrição:

- Controle transcricional: controle de quando e quão freqüente um determinado gene será transcrito;

- Controle no processo de transcrição do mRNA: controle da ocorrência da transcrição;
- Controle no transporte e localização do mRNA: seleção dos mRNAs que serão exportados para o citossol e determinação da localização deles no citossol.

Mecanismos Traducionais:

- Controle de degradação do mRNA: desestabilização de certos mRNAs no citoplasma.
- Controle na etapa da tradução do mRNA em produtos.
- Controle na seleção de quais mRNAs, no citoplasma, serão traduzidos.
- Controle de atividade da proteína: ativação e desativação, degradação ou compartimentalização de determinadas proteínas depois de elas terem sido traduzidas.

A maioria dos controles transcricionais de genes são dominantes. Isto faz sentido porque somente o controle transcricional assegura que a célula não irá sintetizar intermediários desnecessários (ALBERTS *et al.,* 2005; MAYER, 2005).

2.13.2 Promotores transcricionais

A seqüência do DNA a qual a enzima polimerase se liga para iniciar a transcrição de um gene, é chamada de seqüência promotora ou promotor. A seqüência do DNA envolvida na função promotora localiza-se na região à montante (*upstream*) do inicio da transcrição e é formada por dois conjuntos de seqüências que ocorrem em diferentes genes. Estas duas seqüências são compostas de seis nucleotídeos cada uma, e estão localizadas a aproximadamente -10 e -35 pares de bases à montante do local do início da transcrição. Eles são chamados de elemento -10 (também conhecido como *Pribnow-box*) e elemento -35, e a posição de início da transcrição de posição +1, que na maioria dos procariotos ocorre em um único nucleotídeo de purina. As seqüências nas posições -10 e -35 em diferentes genes não são idênticas, mas elas são similares o bastante para estabelecer-se uma seqüência consenso, que são bases mais

freqüentemente encontradas em cada posição. Na Figura 10, é exibida uma parte da seqüência do gene *lac* de *E.coli* e a sua região promotora (VOET *et al.*, 2000).

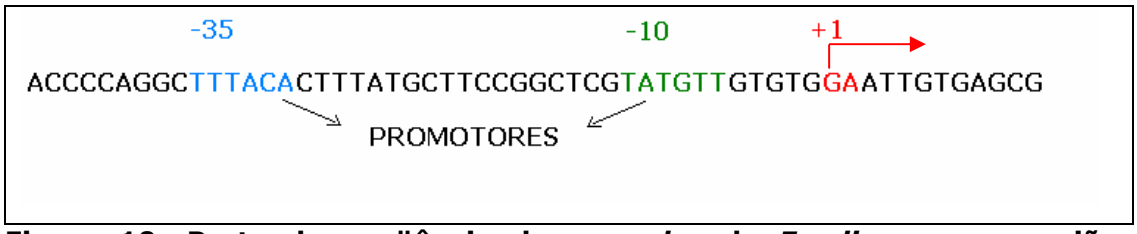


Figura 10: Parte da seqüência do gene *lac* de *E.coli* e a sua região promotora, o elemento -10 (*Pribnow-box*) e o elemento -35. Também é indicada a posição +1, onde se dá o início da transcrição.

Muitos tipos de evidências experimentais suportam a importância funcional dos elementos promotores -10 e -35. Primeiro, genes com promotores que diferem da seqüência consenso são transcritos menos eficientemente do que os mais parecidos. Segundo, mutações introduzidas nestas regiões têm enorme efeito sobre a função promotora. Terceiro, os locais nos quais a RNA polimerase se liga ao promotor tem sido diretamente identificada por experimentos laboratoriais "footprinting", os quais são largamente usados para determinar os locais nos quais as proteínas se ligam ao DNA.

Tais experimentos têm mostrado que a RNA polimerase geralmente se liga às regiões promotoras por aproximadamente 60 pares de base, estendendo-se de -40 a + 20 (isto é de 40 nucleotídeos "upstream" até 20 nucleotídeos "downstream" do local de início da transcrição) (COOPER, 2000).

2.13.3 Operon

Segundo o modelo operon de Jacob-Monod (JACOB e MONOD, 1961), operon é um conjunto de genes contíguos juntamente com os seus elementos regulatórios, ou seja, é um grupo de genes contíguos e transcritos a partir de um único mRNA.

É muito comum, principalmente nas bactérias, a presença de genes diferentes porém, com atuação nos mesmos processos metabólicos, situados de forma contígua. Esses genes, ficam situados num mesmo local do DNA circular bacteriano e compõem um operon. Eles são transcritos juntos pela mesma RNA polimerase. Isso resulta numa resposta mais rápida da bactéria a alguma mudança do meio. Em bactérias os genes com funções relacionadas, são, geralmente, localizados adjacentemente e eles são regulados coordenadamente. Num operon há genes que codificam para proteína e há genes que são reguladores da expressão do operon.

De uma maneira bem geral, pode-se dizer que em operons há dois mecanismos básicos:

Mecanismos de transcrição de operons que são ativados por causa da presença de alguma substância indutora (são comuns em vias metabólicas que resultam no catabolismo desta substância).

Mecanismos de transcrição de operon cujo indutor normalmente é o substrato para a via (MATHEWS *et al.*, 1999).

2.13.4 Terminadores transcricionais

A DNA polimerase continua a transcrição até que encontre um sinal terminador, neste ponto a transcrição cessa, o mRNA é liberado da polimerase, e a enzima se dissocia da fita molde (*template*) de DNA.

O terminador Rho-independente é o modo mais simples e comum de sinal terminador na *E.coli.* Ele consiste de uma região de repetição dos nucleotídeos G e C (diz-se região rica em GC) simetricamente invertidas (palíndromes) podendo haver alguns nucleotídeos A e T no meio desta série de G e C, seguida de quatro ou mais nucleotídeos As (diz-se cauda poli-U). A transcrição de Gs e Cs seguidas e palindrômicas, resultam na formação de um segmento com a estrutura em forma de grampo (chamado de "stemloop" ou "harpin") no mRNA por causa do pareamento das bases. A formação desta estrutura interrompe o caminho para a polimerase continuar a transcrição. Por causa da fraca ligação da ponte de hidrogênio entre A e U, a presença de bases A à jusante (*downstream*) da seqüência

repetida (ligação entre G e C é mais forte) e a seqüência invertida de Gs e Cs parecem ajudar na dissociação do mRNA de seu molde. Outros tipos de sinais de terminação da transcrição podem ocorrer, dependendo da ligação das proteínas que terminam a transcrição para determinadas seqüências de DNA (COOPER, 2000; LODISH *et al.*, 2000). Na Figura 11 é mostrado o esquema gráfico de um terminador Rho-independente.

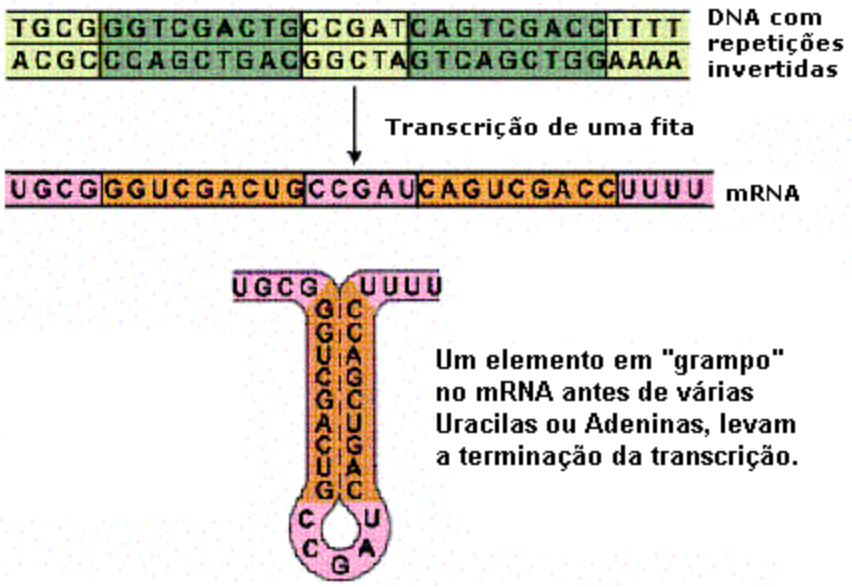


Figura 11: Esquema do teminador Rho-independente, chamado de grampo ("stemloop" ou "harpin") com a cauda poli(U).

Terminador Rho-dependente - O fator Rho, é uma proteína hexamérica, que se liga ao mRNA que está sendo transcrito (cobre de 70 a 80 bases do mRNA). Esta interação ativa a função ATPase do fator Rho que está associada com seu movimento ao longo do RNA na direção 3' até que ele desligue a hibridização do mRNA no DNA no local onde a polimerase está. O fator Rho analisado em *E.coli* não mostrou nenhuma seqüência de consenso e recentes análises no genoma da *E.coli* indicam que este fator regula, relativamente, poucos operons (COOPER, 2000; LODISH *et al.*, 2000).

2.13.5 O efeito da expressão gênica nas vias metabólicas

Em uma única via metabólica pode haver várias interações químicas e estas interações, na sua maioria, precisam de uma enzima para viabilizá-las. Por isto, para a via metabólica ser processada será necessária a expressão de vários genes. Muitas vezes, são necessários vários genes para expressar uma enzima, outras vezes uma reação necessita de várias enzimas (complexo enzimático) para catalisá-la, e outras vezes um mesmo gene pode expressar enzimas que possuem dupla funcionalidade, chamadas de enzimas bifuncionais.

Na Figura 12 é apresentado um exemplo de um gráfico representando a via metabólica da biossíntese de violaceína e seus genes conhecidos atribuídos às reações. As reações são simbolizadas por setas.

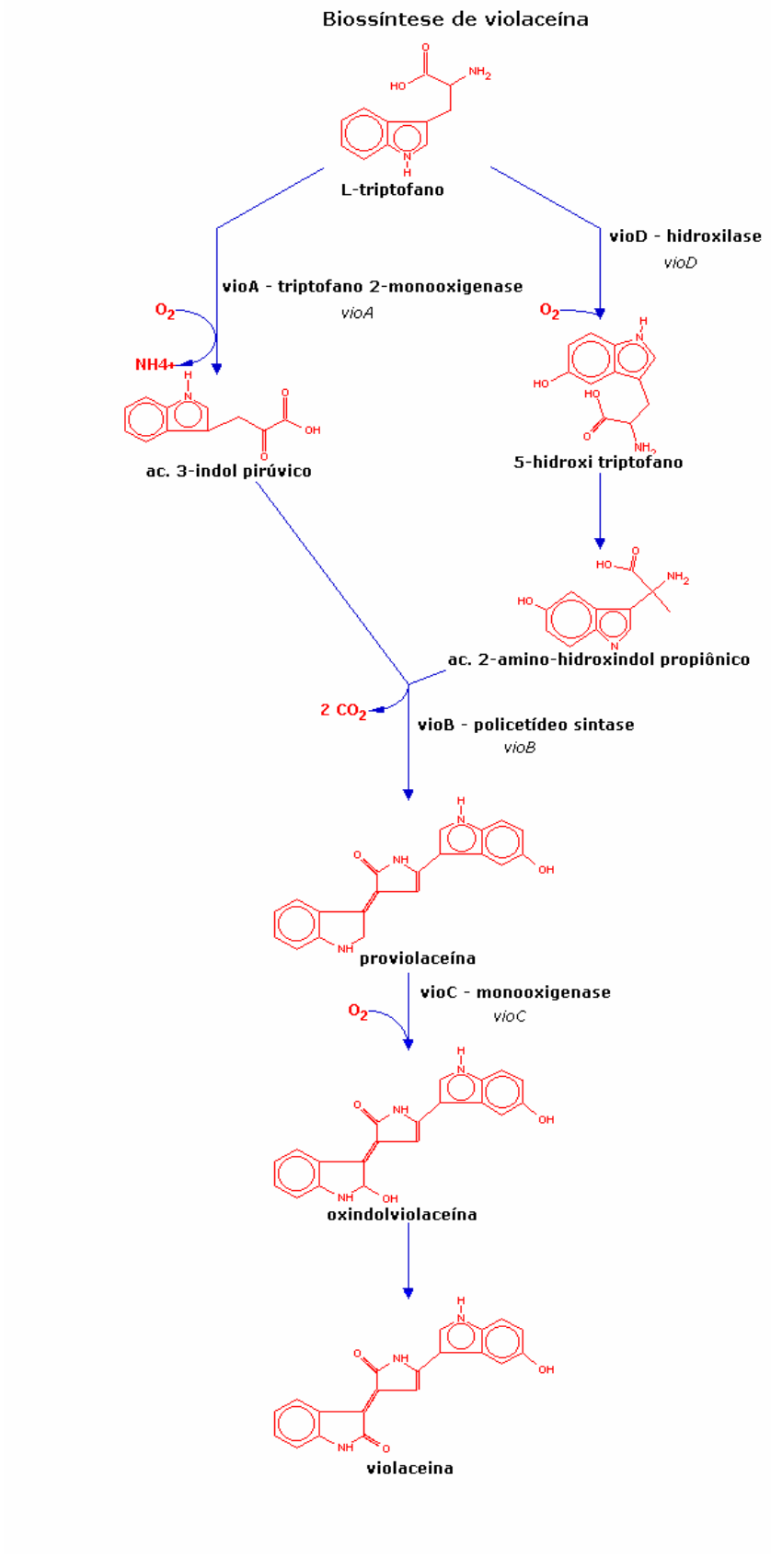


Figura 12: Gráfico da via metabólica da biossíntese de violaceína em *C violaceum*, segundo August *et al.*,2000.

2.14 Bases de dados biológicos

Para biologia molecular, os tipos de dados encontrados em pesquisas são as seqüências de ácidos nucléicos, as seqüências de proteínas, e as estruturas 3D moleculares; por causa do enorme volume de dados que essas informações ocupam, esses resultados não são todos relatados em artigos científicos, mas sim, colocados em base de dados e apenas um sumário destes dados é escrito em artigos para publicação científica (KANEHISA, 2000).

A comunidade científica tem gerado, nas últimas décadas, uma enorme quantidade de dados e criado variadas bases de dados para suportá-los. Muitas destas bases de dados fornecem informações muito úteis sobre diversos tipos de elementos moleculares, tais como ácidos nucléicos, proteínas, locais de conexão no DNA ("binding sites"), motivos ("motifs") em proteínas, domínios em enzimas, e muitos outros.

Algumas destas bases de dados, categorizadas com base em WITTIG e DE BEUCKELAER (2001), serão discutidas a seguir.

2.14.1 Bases de dados de genomas

No fim dos anos 70 ficou aparente que a tecnologia para o seqüenciamento de DNA transformaria a biologia com uma inundação de dados de seqüências. Esforços mundiais para estabelecer um banco de dados de DNA foram iniciados em torno de 1979 tanto na Europa quanto nos Estados Unidos. Em conseqüência a isto, a base de dados GenBank do laboratório LANL ("Los Alamos National Laboratory") e a base de dados EMBL ("European Molecular Biology Laboratory") , do EBI ("European Bioinformatics Institute") localizado em Hinxton, Reino Unido, têm estado em colaboração desde então, e a base de dados DDBJ ("DNA Data Bank of Japan") do Japão uniu-se aos anteriores em 1984. Em 1992 o GenBank inteiro foi assumido pelo NCBI ("National Center for Biotechnology Information") e em 1994 a operação do EMBL foi também movido para o EBI. GenBank, EMBL e DDBJ formam o "International Nucleotide Sequence Database Collaboration". Elas recebem dados diretamente dos autores para

serem submetidos e trocam dados diariamente, de maneira a ficarem todas atualizadas com os mesmos dados (GIBAS E JAMBECK, 2001).

O NCBI fornece acesso gratuito, através do seu sítio na Internet (http://www.ncbi.nlm.nih.gov/), a dados e publicações biomédicas e biológicas, incluindo o PubMed, uma base de dados de mais de 15 milhões de resumos biomédicos.

2.14.2 Bases de dados de proteínas

Conforme GIBAS e JANBECK (2001) ao redor de 1980 foi criada nos Estados Unidos, uma base computadorizada a partir das coleções de seqüências de aminoácidos de Margaret Dayhoff da NBRF ("National Biomedical Research Foundation") em Washington, DC, publicada em "Atlas of protein sequences and structures" (de 1968 – 1978). Esta base de dados foi chamada de base de dados de seqüência de proteínas NBRF ("National Biomedical Research Foundation"). Esta base evoluiu para PIR ("Protein Information Resource") e foi estabelecida em 1984 com suporte do NIH ("National Institutes of Health"). Desde 1988 PIR passou a ser "PIR-International Protein Sequence Database**"** e está sendo mantida em colaboração pelo MIPS ("Munich Information Center for Protein Sequences") na Alemanha e por JIPID ("Japanese International Protein Sequence Database") no Japão.

Outra importante base de dados de proteínas, SWISS-PROT (BAIROCH e APWEILER, 2000), foi criada em 1986 na Universidade de Geneva, é uma outra base de dados de proteína, e logo se tornou a melhor em termos de qualidade dos dados, afirmam GIBAS e JAMBECK (2001) em seu trabalho. SWISS-PROT tem colaboração, desde 1987, do EMBL e a base de dados TREMBL ("Translation do EMBL" é a tradução da base de seqüência de nucleotídeos") tem sido usada para suplementar a SWISS-PROT. A base SWISS-PROT é operada atualmente por SIB ("Swiss Institute of Bioinformatics") e pelo instituto EBI do EMBL (GIBAS e JAMBECK, 2001).

Outra base de dados de Proteína, o PDB ("Protein Data Bank") (BERMAN *et al.*, 2000) foi estabelecido em 1971 no laboratório BNL

("Bookhaven National Laboratory"). Em 1999 o PDB foi movido pra o RCSB ("Research Collaboratory for Structural Bioinformatics").

A base de dados PDB foi a primeira voltada para a bioinformática e foi desenhada para armazenar dados complexos de estruturas moleculares. Revistas científicas que publicam resultados cristalográficos agora requerem o envio a base PDB como condição para publicação, o que significa, praticamente, que os dados de estrutura de proteínas obtidas por pesquisadores são disponibilizados na base PDB em um prazo bastante razoável. Há, até a data deste trabalho, um total de 14.682 enzimas na base PDB (GIBAS e JAMBECK, 2001).

A base de dados COG ("Clusters of Orthologous Groups") é uma tentativa de classificação filogenética de proteínas, um esquema que indica as relações evolucionárias entre organismos. Cada COG é um grupo de 3 ou mais proteínas originadas de genes ortólogos. O COG pode ser usado para: identificar similaridade e diferenças entre espécies, para identificar famílias de proteínas e na predição de novas proteínas (NCBI, 2003).

2.14.3 Bases de dados de enzimas

Algumas Bases de dados mantêm informações principalmente a respeito de enzimas, a base BRENDA (SCHOMBURG *et al.*, 2004*)* mantido pelo "Institute of Biochemistry at the University", é uma destas bases. Esta base de dados mantém uma coleção de dados funcionais de enzimas disponíveis para a comunidade científica. As enzimas estão classificadas de acordo com a lista de enzimas mantidas e publicadas pela "Enzyme Commission*"*. A base de dados BRENDA tem um total de 3.500 enzimas diferentes até a data deste trabalho. (www.brenda.uni-koeln.de)

Outra base de dados de enzimas a **ExPASy-Enzyme** ("Expert Protein Analysis System-Enzyme") (BAIROCH *et al.*, 2000) é uma base de dados de informações relativas a nomes de enzimas. É baseado no IUBMB e contém vários tipos de dados relativos a enzimas caracterizadas para as quais existem EC_numbers atribuídos, como por exemplo, nome recomendado, nomes alternativos, atividade catalítica e outros (BAIROCH *et al.,* 2000).

2.14.4 Bases de dados Via/Genoma

As bases de dados Via/Genoma armazenam informações relativas ao genoma, às vias metabólicas, redes regutatórias, reagentes, e as relações entre os genes, enzimas e reações.

Muitas ferramentas e bases de dados para via/genoma têm sido desenvolvidas principalmente para a análise de dados de "microarrays". Em aplicações web públicas, contudo, as vias complexas têm sido mostradas como arquivos com imagens estáticas que podem não estar atualizadas ou mesmo a reconstrução das mesmas ser muito lenta. Ainda, as análises de expressão de genes, muitas vezes, focam em "probes" (marcações) individuais ou em genes de pouca importância nas vias. Estas abordagens revelam poucas informações sobre as vias que são chaves para um amplo entendimento dos blocos de construção dos sistemas biológicos. Portanto, há necessidade de ferramentas úteis que possam gerar vias sem a construção de imagens manuais e que permitam que os dados de expressão de genes sejam integrados e analisados em nível de vias para os organismos a serem estudados (PAN *et al.*, 2003).

Segundo DEVOS e VALENCIA (2000), quando outras fontes de informação são usadas na anotação genômica, tal como, contexto genômico, vias metabólicas, análise de famílias de proteínas, informação evolucionária ou dados experimentais, os erros de anotação podem ser menos freqüentes. Eles afirmam também que apesar de as análises dos genomas completos terem influência positiva no desenvolvimento da biologia e biomedicina, é importante saber que possíveis erros nas anotações de funções de genes podem ocorrer pela prática de predições padronizadas durante a primeira etapa da análise genômica.

O KEGG é uma base de dados de conhecimento que se encaixa nesta divisão pois integra o conhecimento corrente sobre as redes de interações moleculares tais como as vias e os complexos (base de dados Pathway), informações sobre genes e proteínas geradas pelo projeto genoma (base de dados GENES/SSDB/KO) e informações sobre compostos e reações bioquímicas (base de dados COMPOUND/GLYCAN/REACTION).

A base de dados EMP "Enzyme and Metabolic Pathways" (SELKOV *et al.*, 1998) é uma base de dados que contém informações bioquímicas, e cobre vários aspectos da enzimologia e do metabolismo, ela tem 3000 diagramas metabólicos até 14 de agosto de 2005 (http://www.empproject.com/).

A base de dados EcoCyc (KARP *et al*., 2002; KESELER *et al.,* 2005) é uma base de dados de via/genoma que contém vias metabólicas e vias de sinais-transducionais da *E. coli K-12*, suas enzimas, suas proteínas transportadoras e seus mecanismos de expressão gênica e controles transcricionais. A base de dados MetaCyc (KARP *et al*., 2002) é baseada na mesma ontologia da EcoCyc, possuindo vias de centenas de diferentes espécies. Elas contêm várias citações da literatura sobre enzimas e reações. Estas bases de dados podem ser acessadas através do software Pathway Tools (KARP *et al*., 2002), o qual possui capacidade de consulta, edição, e visualização destas bases.

O sistema BioCyc (KRUMMENACKER *et al*., 2005), é uma coleção de bases de via/genoma vindas de uma grande variedade de organismos, principalmente microrganismos e plantas. O objetivo desta base de dados é conter uma amostra representativa das vias que foram elucidadas experimentalmente.

Na base de dados Metacyc (KARP *et al*, 2002) há informações de vias metabólicas e dados biológicos pertencentes a mais de 300 organismos; sendo que, 90% de suas vias são manualmente curadas com citação da literatura. Os outros 10% das vias, as quais foram originalmente importadas da base de dados WIT (OVERBEEK *et al.,*2000), estão sob cura manual, até a data deste trabalho.

Na Tabela 2 são apresentados os endereços eletrônicos dos sítios da bases de dados citadas neste item.

Tabela 2: Endereços eletrônicos de algumas bases de dados públicos existentes na Internet.

GenBank	http://www.ncbi.nlm.nih.gov
EMBL	http://www.ebi.ac.uk/embl/
DDBJ	http://www.genome.jp/
PIR	http://www-nbrf.georgetown.edu**/**
SWISS-PROT	http://au.expasy.org/
PDB	http://www.rcsb.org/pdb/
COG	http://www.ncbi.nlm.nih.gov/COG/
ExPASy	http://au.expasy.org/enzyme/
BRENDA	http://www.brenda.uni-koeln.de/
KEGG	http://www.genome.jp/kegg/
EMP	http://www.empproject.com/
EcoCyc **BioCyc** **Metacyc**	http://ecocyc.org/ http://biocyc.org http://metacyc.org/

2.15 Programas de bioinformática

A comparação de seqüências de DNA e proteínas é uma das funções básicas da bioinformática. Os programas de bioinformática procuram ter capacidade cada vez maior de executar com precisão, as comparações automatizadas de seqüências, de predição e de construir os modelos estruturais de proteínas, e de compor o projeto e a análise de experiências com expressões gênicas (GIBAS e JAMBECK, 2002).

Teoricamente as proteínas que compartilham uma seqüência similar, geralmente compartilham a mesma estrutura básica. Portanto, por determinação experimental da estrutura da proteína para um membro de uma família de proteínas, pesquisadores podem ter um modelo no qual baseiam a estrutura das outras proteínas da mesma família (NCBI, 2004).

Uma numerosa quantidade de ferramentas, de domínio público, tem sido desenvolvidas para ajudar os pesquisadores a usar as bases de dados de seqüências genômicas disponíveis na WEB. Estas ferramentas diferem-se não só nas técnicas que são usadas na comparação de seqüências, quanto no tamanho da unidade de análise, na qual pode ser o genoma inteiro, um domínio, uma seqüência ou um motivo ("motif") (TURCHIN e KOHANE, 2002).

2.15.1 Programas para alinhamento de seqüências genômicas

Os programas para alinhamento de seqüências foram as primeiras ferramentas disponibilizadas para os biólogos procurar homologias. Este tipo de ferramenta, basicamente, recebe uma consulta ("query") e busca em base de dados de seqüências conhecidas, por exemplo GenBank, encontrar possíveis homólogos.

Dentre muitos programas para busca de alinhamentos entre seqüências, alguns são apresentados como exemplos:

BLASTP (ALTSCHUL *et al*., 1990) – É usado para comparar uma nova seqüência com aquelas contidas em bancos de dados de nucleotídeos ou proteínas. Esta comparação se dá através do alinhamento da nova seqüência com genes previamente caracterizados nestes bancos. A ênfase desta ferramenta é encontrar regiões de similaridade na seqüência de interesse, a qual produzirá indícios da estrutura e função desta nova seqüência. Regiões de similaridade detectadas através deste tipo de programa podem ser também locais, onde a região de similaridade é baseada em segmentos parciais da seqüência, ou global, onde as regiões de similaridade são detectadas de um extremo a outro da seqüência. Algumas variáveis de resulta deste programa, disponibilizam informações para uma análise da validade dos alinhamentos:

"Score": Representa o valor da pontuação alcançada para a qualidade do alinhamento, levando em conta uma matriz de substituição;

"Expect" (E) value: Representa o número de alinhamentos que podem ter acontecido por acaso. Este valor quanto menor melhor. É conhecido como Evalue;

"Identities" (I): número de aminoácidos ou nucleotídeos que são idênticos entre as seqüências alinhadas;

"Positives" (P): número de aminoácidos ou nucleotídeos que são idênticos ou equivalentes entre as seqüências alinhadas.

"Gaps" (G): número de espaçamentos que foram incluídos para possibilitar um melhor alinhamento.

ClustalW (STOCSITS *et al*., 2005) é um programa para múltiplos alinhamentos de seqüências de DNA ou de aminoácidos. Ele produz alinhamentos biologicamente significativos entre seqüências divergentes; o programa calcula o melhor "match" (grau de coincidência entre seqüências) para as seqüências selecionadas e as alinha de acordo com as suas identidades, similaridades e desta forma as divergências podem ser observadas. Além disto, relações evolucionárias podem ser mostradas através de cladogramas ou filogramas

FASTA (LIPMAN e PEARSON, 1985) - Faz comparações com proteínas ou nucleotídeos. Este programa alcança um alto nível de sensibilidade para

pesquisa de similaridade e com alta velocidade. O alto nível de sensibilidade é alcançado pela execução otimizada de pesquisas para alinhamentos locais usando uma matriz de substituição. E a alta velocidade é alcançada pelo uso de um modelo de palavra para identificar potenciais "matches" antes da tentativa de otimizar a consulta. O balanço entre a velocidade e a sensibilidade é controlada pelo parâmetro ktup, o qual especifica o tamanho da palavra. O aumento do ktup, diminui o número total de "hits"(número de seqüências alinhadas). Nem todas as palavras encontradas são investigadas, procura sim, inicialmente, por segmentos contendo vários "hits" próximos.

PSI-BLASTP (ALTSCHUL *et al*., 1997) "Position specific iterative" BLASTP. Este programa usa uma busca interativa nas quais seqüências encontradas em uma passagem de busca são usadas para construir um modelo *score* para a próxima passagem de busca. As posições altamente conservadas recebem altos scores, e posições fracamente conservadas recebem scores perto de zero. Um procedimento ("profile") é usado para executar uma segunda pesquisa BLASTP, e os resultados de cada "interação", são usados para redefinir o procedimento. Esta estratégia de busca interativa resulta em um aumento na sensibilidade.

PHI-BLASTP (ZHANG *et al.,* 1998) "Pattern Hit Initiated" BLASTP - Combina a coincidência ("matching*")* de padrões de expressão regular com uma busca interativa a proteínas em uma posição específica. PHI-BLASTP pode localizar outras seqüências de proteínas as quais ambas contém o padrão e são homólogas para uma consulta *(*"query*")* de seqüência de proteína.

Na Tabela 3 estão listados os endereços eletrônicos dos programas citados neste íntem.

Tabela 3: Endereços eletrônicos dos programas para análise de seqüência, disponíveis publicamente.

BLAST	http://www.ncbi.nlm.nih.gov/blast/ http://www.ebi.ac.uk/blast/
ClustalW	http://www.ebi.ac.uk/clustalw/
FASTA	http://www.ebi.ac.uk/fasta/ http://fasta.genome.jp/
PSI-BLAST	http://www.ncbi.nlm.nih.gov/blast/
PHI-BLAST	http://www.ncbi.nlm.nih.gov/blast/

2.17 A *Chromobacterium violaceum*

Chromobacterium violaceum pertence ao grupo das bactérias Gram-negativas, facultativa anaeróbica e sua temperatura de crescimento pode variar de 15-40°C, sendo 30°C a temperatura ideal para seu cultivo, favorável em regiões tropicais. Elas são consideradas como flora comum de solos e águas tropicais e subtropicais onde podem ter um papel importante na rizosfera. Estas bactérias não estão presentes como parte normal da flora animal ou humana. Este organismo é patogênico oportunista e ocasionalmente causa sérias infecções em mamíferos, inclusive no homem(HUNGRIA *et al*., 2004). A primeira ocorrência de contaminação em humanos foi relatada em 1927, e tem ocorrido em vários continentes, particularmente na Austrália, América do Sul, e Sudeste da Ásia. Tem grande adaptabilidade ao meio em que vive e pode sobreviver sob condições ambientais de estresses (HUNGRIA *et al*., 2004).

A *C. violaceum* tem um grande potencial biotecnológico; sua mais notável característica é a produção de um pigmento, quimicamente bem caracterizado, chamado de violaceína (BROMBERG e DURAN, 2001). O pigmento produzido pela maioria das linhagens dá às colônias uma cor metálica violeta escura. Estudos anteriores indicaram atividades da violaceína com efeitos antibióticos e anti-*Trypanosoma cruzi* (MOMEN e HOSHINO, 2000) (ANDRIGHETTI-FROHNER, 2003), anti-tumoral (MELO *et al*., 2000), e anti-leishemanial (LEON *et al.*, 2001), dentre outros. Há evidências de que a violaceína é eficaz contra células humanas de melanoma da íris (SARAIVA *et al.*, 2004). Muitos estudos têm sido feitos no sentido de usar esta substância como uma alternativa para o tratamento quimioterápico contra o câncer, pois há indícios que a violaceína, em conjunto com outras moléculas, minimizam a deterioração das células saudáveis (SARAIVA *et al.*, 2004).

Há também interesse industrial pela *C. violaceum* por produzir poliesteres que podem ser usados para fazer polihidroxialcanoatos (PHAs), os quais são alternativa para os plásticos feitos por petroquímicos (FORSYTH *et al.*, 1958); e tem também dispertado interesse por produzir cianeto, o qual exerce função para a solubilização do ouro, processo livre de

mercúrio, evitando a consequente contaminação do ambiente por este metal pesado (CAMPBELL *et al.*, 2001).

As informações do genoma da *C. violaceum*, linhagem ATCC 12472, feito pelo *Brazilian National Genome Project Consortium*, estão disponibilizadas no NCBI, (http://www.ncbi.nlm.nih.gov) desde setembro de 2003 e espera-se que, através dos estudos genômicos, funções importantes ainda desconhecidas sejam reveladas.

CAPÍTULO III

Métodos e Recursos Computacionais

3.1 Instalação do programa Pathway Tools

Através de uma licença acadêmica, o software Pathway Tools foi instalado num computador pessoal, com processador Intel Pentium IV, de 2,4 GHz, 512 MB de memória RAM, 40 GB de disco rígido e com o Sistema Operacional Linux, distribuição RedHat 9, previamente instalado. Esta plataforma satisfaz os requisitos mínimos indicados no Manual do Pathway Tools para o sistema operacional Linux: hardware com um processador de 1 GHz, 512 MB de RAM e um mínimo de 100 MB de disco rígido.

É requisito necessário a instalação prévia do Netscape Navigator Web Browser, de um software HTTP Server (o http server que acompanha a distribuição RedHat 9 foi instalado previamente) e bibliotecas Motif versão 2.1.30 (obtidas do endereço http://www.opengroup.org/openmotif/).

No pacote de instalação do Pathway Tools estão incluídos, além do próprio conjunto de programas, mais dois PGDBs, um *da Escherichia coli* K-12, o EcoCyc, e um outro chamado MetaCyc que é composto de dados de vias metabólicas de 240 organismos diferentes compreendendo mais de 500 vias metabólicas (http://biocyc.org, até a publicação deste trabalho). Depois de instalados os bancos de dados podem tanto ser transferidos para o SGBD Oracle — que é um sistema gerenciador de base de dados (SGDB) proprietário da Oracle Corporation ao qual o Pathway Tools possui acesso — quanto ter seus arquivos instalados em formato ASCII (arquivos de texto ou "flatfiles") e, neste último caso irá utilizar o sistema gerenciador de base de dados objeto Ocelot (Ocelot Computer Services Inc., 2002), que é uma distribuição gratuita e já vem integrado ao Pathway Tools (disponível em http://www.ocelot.ca/download.htm). Esta opção foi a escolhida para evitar o uso de software proprietário, plataforma computacional de maior poder de processamento, e de maior complexidade gerencial.

3.2 Pathway Tools

O Pathway Tools é um software desenvolvido para identificar, curar, armazenar, analisar e publicar vias bioquímicas metabólicas de genomas. Serve, também, para criar base de dados para vias metabólicas e para informações genômicas (PGDBs), também chamadas de bases de dados Via/Genoma. Este tipo de base de dados integra dados genômicos com as anotações detalhadas do genoma, tais como descrições dos metabólitos e vias de sinalização, que poderão ser incluídas a qualquer momento na base de dados.

Ele foi inicialmente criado, por volta dos anos 90, para o Projeto da EcoCyc, porém em 1997 iniciou-se um processo de generalizar o software para ser aplicado em outros organismos, inclusive para permitir comparações simultâneas no estudo genômico.

O Pathway Tools foi desenvolvido pelo grupo "Bioinformatics Research Group" do "SRI International", foi licenciado para mais de 600 grupos fora do SRI e que criou 165 PGDBs, disponíveis em: http://biocyc.org (acessado em: 10 de maio 2005).

É possível exportar informações dos PGDBs para arquivos texto ("flat") ou arquivos SBML ("Systems Biology Markup Language") e também importar informações para dentro de um PGDB através destes mesmos tipos de arquivos.

A Figura 13 é mostrada a tela inicial do software, onde é possível escolher qual dos organismos, criados localmente, se quer trabalhar. Não há restrição no número de organismos que se pode criar; isto depende apenas dos recursos computacionais que se tem, principalmente espaço em disco rígido.

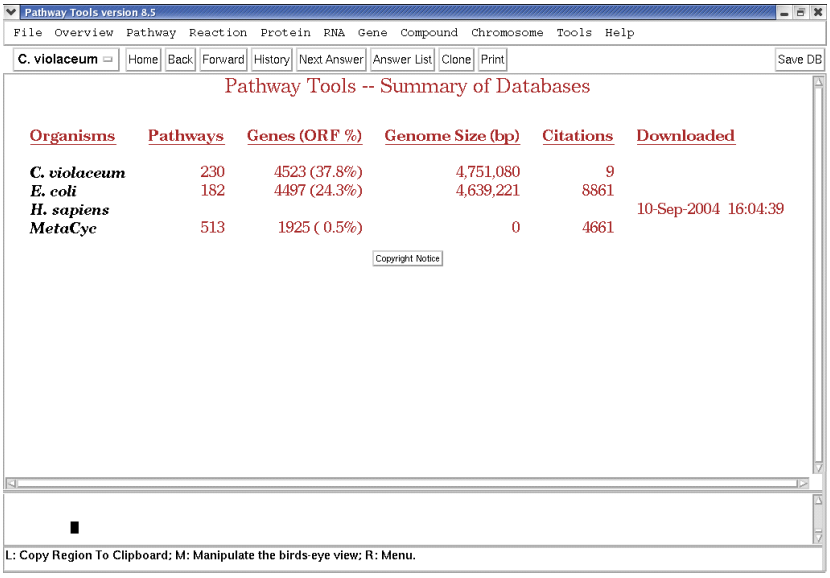


Figura 13: Tela inicial do Pathway Tools que mostra um sumário dos PGDBs que foram criados localmente.

3.2.1 Módulos do Pathway Tools

O Pathway Tools é composto de quatro componentes principais: o PathoLogic, o Pathway/Genome Navigator, o Pathway/Genome Editor, e Pathway Tools Ontology. Serão descritos a seguir, de acordo com o manual do software (PANGEASYSTEMS, 2002), cada um destes módulos, suas funções, e como foram utilizadas para a criação e análise do CvioCyc.

1.**PathoLogic**: É o módulo que cria as novas bases de via/genoma de organismos em geral, os PGDBs, a partir do genoma anotado de um organismo sequenciado. Também executa, através do programa Predictor, (PALEY *et al.,2002)* as predições de vias metabólicas, realizando comparações das vias-metabólicas existentes nas bases de dados EcoCyc e MetaCyc, e também realiza a predição de operons. O PathoLogic avalia as evidências da existência de cada uma das vias metabólicas no organismo que está sendo criado, que estão contidas nas bases de dados EcoCyc e MetaCyc, e as incorpora ao PGDB em criação.

Após a criação do PGDB, o módulo PathoLogic gera relatórios um informando o que o PathoLogic encontrou no arquivo de anotação, outro

sobre as coincidências (“matchings”) das enzimas do organismo de interesse em relação ao da base que o PathoLogic usa para pesquisa (`pangea-enzyme-mappings.dat`), e em relação a um arquivo criado pelo usuário (`local-enzyme-mappings.dat`), e também em relação aos nomes de enzimas contidos no MetaCyc.

Quando um nome de função enzimática é comparado na base de dados, há duas possibilidades:

a) “Matching” não ambíguo: neste caso uma conexão reação-enzima é criada automaticamente, ou

b)“Matching” ambíguo: neste caso não são criadas conexões automáticas e sim deixadas para que o usuário as inclua posteriormente. Isto significa que mais de uma reação bioquímica foi identificada usando este nome de enzima vindo de diferentes fontes, além daquelas que estão na base de dados do PathoLogic.

Quando um nome de função não é encontrado, o usuário deve tentar fazer pesquisas por partes do nome (“substrings”) para encontrar potenciais correspondências que não são encontradas automaticamente. Onde as reações já estão associadas com o produto de genes através de seu número EC (Enzyme Commission), o “matching” por nome não altera essas correspondências (elas têm precedência). Porém, quando o número EC e nome de função são informados, o relatório também vai informar se há conflitos ou não.

A criação de um novo PGDB necessita de algumas etapas iniciais:

- a criação de um arquivo texto chamado `genetic-elements.dat`, que conterá as propriedades do elemento genético (cromossomo ou plasmídeo) referentes ao organismo em questão;
- um arquivo texto no formato FASTA[1], contendo a seqüência completa de nucleotídeos. (Este arquivo deverá ter sufixo `.fsa` ou `.fna`);

[1]FASTA – É um formato de arquivo texto que inicia com uma linha de descrição da seqüência, seguida de linhas com os dados da seqüência propriamente dita. (Disponível em http://www.ncbi.nlm.nih.gov/BLAST/fasta.shtml, acessado em 18 maio 2005).

- um arquivo de anotação descrevendo os genes preditos. Este deve estar no formato Genbank (`.gbk`) ou PathoLogic (`.pf`).

Apenas as vias metabólicas de pequenas moléculas são encontradas pelo Predictor. O Predictor não infere as vias envolvidas no metabolismo macromolecular, nem nas vias de transporte, e nem nas vias de transdução de sinais (ativadas por moléculas sinalizadoras).

Na Figura 14 é mostrada a tela inicial do módulo PathoLogic do Pathway Tools. É a partir desta tela que são executadas as diversas funções descritas anteriormente.

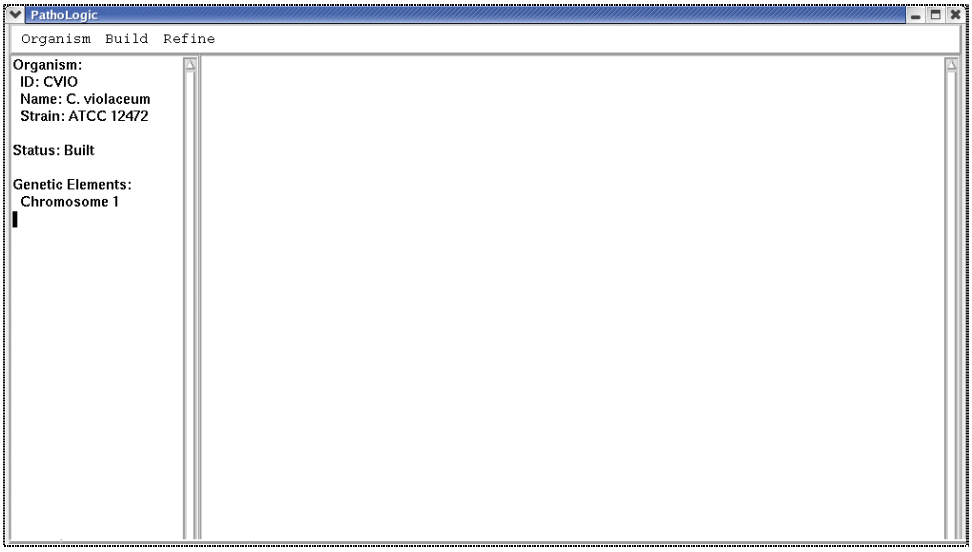


Figura 14: Tela inicial do módulo PathoLogic. Fonte: Software Pathway Tools.

2.**Pathway/Genome Navigator**: É o módulo que faz o processamento de consultas, visualizações e serviços de "Web-publishing" (publicação na Internet) para estes PGDBs.

3.**Pathway/Genome Editors**: Os editores Pathway/Genome estão associados ao processo de suporte interativo com o usuário e as atualizações nos PGDBs. Possuem ferramentas interativas gráficas para todo o tipo de dado manipulado pelo Pathway Tools, incluindo cromossomos, genes, proteínas, reações bioquímicas e vias metabólicas, metabólitos (compostos de pequenas moléculas), operons, íntrons, éxons e "splicings", além de visualizações especiais para enzimas, transportadores, e para fatores de transcrição (mostrando todos os operons controlados pelo

fator de transcrição). Dois exemplos da interface de Edição são mostrados abaixo nas figuras Figura 15 e Figura 16.

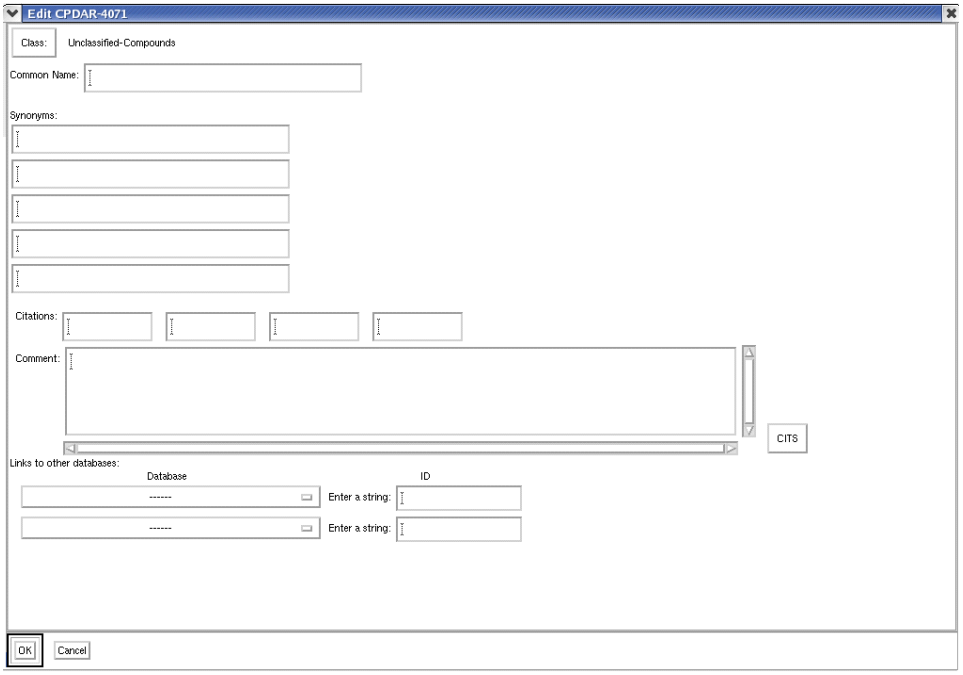


Figura 15: Tela do Pathway Tools para a criação de um composto.

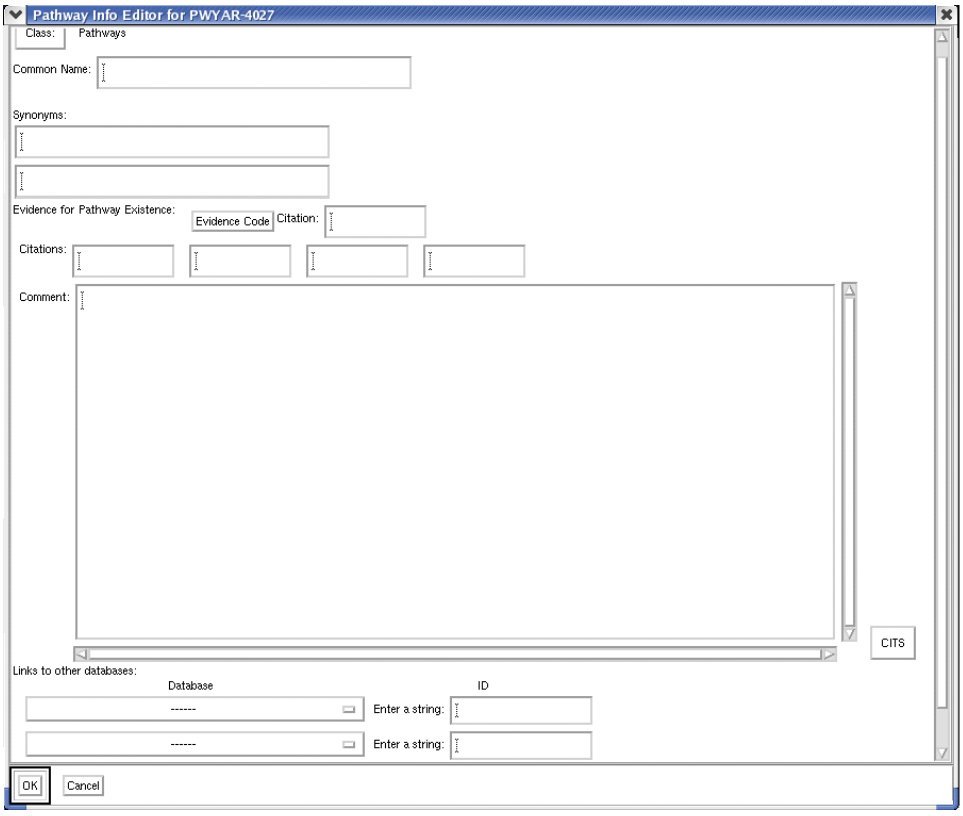


Figura 16: Tela do PathwayTools para a criação de via metabólica.

Outra importante função deste módulo é a apresentação do diagrama "Overview Map" que disponibiliza uma visão geral do metabolismo completo do organismo. Pode ser visto um exemplo deste diagrama completo na Figura 34, mostrada no capítulo IV, ítem 4.2.1. Na Figura 17 é mostrada uma pequena parte do "Overview Map" e uma aproximação do mapa mostra os significados dos diferentes símbolos usados.

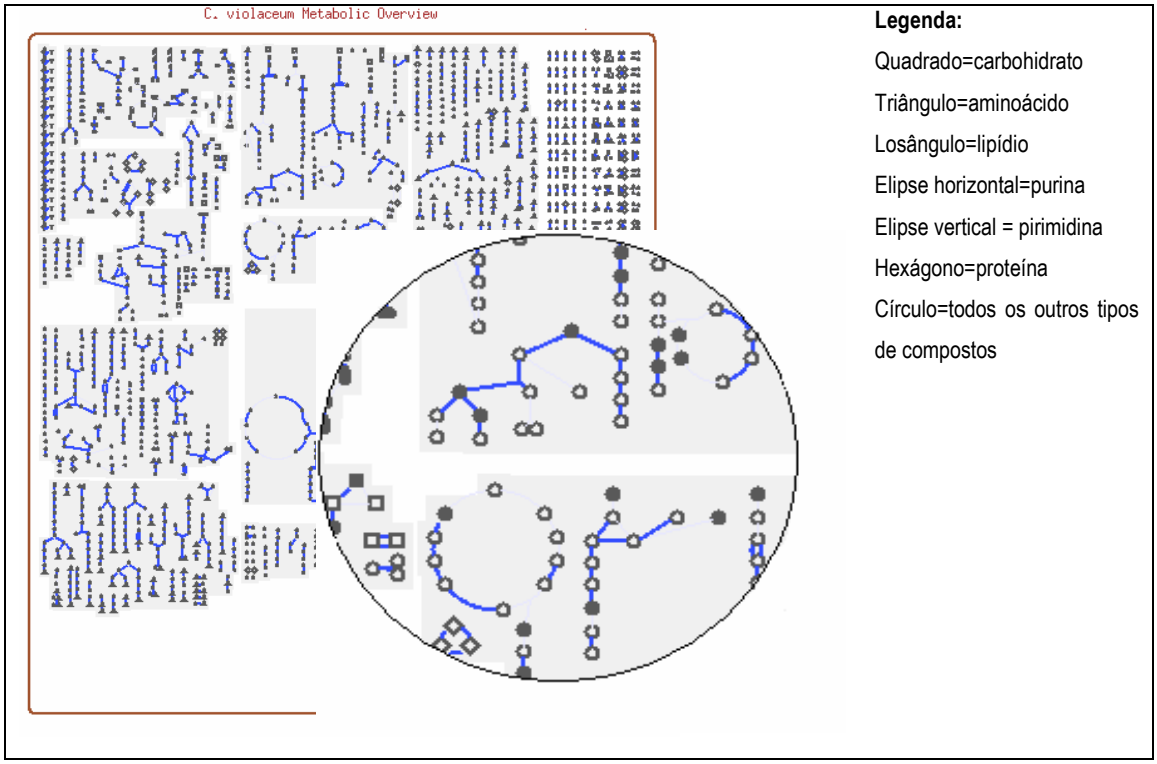


Figura 17: Aproximação do "Overview Map" mostrando os símbolos usados para os metabólitos. Adaptada do software Pathway Tools. O preenchimento em preto dos símbolos indica que os metabólitos estão fosforilados.

Usando o "Overview Map" é possível fazer várias pesquisas comparativas importantes, algumas delas serão descritas utilizando o comando "Highlight" a seguir.

Através do comando "Highlight" é possível visualizar em destaque qualquer entidade do "Overview Map". As pesquisas utilizando o comando podem ser feitas por:

- Compostos: Usando nome do composto, parte dos nomes, estrutura SMILES (é uma linguagem para escrever estruturas químicas usando caracteres ASCII) ou classe do composto;
- Reações: Usando o nome da enzima, parte do nome da enzima, o EC_number, apenas a classe do EC_number, todas as classes no maior nível do EC (colorindo todo o diagrama para mostrar o tipo de cada reação química), todas as reações em que faltam os EC_numbers, vias metabólicas, substratos, efeitos dos compostos sobre as enzimas (esta opção permite destacar reações de acordo com a modulação da(s) enzima(s) que catalisam a(s) reação(ões), isto é, destacar todas as reações nas quais a enzima é ativada pelo ADP), localização celular da enzima, todas as reações com múltiplas isoenzimas e todas as reações que ocorrem em várias vias.
- Vias Metabólicas. Usando o nome da via, parte do nome, sua classe, todas as classes (colore todas as vias de acordo com sua função, por exemplo todas as vias de biossíntese de aminoácidos), agrupamento genômico (colore vias de acordo com o agrupamento dos genes que codificam enzimas desta via, uma janela "pop-up" descreve o esquema de cores).
- Genes: Usando o nome do gene, parte do nome, sua classe, a proteína reguladora do gene (permite que se selecione um fator de transcrição, desta maneira serão destacadas todas as reações as quais os genes estão em operons e são reguladas pelo fator de transcrição informado), por uma lista de nomes de genes vinda de um arquivo, todos os genes por Replicon (reações em cores de acordo com o replicon, cromossomo ou plastídeo, nos quais os seus genes estão localizados).

As opções usadas para visualizar as características desejadas podem ser salvas em arquivos para futura reutilização. Outra função disponível no "Overview Map" é a comparação com outros organismos, onde as entidades destacadas podem ser também pesquisadas em outro organismo criado localmente.

Além disso, operações adicionais em relação às entidades destacadas estão disponíveveis através do botão-direito do mouse. Por exemplo, o

botão-direito sobre um composto possibilita o acesso tanto ao menu de compostos quanto ao menu da via metabólica,.ou ainda, o botão-direito sobre uma reação possibilita o acesso ao menu de reações, de via ou dos compostos envolvidos na reação. Esta tela pode ser vista na Figura 18.

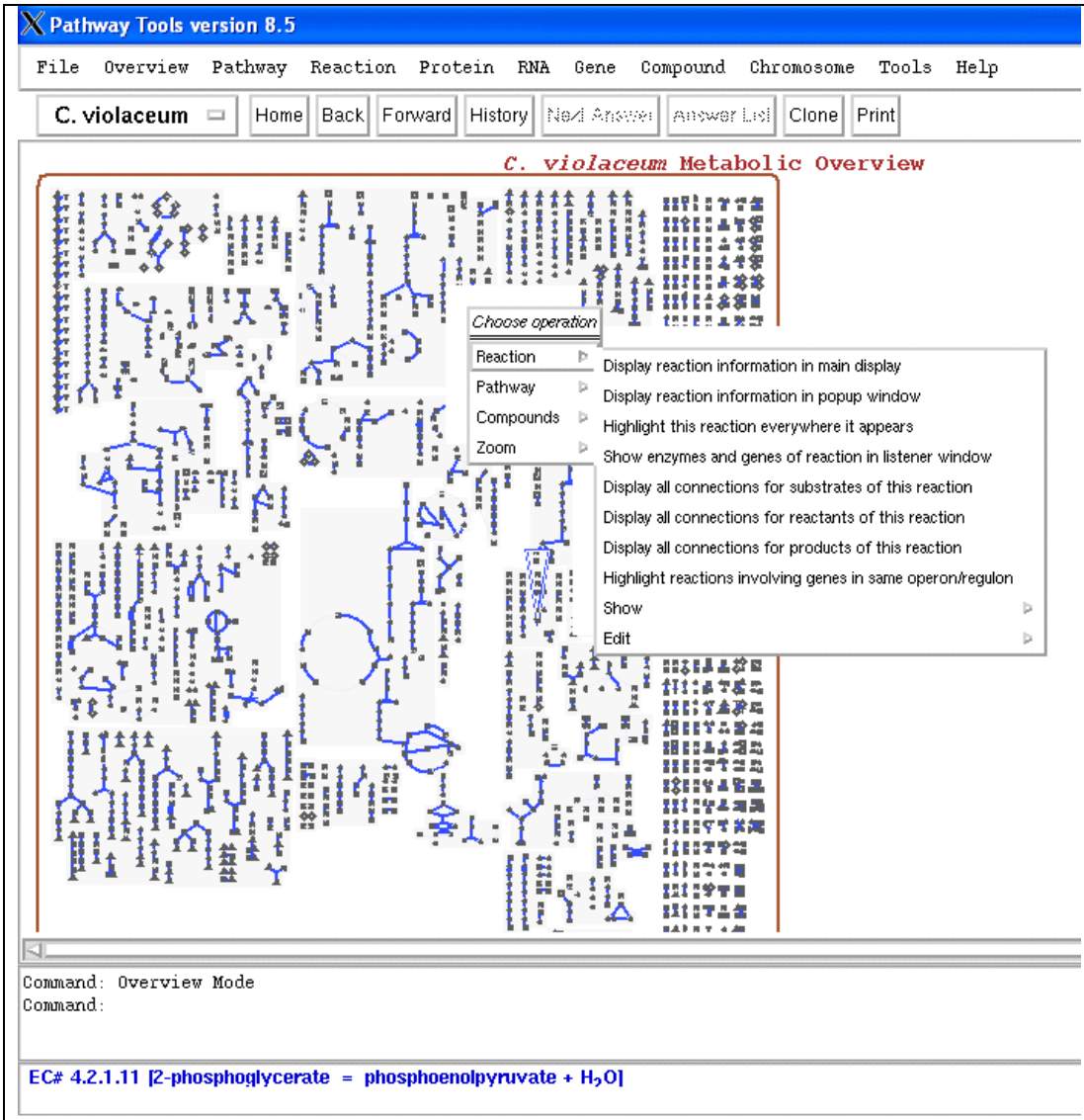


Figura 18: As janelas com menus sobrepondo o "Overview Map" servem para pesquisas a partir das entidades deste diagrama. Este exemplo mostra as opções para pesquisa a partir de uma reação de uma via qualquer. Fonte: Software Pathway Tools.

Uma visualização bastante interessante, também da mesma forma que a anterior, isto é, através de um diagrama geral, permite ao usuário variar os níveis de expressão de genes para enzimas metabólicas,

sobrepondo o diagrama com cores de acordo com maior ou menor nível de expressão genética. Isto pode ser usado para:

- Mostrar como os níveis de expressão genética variam durante o curso de um experimento, comparando um ponto anterior no tempo com um ponto num tempo mais recente;
- Comparar níveis de expressão entre múltiplos conjuntos de condições experimentais de crescimento, por exemplo um meio de cultura rico e um meio mínimo;
- Comparar níveis de expressão entre duas linhagens de um mesmo organismo;
- Comparar estados patológicos em relação ao estados de não doença (para organismos superiores);
- Comparar um tipo de tecido com outro (para organismos superiores).

Dados de expressão podem ser importados diretamente da base de dados de "Stanford Microarray Database" (SMD), ou de Planilhas SAM (Statistical Analysis of Microarrays), ou pode-se criar um arquivo texto próprio contendo expressões de genes.

4.**Pathway Tools Ontology**: É o processo que define os esquemas dos PGDBs, usa o sistema gerenciador de banco de dados Ocelot para os serviços de gerenciamento dos dados nos PGDBs. Um PGDB descreve o genoma de um organismo (seus cromossomo(s), genes e seqüência do genoma), o produto de cada gene, as reações bioquímicas catalisadas por cada produto do gene, os substratos de cada reação, e a organização das reações dentro das vias. Também pode descrever a rede genética de um organismo: seus promotores, operons, fatores de transcrição, e sítios de ligação dos fatores de transcrição.

3.2.2 Directons, operons e unidades transcricionais (TUs)

Um operon é um conjunto de genes contíguos transcritos sob a regulação de um promotor que dá origem a um mRNA policistrônico.

As unidades de transcrição (TUs) são o mesmo que operons, porém podem conter um ou mais genes. Para todos os fins práticos, considera-se neste trabalho que um operon é uma TU que inclui mais de um gene. As TUs incluem também os elementos regulatórios tais como promotores e terminadores. A organização de genes numa TU depende mais de sua ordem espacial, isto é, posições dos genes dentro de um cromossomo, do que das formas de organização do gene. A ordem espacial é conseqüência dos relacionamentos funcionais e regulatórios. Portanto, saber a organização do gene pode ser uma pista importante para inferir os arranjos organizacionais no genoma.

O método da distribuição da distância intergênica (IDD) proposto por SALGADO e colaboradores (2000) é utilizado pelo Pathway Tools como ponto de partida para um preditor de TUs mais preciso, incluindo as informações funcionais de classes de cada gene, quando conhecida (RILEY, 1993).

Riley em 1993 criou categorias para os genes de acordo com a sua função fisiológica celular. É destacado que esta classificação foi arbitrária e que poderia ser feita de outras tantas maneiras, pois não há apenas uma possibilidade para este tipo de atribuição. Os genes podem ser classificados dentro de até quatro categorias, num total de 118 categorias (RILEY e SPACE, 1996). Algumas das categorias usadas: Metabolismo Intermediário ("Intermediary metabolism"), Biossíntese de pequenas moléculas ("Biosynthesis os small molecules"), Metabolismo de macromoléculas ("Macromolecules metabolism").

A distância intergênica é uma das evidências que o gene pertence ou não ao operon. O Pathway Tool utiliza a base de dados RegulonDB (SALGADO *et al.*, 2004), que contém as informações da posição relativa dos

sítios regulatórios, o início da transcrição, a distância do começo do gene transcrito, e suas coordenadas no genoma.

O PGDB de referência, por excelência, é o banco de dados de via/genoma da *Escherichia coli*, o EcoCyc. No EcoCyc existem 983 unidades de transcrição, 99 terminadores, 1047 promotores, e 1327 sítios de ligação de fatores de transcrição, todos experimentalmente determinados (SRI International, 2005). Os processos e entidades biológicos da *E. coli*, tais como vias metabólicas, reações bioquímicas, enzimas, proteínas de transporte, e regulação da transcrição da expressão do gene, incluídos e descritos no EcoCyc, tornam este PGDB uma base de dados de genoma anotado de alto nível.

O algorítmo que faz predições de possíveis operons ou TUs, o Predictor, começa por particionar o genoma em Directons (SALGADO *et al.*, 2000). Um directon é um conjunto de genes contíguos que são transcritos na mesma direção e que são delimitados nas suas fronteiras por genes transcritos na direção oposta (Figura 19). Cada directon é separado em pares de genes. Um directon produzirá n-1 pares de genes. Os pares de genes que estão em um directon são chamados WD ("Within Directon").

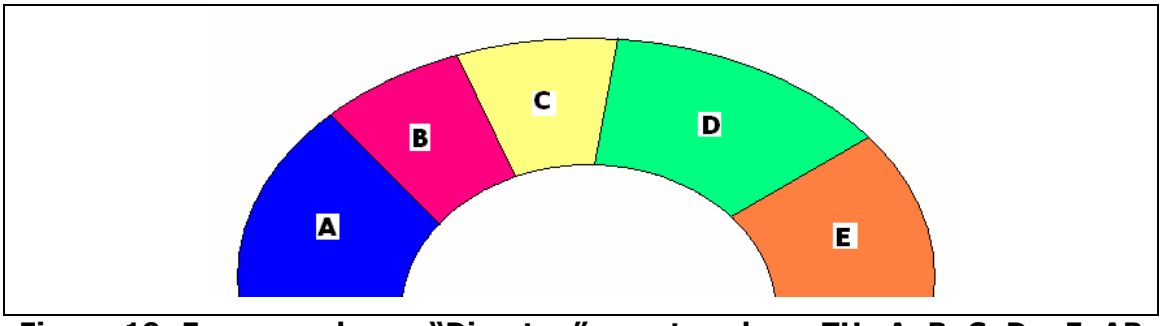


Figura 19: Esquema de um "Directon", mostrando as TUs A, B, C, D e E. AB = TUB, BC = WO, CD = WO, DE = TUB (Veja descrição no texto).

Usando a informação da TU do EcoCyc, um conjunto de pares de genes pode ser marcado para formar um conjunto para treinamento. Estes genes localizados dentro de um directon são nomeados por pares WO ("Within Operon"), e pares de genes situados entre os limites de TUs são chamados pares TUB ("Transcription Unit Boundary"). TUs de RNA são eliminadas porque elas diferem das características de TUs que codificam proteínas.

SALGADO e colaboradores (2000) sugerem que a distribuição da distância de pares WO difere dos pares TUB. As duas distâncias mais freqüentes entre pares WO são sobreposições de 4 a 1 base e não há distância prevalescente entre os genes vizinhos que pertençam a diferentes TUs. Estes estudos indicam que a seguinte fórmula pode ser aplicada para o valor mais provável de que um par de genes seja um par WO:

$$LL(dist) = \log \frac{Nwo(dist) \big/ TNwo}{Ntub(dist) \big/ TNtub}$$

onde pares WO e TUB são coletados em grupos de 10 bp. Assim, *Nwo(dist)* e *Ntub(dist)* correspondem, respectivamente, ao número de pares WO e TUB cujos genes estão separados por uma distância *dist* num intervalo de 10 pares de bases (10, 20, 30...). *TNwo* e *TNtub* são os números totais de pares WO e TUB, respectivamente. *LL(dist)* é a probabilidade de que um par de genes com uma distância *dist* seja um par WO. Além disso, SALGADO e colaboradores (2000) usaram uma segunda característica para desenvolver seu programa Predictor, a classe funcional de Riley (RILEY, 1993), considerando dois genes de mesma classe funcional quando coincidem no segundo nível; neste caso, só haveria predição se houvesse classe definida para o gene, o que acontece para apenas 50% dos genes do genoma da *E. coli*. O Predictor do Pathway Tools é mais robusto nos níveis hierárquicos, e considera o nível 4 para o compartilhamento de classes. O preditor classifica os pares de genes em WO ou TUB baseado nos seguintes critérios:

1. A distância intergênica entre um par de genes;
2. Se as classes funcionais dos genes são equivalentes (quando forem conhecidas);
3. Se ambos os produtos dos genes são ou não enzimas da mesma via metabólica;
4. Se ambos os produtos dos genes são ou não monômeros do mesmo complexo protéico;
5. Se o produto do gene transporta ou não um substrato para uma via metabólica na qual o outro produto do gene está envolvido como uma enzima;

6. Se um gene à montante (“upstream”) ou à juzante (“downstream”) de um par de genes (e dentro do mesmo ”directon”) está ou não relacionado a um dos genes no par de genes pelas condições 1, 2 ou 3;
7. Se há similaridade entre os códons de um par de gene (conservação por filogenia).

O Predictor do Pathway Tools foi testado para a base de dados da *Bacillus subtilis*, o BsubCyc, PGDB gerado pelo Pathway Tools computacionalmente, sem o uso das classes funcionais de Riley, e sem o uso das descrições estruturadas das informações de transporte, que poderiam ter sido usadas pelo método de predição.

Para a BsubCyc o desempenho do Predictor caiu para 46% com relação à predição de operons, enquanto que para a *E. coli* essa predição foi de 62%. Os resultados da análise dos pares dos genes considerados (com relação a sua eventual organização em operons) tiveram 81% de sensibilidade (isto é, taxa de falso-negativos) e 48% de especificidade (isto é, taxa de falso-positivos), com uma precisão de 64%, que é a média da sensibilidade mais a especificidade. Baixa especificidade nas predições de pares de genes significa que o preditor gera uma alta taxa de genes falso-positivos, isto é, muitos pares TUB foram preditos como sendo pares WO. Mas os pares WO foram preditos com alta sensibilidade. Isto indica que o preditor é mais sensitivo do que específico (ROMERO e KARP, 2004).

ROMERO e KARP (2004) demonstraram tendências esperadas de que genes WO têm, freqüentemente, distâncias mais curtas entre eles do que os genes TUBs. Também demonstraram que os genes no mesmo operon (WO) tendem a ter a mesma classe funcional fisiológica. Os resultados destas análises foram usados pelos desenvolvedores do Pathway Tools para implementar um método para predição da organização genômica dos genes em TUs. O método tem uma precisão de 88% em relação à correta identificação de que os pares de genes adjacentes estão em um operon, ou nas fronteiras de uma TU, e identifica 75% das TUs conhecidas quando usadas para predizer a organização das TUs do genoma da *E. coli*. Baseados

na freqüência das distribuições da distância foram estimados 630 de 700 operons em *E. coli*.

Os métodos descritos acima foram utilizados neste trabalho para a análise dos prováveis operons da *C. violaceum*, a exemplo do operon da biossíntese de violaceína (ANTONIO e CRECZYNSKI-PASA, 2004 ; AUGUST *et al.*, 2000).

3.3 Criação do PGDB CvioCyc

Instalado o software Pathway Tools, foram executados os procedimentos para a criação do Banco de Dados Via/Genoma, ou PGDB (Pathway/Genome Database) do organismo de interesse, a *Chromobacterium violaceum*. O nome dado a esta base de dados foi CvioCyc, em conformidade com a nomenclatura usualmente adotada pelo SRI International.

Para isto foi necessário criar um arquivo com o nome `genetic-elements.dat`, conforme mostrado na Figura 20, com as informações a respeito dos elementos genéticos do organismo em questão. Esses dados compreendem o único cromossomo da *C. violaceum*, sua estrutura circular, o correspondente arquivo de anotação (`cvio.gbk`) descrevendo as predições dos genes para este cromossomo, e um outro arquivo (`cvio.fna`) com a seqüência completa dos ácidos nucléicos do cromossomo, no formato FASTA, ambos adquiridos do endereço (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Chromobacterium_violaceum/).

```
;;============
;
ID    CHROM-1
NAME  Chromosome 1
TYPE  :CHRSM
CIRCULAR?   Y
ANNOT-FILE  cvio.gbk
SEQ-FILE    cvio.fna
//
```

Figura 20: Estrutura e conteúdo do arquivo `elements_genetic.dat` criado para inicialização da base de dados CvioCyc. Este arquivo contém informações genéticas do microrganismo, como por exemplo o número de cromossomas e número de plasmídeos.

O genoma da *C. violaceum* contém 4.751.080 pares de bases (ou *basepairs,* bp), com 4.431 ORFs codificando proteínas, cobrindo 89% do genoma. Destas 4.431 ORFs, 61,3% foram anotadas como ORFs de funções conhecidas, 21,6% foram anotadas como proteínas conservadas hipotéticas, e 17,1% foram designadas proteínas hipotéticas (Brazilian National Genome Project Consortium, 2003).

Muitas das vias metabólicas incluídas automaticamente no CvioCyc, pelo Pathway Tools, não estão completas. Através dos relatórios "Evidências das Vias Metabólicas" (`Pathway_Evidence.html`) e "Identificação das Reações Faltantes nas Vias Metabólicas da *C. violaceum*" (`Identification of reactions from pathway holes in C. violaceum.html`) foram realizadas remoções das vias não prováveis de ocorrer, seja por falta de evidências da existência das enzimas faltantes, ou por já estarem incluídas em uma outra via.

O módulo PathoLogic do Pathway Tools contém a função que cria as PGDBs, e foi então utilizado para a criação da base de dados da *C. violaceum*, o CvioCyc. O software cria uma estrutura de diretórios e vários arquivos dentro destes diretórios. A estrutura de diretórios criada está mostrada na Figura 21.

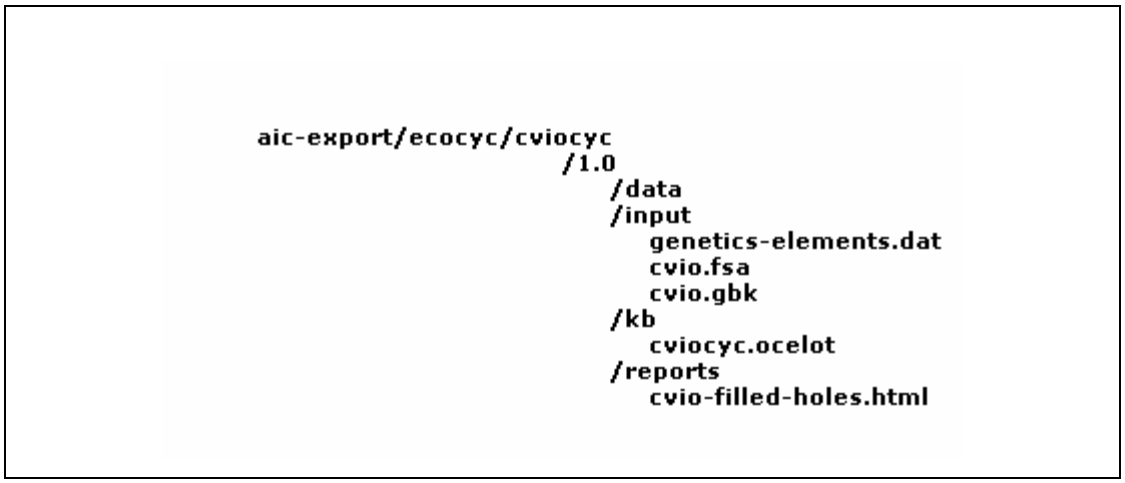


Figura 21: Estrutura de diretórios gerada pelo Pathway Tools, durante a criação da base CvioCyc.

O PathoLogic faz uma série de análises das vias que podem existir no PGDB em criação. A abordagem do algorítmo de avaliação das vias é conservativa pois procura trazer mais vias potenciais e deixar para o usuário fazer as devidas exclusões. As vias são excluídas do novo PGDB com base nas seguintes condições:

- Nenhuma reação contida na via é conhecida no organismo OU
- Nenhuma das reações na via que são catalisadas pelo organismo são reações que apenas esta via utiliza E
    - a via consiste de mais de duas reações mas contém evidência para somente uma reação, OU
    - o conjunto de reações que havia evidências era a mesma que o conjunto de reações que havia evidencia para outra via, mas a via contém reações adicionais (para as quais não há evidências) não contidas na outra via, OU
    - O conjunto de reações para o qual há evidência é um subconjunto do conjunto de reações que há evidência em alguma outra via, E
        - A via em análise e uma outra via são membros da mesma classe variante. OU
        - A via está classificada como via biossintética e está ausente no mínimo as duas últimas reações do final da via. A explicação para este critério é que se as reações finais estão ausentes, então a via não está degradando o

seu suposto substrato, assim a outra via contendo as reações em comum está mais em conformidade para refletir a utilização destas reações no organismo. OU

- A via está classificada como via de degradação e está faltando um ou mais passos no início da via. A explicação para este critério é que se as reações iniciais estão ausentes, então a via não degrada seu suposto substrato, assim a outra via contendo as reações em comum está mais em conformidade para refletir a utilização destas reações no organismo. OU
- A via foi classificada como uma via de metabolismo de energia e está faltando no mínimo a metade de suas reações. A razão para este critério é que que há uma grande negociação de sobreposição entre as reações do metabolismo de energia com muitas reações aparecendo em muitas vias. Portanto, muitas reações falso-positivas podem ser inferidas vindas destas vias. Um limite arbitrário mas confiável para excluir estas falso-positivas é de 50%.

Após a criação do PGDB, o Pathway Tools permite que seja gerado o relatório `cvio_filled-holes.html`, através do PathoLogic. Este relatório é gerado por um algoritmo (GREEN e KARP, 2004) que indicará qual a probabilidade (P) de que uma enzima anotada seja a que está ausente na via.

O algorítmo para a escolha das seqüências prováveis a catalisarem as reações ausentes (*missing-reactions*), baseia-se em:

I. Identificação das candidatas:
    1. Busca das seqüências: Busca no SWISS-PROT (BOECKMANN *et al.*, 2003) e no PIR (WU *et al.*, 2003) de seqüências para enzimas que catalisam esta mesma reação em outros organismos. Como as seqüências podem não ser homólogas, estas enzimas serão citadas como isoenzimas.

2. Busca por homologia: Usa o BLASTP para cada uma das isoenzimas contra o genoma do organismo de interesse.
3. Consolidação dos dados através da análise dos *hits*[2] encontrados pelo BLASTP para verificar as seqüências que alinham com uma ou mais isoenzimas.

II. Avaliação da candidata: Calcula a probabilidade que a proteína candidata tem de ser a seqüência requerida pela reação faltante. Como a atividade da reação ausente é conhecida para a via inferida, o software usa uma isoenzima com esta função para procurar as seqüências no organismo-alvo que tenha similaridade com esta isoenzima.

Neste trabalho foram analisados, além do relatório "`cvio_filled-holes.html`", os relatórios "`Name-matching-report.txt`" e "`PathwayEvidence.html`". Diversas pesquisas foram feitas para análise e confirmação dos resultados usando ferramentas de bioinformática e disponibilizados na Internet, como:

- BLASTP: Verificação da anotação feita pelo BRgene, através da verificação da candidata em relação a outros genomas. (disponível em http://www.ncbi.nlm.nih.gov/BLAST/).
- BLASTP contra o genoma da *C. violaceum*: Verificação da identidade e similaridade, através do BLAST, de uma isoenzima em relação ao genoma da *C. violaceum*. Para a análise do genoma da *C. violaceum*, utilize o endereço (http://www.ncbi.nlm.nih.gov/genomes/geblast.cgi?gi=321).
  Observe que foi usado como parâmetro, no BLASTP, a base de dados "**nr**" (Non-redundant), que compreende as informações não redundantes da tradução das CDS nas bases de dados do "GenBank", "PDB", "SwissProt", "PIR" e "PRF", que são bases de dados de seqüências já anotadas de proteínas.

[2] "hits" refere-se ao número de seqüências encontradas na base de dados que são similares à seqüência procurada (query).

- KEGG: Verificação da via metabólica em que o KEGG colocou a reação ausente. Também a verificação do caso em que a candidata tenha sido atribuída à função ausente, pois o KEGG não usa a anotação dos genes existentes, ele faz uso apenas das seqüências de nucleotídeos e faz as suas próprias pesquisas de similaridades. Disponível em http://www.genome.jp/kegg/.
- BIOCYC: Verificação da via metabólica em outros organismos. Disponível em: http://biocyc.org.
- PUBMED: Consultas às publicações científicas existentes até o momento. Disponível em: http://www.ncbi.nlm.nih.gov.
- NCBI: Pesquisa a seqüências em outros organismos relativas às enzimas inexistentes na via metabólica da *C. violaceum*. Disponível em: http://www.ncbi.nlm.nih.gov.
- BRENDA: Verificação dos nomes e EC_number de enzimas relativas à via metabólica, e verificação da via metabólica na qual a enzima está incluída. Disponível em: http://www.brenda.uni-koeln.de/.
- BRGENE: Verificação das informações detalhadas da anotação da enzima relativa à via metabólica. Disponível em: http://www.brgene.lncc.br/cviolaceum/.
- CDD–NCBI(MARCHLER-BAUER e BRYANT, 2004): Verificação da existência de domínio. Disponível em: ttp://www.ncbi.nlm.nih.gov.

3.4 Estratégia usada para analisar as vias metabólicas da base de dados da *C. violaceum* - CvioCyc

Conforme o trabalho de YEH e colaboradores (2004), algumas regras podem ser usadas para remover vias metabólicas que o Pathway Tools infere: (1) Nenhuma via que contenha reação que seja única para esta via deve ser removida; (2) Se o conjunto de reações de uma via é um subconjunto de reações em uma outra via, a via menor deve ser eliminada; (3) No caso de duas vias conterem o mesmo conjunto de reações, elimina-se a via em que há mais reações ausentes e sem evidências experimentais.

As regras acima foram adaptadas e, usando condições adicionais, foram feitas eliminações de algumas das vias metabólicas que foram inferidas pelo PathoLogic. Foram consideradas 61 vias que, conforme o fluxograma mostrado na Figura 22 (item 3.4.1), foram analisadas minuciosamente, tendo como base o relatório de evidência das vias metabólicas, no arquivo "`PathwayEvidence.html`", e do relatório de preenchimento, no arquivo "`cvio_filled-holes.html`".

3.4.1 Descrição do algoritmo de análise de vias metabólicas

Algumas condições para a eliminação das vias metabólicas inferidas pelo software Pathway Tools foram testadas pelo algoritmo apresentado no fluxograma da Figura 22 e serão comentadas a seguir.

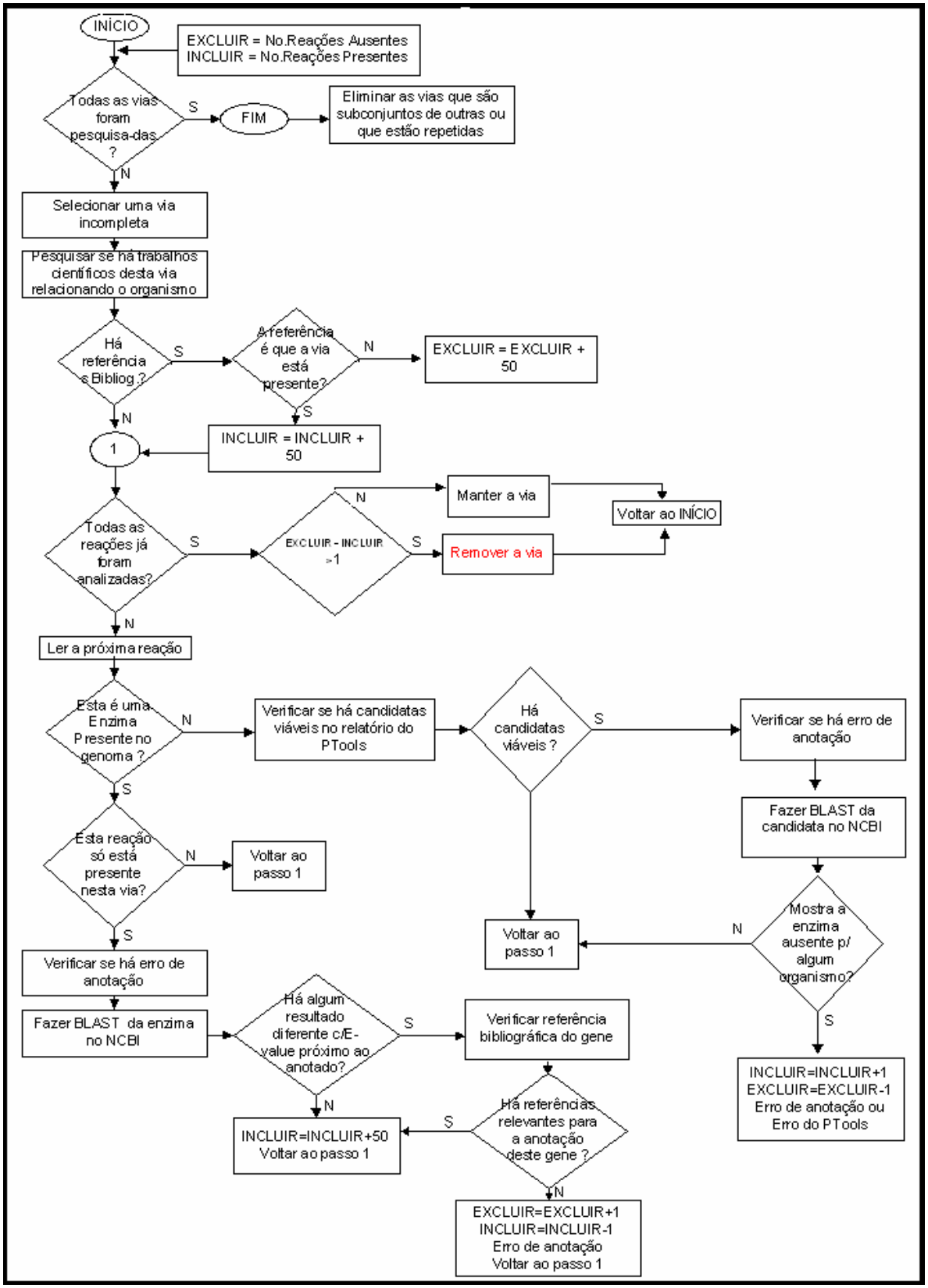


Figura 22: Fluxograma para a análise de consistência das vias metabólicas.

As condições observadas no algoritmo são:

As variáveis de decisão EXCLUIR e INCLUIR do fluxograma são verificadas ao final do algoritmo. Se o valor da variável EXCLUIR for menor do que o da variável INCLUIR em, no mínimo, duas unidades, a via metabólica será excluída da base de dados CvioCyc. Este critéiro é mais conservativo do que o critério usado pelo PathoLogic para selecionar as vias que são incluídas para o organismo de interesse.

A variável EXCLUIR refere-se ao número de reações ausentes na via, mais os pesos que vão sendo acrescentados de acordo com as ocorrências de condições que são atendidas. E a variável INCLUIR refere-se ao número de reações presentes na via, mais os pesos que vão sendo acrescentados de acordo com as ocorrências de condições que são atendidas.

- Número de reações ausentes: o número de reações ausentes na via deverá ser, no mínimo, a metade mais duas do que as reações existentes. Por exemplo, se INCLUIR for igual a dez, EXCLUIR deverá ser no mínimo igual a doze para que seja considerada a eliminação desta via, caso ela atenda todas as condições necessárias. Porém, não será eliminada a via na qual pelo menos uma enzima faça parte exclusivamente desta via.
- Existência de referência bibliográfica para a via metabólica: a primeira condição testada é se a via metabólica de interesse tenha sido citada em alguma referência bibliográfica em relação ao organismo (no caso deste trabalho, a *Chromobacteirum violaceum)*. Se existir, então esta via não deverá ser eliminada, não importando quantas reações estão presentes ou ausentes. Mas todas as reações serão analisadas para verificação de possíveis erros de anotação.
- Existência de citação bibliográfica para o gene: a cada reação analisada, é verificado o caso de haver alguma referência bibliográfica a respeito desta reação, em relação ao organismo estudado.
- Reação só presente em uma via: é verificado o caso da reação existir em uma única via; neste caso, é verificado se não há erro de anotação; se não houver, esta via não deverá ser eliminada, não

importando quantas reações estão presentes ou ausentes, conforme YEH e colaboradores (2004).

Verificação de erro de anotação:

- Erro de anotação em reação presente: é verificado se o gene foi anotado corretamente, quando da busca de similaridade para reações exclusivas da via em estudo.
- Erro de anotação em reação ausente: é verificado se a reação ausente, na via em estudo, tem uma candidata no genoma e esta pode ser a reação em análise, verificando-se resultados do BLAST.

CAPÍTULO IV

Resultados e Discussão

4.1 As Interfaces web e local da CvioCyc

O Pathway Tools possui duas formas de acesso a sua interface gráfica, uma através de interface WEB e outra através de interface local, tanto usando o sistema operacional Windows quanto usando o sistema operacional UNIX. A interface WEB é especialmente desenhada para se fazer apenas consultas à base PGDB, enquanto que a interface local é usada para qualquer tipo de operação.

É interessante lembrar que o acesso livre através da Internet é muito importante pois os dados são disponibilizados para qualquer interessado a fazer consultas aos mesmos, desta forma as informações da base de dados CvioCyc ficam disponíveis universalmente.

A Figura 23 apresenta a "homepage" da base de dados CvioCyc, que está acessível através da Internet, usando-se qualquer "browser", no endereço: http://cviocyc.intelab.ufsc.br. As opções de consulta estão listadas ao lado esquerdo da janela, porém, também é possível pesquisar em todas as opções de consulta ao mesmo tempo; por exemplo, digita-se "*violacein* " e clica-se com o botão-esquerdo do mouse sobre o botão *go*. Os resultados são mostrados na Figura 24. No campo "*Quick Search*', pode digitar-se os caracteres a serem pesquisados e clica-se em uma das opções de consulta: Database Search, onde faz-se pesquisa em todos os objetos da base de dados; Advanced Database Search, onde uma tela com mais opções de pesquisa é aberta e pode-se usar composições de valores para alguns dados, usando-se os conectores lógicos AND ou OR; BLAST, nesta opção uma tela é aberta para informar qual organismo e qual a seqüência se deseja consultar e o programa BLASTP ou BLASTX faz a pesquisa, dependendo se foi informada seqüência de aminoácidos ou seqüência de nucleotídeos; "Pathways", onde a pesquisa fica restrita apenas às vias; Genes, onde a pesquisa fica restrita apenas aos genes; "Reactions", onde a pesquisa fica restrita apenas às reações bioquímicas (ou enzimas, na

maioria dos casos); “Compounds” , onde a pesquisa fica restrita apenas aos compostos químicos; “Metabolic Chart”, mostra o diagrama de todas as vias metabólicas do organismo, chamado de “The Omics Viewer”, que é uma ferramenta para a análise e visualização de informações.

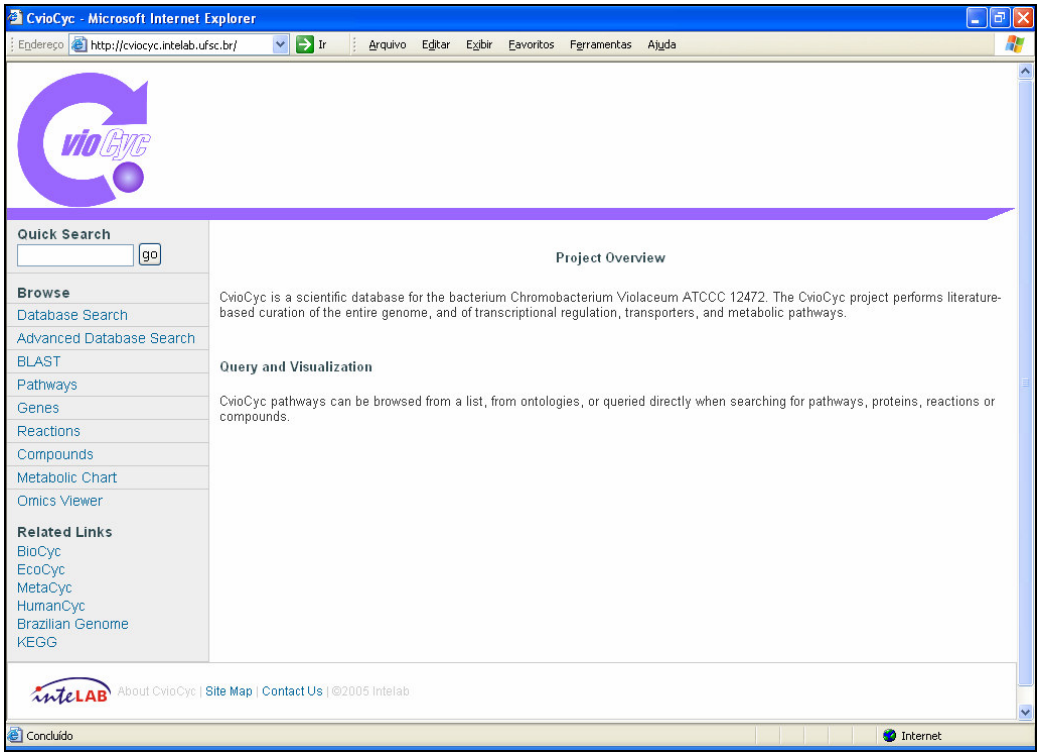


Figura 23: Página de acesso à base de dados via/genoma CvioCyc via Internet. Endereço WEB http://cviocyc.intelab.ufsc.br.

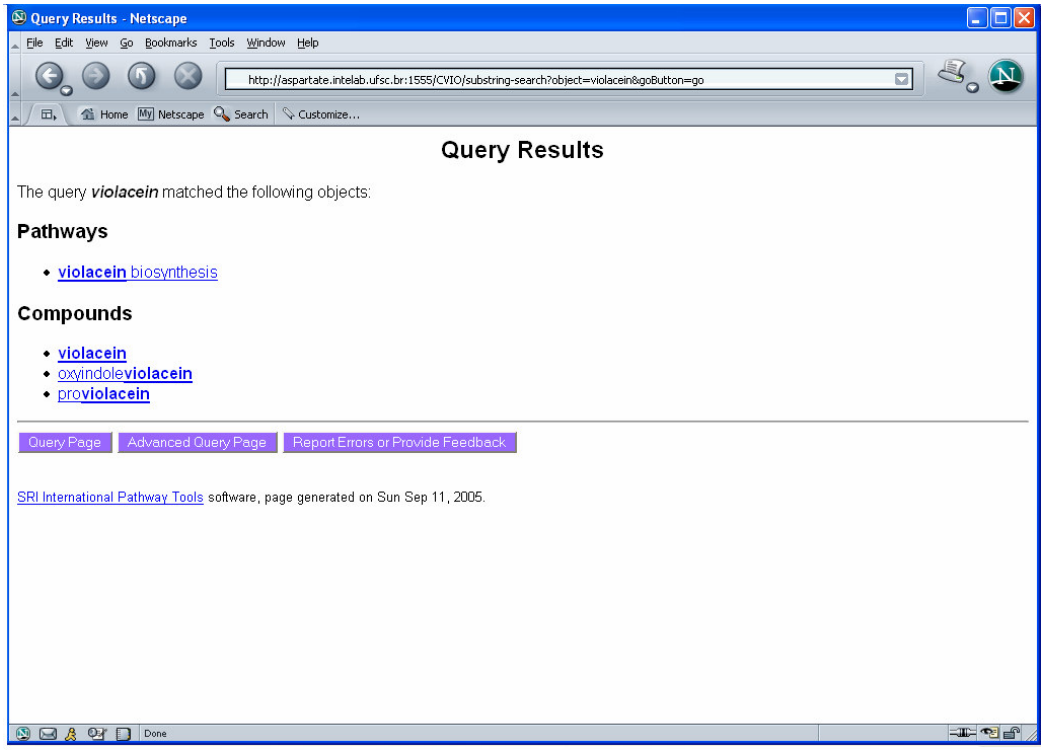


Figura 24: Página de resultados da pesquisa ao nome *violacein* a partir da homepage da CvioCyc (http://cviocyc.intelab.ufsc.br).

É mostrado na Figura 25 o resultado da consulta às vias metabólicas na interface WEB, digitando-se *violacein biosynthesis* e clicando-se o botão-esquerdo do mouse sobre o botão *go* ; desta forma obtém-se o diagrama da via. Todas as entidades deste diagrama podem ser consultadas através dos links que esta tela disponibiliza; para isto clica-se o botão-esquerdo do mouse sobre a entidade desejada.

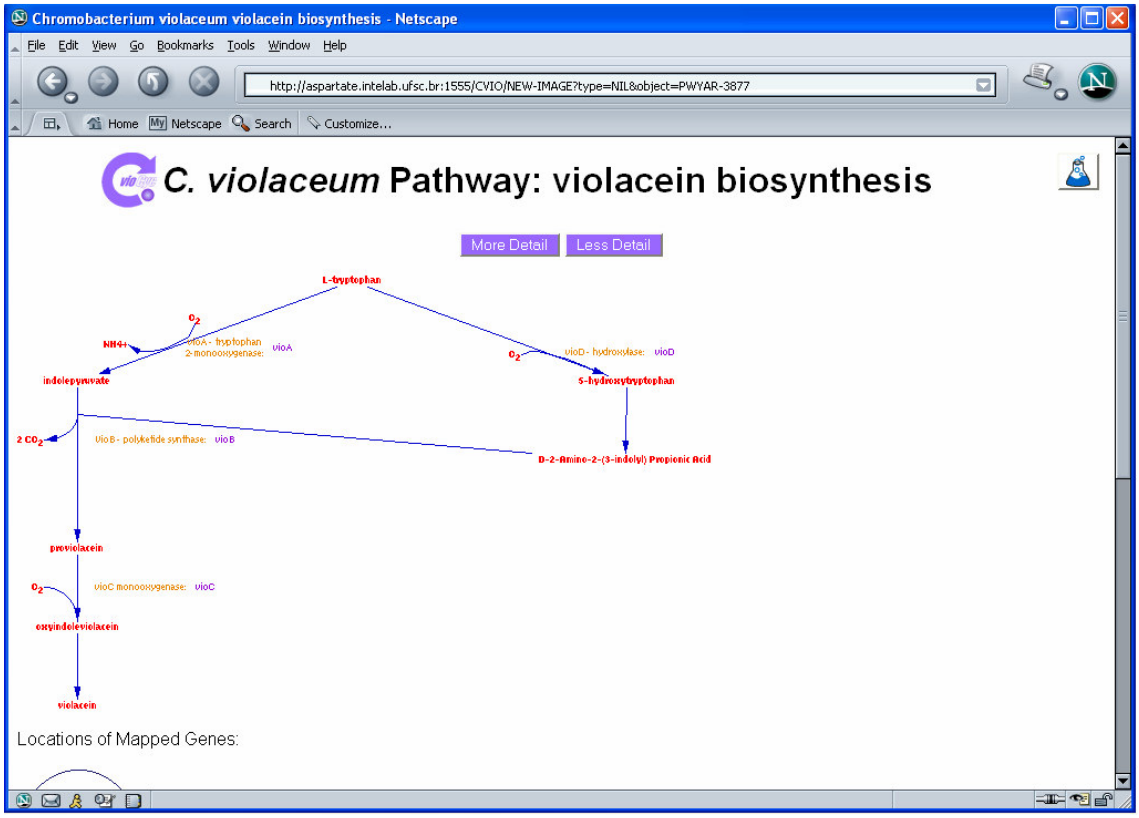


Figura 25: Página resultado da consulta a via de biossíntese da violaceína. Nesta tela estão disponíveis links para a apresentação de informações de todas as entidades, da CvioCyc, que fazem parte da via.

Através da interface local ficam acessíveis todas as funções que o software Pathway Tools oferece, diferente da iterface WEB que apenas as funções de consulta estão disponíveis. Como exemplos são mostradas algumas das telas de maior importância para a manutenção da base de dados. As telas dos exemplos são chamadas a partir da janela mostrada na Figura 26 que é a janela de abertura do PathwayTools, onde todas as bases de dados existentes no ambiente local são mostradas e estão disponíveis na tela. A partir do momento em que é selecionada uma destas bases (clicando-se em cima do nome de uma delas) é possível obter-se qualquer uma das opções do menu desta tela, só então um outro menu é aberto e as opções disponíveis para esta pesquisa são mostradas na Figura 27, onde foi escolhida "Reaction". Essas opções irão variar um pouco de acordo com a pesquisa escolhida; na Figura 28 a tela de edição de uma reação já existente, a reação de "EC_number" 1.2.7.3, é mostrada na janela. Já a tela de criação de uma reação está mostrada na Figura 29, onde uma nova

reação será incluída no PGDB após a validação destas informações através do botão OK. O mesmo método de acesso usado para a criação de uma reação é usado para qualquer uma das entidades da base de dados.

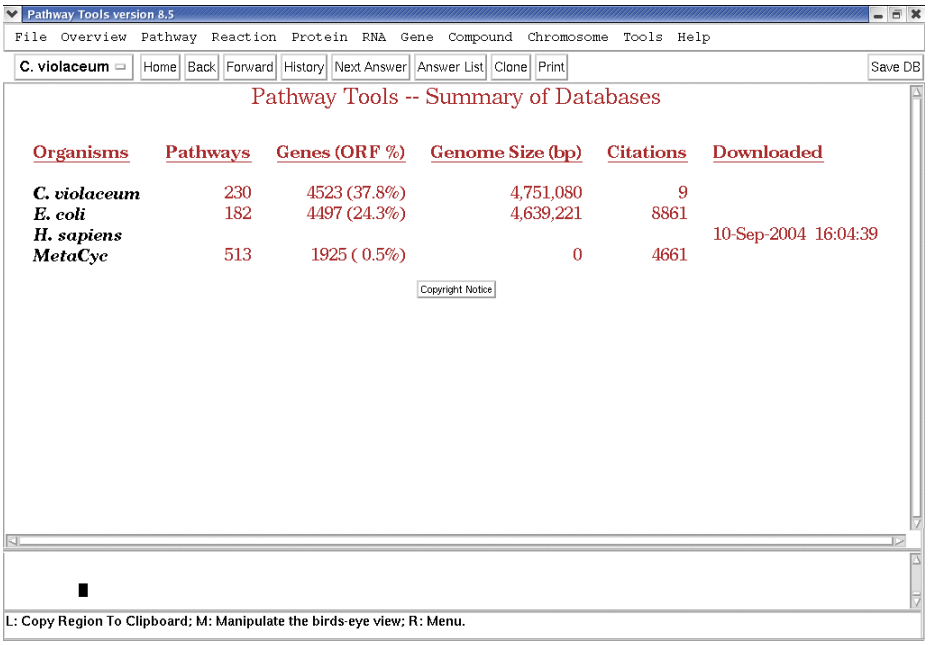


Figura 26: Tela principal para a manutenção do PGDB em questão.

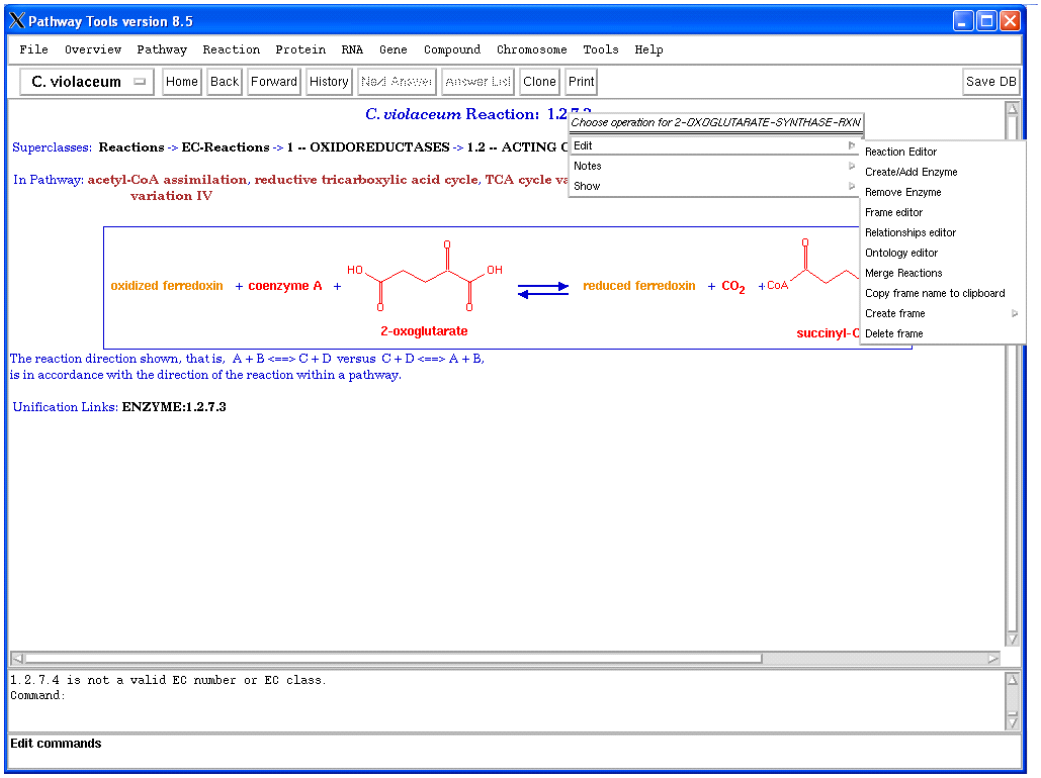


Figura 27: Tela dos menus de opções para a manutenção da reação 1.2.7.3 da Cviocyc, onde podem ser feitas edição, deleção e criação. Também pode-se mostrar as informações da reação ao abrir a opção *show* e compará-la com qualquer organismo que tenha sido criado localmente.

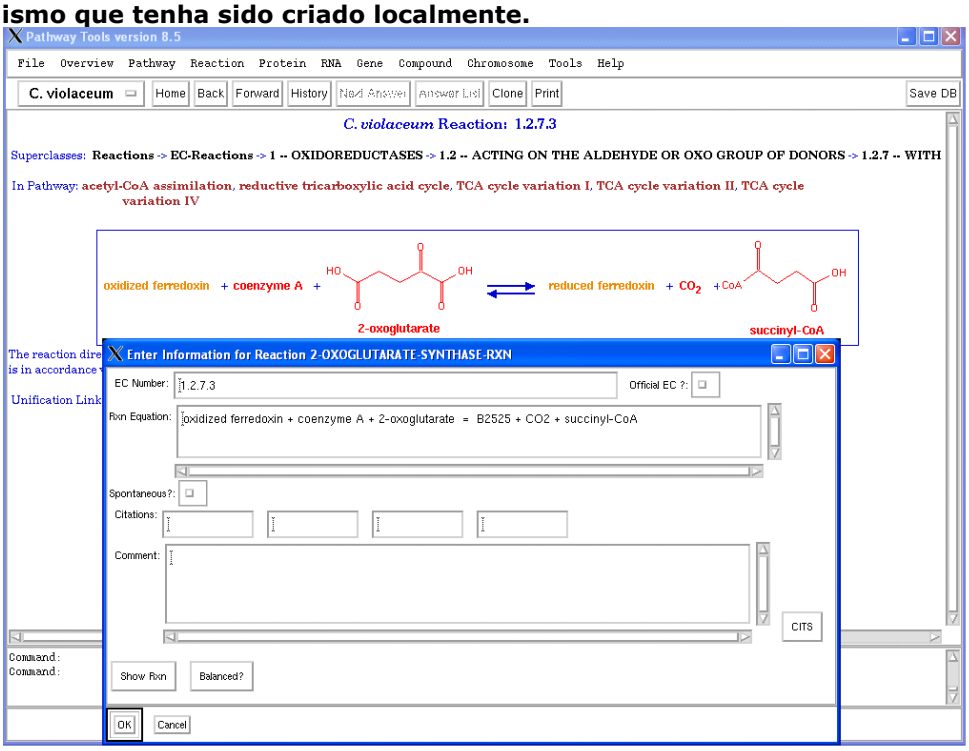


Figura 28: Tela para a edição da reação 1.2.7.3 da Cviocyc, onde se pode fazer as alterações desejadas.

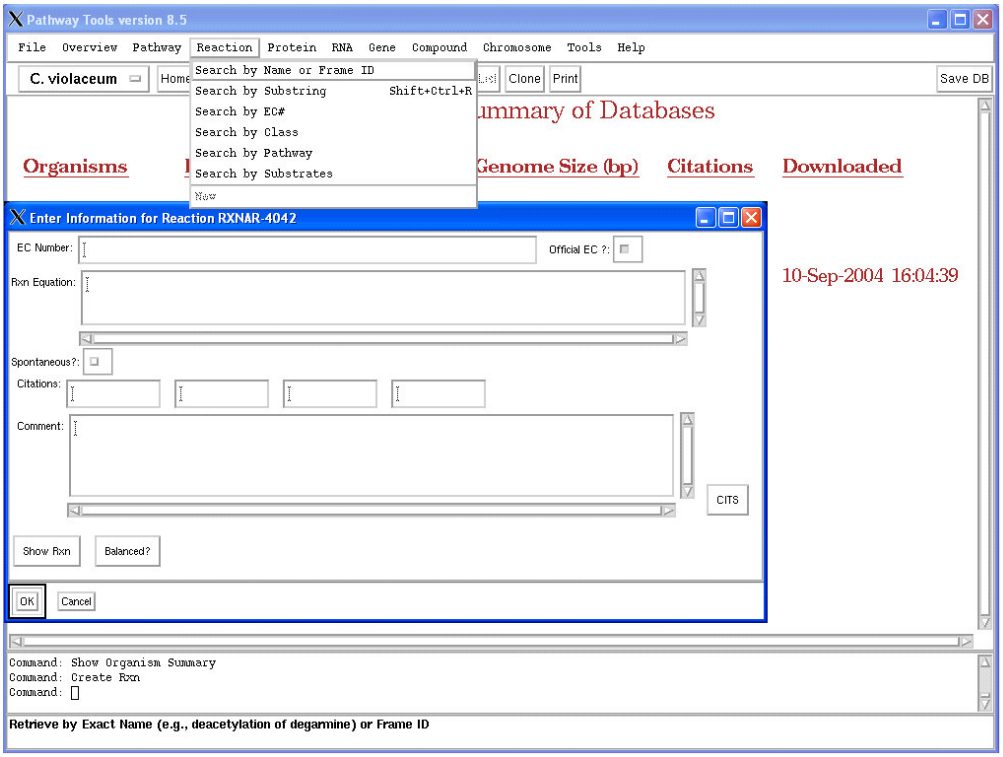


Figura 29: Tela para a criação de uma reação na CvioCyc. Ela é acionada através do clique sobre a opção *New* do menu de opções da entidade *Reaction*. Fonte: Software Pathway Tools.

4.2 Os resultados da geração automática e a cura de vias na CvioCyc

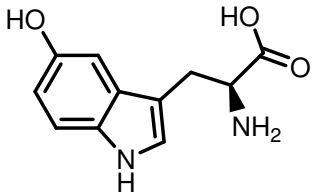
Inicialmente o Pathway Tools inferiu 233 vias metabólicas que estão sendo refinadas através de pesquisas com ferramentas de Bioinformática. Neste trabalho, a principal via metabólica de interesse, no caso da *C. violaceum*, é a via metabólica da biossíntese de violaceína. Através das anotações e dos trabalhos da literatura, essa via foi incluída manualmente na CvioCyc, a base de dados Via/Genoma criada para esse organismo apresentado na Figura 25. Foram criados cinco novos compostos desta via, que não existiam em nenhum dos organismos do MetaCyc. As estruturas moleculares desses compostos estão mostradas na Figura 30. A via de biossíntese de violaceína não foi incluída automaticamente pelo software porque a mesma não está presente em nenhum organismo das bases de dados que o software consulta. Isto não foi uma surpresa, porque a

biossíntese de violaceína é uma característica de apenas algumas poucas espécies, em geral pertences ao gênero *Chromobacterium* (MATZ, 2004)

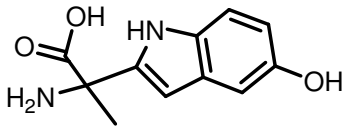
As 61 vias metabólicas analisadas estão listadas na Tabela 4 e se chegou a algumas situações:

- algumas possíveis anotações de ORFs equivocadas,
- existência de complexos para completar vias metabólicas que o Pathway Tools não encontrou,
- vias incompletas porém com muitas enzimas presentes para esta via,
- vias incompletas com poucas enzimas presentes .

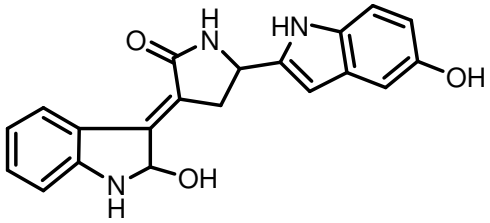
As vias metabólicas que estão completas não foram checadas neste primeiro momento.



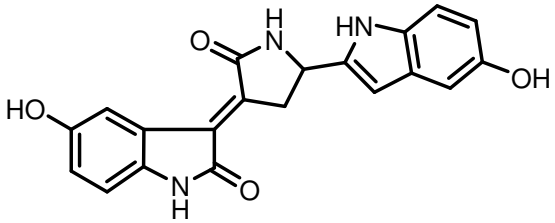
5-HIDROXITRIPTOFANO



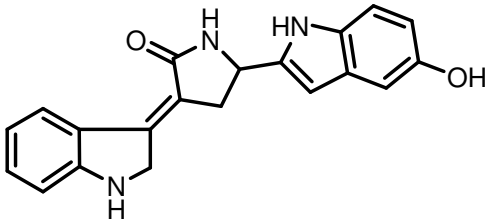
ÁCIDO D-2-AMINO-2(3-INDOLIL) PROPIÔNICO



OXINDOLVIOLACEÍNA



OXIVIOLACEÍNA



PROVIOLACEÍNA

Figura 30: Compostos criados na CvioCyc para compor a via metabólica da biossíntese de violaceína (Estruturas apresentadas por August et al., 2000).

Foi identificado um total de 458 reações ausentes. Para a depuração das vias metabólicas foi utilizado como ponto de partida o relatório `cvio_filled-holes.html` que pode ser visto, em parte, na Figura 31.

alanine biosynthesis I

Total # of reactions in pathway = 3

Present reactions: 2

Reaction	Protein(s)
ALARACECAT-RXN	(DADB-MONOMER ALR-MONOMER)
BRANCHED-CHAINAMINOTRANSFERVAL-RXN	(ILVE-MONOMER)

Missing reactions: 1

VALINE-PYRUVATE-AMINOTRANSFER-RXN

Evidence for VALINE-PYRUVATE-AMINOTRANSFER-RXN (Valine--pyruvate aminotransferase)

Hit	Common	P	#Q	Best Qry	Best Eval	Avg Rank	Aln Len	Qry Len	IDOP?	adj?
GAR-2843-MONOMER	probable transcriptional	0.124744974	14.0	E75208	1.0d-39	2.357143	0.6379343	410	NIL	NIL
GAR-1610-MONOMER	probable transcriptional	0.10764715	14.0	E75208	1.0d-40	3.0	0.70831954	410	NIL	NIL
GAR-168-MONOMER	probable transcriptional	0.08205709	12.0	E75208	4.0d-37	4.6666665	0.7237324	410	NIL	NIL
GAR-1888-MONOMER	probable transcriptional	0.038375832	12.0	E75208	8.0d-41	4.25	0.7159851	410	NIL	NIL
GAR-1635-MONOMER	probable transcriptional	0.02865158	14.0	E75208	2.0d-34	4.214286	0.7532811	410	NIL	NIL

Figura 31: Parte do relatório de saída gerado pelo PathoLogic, `cvio_filled-holes.html`.

4.2.1 Vias metabólicas inferidas pelo software Pathway Tools

Através do relatório "`PathwayEvidence.html`" obteve-se dados quanto às vias metabólicas inferidas, conforme mostrado na Figura 32.

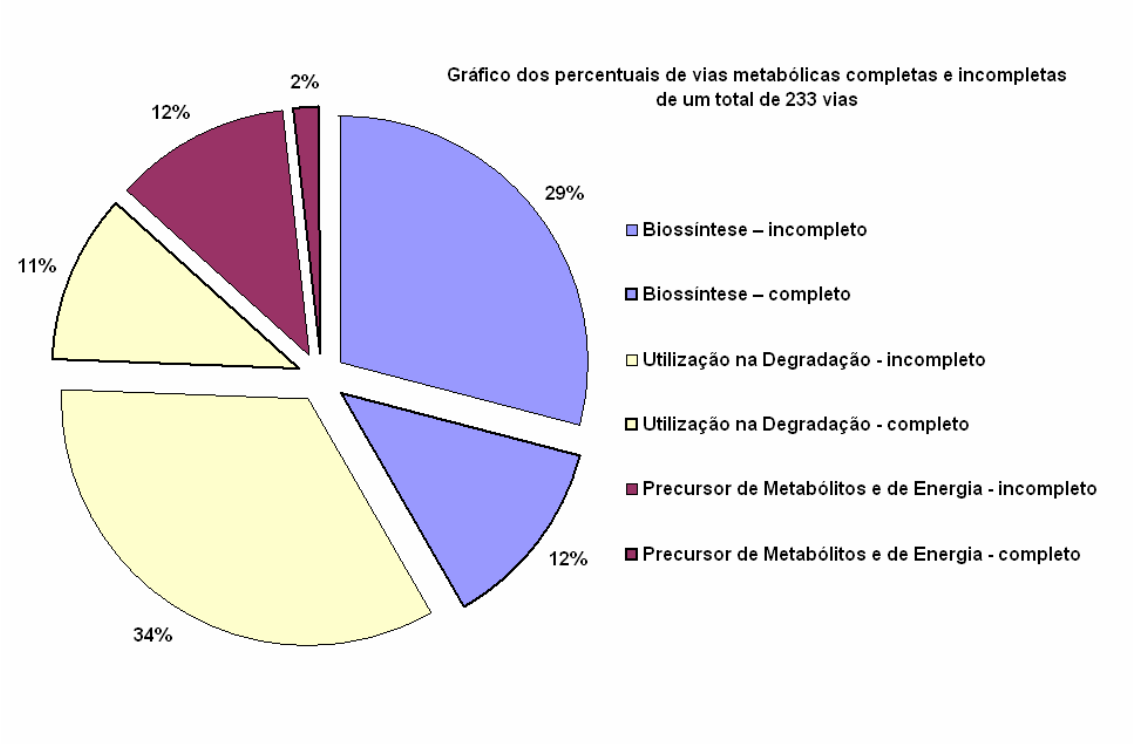


Figura 32: Percentuais de vias metabólicas completas e incompletas agrupadas pelo tipo de via metabólica.

Um subconjunto de 61 vias das vias incluídas pelo PathwayTools na base de dados CvioCyc foram analisadas. O critério usado para a escolha destas vias foi, em primeiro lugar, que as vias fossem incompletas. Na seleção destas vias foi usado o relatório `Pathway_Evidence.html`. Dentre as vias incompletas, foram escolhidas as vias mais factíveis de confirmação, ou seja, as vias com relativamente mais, ou menos, enzimas presentes. No caso de vias quase completas, procurou-se encontrar enzimas falso-negativas, que são as enzimas que não foram anotadas apesar de existirem. No segundo caso, quando há relativamente poucas enzimas presentes na via, procurou-se por enzimas falso-positivas, que são enzimas anotadas porém sem suficiente evidência das mesmas existirem na *C. violaceum*.

Das 61 vias metabólicas analisadas, 17 foram removidas. A Figura 33 mostra o mapa metabólico ("Overview Map") antes da remoção das vias e a Figura 34 mostra o mapa metabólico após as exclusões das vias. Os fragmentos de vias metabólicas, que aparecem à direita nessas figuras,

referem-se a reações isoladas, que ainda que não foram enquadradas em nenhuma via metabólica.

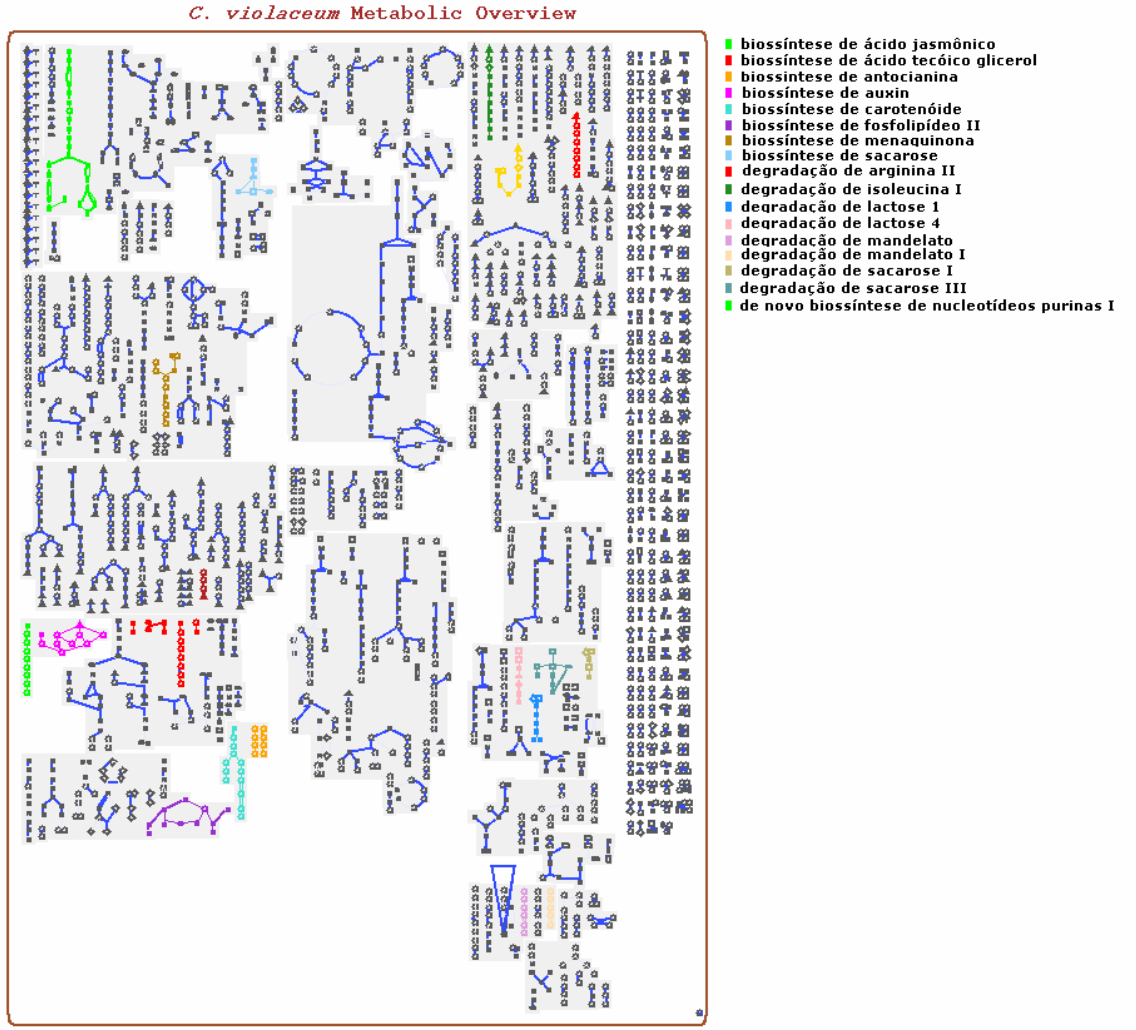


Figura 33: Mapa metabólico da *C. violaceum*, gerado automaticamente pelo Pathway Tools, antes da análise das vias. As vias em cores mostradas na legenda são as vias que foram identificadas como inexistentes por falta de evidências. Adaptada do Software Pathway Tools.

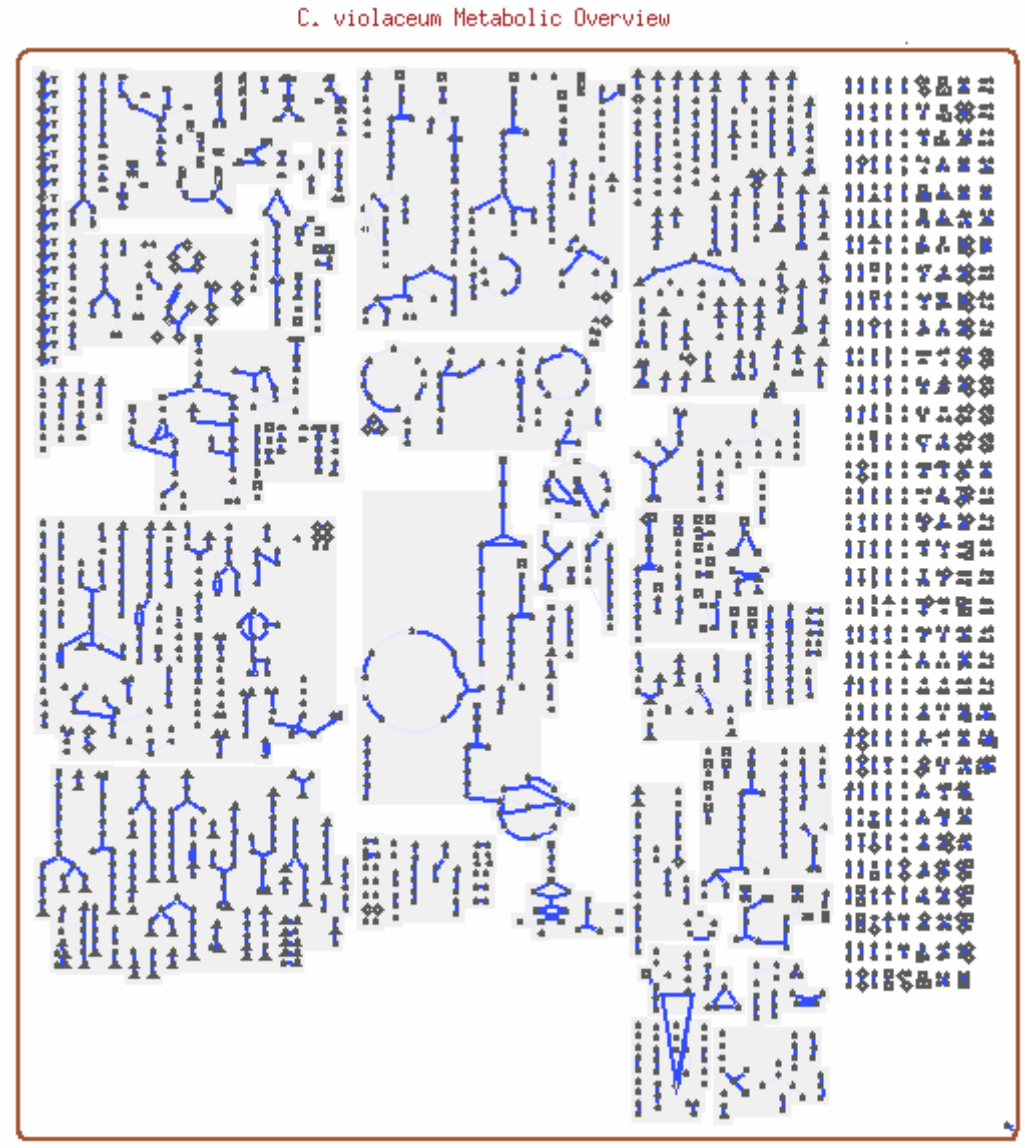


Figura 34: Mapa metabólico da *C. violaceum* após a remoção de 17 das 61 vias analisadas. Fonte: Software Pathway Tools.

Durante a análise das vias que estão incompletas foram encontrados diversos erros de anotação os quais serão descritos no item 4.2.2.

A Tabela 4 lista em ordem alfabética as vias metabólicas analisadas, com as seguintes informações: nome da via, número total de reações da via (NR), número de reações que estão presentes (NP), valor final da variável INCLUIR (VI), valor final da variável EXCLUIR (VE), a ação que a via metabólica sofreu, e o número das ORFs alteradas. Os valores de INCLUIR e de EXCLUIR foram retirados da execução do algoritmo apresentado no fluxograma da Figura 22. Os tipos de ações tomadas podem ser: manter a

via metabólica (M), remover a via metabólica (R), alterar CvioCyc (A) ou modificar ORF (O).

Tabela 4: Listagem, em ordem alfabética, das vias metabólicas analisadas no CvioCyc.

#	Nome da via	NR	NP	VI	VE	Ação	ORFs alt.
1	Anabolismo de Trealose	4	2	2	3	M	
2	Assimilação de Carbono C4 Fotossíntético	4	3	3	2	M	
3	Assimilação de Sulfato	5	3	4	1	M A	
4	Assimilação de Sulfato II	5	3	2	3	M	
5	Biossíntese de Ácido Jasmônico	8	1	1	7	R	
6	Biossíntese de Ácido Glicerol Tecóico	11	2	2	1	R	
7	Biossíntese de Alanina I	3	2	2	1	M	
8	Biossíntese de Antígeno Enterobacterial Comum	9	5	205	8	M O	CV4033 CV4034
9	Biossíntese de Antígeno O	9	6	6	3	M	
10	Biossíntese de Antocianina	6	6	1	0	6	CV0690
11	Biossíntese de Ascorbato	7	4	4	3	M	
12	Biossíntese de Auxina	12	1	1	11	R	
13	Biossíntese de Blocos Construtores do Ácido Colânico	11	5	56	6	M	
14	Biossíntese de Carotenóide	14	1	1	13	R	
15	Biossíntese de Cisteína II	5	4	4	1	M	
16	Biossíntese de Cobalamina, via aeróbica	18	8	13	1	M A O	CV1562 CV1568 CV1556 CV1555 CV1576 CV1575 CV0491

Onde NR= Número total de reações existentes na via; NP= Número de reações identificadas no genoma; VI = Valor final da variável INCLUIR; VE = Valor final da variável EXCLUIR; Ação = Ação que foi tomada sobre a via metabólica, seu valor pode variar entre: M = Manter a via, A = Alterar a via na base de dados CvioCyc e R = Remover a via; ORFs alt = Número da(s) ORF(s) que foi alterada.

#	Nome da via	NR	NP	VI	VE	Ação	ORFs alt.
17	Biossíntese de Difosfato de Isopentanil – Mevalonato Independente	7	2	2	6	M A	CV1258 CV4059 CV1259 CV3536 CV3567
18	Biossíntese de Enterobactina	4	3	154	0	M A	
19	Biossíntese de Fenilalanina	4	3	3	1	M	
20	Biossíntese de Fosfolipídio	9	7	207	2	M O	CV4039
21	Biossíntese de Fosfolipídio II	9	4	4	8	R	
22	Biossíntese de Glutamato I	2	1	51	1	M	
23	Biossíntese de Homoserina e Metionina	8	6	6	2	M O	CV1568
24	Biossíntese de KDO -- incluindo a transferência de lipídio IV A	6	4	56	2	M A O	CV3328 CV0225
25	Biossíntese de Menaquinona	8	3	3	5	R	
26	Biossíntese de Peptidioglicana	11	9	59	1	M A O	CV4343
27	Biossíntese de Precursor do Lipídio-A	8	6	52	2	M A O	CV4337 CV2206
28	Biossíntese de Prolina I	4	3	3	1	M	
29	Biossíntese de Protoheme e Siroheme	15	2	14	1	M A O	CV3433 CV2819 CV0813

Onde NR= Número total de reações existentes na via; NP= Número de reações identificadas no genoma; VI = Valor final da variável INCLUIR; VE = Valor final da variável EXCLUIR; Ação = Ação que foi tomada sobre a via metabólica, seu valor pode variar entre: M = Manter a via, A = Alterar a via na base de dados CvioCyc e R = Remover a via; ORFs alt = Número da(s) ORF(s) que foi alterada.

#	Nome da via	NR	NP	VI	VE	Ação	ORFs alt.
30	Biossíntese de Riboflavina e FMN e FAD	9	5	205	1	M A	
31	Biossíntese de Sacarose	5	2	2	4	R	
32	Biossíntese de Treonina a partir de Hemoserina	2	1	51	0	M O	CV0776
33	Biossíntese de UDP-N-Acetilglucosamina	4	2	52	2	M O	CV2172 CV0674 CV3103
34	Biossíntese e Degradação de Trehalose – baixa osmolaridade	4	2	52	2	M	
35	Ciclo do TCA - Respiração Aeróbica	10	9	9	1	M O A	CV1074 CV1072
36	Ciclo do TCA variante VIII	17	0	13	4		
37	Degradação de Alanina I	2	1	2	0	M A	
38	Degradação de Arginina II	7	3	3	5	R	
39	Degradação de Arginina VI	5	3	4	1	M A O	CV1499
40	Degradação de Frutose - Anaeróbica	10	9	59	1	M	
41	Degradação de Glutamato I	3	2	2	2	M	
42	Degradação de Glutamato IV	4	2	2	2	M	
43	Degradação de Glutamato VII	7	5	5	2	M	
44	Degradação de Glutamato VIII	7	5	5	2	M	
45	Degradação de Isoleucina I	10	4	4	6	R O	CV1764 CV1762
46	Degradação de Isoleucina III	9	4	4	5	M	

Onde NR= Número total de reações existentes na via; NP= Número de reações identificadas no genoma; VI = Valor final da variável INCLUIR; VE = Valor final da variável EXCLUIR; Ação = Ação que foi tomada sobre a via metabólica, seu valor pode variar entre: M = Manter a via, A = Alterar a via na base de dados CvioCyc e R = Remover a via; ORFs alt = Número da(s) ORF(s) que foi alterada.

#	Nome da via	NR	NP	VI	VE	Ação	ORFs alt.
47	Degradação de Lactose 1	5	1	1	54	R	
48	Degradação de Lactose 4	6	3	3	53	R	
49	Degradação de Mandelato	6	1	1	51	R O	CV3101
50	Degradação de Mandelato I	5	1	1	5	R	
51	Degradação de Sacarose I	3	1	1	3	R	
52	Degradação de Sacarose III	7	3	3	6	R O	CV3795
53	"De novo" Biossíntese de Nucleotídeos de Purina I	28	0	20	8	R	
54	Fermentação	12	9	9	1	M A	
55	Fermentação de Glutamato – A via Hidroxiglutarato	5	1	51	4	M	
56	Fotorespiração	9	4	53	7	M O	CV4235 CV0901
57	Gluconeogênese	13	1	11	2	M	
58	Ramo Não-Oxidativo de Pentose Fosfato	5	4	4	2	M O A	CV0191
59	Supervia da Biossíntese de Serina e Glicina	5	2	3	2	M O	CV1094
60	Via de Carregamento de tRNA	20	9	19	1	M A	
61	Via de Salvação de Ribonucleotídeos Pirimidínicos	8	4	4	4	M O A	CV3471

Onde NR= Número total de reações existentes na via; NP= Número de reações identificadas no genoma; VI = Valor final da variável INCLUIR; VE = Valor final da variável EXCLUIR; Ação = Ação que foi tomada sobre a via metabólica, seu valor pode variar entre: M = Manter a via, A = Alterar a via na base de dados CvioCyc e R = Remover a via; ORFs alt = Número da(s) ORF(s) que foi alterada.

4.2.2 Alterações sugeridas para a re-anotação genômica

Algumas alterações nas anotações genômicas são propostas neste trabalho e estão listadas na Tabela 5 , a qual contém o número da ORF, a anotação original e as alterações propostas. As argumentações e comentários a respeito destas alterações estão discutidos na seção 4.2.2.1.

Tabela 5: Listagem das ORFs alteradas

#	ORF	Anotação Original	Anotação Proposta
8	CV4033	/product=”N-acetylglucosamine-6-phosphate 2-epimerase/N-acetylglucosamine-6-phosphatase”	/product=”UDP-N-acetylglucosamine 2-epimerase”
8	CV4034	/product=”Probable aminotransferase” Sem /gene	/product=”Predicted pyridoxal phosphate-dependent enzyme apparently involved in regulatory of cell wall biogenesis /gene=WecE
10	CV0690	/product=“dihydrokaempferol 4-reductase” /ECumber=”1.1.1.219” /gene=hpnA	/product=“Putative oxidoreductase” Sem /EC_number Sem /gene
16	CV1562	/product="precorrin-3B C17-methyltransferase" /gene=”cbiJ”	/product="Probable bifunctional: precorrin-3 methyltransferase and precorrin-6 reductase " /gene=”cbiJH”
16	CV1568	/product=”precorrin 6y methylase methyltransferase protein” /EC_number="2.1.1.13”	/product=”Precorrin-6Y C(5,15)-methyltransferase (decarboxylating) /EC_number="2.1.1.132”
16	CV1556	/product=”cobyrinic acid A,C-diamide synthase” /EC_number="6.3.1.-"	/product=”Hydrogenobyrinic acid a,c-diamide synthase (glutamine-hydrolyising) /EC_number="6.3.5.9”
16	CV1555	/product="probable oxidoreductase protein” Sem EC_number	/product="Putative cob(II)yrinic acid a,c-diamide reductase" /EC_number="1.16.8.1"
16	CV1576	/product=”cobyrinic acid A,c-diamide synthase /EC_number="6.3.1.-"	/product=”Adenosylcobyric acid synthase (glutamine-hydrolysing)” /EC_number="6.3.5.10”
16	CV1575	/product=”cobalamin biosynthetic protein” /EC_number="2.6.1-“	/product=”Adenosylcobinamide-phosphate synthase” /EC_number="6.3.1.10”
16	CV0491	/product=”cobalamin (5´phosphate) synthase Sem EC_number	/product=”Adenosylcobinamide-GDP ribazoletransferase” /EC_number="2.7.8.26”
17	CV1258	/product="4-diphosphocytidyl-2C-methyl-	/product="2-C-methyl-D-erythritol 4-

#	ORF	Anotação Original	Anotação Proposta
		D-erythritol synthase" Sem EC_number	phosphate cytidylyltransferase" /EC_number="2.7.7.60"
17	CV4059	/product="probable isopentenyl monophosphate kinase(4-diphosphocytidyl-2C-methyl-D-erythritol kinase)"	/product="4-diphosphocytidyl-2-C-methyl-D-erythritol kinase"
17	CV1259	/product="2-C-methyl-d-erythritol-2,4-cyclodiphosphate synthase" Sem EC_number	/product="2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase" /EC_number="4.6.1.12"
17	CV3536	/product="hypothetical protein" Sem EC_number Sem gene	/product="4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase" /EC_number="1.17.4.3" /gene="ispG"
17	CV3567	/product="lytB protein" Sem EC_number /gene="lytB"	/product="4-hydroxy-3-methylbut-2-enyl diphosphate reductase" /EC_number="1.17.1.2" /gene="ispH"
20	CV4039	/product=”probable phosphatidylglycerol phosphate synthetase”	/product=”CDP-diacylglycerol-glycerol-3-phosphate 3-phosphatidyltransferase”
23	CV1568	/product=”precorrin-6Y methylase methyltransferase protein” /EC_number = “2.1.1.13”	/product=”precorrin-6Y C5,15-methyltransferase (decarboxylating) /EC_number = “2.1.1.132”
24	CV3328	/product=“hypothetical protein" Sem EC_number Sem gene	/product=“3-deoxy-manno-octulosanate-8-phosphate phosphatase” /EC_number=”3.1.3.45” /gene=”KdsC”
24	CV0225	Sem gene definido	/gene=”kdtA”
26	CV4343	/product=”peptidoglycan glycosyltransferase” /EC_number=”2.4.1.129”	/product=”UDP-N-acetylglucosamine: N-acetylmuramoyl-(pentapeptide)pyrophosphoryl-undecaprenol N-acetylglucosamine transferase” /EC_number=”2.4.1.227”
27	CV4337	/product=”UDP-3-O-[3-hydroxymyristoyl]	/product=”UDP-3-O-acyl N-

#	ORF	Anotação Original	Anotação Proposta
		N-acetylglucosamine deacetylase”	acetylglucosamine deacetylase”
27	CV2206	/product=”UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acetyltransferase (firA protein	/product=”UDP-3-O-[3- hydroxymyristoyl] glucosamine N-acetyltransferase”
29	CV3433	/product=”probable glutamyl-tRNA synthetase- synthetase-related protein” /EC_number=”6.1.1.17”	/product=”COG0008: Glutamyl- and glutaminyl- tRNA synthetases” Sem /EC_number
29	CV2819	/product="probable glutamate-1-semialdehyde aminotransferase” /EC_number=”5.4.3.8”	/product=”COG0001: Glutamate-1-semialdehyde aminotransferase" Sem /EC_number
29	CV0813	/product="uroporphyrin-III C-methyltransferase" /EC_number="2.1.1.107" /gene="cobA2"	/product="Siroheme synthase [Includes: Uroporphyrin-III C-methyltransferase; precorrin-2 oxidase; ferrochelatase]” /EC_number="2.1.1.107" /EC_number=”1.3.1.76” /EC_number="4.99.1.4" /gene=”cysG”
32	CV0776	/product=”ketohexokinase” /EC_number=”2.7.1.3”	/product=“Hemoserine kinase” /EC_number=”2.7.1.39”
33	CV2172	/EC_number=”5.4.2.8” /EC_number=”5.4.2.2”	/EC_number=”5.4.2.8”
33	CV0674	/product=”bifuncional:UDP-N-acetylglucosamineglucose-1-phosphate thymidylyltransferase; Glucosamine-1-phosphate" /EC_number=“2.7.7.23” /EC_number=“2.7.7.24”	product=”UDP-N-acetylglucosamine pyrophosphorylase” /EC_number=”2.7.7.23”
33	CV3103	/product="probable Bifuncional:UDP-N-acetylglucosamineglucose-1-phosphatethymidylyltransferase; Glucosamine-1-phosphate” /EC_number="2.7.7.23” /EC_number="2.7.7.24"	/product=”Probable glucose-1- phosphate thymidylyltransferase” Sem /EC_number
35	CV1074	/product=”dihydrolipoamide	/product=”Putative dihydrolipoamide

#	ORF	Anotação Original	Anotação Proposta
		dehydrogenase” /gene=”lpdA2”	dehydrogenase E3 component” /gene=”lpdA”
35	CV1072	/EC_number=”2.3.1.6”	/EC_number=”2.3.1.61”
39	CV1499	/product="succinylglutamate 5-semialdehyde dehydrogenase" /gene=”aruD”	/product=”Succinylglutamic semialdehyde dehydrogenase” /gene=”astD”
45	CV1764	/product=“probable propionyl-CoA carboxylase (beta subunit)” /EC_number=”6.4.1.3”	/product=” Acetyl-CoA carboxylase, carboxyltransferase component (subunits alpha and beta)” Sem /EC_number
45	CV1762	/product=”probable acyl-CoA carboxylase subunit” /EC_number=”6.3.4.14”	/product=”COG4770: Acetyl/ propionyl-CoA carboxylase, alpha subunit” Sem /EC_number
49	CV3101	/product=”benzoylformate decarboxylase” /EC_number=”4.1.1.7”	/product=”COG0028: Thiamine pyrophosphate-requiring enzymes [acetolactate synthase, pyruvate dehydrogenase(cytochrome), glyoxylate carboligase, phosphonopyruvate decarboxylase]" Sem /EC_number
52	CV 3795	/product=”phosphoglucosamine mutase”	/product=”Phosphoglucomutase”
56	CV 4235	/product=”probable (S)-2-hydroxy-acid oxidase” /EC_number=“1.1.3.15”	/product=”Putative oxidoreductase” Sem /EC_number
56	CV 0901	/product=”probable oxidoreductase” /EC_number=”1.1.3.15”	/product=”Putative oxidoreductase” Sem /EC_number
58	CV 0191	Sem /EC_number	/EC_number=”2.2.1.1”
59	CV 1094	/product=”pyridoxal-phosphate-dependent aminotransferase” /EC_number=”2.6.1.44” /EC_number=”2.6.1.51”	/product=”COG1104: Cysteine sulfinate desulfinase//cysteine desulfurase and related enzymes” Sem /EC_number
61	CV 3471	/product=”probable cytidine deaminase” /EC_number=”3.5.4.5”	/product=”Cytosine deaminase” /EC_number=”3.5.4.1”

4.2.2.1 Descrição das alterações nas ORFs e na base de dados CvioCyc

Observe-se que nos diagramas das vias, que são mostrados juntamente com as descrições das alterações, quando os nomes das enzimas referentes às reações aparecem, é porque elas foram encontradas pelo Pathway Tools no genoma da *Chromobacterium violaceum*, do contrário o software não as encontrou e não são mostradas nas linhas que indicam a reação.

A ferramenta BLASTP foi utilizada sistematicamente para a análise das ORFs. Os resultados de BLASTP apresentam variáveis estatísticas "Expect"(E), "Identities"(I), "Positives"(P), "Gaps"(G), que são referenciadas nas descrições abaixo apenas por suas iniciais, previamente descritas no capítulo de revisão bibliográfica (item 2.15.1).

• Biossíntese de Antígeno Enterobacterial Comum (Via # 8).

A Figura 35 apresenta o desenho da via metabólica Biossíntese de Antígeno Enterobacterial Comum juntamente com as enzimas encontradas pelo Pathway Tools.

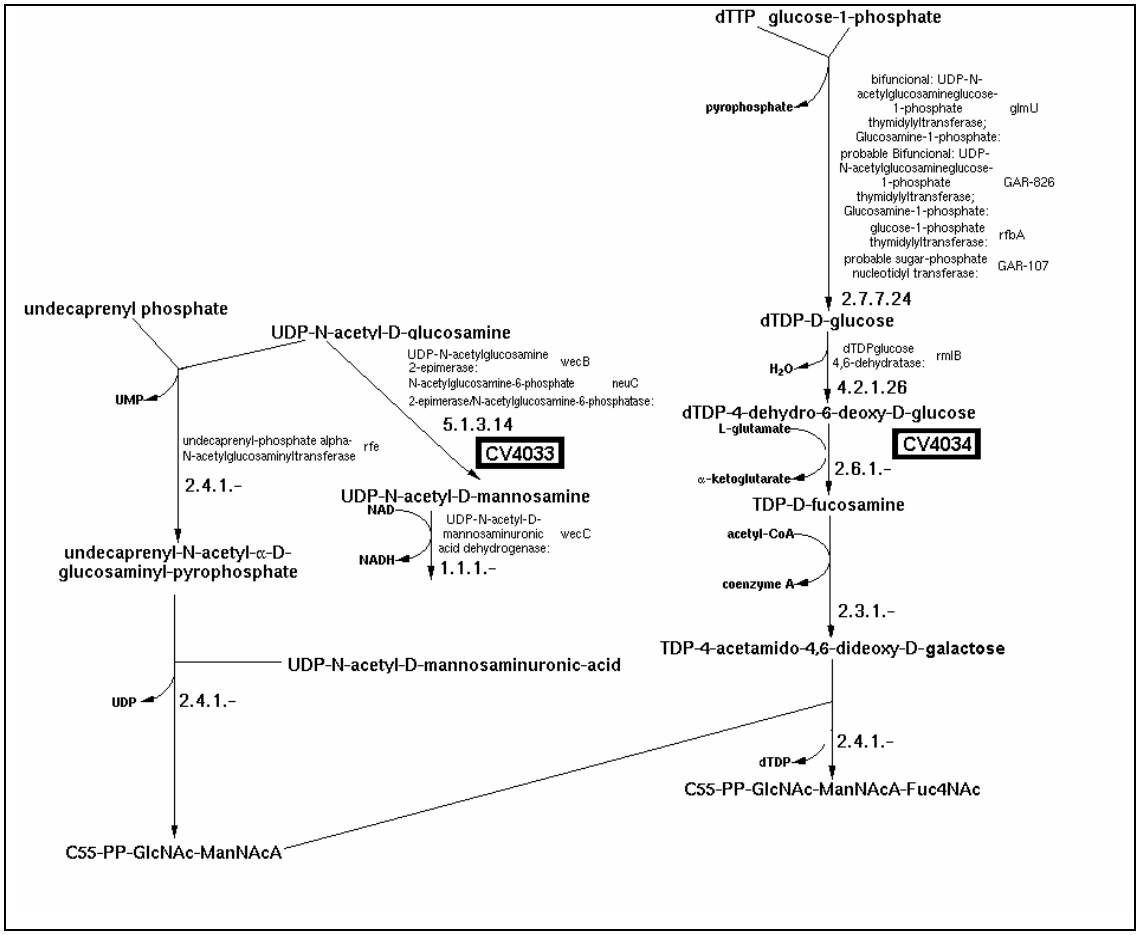


Figura 35: Diagrama da Biossíntese de Antígeno Enterobacterial Comum e as ORFs alteradas em destaque. Adaptado do software Pathway Tools.

➢ ORF CV4033:

Esta anotação não está condizente pois o produto anotado não se refere ao ECnumber anotado, conforme constatado a partir do exame das bases BRENDA e ENZYME. O EC_number 5.1.3.14 é referente à enzima "UDP-N-acetylglucosamine 2-epimerase", produto sugerido. Os quatro melhores resultados do BLASTP referenciam o produto sugerido. O melhor resultado foi uma seqüência de 390 aminoácidos (YP_148979.1), I=200/383 (52%), P=277/382 (72%), G=0/383 (0%) e E=$2e^{-107}$. A anotação usou o quinto melhor resultado do BLAST, uma seqüência de 390 aminoácidos

(CAE06964.1), onde I=188/382 (49%), P=257/382 (67%), G=1/382 (0%) e E=$1e^{-96}$. Outra indicação é que há muitas seqüências resultantes com o produto sugerido, enquanto que para o anotado há poucas.

- ORF CV4034:

O produto sugerido, "Predicted pyridoxal phosphate-dependent enzyme apparently involved in regulatory of cell wall biogenesis" foi listado no relatório "Identification of reactions from pathway holes in C. violaceum.html", para a ORF CV4034 com probabilidade de 70% (P=0,70). Um dado relevante é que a CV4034 está no mesmo operon de outros genes desta via. Além disso, o BLASTP contra os dados do genoma da *C. violaceum*, utilizando a seqüência AAO09314 (380 aminoácidos), uma isoenzima descrita como "pyridoxal phosphate-dependent enzyme apparently involved in regulatory of cell wall biogenesis", produziu resultados relevantes para a seqüência CV4034 (383 aminoácidos), I=231/378 (61%), P=291/378 (76%), G=2/378 (0%) e E=$6e^{-129}$. Esta ORF CV4034 faz a catálise da reação 2.6.1.- nesta via.

• Biossíntese de Antocianina (Via #10).

O Diagrama da via metabólica de Biossíntese de Antocianina é apresentado na Figura 36 juntamente com a enzima encontrada pelo Pathway Tools.

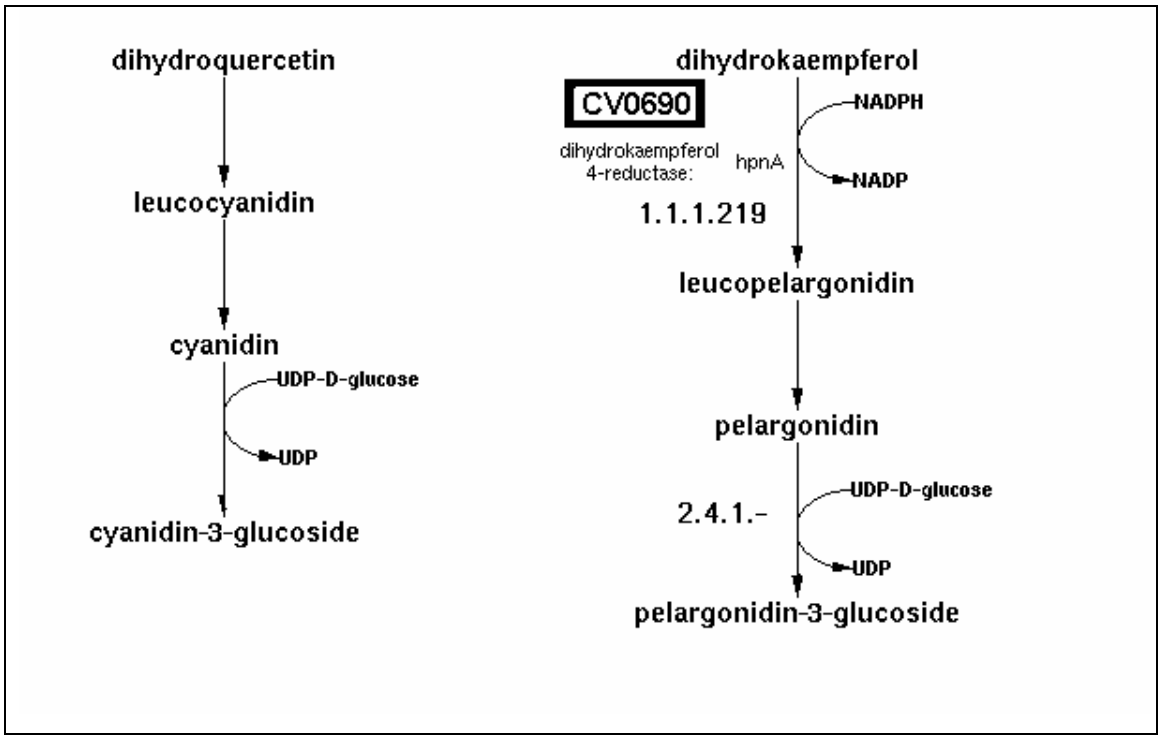


Figura 36: Diagrama da Biossíntese de Antocianina e a ORF alterada em destaque. Adaptado do software Pathway Tools.

➢ ORF CV0690:

Esta anotação não está condizente, pois há resultados melhores do BLAST. O melhor resultado do BLASTP foi uma seqüência encontrada na *Pseudomonas putida* KT2440, de 349 aminoácidos; destes, 329 foram alinhados, e 169 são idênticos (51%); 230 foram positivos (69%) com 4 espaçamentos ("gaps*"),* significando 1%, cujo produto foi anotado como "oxidoreductase, putative", enquanto que a seqüência escolhida para anotação apresentou 338 aminoácidos; destes, 335 puderam ser alinhados e 140 foram idênticos (41%), e 203 foram positivos (60%) com 9 espaçamentos (2%) do organismo *Gloeobacter violaceus* PCC 7421, filogeneticamente mais distante da *C. violaceum* do que a *P. putida*. Além disso foi retirada a palavra "probable" do original, e não há especificação de EC_number e nem de nome de gene.

A via de biossíntese da antocianina é muito bem anotada para organismos eucariotos (*eukariota viridiplantae*). Há seis anotações para bactérias, todas não experimentais, e 42 anotações para eucariotos. A enzima anotada só está presente na via Biossíntese de Antocianina, sendo a única identificada no genoma para esta via. Com esta re-anotação esta via não mais estará presente para a *C. violaceum*, o que parece bastante provável.

- Biossíntese de Cobalamina, via aeróbica (Via #16).

O Diagrama da via metabólica de Biossíntese de Cobalamina é apresentado na Figura 37 juntamente com as enzimas encontradas pelo Pathway Tools.

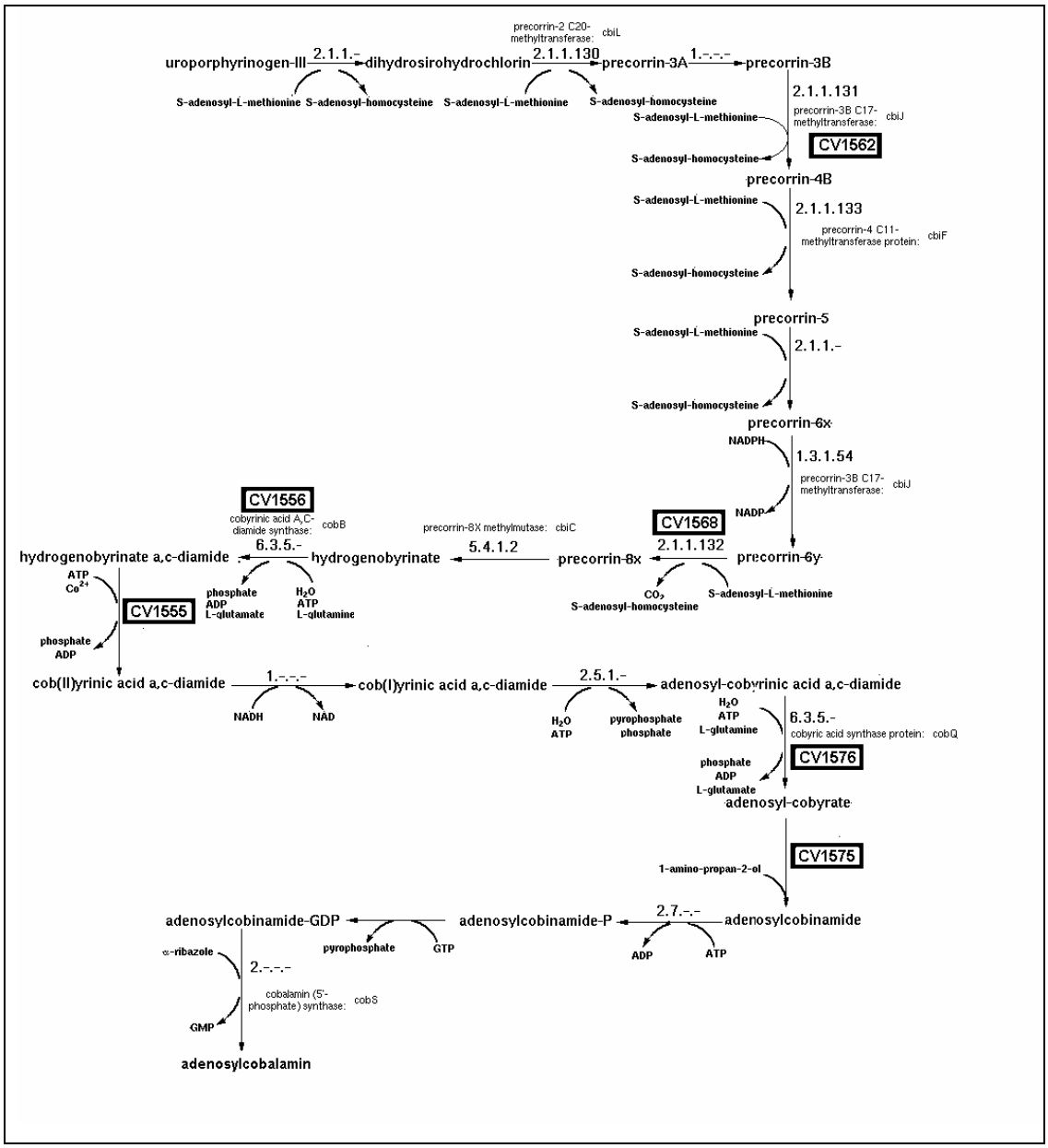


Figura 37: Diagrama da Biossíntese de Cobalamina, via aeróbica e as ORFs alteradas. Adaptado do Software Pathway Tools.

- ORF CV1562:

A anotação tem dois EC_numbers, porém no /product não faz referência à bifuncionalidade da enzima. Os dois melhores resultados do BLASTP são referentes a este produto bifuncional. O primeiro melhor resultado é uma

seqüência de 519 aminoácidos, onde 508 foram alinhados com a seqüência em questão, de 514 aminoácidos. Os seguintes resultados foram encontrados: I=383/508 (75%), P=426/508 (83%), G=0/508 (0%) e E=0. Há também dois nomes de genes por causa da bifuncionalidade desta proteína.

- ORF CV1568:

O BLASTP executado para a CV1568 resultou na enzima com EC_number 2.1.1.132, cujo produto é o anotado, "precorrin-6Y methylase methyltransferase". Na própria anotação no BRgene onde está o resultado do BLASTP usado, NP_522185 do organismo *Ralstonia solanacearum*, o EC_number está anotado como 2.1.1.132. Este é um provável erro de digitação da anotação.

- ORF CV1556:

Há uma entrada nova no BRENDA para o nome de enzima "cobyrinic acid a,c-diamide synthase", com o EC_number 6.3.5.9. Todavia, as informações no BRENDA ainda estão incompletas ("full data in preparation"). É possível que quando o gene (ORF CV1556) foi originalmente anotado, ainda não havia uma classificação EC para esta enzima.

- ORF CV1555:

O relatório "Identification of reactions from pathway holes in C. violaceum.html" da *C. violaceum* sugere que a CV1555 poderia codificar a enzima catalisadora da 11ª reação da via de biossíntese de cobalamina, a enzima "cob(II)yrinic acid a,c-diamide reductase". Há uma nova entrada no BRENDA com a definição desta enzima. Para o Pathway Tools, ela é a enzima de EC_number 1.-.-.-. Como esta enzima não está anotada no genoma da *C. violaceum*, foi utilizada uma seqüência desta enzima porém de um outro organismo (a seqüência Q52685, para a qual existe evidência experimental) anotada no GenBank para se realizar um BLASTP contra o genoma da *C violaceum*. O melhor resultado do BLASTP desta enzima contra o genoma da *C. violaceum* resulta na ORF CV1555 com I=74/194 (36%), P=99/194 (47%), E=9e-24 e Gaps=2/194 (1%), resultado considerado aceitável; além disso, esta ORF faz parte do mesmo operon de outros genes desta via, o que é uma forte evidência (computacional) de que fazem parte da mesma via.

- ORF CV1576:

Há uma nova entrada no BRENDA para esta enzima definindo o EC_number 6.3.5.10. Contudo, novamente, as informações no BRENDA ainda estão incompletas ("full data in preparation").

- ORF CV1575:

O BLASTP para esta ORF tem como melhores resultados o produto anotado, porém não estão relacionados ao EC_number anotado. No BRENDA a enzima CbiB, que é sinônimo da enzima anotada, consta em preparação com o EC_number relacionado 6.3.1.10.

- ORF CV0491:

No BLASTP foi encontrado 1 domínio "COG0368, CobS, Cobalamin-5-phosphate synthase [Coenzyme metabolism]", com 126 aminoácidos alinhando 97,6% com a seqüência pesquisada. O resultado condiz com o produto anotado, porém no BRENDA há uma nova entrada para esta enzima com o EC_number 2.7.8.26.

- Alteração na via da base CvioCyc:

Foram incluídas na CvioCyc as seguintes reações: a reação 2.1.1.- (CV0813 que foi re-anotada depois da CvioCyc ser criada) que é sexta reação desta via, a reação 2.5.1.17 (CV1557), que é a 13ª reação, a reação 6.3.5.10 (CV1576) que é a 14ª reação. Foram também incluídas as seguintes enzimas: a CV0495, que catalisa a 16ª e 17ª reações desta via, e a CV0491 catalisadora da 18ª reação desta via.

- Biossíntese de Difosfato de Isopentanil – Mevalonato Independente (Via #17).

O Diagrama da via metabólica de Biossíntese de Difosfato de Isopentanil – Mevalonato Independente é apresentado na Figura 38 juntamente com a enzimas encontradaB pelo Pathway Tools.

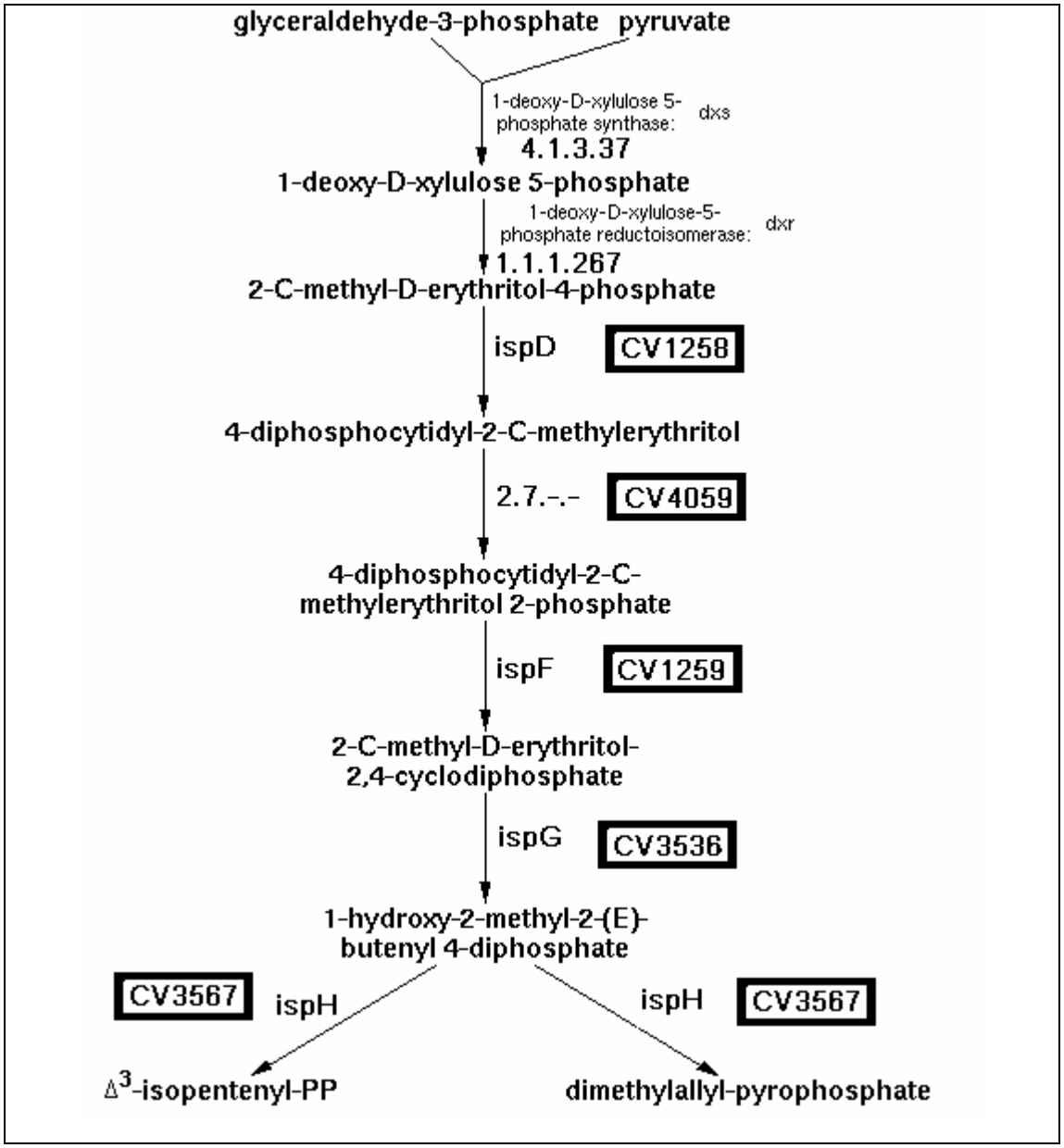


Figura 38: Biossíntese de Difosfato de Isopentanil – Mevalonato Independente. Adaptado do Software Pathway Tools.

- ORFs CV1258 CV4059 CV1259 CV3536 CV3567:

Após a análise das ORFs, verificou-se que havia novas entradas para as mesmas e que combinavam com as alterações que seriam propostas neste trabalho. As ORFs ausentes na via foram listadas no relatório "Identification of reactions from pathway holes in C. violaceum.html". Estas ORFs foram re-anotadas em 25 outubro de 2004 através das seqüências no NCBI : Q7NYL6, Q7NQS8, Q7NYL5, Q7NS88, Q7NS59, respectivamente.

- Alteração na via da base CvioCyc:

Foi incluída a re-anotação realizada para que esta via ficasse completa. O Pathway Tools tinha encontrado duas reações presentes, de um total de sete. Com a re-anotação das enzimas no NCBI, todas as enzimas desta via passaram a figurar como existentes no genoma da *C. violaceum*.

• Biossíntese de Enterobactina (Via #18).

O Diagrama da via metabólica de Biossíntese de Enterobactina é apresentado na Figura 39 juntamente com as enzimas encontradas pelo Pathway Tools.

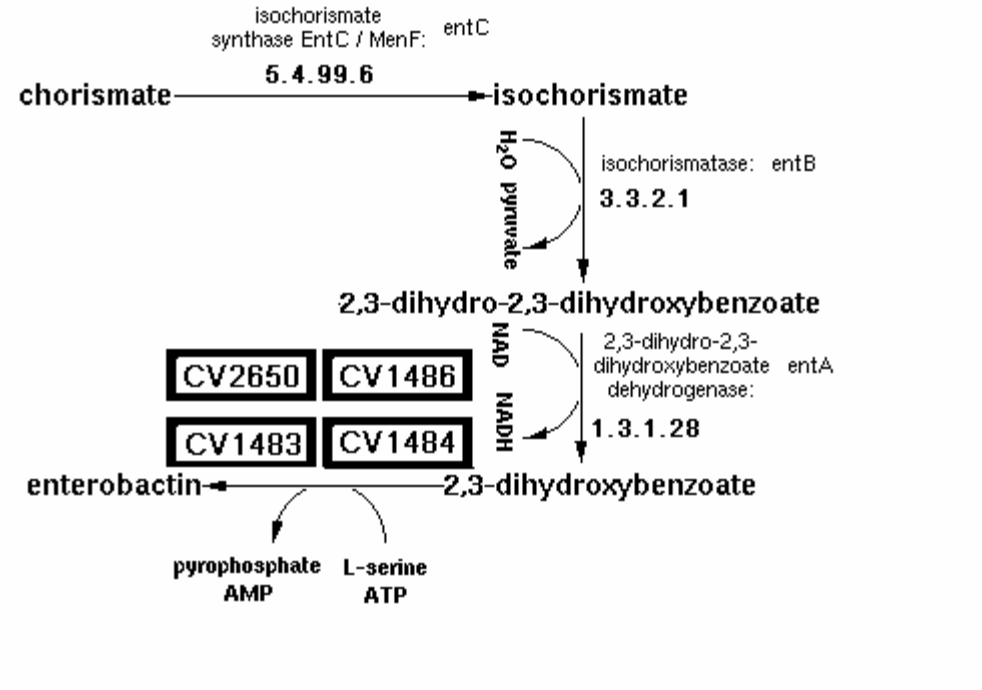


Figura 39: Diagrama da Biossíntese de Enterobactina com as ORFs alteradas. Adaptado do Software Pathway Tools.

➢ Alteração na via da base CvioCyc:

Foi incluído o complexo EntB, EntD, EntF, EntE através das ORFs correspondentes já anotadas, CV1483, CV2650, CV1486, CV1484. Com esta inclusão, a via metabólica da Biossíntese de Enterobactina ficou completa. É importante observar que o Pathway Tools não faz a inclusão quando a reação é catalisada por um complexo enzimático; deste modo, cabe aos usários do software, fazê-la manualmente.

• Biossíntese de Fosfolipídio (Via # 20).

O Diagrama da via metabólica de Biossíntese de Fosfolipídio é apresentado na Figura 40 juntamente com as enzimas encontradas pelo Pathway Tools.

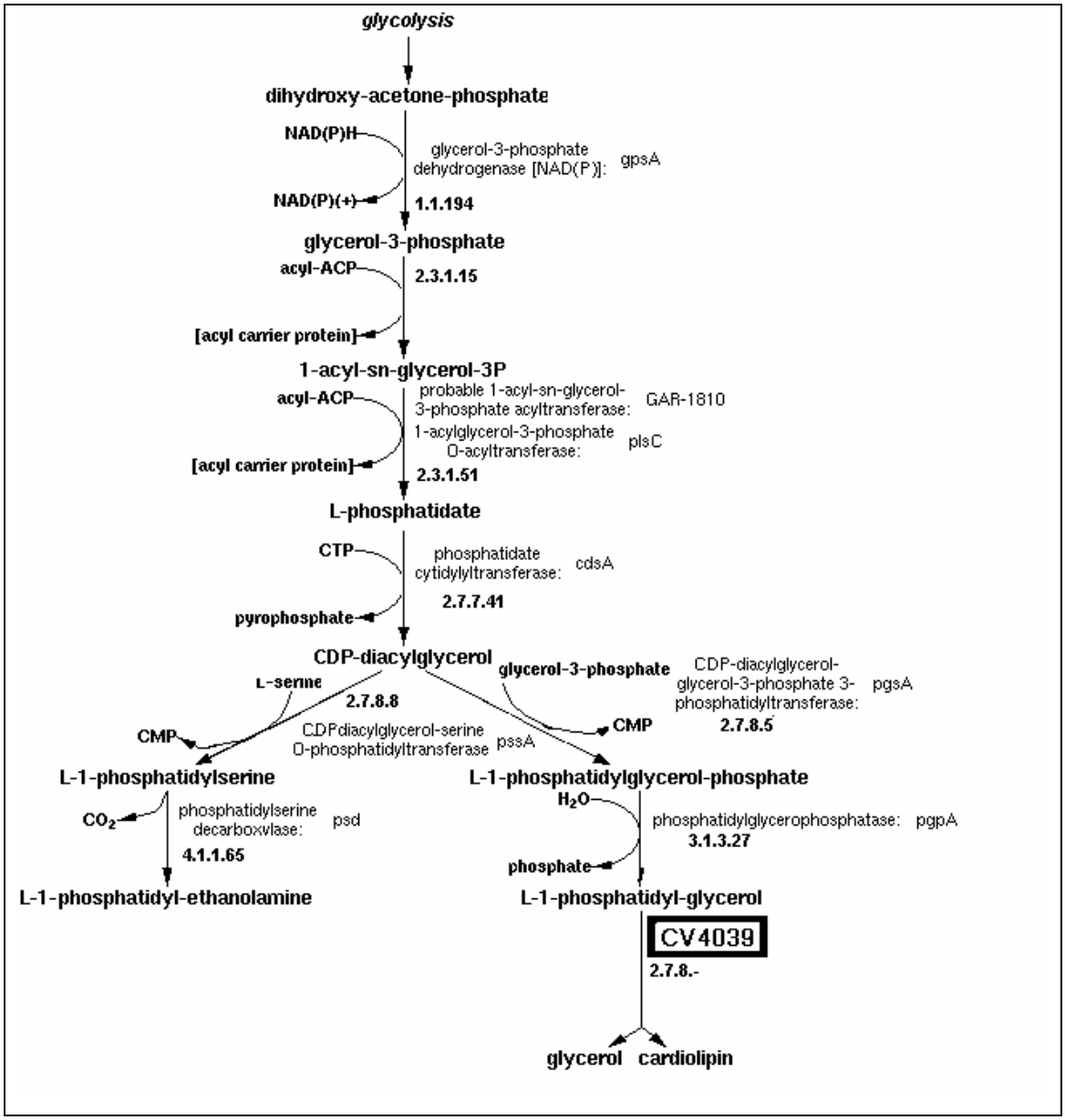


Figura 40: Diagrama da Biossíntese de Fosfolipídio e a ORF alterada. Adaptado do software Pathway Tools.

➢ ORF CV4039:

No resultado do BLASTP foram encontrados 3 domínios, 2 domínios "cd00138, PLDc, Phospholipase D. Active site motifs", os dois com 176

aminoácidos cada e com 64% e 61% de alinhamento cada um, e 1 domínio "COG1502,Cls, Phosphatidylserine/phosphatidylglycerophosphate/cardiolipin synthases and related enzymes [Lipid metabolism]" de 438 aminoácidos que alinhou 92,7% com a seqüência consultada. O domínio COG1502 tem o alinhamento mais significativo e abrange as duas funções, tanto a anotada quanto a sugerida. Porém, na análise dos resultados de similaridade das sequências, verifica-se que não há nenhuma seqüência idêntica à anotação original, e a maioria dos resultados é anotada com o COG1502 ou com o produto sugerido, ou seja, "cardiolipin synthase" Para o resultado, "cardiolipin synthase", os valores estatísticos foram: I=128/379 (33%), P=199/379 (52%), Gaps=9/379 (2%), e $E=8e^{-45}$; estes valores são melhores do que os obtidos para a seqüência similar a anotada originalmente, "probable phosphatidylglycerol phosphate synthase", que provavelmente foi baseada nas seqüências resultantes anotadas com o domínio "COG1502" acima citado, pois não há nenhuma outra seqüência anotada como a anotação original.

• Biossíntese de Homoserina e Metionina (Via # 23).

O Diagrama da via metabólica de Biossíntese de Homoserina e Metionina é apresentado na Figura 41 juntamente com as enzimas encontradas pelo Pathway Tools.

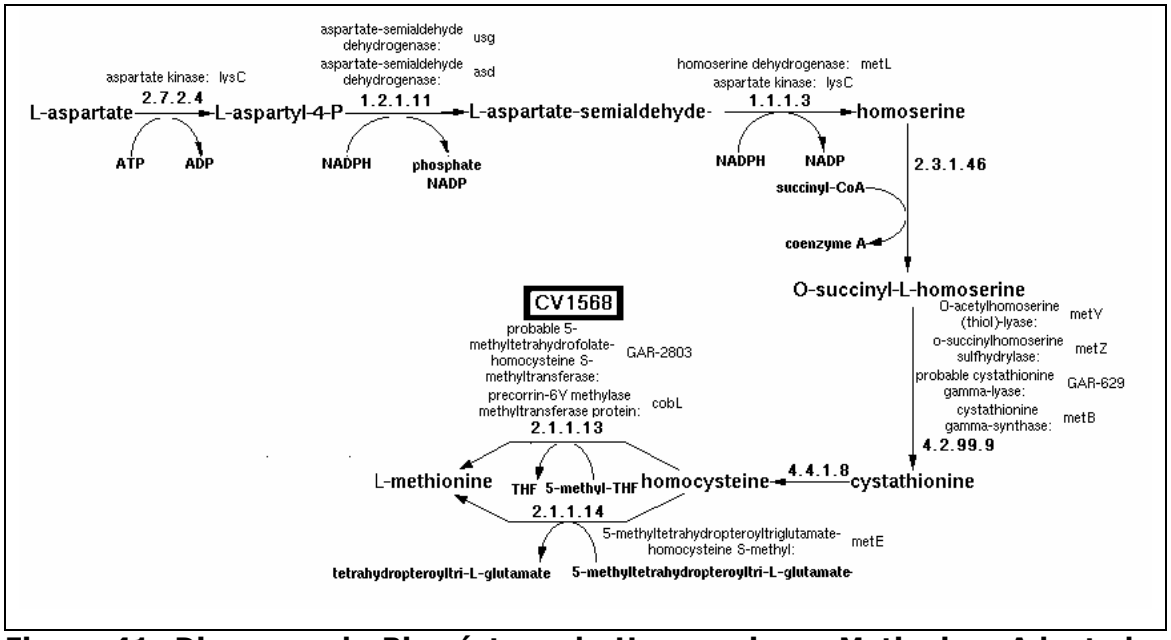


Figura 41: Diagrama da Biossíntese de Homoserina e Metionina. Adaptado do software Pathway Tools.

➢ ORF CV1568:

Foi observado um erro de anotação, que já foi detectado durante a análise da via metabólica de Biossíntese de Cobalamina - via aeróbica; porém, ele aparece novamente na análise desta via, pois a enzima com o EC_number 2.1.1.13, anotada erroneamente na CV1568, é uma das enzimas que faz catálise nesta via metabólica.

- Biossíntese de KDO -- incluindo a transferência de lipídio IV A (Via # 24).

O Diagrama da via metabólica de Biossíntese de KDO -- incluindo a transferência de lipídio IV A é apresentado na Figura 42 juntamente com as enzimas encontradas pelo Pathway Tools.

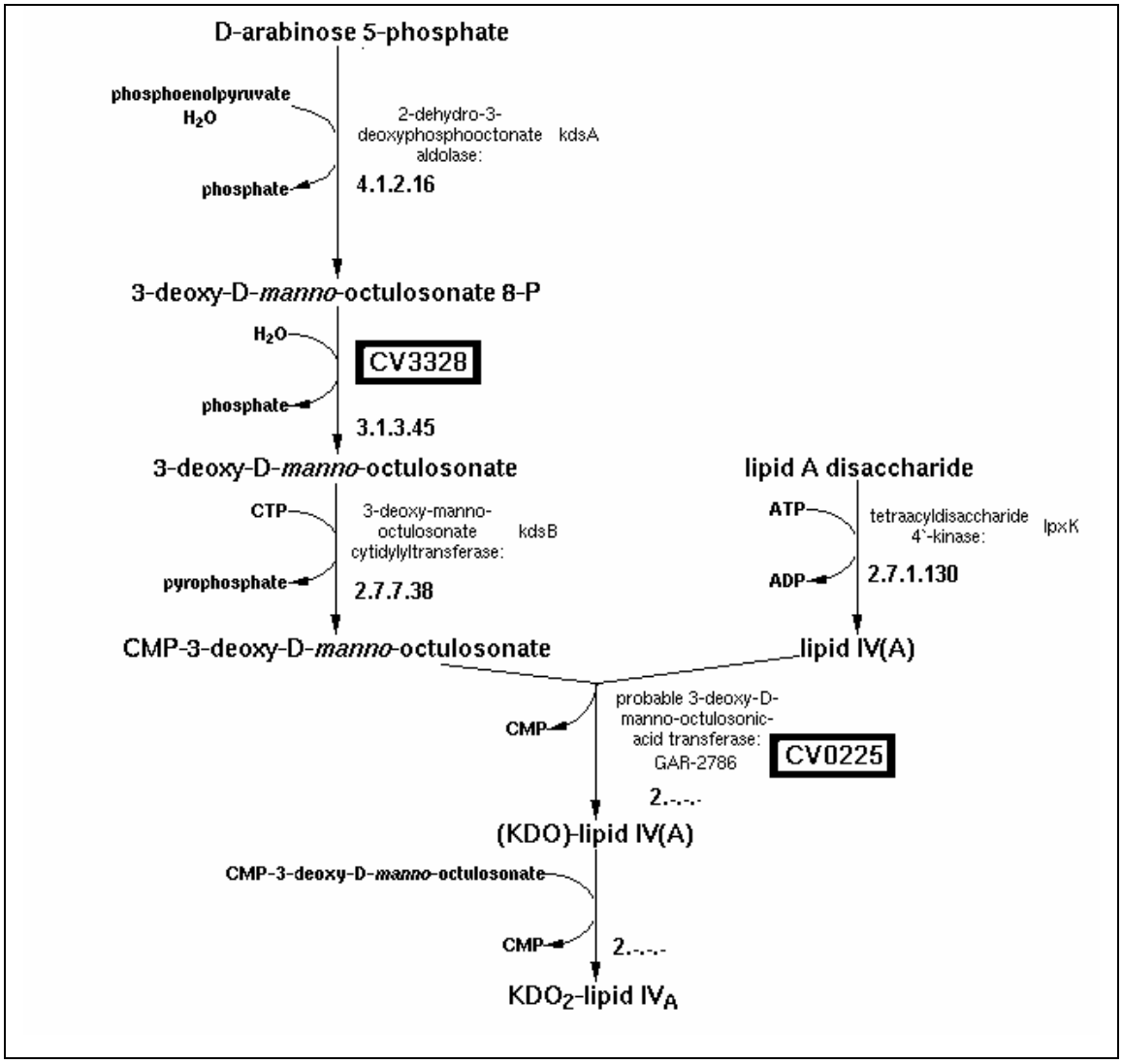


Figura 42: Diagrama da Biossíntese de KDO -- incluindo a transferência de lipídio IV A e as ORFs alteradas. Adaptado do Software Pathway Tools.

- ORF CV3328:

Esta ORF foi sugerida pelo Pathway Tools como possível candidata a ser a enzima "3-deoxy-manno-octulosanate-8-phosphate phosphatase". Os resultados do BLASTP são bastante aceitáveis para seqüências com este produto de gene; Foram alinhados 159 de 174 aminoácidos da CV3328, onde: I=45%, P=67%, E=$7e^{-35}$ e Gaps=0%. Com esta ORF alterada, fica

faltando apenas uma reação para esta via. Várias seqüências resultantes, que apresentaram o produto de gene conforme sugerido, apresentaram também o ECnumber 3.1.3.45 e o nome de gene *KdsC.*

- ORF CV0225:

Esta ORF tem como produto do gene anotado "probable 3-deoxy-D-manno-octulosonic-acid transferase", a enzima que catalisa a quinta reação desta via (2.-.-.-). O nome do gene não foi anotado apesar de que o BLASTP mostra vários bons resultados com seqüências com o mesmo produto do gene, anotado porém está definido o nome do gene, que é *kdtA*.

- Alteração na base CvioCyc:

Como o NC-IUBMB transferiu a reação 4.1.2.16 para 2.5.1.55 (3-deoxy-8-phosphooctulonate synthase), este EC_number foi devidamente re-anotado. Nota-se que, para a última reação, foi necessário informar ao Pathway Tools que quem a catalisa é a mesma enzima "3-deoxy-D-manno-octulosonic acid transferase" da reação anterior, de acordo com as informações obtidas por consulta à base BRENDA e também da base EcoCyc.

• Biossíntese de Peptidioglicana (Via # 26).

O Diagrama desta via metabólica é apresentado abaixo na Figura 43.

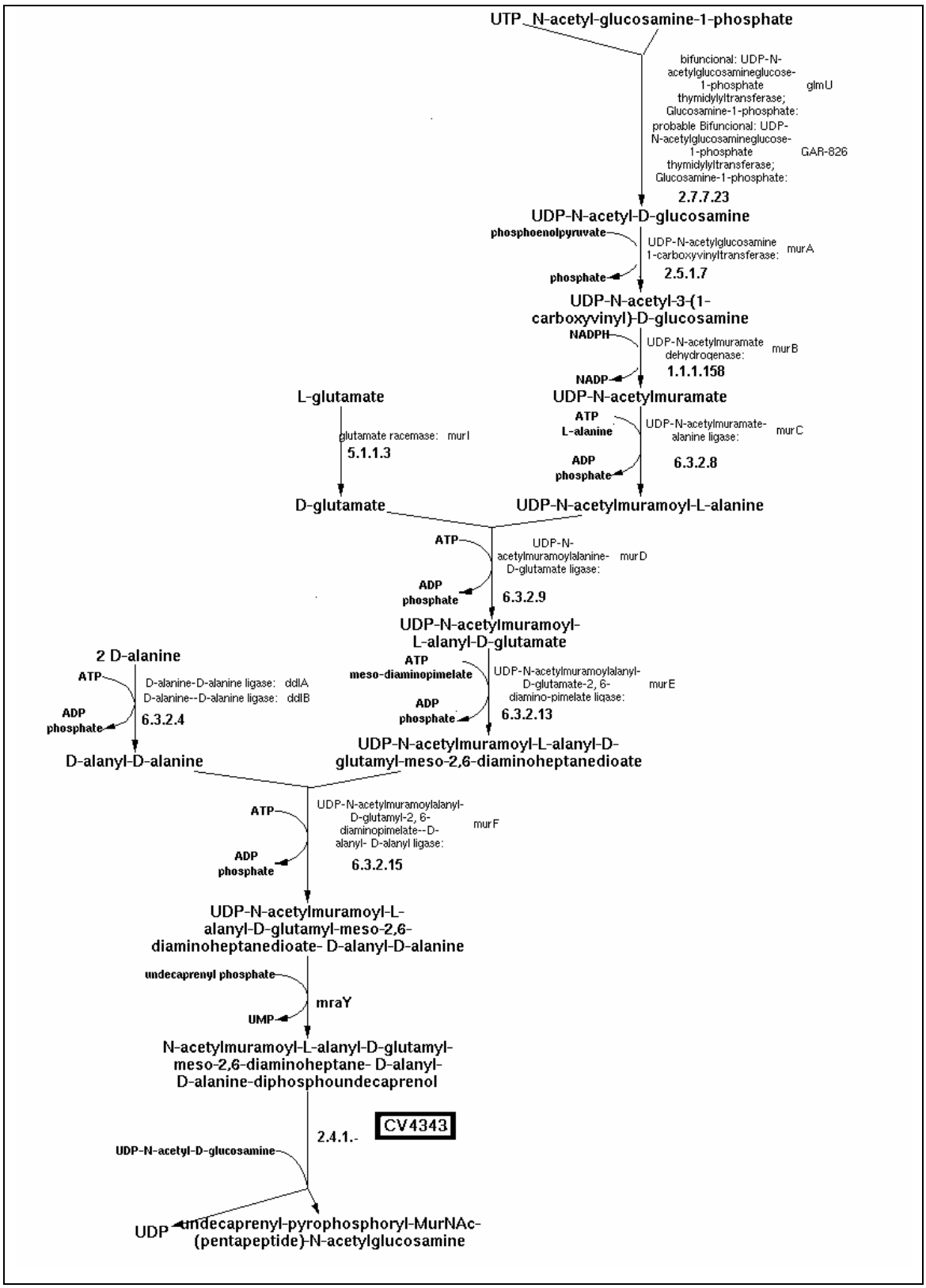


Figura 43: Diagrama da Biossíntese de Peptidioglicana e as ORFs alteradas. Adaptado do software Pathway Tools.

- ORF CV4343:

Os resultados do BLASTP, na grande maioria, têm como produto ou "COG0707:UDP-N-acetylglucosamine:LPS N-acetylglucosamine transferase [Cell envelope biogenesis, outer membrane]", ou o produto sugerido: "UDP-N-acetylglucosamine-N-acetylmuramoyl-(pentapeptide) pyrophosphoryl-undecaprenol N-acetylglucosamine transferase". Estes são, respectivamente, o primeiro e o segundo melhores resultados do BLASTP. Várias seqüências com estes produtos têm a anotação do EC_number 2.4.1.227 e também de gene, *MurG*. O melhor resultado de BLASTP é uma seqüência com 359 aminoácidos, enquanto a seqüência consulta tem 360 aminoácidos, com I=200/309 (64%), P=240/309 (77%), Gaps=4/309 (1%), e $E=1e^{-92}$. O BLASTP apresentou três domínios para glycosyltransferase e MurG com alinhamento bastante significativo para o último, 86,36% em um domínio de comprimento de 357 aminoácidos, com $E=2e^{-62}$. Não foi encontrado resultado cujas anotações coincidam com o produto e o EC_number anotados. No site BRENDA foi confirmada a relação entre o ECnumber 2.4.1.227 com o nome sugerido, além disso MurG aparece como sinônimo desta enzima.

Observa-se que no artigo de BRITO e colaboradores *de* 2004, há referência desta via metabólica relacionada à *C. violaceum*.

- Alteração na base CvioCyc:

Foi incluída a enzima "Phospho-N-acetylmuramoyl-pentapeptide-transferase" (CV4346) que catalisa a décima reação desta via metabólica.

• Biossíntese de Precursor do Lipídio-A (Via # 27).

O Diagrama da via metabólica de Biossíntese de Precursor do Lipídio-A é apresentado na Figura 44 juntamente com as enzimas encontradas pelo Pathway Tools.

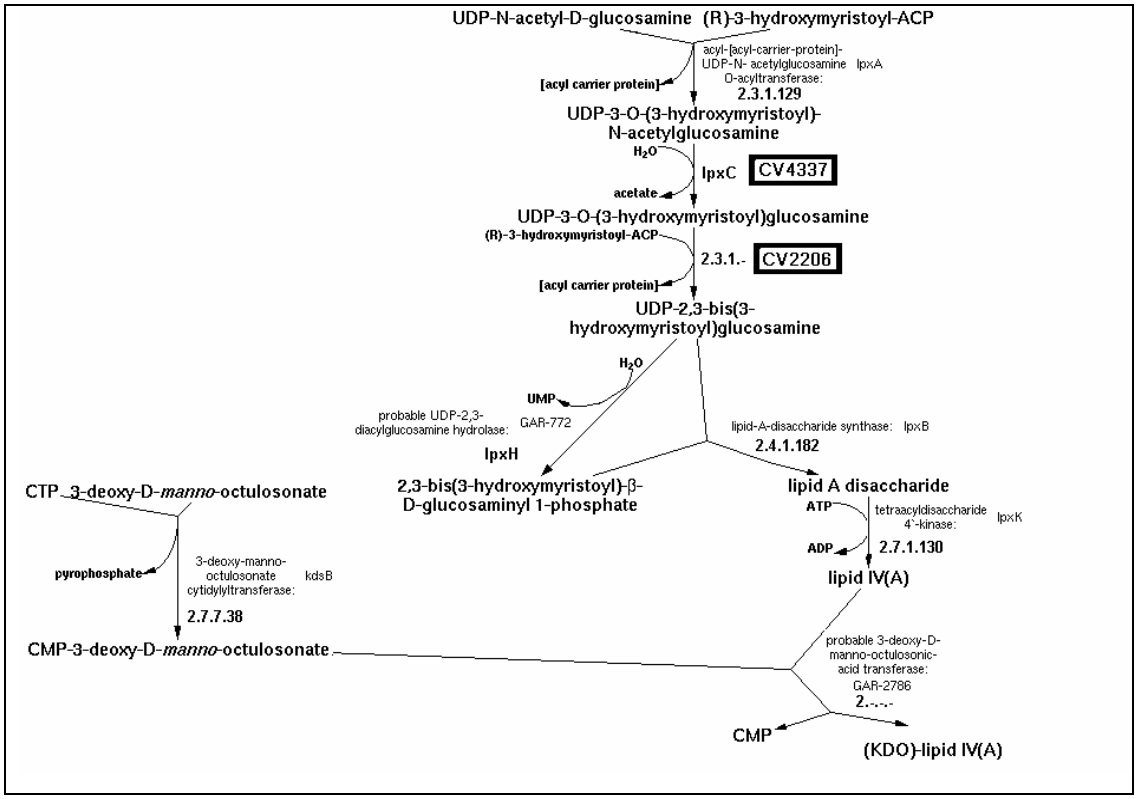


Figura 44: Diagrama da Biossíntese de Precursor do Lipídio-A e as ORFs alteradas. Adaptado do software Pathway Tools.

➢ ORF CV4337 e CV2206:

Os nomes de produto anotado e o sugerido são nomes sinônimos para a enzima, porém o nome sugerido, ”UDP-3-O-acyl N-acetylglucosamine deacetylase” é mais freqüentemente encontrado nos artigos científicos publicados no PubMed do NCBI, ao contrário do nome anotado, ”UDP-3-O-[3- hydroxymyristoyl] glucosamine N-acetyltransferase”. Sair do padrão é um problema para os softwares encontrarem dados congruentes, conforme já comentado no capítulo revisão bibliográfica.

➢ Alteração na base CvioCyc:

Foi incluída a CV4337 na CvioCyc, que catalisa a segunda reação da via. O Pathway Tools não a encontrou, provavelmente devido à falta de padronização no nome do produto. Com estas alterações a via fica completa.

• Biossíntese de Protoheme e Siroheme (Via # 29).

O Diagrama da via metabólica de Biossíntese de Protoheme e Siroheme é apresentado abaixo na Figura 45.

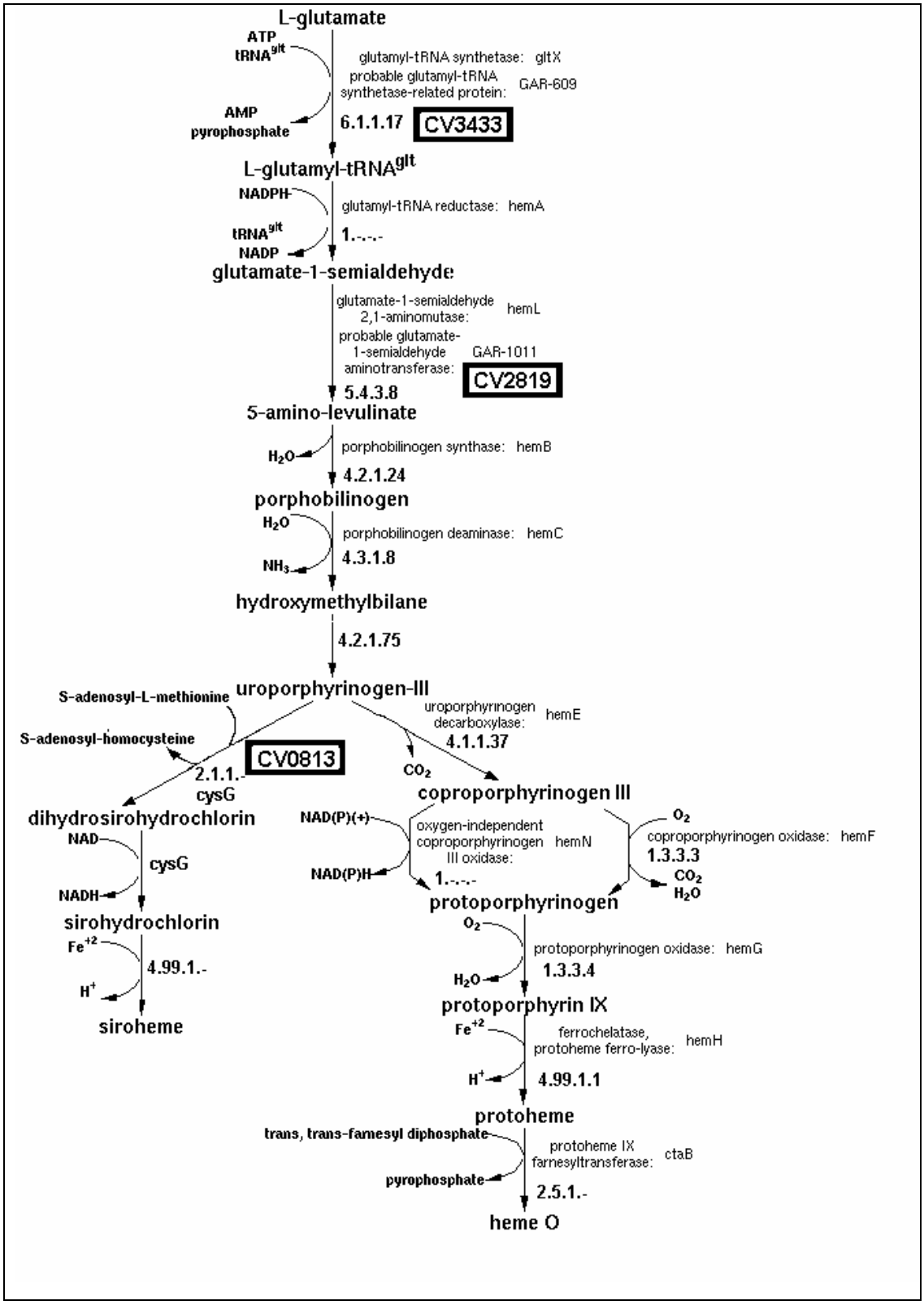


Figura 45: Diagrama da Biossíntese de Protoheme e Siroheme e as ORFs alteradas. Adaptado do software Pathway Tools.

- ORF CV3433:

O BLASTP encontrou um domínio de 326 aminoácidos, o GluRS_core, cujo alinhamento com a seqüência consulta de 298 aminoácidos foi de 93,9%. Os resultados do BLASTP apresentam muitas seqüências com os valores estatísticos muito próximos. Além de ser o melhor resultado do BLAST, há várias seqüências com o produto "COG0008: Glutamyl- and glutaminyl-tRNA synthetases". Os valores para o melhor resultado do BLASTP são I=178/297 (59%), P=207/297 (69%), Gaps=4/297 e E=6e-91. O EC_number para prováveis enzimas (reações) pode levar a resultados futuros equivocados, conforme já comentado.

- ORF CV2819:

Os resultados do BLASTP apresentam melhores resultados para o produto sugerido, "COG0001: Glutamate-1-semialdehyde aminotransferase". O melhor resultado de BLASTP é da seqüência ZP_00497428, cujo I=298/444 (67%), P=341/444 (76%), Gaps=0/444 (0%) e E=$1e^{-158}$. Em contrapartida, o produto anotado foi encontrado no nono e no décimo terceiro resultados. O EC_number para prováveis enzimas pode, novamente, induzir a resultados equivocados, por vezes falso-positivos.

- ORF CV0813:

Há muitos resultados próximos para esta ORF. Na grande maioria, os resultados são anotados como COG0007: Uroporphyrinogen-III methylase ou siroheme synthase. As entradas P11098 e P57500 foram disponibilizadas no NCBI em janeiro e maio de 2005, respectivamente; ambas têm evidência experimental. O BLAST, da entrada do NCBI P57500 de 473 aminoácidos, contra o genoma da *C. violaceum*, produz como o melhor resultado a CV0813 com 470 aminoácidos, com I=196/462 (42%) e P=291/462 (62%), G=4/462 (0%) e E=$6e^{-102}$. O gene *cobA2* só está anotado para a *C. violaceum*. Aparentemente o nome de gene *cobA2* está em desuso, pois no NCBI esta é a única entrada. Para estas enzimas o mais comum é a referência ao gene *cysG*. Com estas alterações mais duas reações desta via foram introduzidas, faltando apenas uma reação para completá-la. A ocorrência de nomes diferentes para o mesmo gene já foi comentada na revisão bibliográfica quando se tratou dos erros de anotação.

Como a enzima é multifuncional, há três reações (EC_numbers) as quais ela catalisa.

- Alteração na base CvioCyc:

Foi alterada a reação de EC_number 4.3.1.8, pois ela foi transferida para EC_number 2.5.1.61 pelo NC-IUBMB.

• Biossíntese de Riboflavina e FMN e FAD (Via # 30).

O Diagrama desta via metabólica é apresentado abaixo na Figura 46.

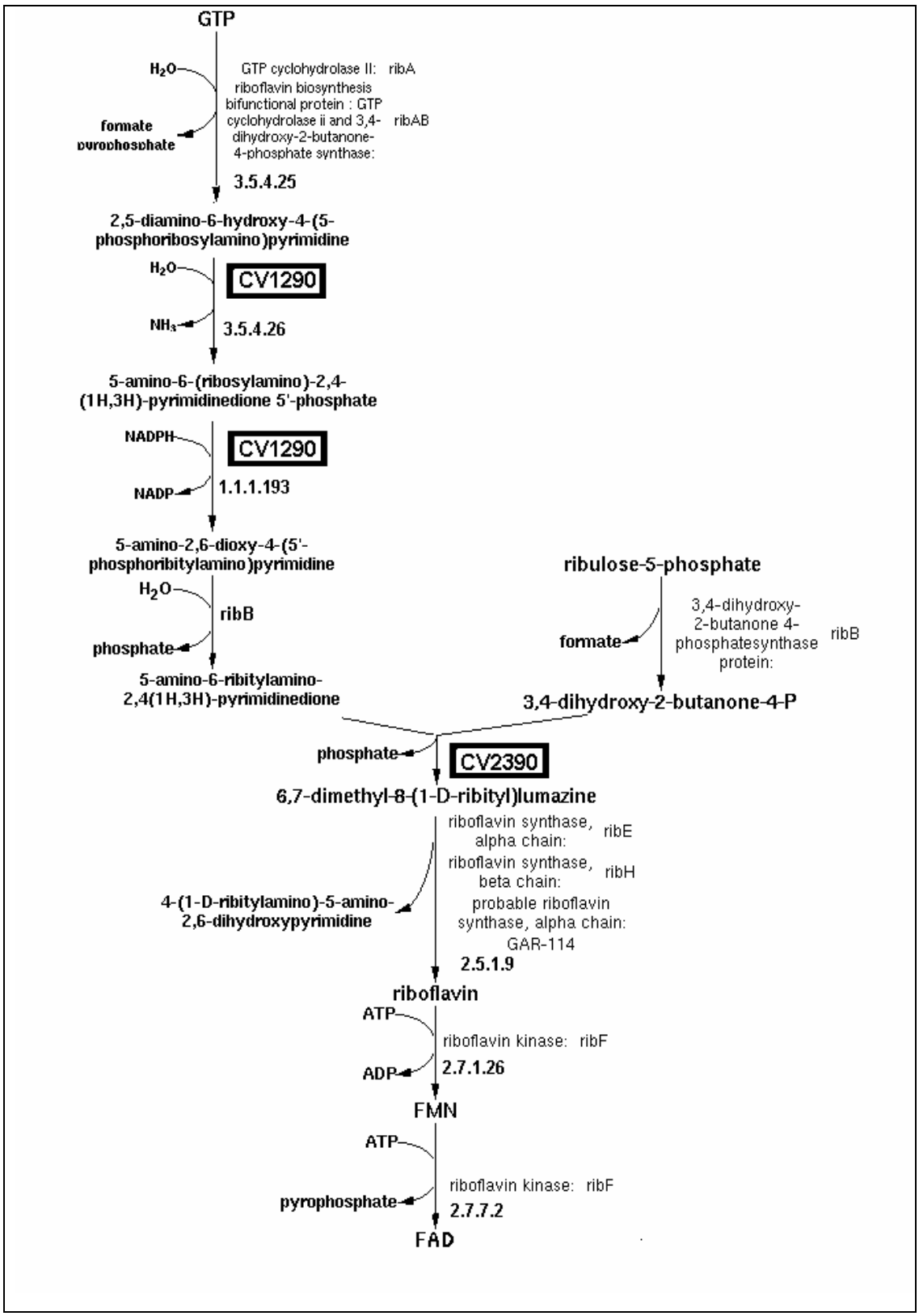


Figura 46: Diagrama da Biossíntese de Riboflavina e FMN e FAD. Adaptado do Software Pathway Tools

- Alteração na base CvioCyc:

Foram incluídas as enzimas referentes à segunda e terceira reações, que correspondem às enzimas de EC_number 3.5.4.26 e EC_number 1.1.1.193 que estão anotadas na ORF CV1290; porém o PathoLogic não identificou provavelmente devido à nomenclatura da enzima anotada. A enzima 2.5.1.9 (anotada na ORF CV2390) é bifuncional e também pode ser chamada de "lumazine synthase", e catalisa também a sexta reação. Desta forma, a via fica com oito reações catalisadas por enzimas já anotadas. Nesta via há apenas uma reação ausente.

- Biossíntese de Treonina a partir de Homoserina (Via # 32).

O Diagrama da via metabólica de Biossíntese de Treonina a partir de Homoserina é apresentado na Figura 47 juntamente com as enzimas encontradas pelo Pathway Tools.

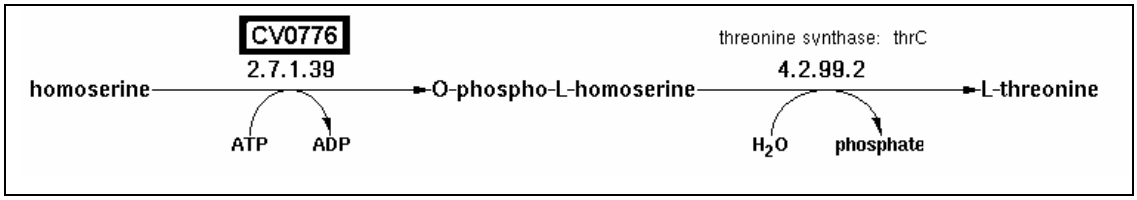


Figura 47: Diagrama da Biossíntese de Treonina a partir de Homoserina e a ORF alterada. Adaptado do Software Pathway Tools.

- ORF CV0776:

  A enzima anotada, "ketohexokinase", não condiz com os resultados do BLAST. Nenhum resultado tem o produto anotado. Outra evidência que a anotação está equivocada é que no NCBI só há entradas para organismos eucariotos referentes à enzima "ketohexokinases". No BRgene há referência do KEGG que define a "hemoserine kinase" como sendo o melhor resultado, o produto sugerido neste trabalho. Para a enzima "hemoserine kinase", o EC_number corresponde ao 2.7.1.39, conforme BRENDA e não o anotado 2.7.1.3. O gene anotado, *thrB*, foi pesquisado nas seqüências do NCBI e 104 seqüências foram encontradas. Dentre estas, somente a *C. violaceum* foi anotada com o produto igual a "ketohexokinase", sendo que 91 outras seqüências foram anotadas como produto igual a "homoserine kinase" ou "probable homoserine kinase", ou seja, para o gene thrB o produto é relativo a "homoserine kinases". Com esta alteração esta via fica completa.

• Biossíntese de UDP-N-Acetilglucosamina (Via # 33).

O Diagrama da via metabólica de Biossíntese de UDP-N-Acetilglucosamina é apresentado na Figura 48 juntamente com as enzimas encontradas pelo Pathway Tools.

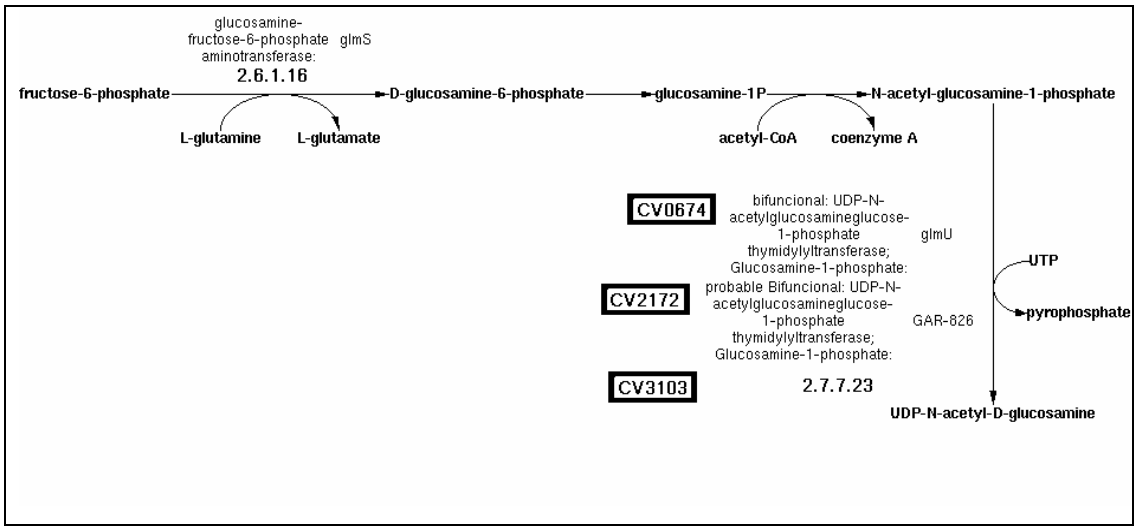


Figura 48: Diagrama da Biossíntese de UDP-N-Acetilglucosamina e as ORFs alteradas. Adaptado do Software Pathway Tools.

➢ ORF CV2172:

O produto anotado, "phosphomannomutase", tem como EC_number 5.4.2.8, de acordo com as bases de enzimas BRENDA e ENZYME. O EC_number 5.4.2.2 é relativo à enzima "phosphoglucomutase". É provável que o erro tenha sido feito por erro de leitura, ou na intenção de colocar como produto ambas as enzimas, "phosphomannomutase" e "phosphoglucomutase", porém os melhores resultados e mais freqüentes no BLASTP são apenas para a enzima "phosphoglucomutase ", conforme sugerido.

➢ ORF CV0674:

O BLASTP forneceu melhores resultados para "UDP-N-acetylglucosamine pyrophosphorylase", o produto sugerido neste trabalho para esta reação. Este produto foi encontrado no melhor resultado obtido, uma seqüência de 452 aminoácidos, YP_283449, com I=286/449 (63%), P=351/449 (78%), G=0/449 (0%) e E=8e-164. Há muitas ocorrências nos resultados do BLASTP para o produto sugerido. Muito poucas são as seqüências resultantes anotadas como bifuncional. Dentre estas, a melhor ocorrência é

o nono resultado, a seqüência Q50986, com 456 aminoácidos, I=270/451 (59%), P=333/451 (73%), G=1/451 (0%) e E=$6e^{-151}$. Apesar dos valores estatísticos dos resultados serem próximos, mostrando que as duas opções são bastante similares, as poucas ocorrências do produto "bifuncional:UDP-N-acetylglucosamineglucose-1-phosphate thymidylyltransferase; Glucosamine-1-phosphate" dá suporte à anotação sugerida.

- ORF CV3103:

Os resultados do BLASTP são de baixa similaridade para esta seqüência. Apenas 13 seqüências resultam em alinhamento entre 80 e 200 aminoácidos, todos os outros 541 resultados são menores que 80. A anotação original parece ter sido baseada em um resultado muito ruim do BLASTP, o 18º (YP_066659), onde a CV3103 (com 204 aminoácidos) e a seqüência encontrada (com 265 aminoácidos) obtiveram os seguintes valores estatísticos: I=51/151 (33%), P=77/151 (50%), Gaps=24/151 (15%) e E=$3e^{-13}$. Já o melhor resultado do BLASTP foi uma seqüência de 401 aminoácidos cujo produto do gene é o sugerido. Os valores estatísticos desse resultado foram: I=72/191 (37%), P=103/191 (53%), Gaps=6/191 (3%) e E=$1e^{-21}$. Observa-se os baixos valores para a similaridade. A maioria dos melhores resultados é de transferases, mais freqüentemente da glucose-1-phosphate thymidylyltransferase, e sem EC_number definido. Mais uma vez, o EC_number para prováveis enzimas pode levar a interpretações equivocadas, com falso-positivos, conforme já discutido anteriormente.

• Ciclo do TCA - Respiração Aeróbica (Via # 35).

O Diagrama da via metabólica de Biossíntese de TCA - Respiração Aeróbica é apresentado na Figura 49 juntamente com as enzimas encontradas pelo Pathway Tools.

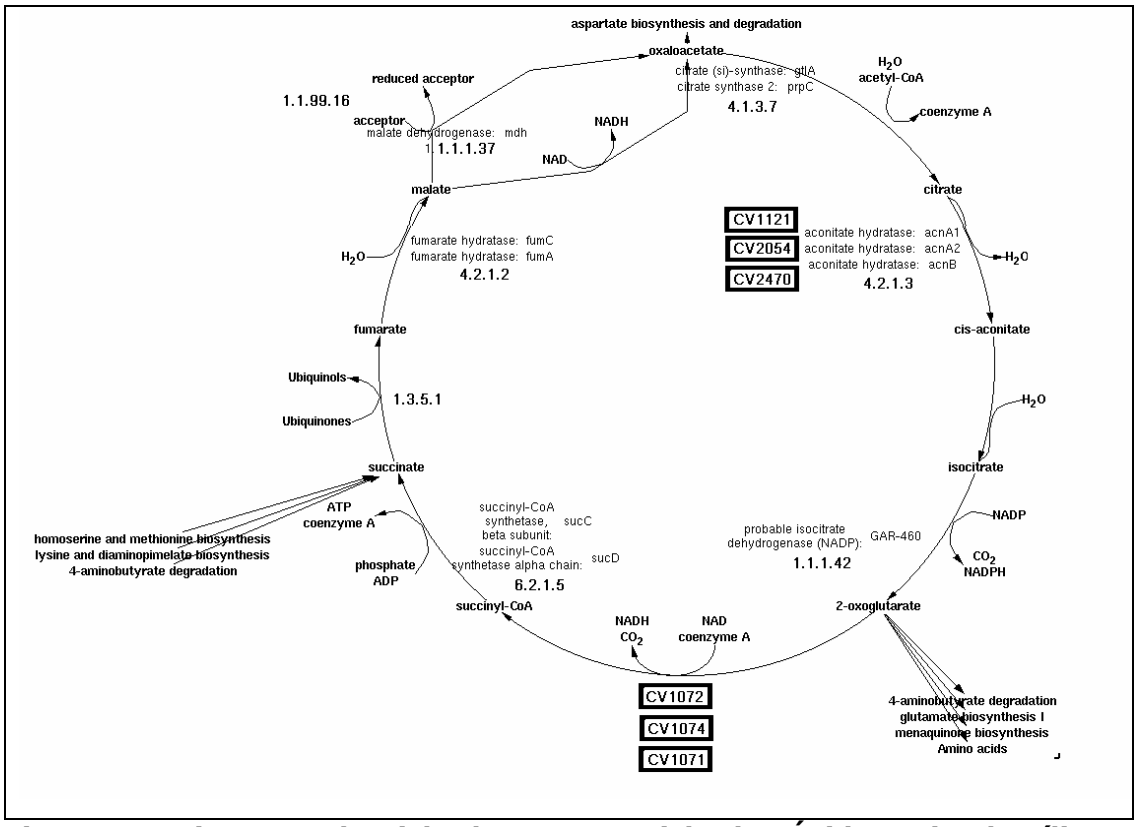


Figura 49: Diagrama do Ciclo do TCA, ou Ciclo dos Ácidos Tricarboxílicos (Ciclo de Krebs) - Respiração Aeróbica, e as ORFs alteradas. Adaptado do Software Pathway Tools.

➢ ORF CV1074:

Os melhores resultados do BLASTP para esta ORF são para o produto ”putative dihydrolipoamide dehydrogenase E3 component”, que é semelhante ao anotado; contudo nota-se que, ao contrário do anotado, há o adjetivo “putative”. O gene anotado foi o *lpdA2*; porém, há uma maior ocorrência nos resultados BLASTP do nome *lpdA,* sendo recomendado que se use o nome mais freqüente.

➢ ORF CV1072:

A anotação do produto “dihydrolipoamide dehydrogenase” confere com os resultados do BLAST, porém o EC_number 2.3.1.6 não. De acordo com

dados do BRENDA, a enzima citada refere-se ao EC_number 2.3.1.61. É possível que, neste caso, tenha havido erro de digitação.

- Alteração na base CvioCyc:

Foi incluído o complexo já anotado para a reação de EC_number 1.3.5.1, "succinato dehidrogenase" , composta de quatro subunidades, com os seguintes genes e ORFs: *sdhA* codificado por CV1067, *sdhB* codificado por CV1068, *sdhC* codificado por CV1065, e *sdhD* codificado por CV1066.

Foi também incluído o complexo já anotado para a reação de EC_number 4.2.1.3, "aconitato hidratase", composto de três subunidades, com os seguintes genes e ORFs: *acnA1* codificado por CV1121, *acnA2* codificado por CV2054, e *acnB* codificado por CV2470.

Além disso, incluiu-se ainda o complexo anotado para a reação que não tem EC_number, "2-oxoglutarate decarboxylase complex". Este complexo se compõe de três enzimas, com os seguintes genes e ORFs: *sucA* codificado por CV1071, *sucB* codificado por CV1072, e *lpd* codificado por CV1074

Com estas inclusões, a via fica com nove reações identificadas. Embora a reação 1.1.99.16 não esteja presente, conforme sugerido pelo Pathway Tools, foi identificada a reação 1.1.1.37, que é uma reação alternativa, existente na *C. violaceum*.

• Degradação de Alanina I (Via # 37).

O Diagrama da via metabólica de Degradação de Alanina I é apresentado na Figura 50 juntamente com as enzimas encontradas pelo Pathway Tools.

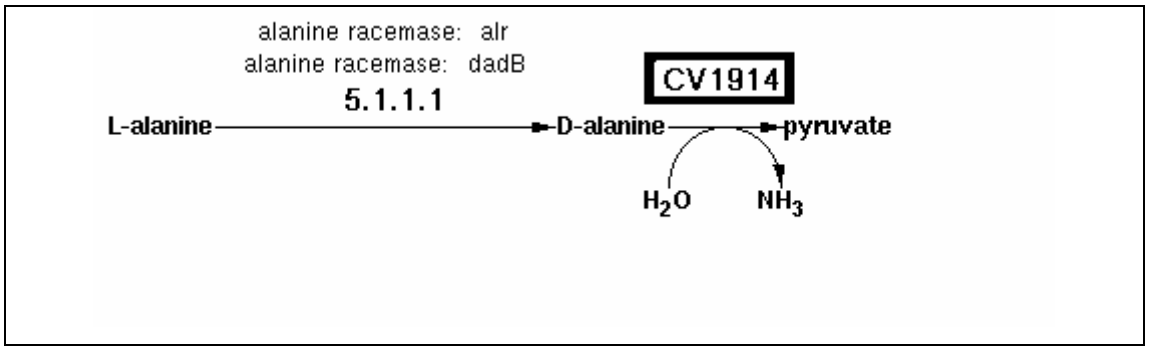


Figura 50: Diagrama da Degradação de Alanina I e a ORF incluída na CvioCyc. Adaptado do Software Pathway Tools.

➢ Alteração na base CvioCyc:

É sugerida a inclusão da ORF CV1914 – "D-amino acid dehydrogenase" na segunda reação desta via. Esta enzima não possui EC_number e o nome do gene foi anotado como *dadA2*; acredita-se que o Pathway Tools não tenha identificado esta enzima devido ao nome do gene, que no MetaCyc é *dadA*.

• Degradação de Arginina VI (Via # 38).

O Diagrama da via metabólica de Degradação de Arginina VI é apresentado na Figura 51 juntamente com as enzimas encontradas pelo Pathway Tools.

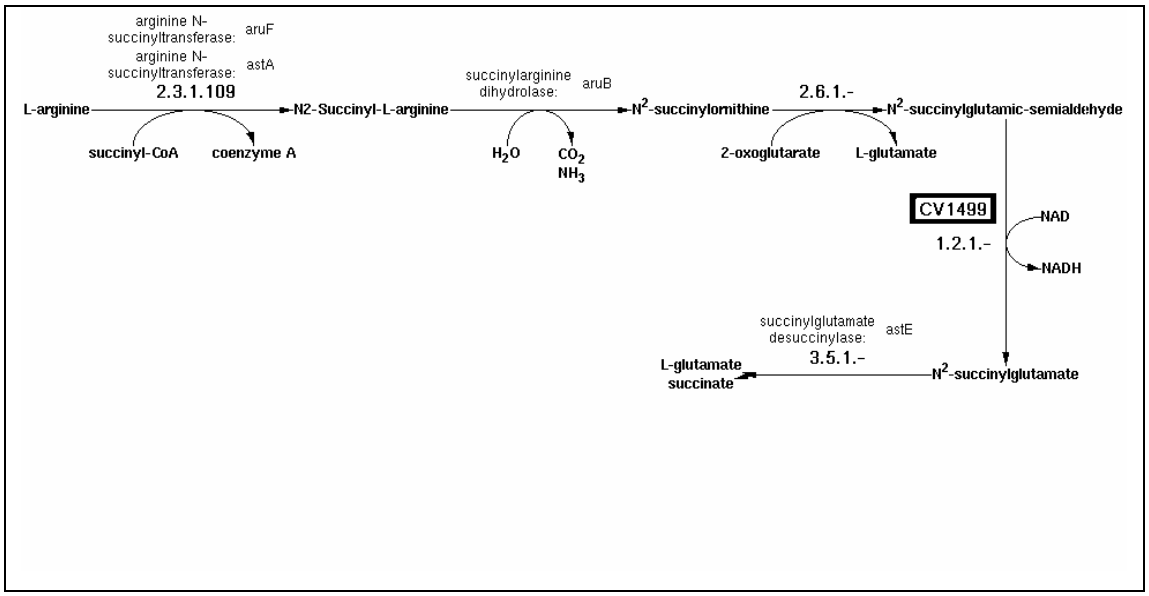


Figura 51: Diagrama da Degradação de Arginina VI e a ORF incluída na CvioCyc. Adaptado do Software Pathway Tools.

➢ ORF CV1499

Apesar de a re-anotação sugerida e a anotação original serem sinônimos, o produto de gene e o nome de gene sugeridos, "succinylglutamic semialdehyde dehydrogenase" e *astD,* são encontrados mais freqüentemente nas bases de dados biológicas. Para que os softwares automáticos possam encontrar informações nas bases de genoma, é indicada a utilização de nomes mais usuais.

➢ Alteração na base CvioCyc:

Foi incluída a ORF CV1499 para a quarta reação dessa via. É provável que o Pathway Tools não tenha reconhecido esta anotação por não ter encontrado o nome do gene e da enzima, já que os nomes obtidos do MetaCyc são *astD* para o gene, e "succinylglutamic semialdehyde dehydrogenase" para a enzima.

- Degradação de Isoleucina I (Via # 45).

O Diagrama da via metabólica de Degradação de Isoleucina I é apresentado na Figura 52 juntamente com as enzimas encontradas pelo Pathway Tools.

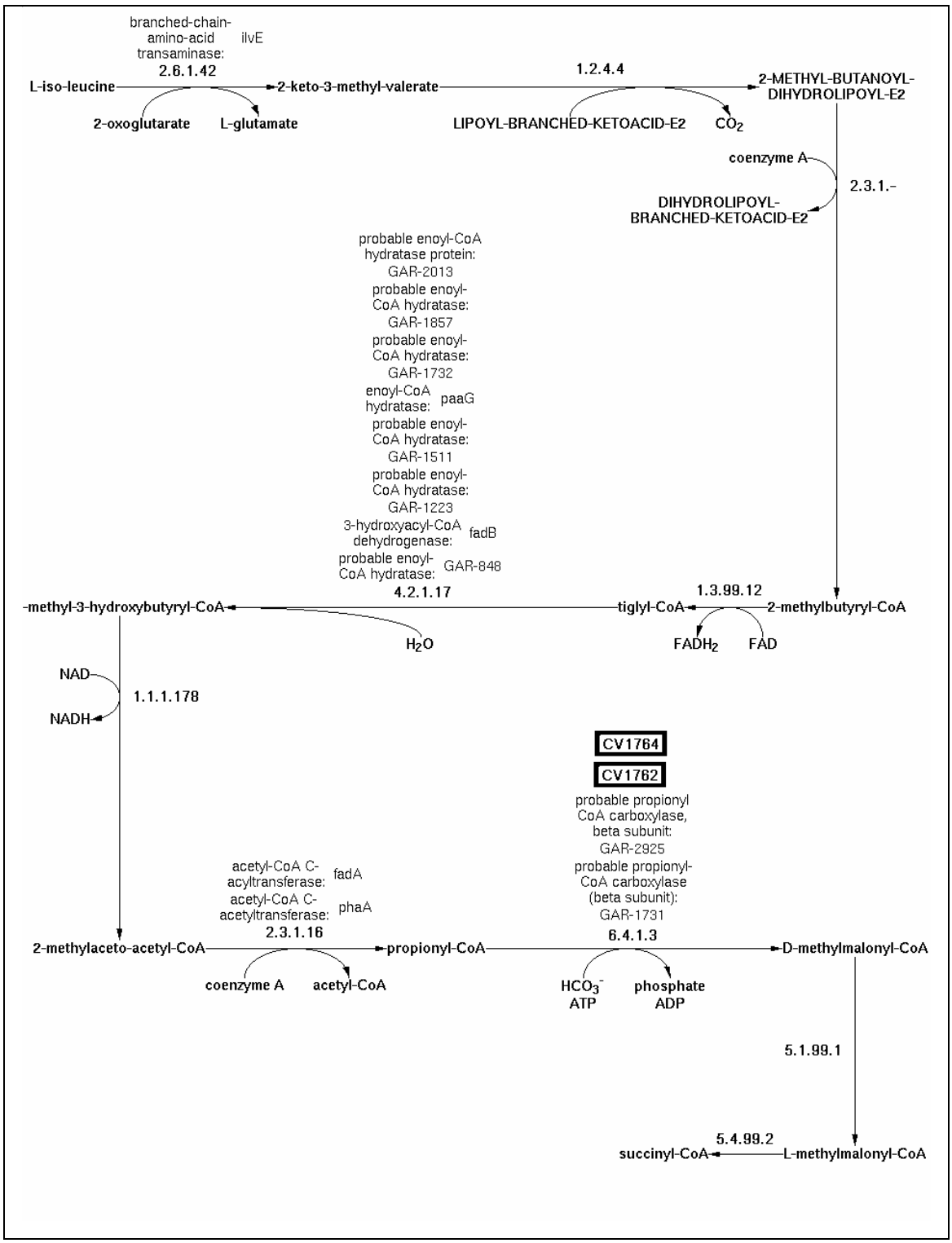


Figura 52: Diagrama da Degradação de Isoleucina I e as ORFs alteradas. Adaptado do Software Pathway Tools.

- ORF CV1764:

Os resultados de BLASTP para a seqüência CV1764, de 532 aminoácidos, apresentaram melhores resultados para ”Acetyl-CoA carboxylase, carboxyltransferase component (subunits alpha and beta)” dado pela seqüência ZP_00152852, de 535 aminoácidos. Obteve-se os seguintes valores: I=412/535 (77%), P=465/535 (86%), Gaps=3/535 (0%) e E=0. A maioria dos resultados do BLASTP tem como anotação o produto sugerido neste trabalho. Esse é um possível caso de anotação falso-positiva pois o EC_number 6.4.1.3 foi atribuído a uma enzima provável.

- ORF CV1762:

Para o caso da ORF CV1762, optou-se pelo segundo melhor resultado do BLASTP, relativo à seqüência ZP_00152849, de 666 aminoácidos, que apresentou valores I=401/666 (60%), P=481/666 (72%), Gaps=16/666 (2%) e E=0. Isso foi devido ao fato de que o “melhor” resultado era de baixa ocorrência quando comparado aos demais. O BLASTP para a seqüência CV1762, de 651 aminoácidos, apresentou vários resultados com a anotação sugerida, ”COG4770: Acetyl/propionyl-CoA carboxylase, alpha subunit”. Além disso, os resultados estatísticos são praticamente equivalentes ao primeiro melhor resultado. Como no caso da ORF discutida anteriormente, trata-se possivelmente de um erro de anotação, um falso-positivo, pois o EC_number foi atribuído para uma enzima provável.

• Degradação de Mandelato (Via # 49).

O Diagrama da via metabólica de Degradação de Mandelato é apresentado na Figura 53 juntamente com as enzimas encontradas pelo Pathway Tools.

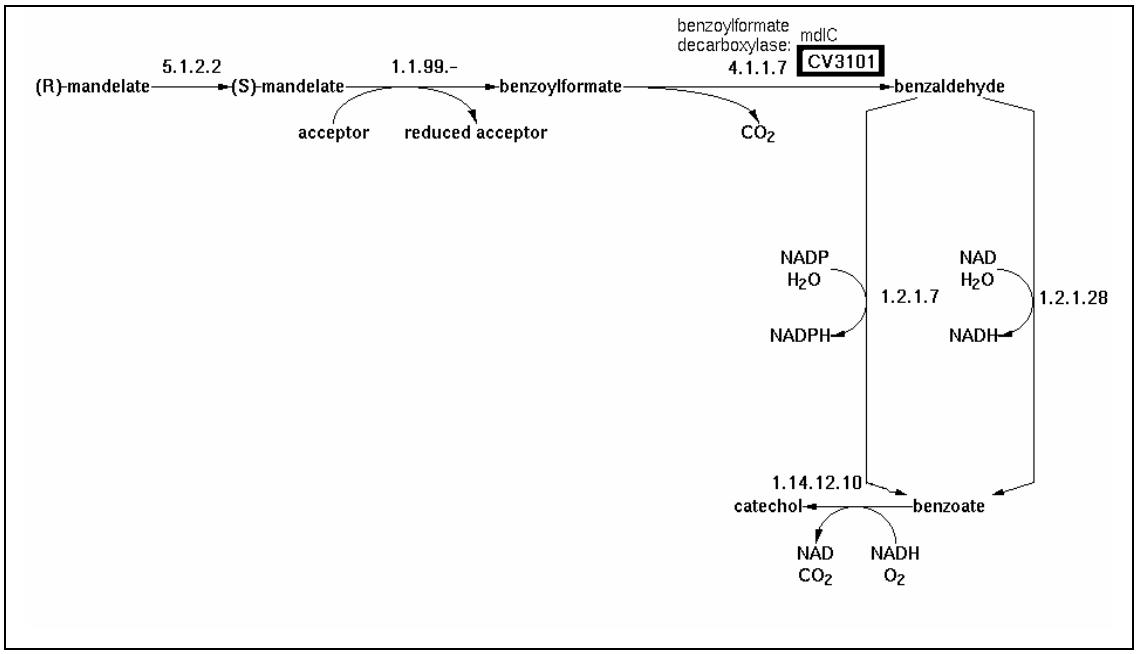


Figura 53: Diagrama da Degradação do Mandelato e a ORF alterada. Adaptado do Software Pathway Tools.

➢ ORF CV3101:

O BLASTP apresentou domínios "TPP_enzyme_Thiaminepyrophosphate". Os resultados do BLASTP não são bons, com apenas seis seqüências com alinhamentos com mais de 200 aminóacidos contra os 638 aminoácidos da seqüência consulta. A anotação original utilizou o quinto melhor resultado, " probable benzoylformate decarboxylase" para o produto do gene, porém retirou a palavra "probable" e não anotou o EC_number relativo à enzima. O melhor resultado de BLASTP, "COG0028: Thiamine pyrophosphate-requiring enzymes [acetolactate synthase, pyruvate dehydrogenase (cytochrome), glyoxylate carboligase, phosphonopyruvate decarboxylase]", a seqüência ZP_00052667 de 562 aminoácidos, produziu os valores I=174/575 (30%), P=274/575 (47%), G=48/575 (8%) e E=$2e^{-61}$. Embora, a provável seqüência usada para a anotação original, a seqüência NP_7742243, de 553 aminoácidos, tenha produzido resultados estatísticos

próximos, I=159/555 (28%), P=264/555 (47%), G=39/555 (7%) e E=$4e^{-50}$, não há nada que sugira que ela deva ser validada, pelo contrário, o domínio encontrado tende a validar a anotação sugerida. Alterando esta ORF, conforme aqui proposto, esta via metabólica é eliminada da anotação automática gerada pelo Pathway Tools para a *C. violaceum*, e não estará presente na base de dados CvioCyc, já que esta era a única enzima anotada no genoma para esta via.

- Degradação de Sacarose III (Via # 52).

O Diagrama da via metabólica de Degradação de Sacarose III é apresentado na Figura 54 juntamente com as enzimas encontradas pelo Pathway Tools.

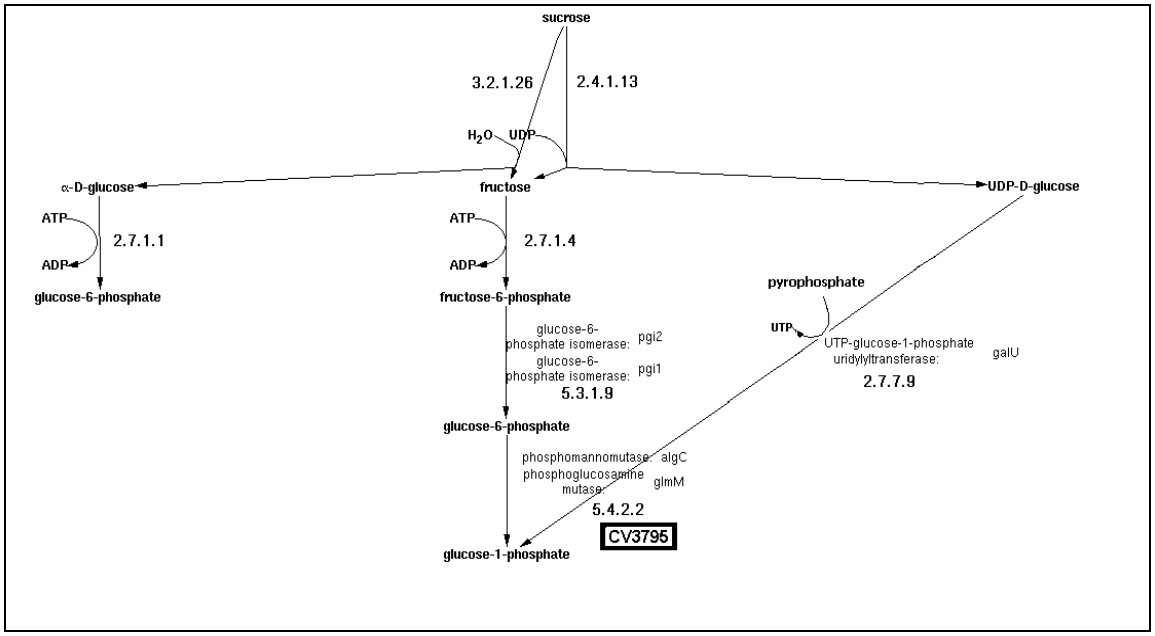


Figura 54:Diagrama da Degradação de Sacarose III e a ORF alterada. Adaptado do software Pathway Tools.

- ORF CV3795:

Os melhores resultados do BLASTP para essa ORF mostram o EC_number 5.4.2.2, que foi anotado corretamente, porém o produto encontrado pelo BLASTP é outro, é "phosphoglucomutase". Foi anotada como produto do gene a enzima "phosphoglucosamine mutase", e para esta enzima o EC_number é 5.4.2.10, conforme o BRENDA. O BLASTP apresenta domínios "PGM_PMM_ I II e III (Phosphoglucomutase/phosphomannomutase, alpha, beta, alpha)". Esse é um provável erro de digitação na anotação original.

• Fermentação (Via # 54).

O Diagrama da via metabólica de Fermentação é apresentado na Figura 55 juntamente com as enzimas encontradas pelo Pathway Tools.

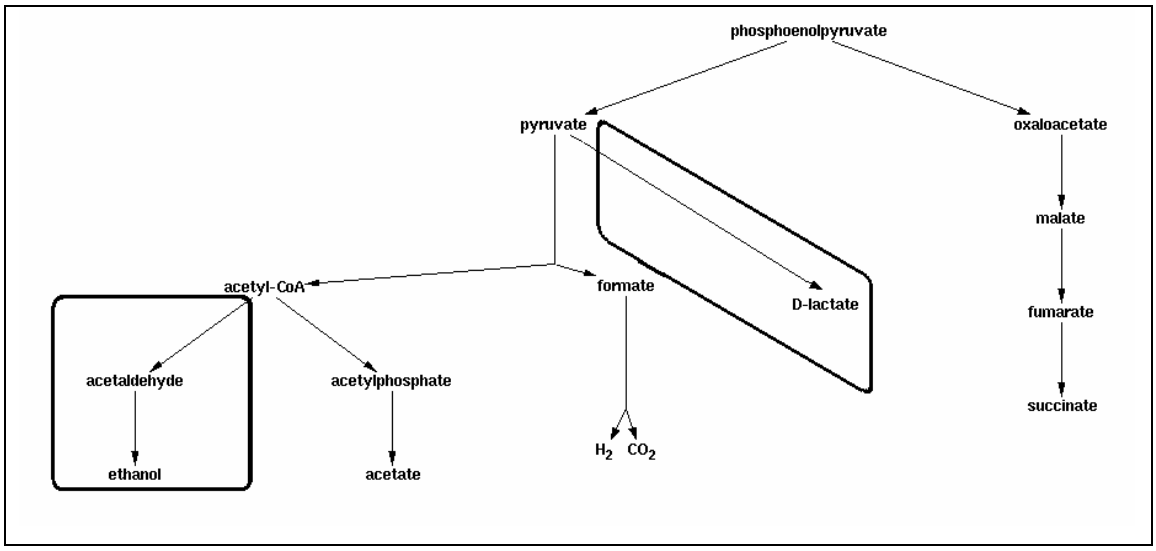


Figura 55: Diagrama da Fermentação. Os quadrados marcam os ramos de devem ser removidos desta via. Adaptado do Software Pathway Tools.

➢ Alteração na base CvioCyc:

Apesar da via estar quase completa, conforme CRECZYNSKI-PASA e ANTÔNIO (2004), a *C. violaceum* em condições anaeróbicas não produz ácido lático nem etanol; por isto devem ser removidos os ramos para estes dois produtos nesta via na base CvioCyc.

• Fotorrespiração (Via # 56).

O Diagrama da via metabólica de Fotorrespiração é apresentado na, Figura 56 juntamente com as enzimas encontradas pelo Pathway Tools.

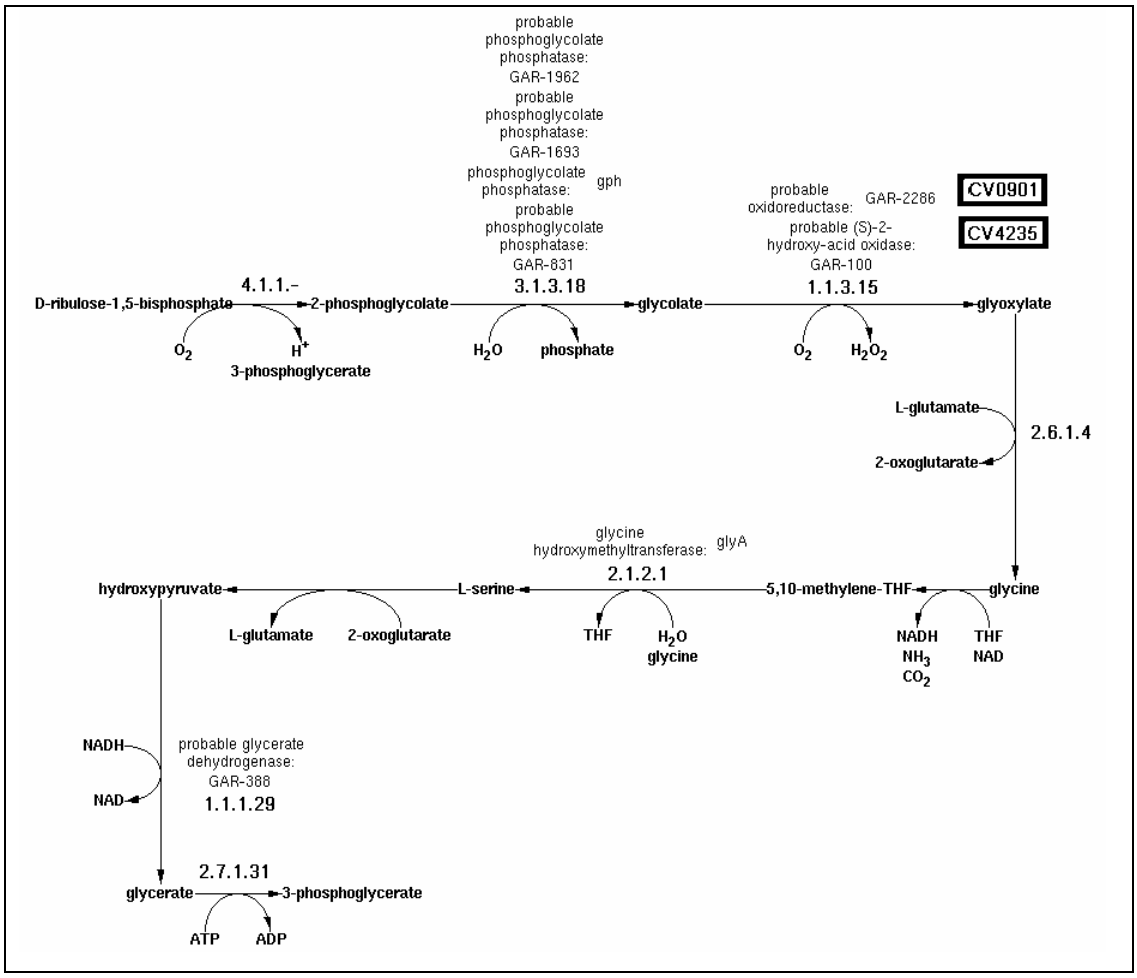


Figura 56:Diagrama da Fotorespiração e as ORFs alteradas. Adaptado do Software Pathway Tools.

➢ ORF CV4235:

No BLASTP para a ORF CV4235 encontrou-se a seqüência NP_283972, com 1.284 aminoácidos, como melhor resultado, cujos valores estatísticos são: I=874/1286 (67%), P=1033/1286 (80%), G=11/1286 (0%) e E=0.0. Esta seqüência tem anotado ”putative oxidoreductase” para produto do gene. A seqüência com resultado similar ao produto do gene anotado originalmente, a YP_004102 (“probable (S)-2-hydroxy-acid oxidase”), tem um comprimento de apenas 418 aminoácidos, contra 1.284 aminoácidos da seqüência pesquisada, e valores estatísticos I=98/418 (23%), P=147/418

(35%), G=81/418 (19%) e E=$5e^{-12}$, bem inferiores à anotação sugerida. Encontra-se aqui uma anotação falso-positiva.

- ORF CV0901:

Para esta ORF, o BLASTP gerou como melhores resultados seqüências com o mesmo produto do gene anotado, porém sem EC_number. No BRENDA, a enzima correspondente ao EC_number 1.1.3.15 é a "(S)-2-hydroxy-acid oxidase". Novamente é o caso de anotação falso-positiva.

• Ramo Não-Oxidativo de Pentose Fosfato (Via # 58).

O Diagrama da via metabólica do Ramo Não-Oxidativo de Pentose Fosfato é apresentado na Figura 57 juntamente com as enzimas encontradas pelo Pathway Tools.

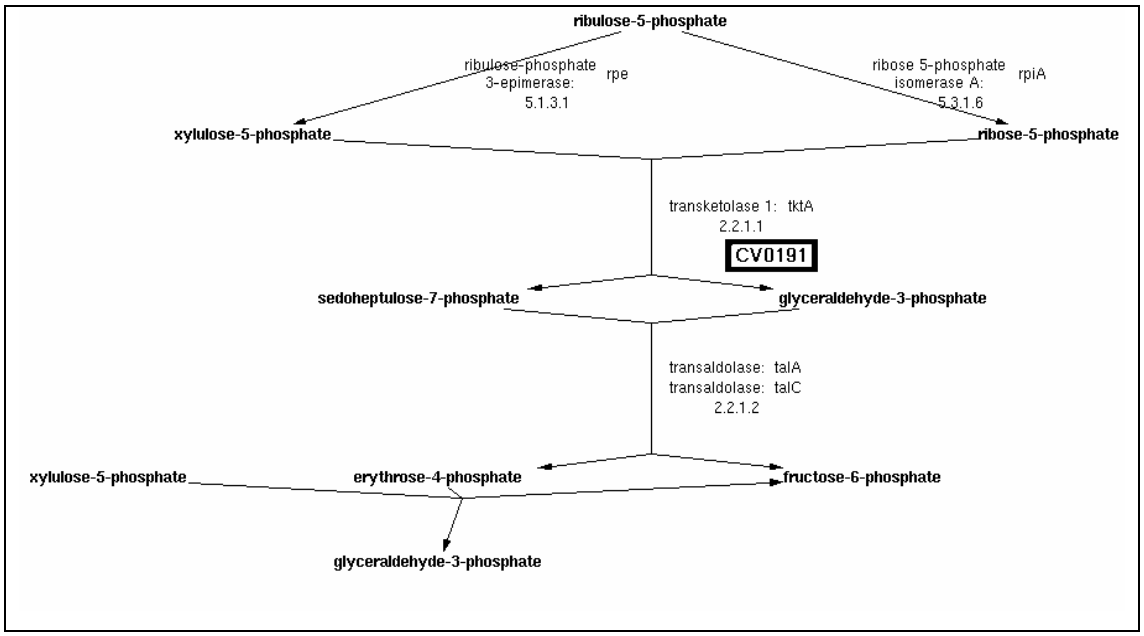


Figura 57: Diagrama Do Ramo Não-Oxidativo de Pentose Fosfato e a ORF alterada. Adaptado do Software Pathway Tools.

➢ ORF CV0191:

Há inúmeros bons resultados de BLASTP com E=0, cujos produtos de expressão podem ser "transketolase1", "transketolase" ou "COG002: Transketolase". Há 444 anotações no GenBank com o produto de gene sugerido para esta ORF, a enzima "transketolase", com EC_number 2.2.1.1. A anotação original desta ORF é a enzima "transketolase1", porém não foi anotado o EC_number.

➢ Alteração na base CvioCyc:

Incluiu-se a ORF CV0191 como catalisadora da reação ausente EC_number 2.2.1.1. O Pathway Tools não a encontrou na anotação original, provavelmente devido à inexistência do EC_number e ao fato do nome da enzima não estar nas bases consultadas pelo software.

• Supervia da Biossíntese de Serina e Glicina (Via # 59).

O Diagrama da via metabólica da Supervia da Biossíntese de Serina e Glicina é apresentado na Figura 58 juntamente com as enzimas encontradas pelo Pathway Tools.

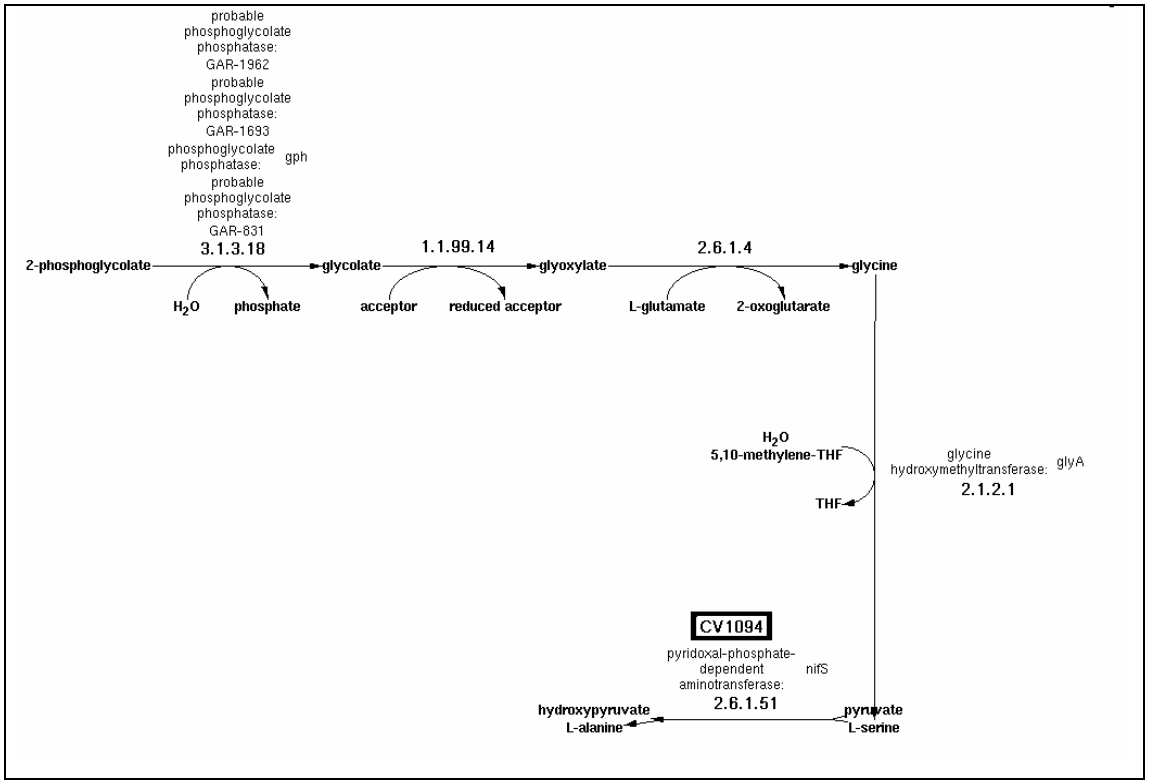


Figura 58:Diagrama da Supervia da Biossíntese de Serina e Glicina e a ORF alterada. Adaptado do Software Pathway Tools.

➢ ORF CV1094:

Os melhores resultados do BLASTP para essa ORF, que possui 405 aminoácidos, são seqüências com a anotação sugerida, ”COG1104: Cysteine sulfinate desulfinase//cysteine desulfurase and related enzymes”; outros resultados também foram encontrados sugerindo a “L-cysteine desulfurase”. Só foi encontrada uma entrada que relaciona o gene *nifS,* anotado, com o EC_number 2.6.1.51 ou ao 2.6.1.44. A maioria das entradas para o gene *nifS* está relacionada à “cysteine desulfurase” cujo EC_number é 2.8.1.7, de acordo com as bases de dados BRENDA e ENZYME. O BLASTP sugere um domínio “COG1104, NifS, Cysteine sulfinate desulfinase/cysteine desulfurase and related enzymes [Amino acid transport and metabolism]”, com 386 aminoácidos e 99,7% de alinhamento. O

melhor resultado de BLASTP foi para a enzima sugerida, na seqüência ZP_00170920 de 405 aminoácidos, que resultou I=340/405 (83%), P=372/405 (91%), G=0/405 (0%) e E=0. O domínio "pfam00282, Pyridoxal_deC, Pyridoxal-dependent decarboxylase conserved domain" de 372 aminoácidos, também resultante do BLASTP, alinhou apenas 52,5%, levando a crer que a anotação original "pyridoxal-phosphate-dependent aminotransferase" não seja apropriada para esta ORF.

• Via de Carregamento de tRNAs (Via # 60).

O Diagrama da via metabólica da Via de Carregamento de tRNAs é apresentado na Figura 59 juntamente com as enzimas encontradas pelo Pathway Tools.

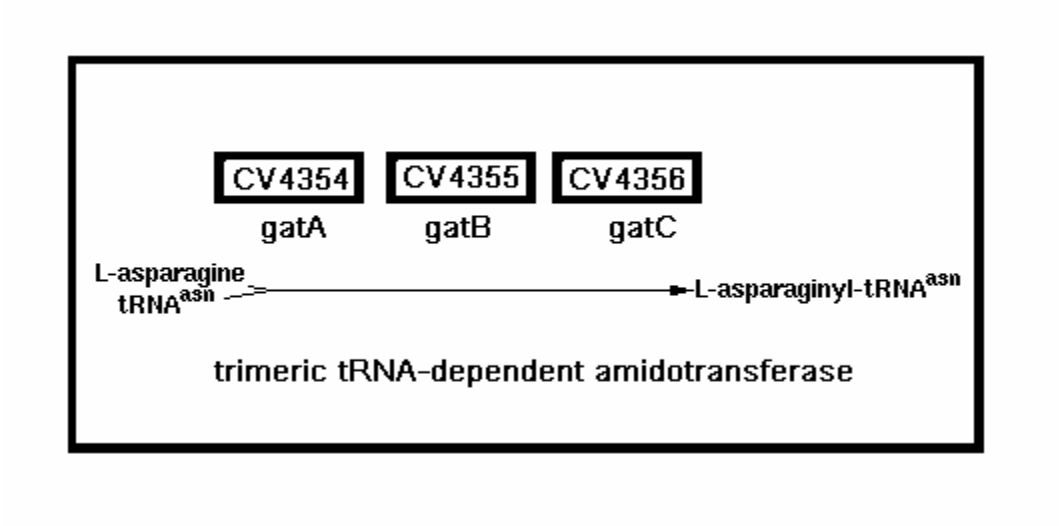


Figura 59: Diagrama proposto para a via indireta da Biossíntese de asparaginil-tRNA na *C. violaceum*. Adaptado do Software Pathway Tools.

➢ Alteração na base CvioCyc:

A via de carregamento do tRNA-glutamina pode ser feita através de uma via indireta, segundo CURNOW e colaboradores (1998) e AKOCHY e colaboradores (2004) e também com a publicação do "Brazilian National Genome Project Consortium" (2003). É possível que a *C. violaceum* também use esta via, já que no seu genoma não foi encontrada a enzima asparaginil-tRNA sintase, que sintetiza o tRNA-asparagina, aminoácido essencial na composição da síntese de proteínas (CURNOW *et al.*, 1998). No genoma da *C. violaceum* há a anotação do Adt, "trimeric tRNA-dependent amidotransferase", que tem papel essencial na formação do Asn-tRNA em *P. aueroginosa* (AKOCHY *et al.*, 2004). Essa enzima está anotada nas ORFs CV4354 (gene *gatB*), CV4355 (gene *gatA*) e CV4356 (gene *gatC*). Esta suposição deve ser mais bem pesquisada via experimentos com transposições para poder consolidar ou não esta hipótese. A via sugerida é Asp-tRNAasn → Asn-tRNAasn, catalisada por AdT, operon gatCAB.

• Via de Salvação de Ribonucleotídeos Pirimidínicos (Via # 61).

O Diagrama da via metabólica da Via de Salvação de Ribonucleotídeos Pirimidínicos é apresentado na Figura 60 juntamente com as enzimas encontradas pelo Pathway Tools.

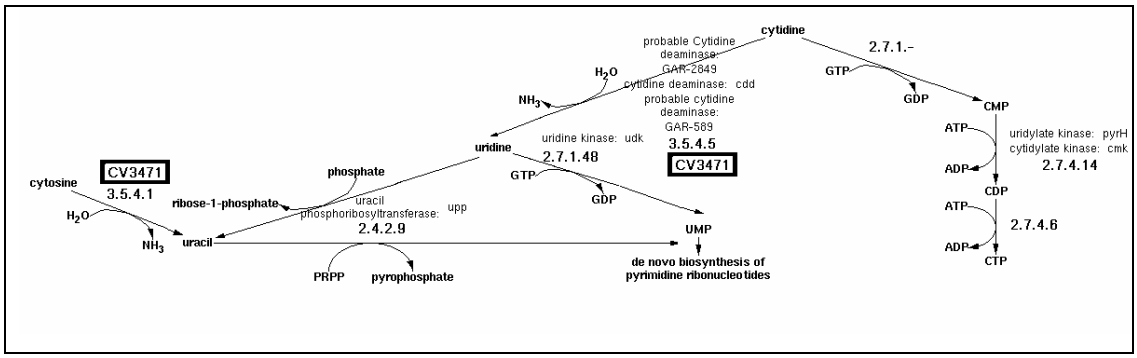


Figura 60: Diagrama proposto para a Via de Salvação de Ribonucleotídeos Pirimidínicos e as ORFs alteradas. Adaptado do Software Pathway Tools

➢ ORF CV3471:

Os resultados do BLASTP são ruins para a ORF CV3471, de 244 aminoácidos. O BLASTP sugere os domínios "COG0590, CumB, Cytosine/adenosine deaminases [Nucleotide transport and metabolism/Translation, ribosomal structure and biogenesis]", de 152 aminoácidos, com 96,7% de alinhamento, e o "cd01285, nucleoside_deaminase, Nucleoside deaminases include adenosine, guanine and cytosine deaminases" de 109 aminoácidos, com 85,3% de alinhamento. A enzima sugerida, "cytosine deaminase", está baseada na seqüência AAW89636, de 239 aminoácidos, cujos parâmetros estatísticos são: I=108/219 (49%) P=149/219 (68%), G=0/219 (0%) e E=$2e^{-47}$. Já existe duas ORFs, CV0127 e CV2570 com a anotação para a enzima "cytidine deaminase".

➢ Alteração na base CvioCyc:

Conforme VALENTIN-HANSEN (1978), a enzima "uridine-cytidine kinase" (2.7.1.48) fosforila, em *E. coli*, tanto uridina quanto citidina, e GTP e dGTP são os mais eficientes doadores de fosfato. Em função desta possibilidade, a enzima correspondente à reação 2.7.1.48 pode catalisar duas reações nesta via. Propõe-se, por isso, incluir esta alteração no CvioCyc criado.

Observa-se da discussão anterior que 17 vias metabólicas foram removidas, em um total de 61 vias analisadas; e 39 ORFs foram re-anotadas num total de 160 ORFS analisadas.

Constataram-se erros que, apesar de não serem anotações falso-positivas, eles geram falso-positivos. Verifica-se 14 ORFs possuem esta característica, CV0690, CV3433, CV2819, CV2172, CV0674, CV23103, CV1072, CV1762, CV1764, CV3101, CV3795, CV0901, CV1094 e a CV3471. Por exemplo, a ORF CV0690 está anotada como "dihydrokaempferol 4-reductase" com ECnumber 1.1.1.219 e nome de gene hpnA, porém deveria ser anotada como "putative oxidoreductase" e sem EC_number, nem nome de gene. Devido a este tipo de erro, o Pathway Tools incluiu esta reação na via de Biossíntese de Antocianina, reação esta que era a única anotada originalmente no genoma para esta via.

As ORFs, CV4034, CV1555, CV3536 e a CV3328, são casos de anotações falso-negativas, ou seja, foram anotadas como "hypothetical proteins", porém há sugestão da atribuição de função para estas como pode ser visto na Tabela 5 e na discussão das alterações. Nestes casos o Pathway Tools pode não ter incluído alguma via cujas reações se encontram em uma destas ORFs.

As outras 21 ORFS, CV0191, CV4033, CV1259, CV0225, CV4337, CV1499, CV 1562, CV1568, CV1556, CV1576, CV1575, CV0491, CV1258, CV1259, CV3567, CV4039, CV4343, CV0813, CV0776, CV1074 e a CV3471, ou podem ter sido anotadas com descrições de funções de genes não usuais ou foram anotadas equivocadamente, ou seja, atribuiu-se uma função para o produto do gene que não corresponde aos resultados computacionais encontrados. Também pode implicar para o Pathway-Tools a possível inclusão de vias impróprias ou a não inclusão de vias que deveriam fazer parte do organismo modelo de interesse, a *Chromobacterium violaceum*.

A Tabela 6 resume, em ordem alfabética, as vias metabólicas que foram validadas, a partir das análises realizadas após a geração automática de vias geradas pelo Pathway Tools. A Tabela 7 resume, também em ordem alfabética, as vias metabólicas que foram removidas da base de dados CvioCyc.

Tabela 6: Listagem das vias metabólicas que foram mantidas na base de dados CvioCyc.

#	Nome da Via Metabólica
1	Anabolismo de Trealose
2	Assimilação de Carbono C4 Fotossíntético
3	Assimilação de Sulfato
4	Assimilação de Sulfato II
5	Biossíntese de Alanina I
6	Biossíntese do Antígeno Comum Enterobacterial
7	Biossíntese de Antígeno O
8	Biossíntese de Ascorbato
9	Biossíntese de Blocos Construtores do Ácido Colânico
10	Biossíntese de Cisteína II
11	Biossíntese de Cobalamina – via aeróbica
12	Biossíntese de Enterobactina
13	Biossíntese Fenilalanina
14	Biossíntese de Fosfolipídio
15	Biossíntese de Glutamato
16	Biossíntese de Homoserina e Metionina
17	Biossíntese de KDO – inlcuindo tranferência para lipídio IV
18	Biossíntese de Peptidoglicana
19	Biossíntese de Precursor do Lipídio-A
20	Biossíntese de Prolina I
21	Biossíntese de Protoheme e Siroheme
22	Biossíntese de Riboflavina e FMN e FAD
23	Biossíntese de Treonina a partir de Homoserina
24	Biossíntese de UDP-N-Acetilglucosamina
25	Biossíntese e Degradação de Trehalose – baixa osmolaridade
26	Ciclo do TCA, variante VIII
27	Ciclo do TCA, Respiração aeróbica
28	Degradação de Alanina I
29	Degradação de Arginina VI
30	Degradação de Frutose – Anaeróbica
31	Degradação de Glutamato I

#	Nome da Via Metabólica
32	Degradação de Glutamato IV
33	Degradação de Glutamato VII
34	Degradação de Glutamato VIII
35	Fermentação
36	Fermentação de Glutamato – A via hidroxiglutarato
37	Fotorespiração
38	Gluconeogenese
39	Ramo Não-Oxidativo de Pentose Fosfato
40	Supervia de Degradação do Gluconato
41	Supervia de Biossíntese de Serina e Glicina
42	Via de Carregamento de tRNA
43	Via de Salvação de Ribonucleotídeos Pirimidínicos

Tabela 7: Listagem das vias metabólicas que foram removidas da base de dados CvioCyc

#	Nome da Via Metabólica
1	Biossíntese de Ácido Jasmônico
2	Biossíntese de Ácido Glicerol Teicóico
3	Biossíntese de Antocianina
4	Biossíntese de Auxina
5	Biossíntese de Carotenóide
6	Biossíntese de Fosfolipídio II
7	Biossíntese de Menaquinona
8	Biossíntese de Sacarose
9	Degradação de Arginina II
10	Degradação de Isoleucina I
11	Degradação de Lactose 1
12	Degradação de Lactose 4
13	Degradação de Mandelato
14	Degradação de Mandelato I
15	Degradação de Sacarose I
16	Degradação de Sacarose III
17	"De novo" Biossíntese de nucleotídeos purinas I

4.3 Operon da biossíntese de violaceína

Além das vias metabólicas, foi mais detalhadamente estudado o operon da via metabólica da biossíntese de violaceína. De acordo com os dados da literatura (AUGUST *et al.*, 2000; Brazilian Consortium, 2003), este operon é constituído de quatro genes: *vioA*, *vioB*, *vioC*, *vioD*. Através da função do PathoLogic, "Predict transcription units", foi gerado o relatório "`operon-prediction-report.txt`" e obteve-se o Directon no qual o operon da violaceína está contido. Este operon (TU1) é mostrado com um gene a mais, a ORF CV3270 (Figura 61).

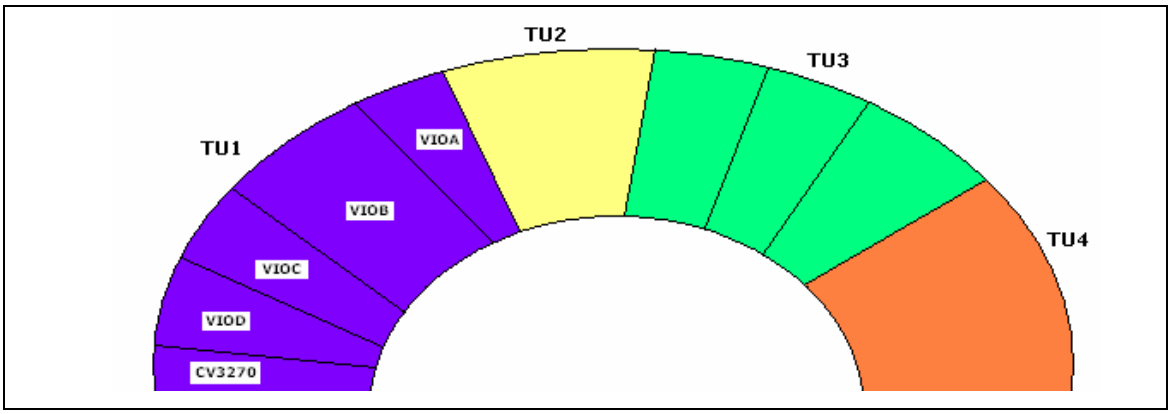


Figura 61: Esquema gráfico do Directon inferido pelo Pathway Tools onde o operon da violaceína, TU1, está inserido.

Foram feitas pesquisas na CV3270, que foi anotada como proteína hipotética ("hypothetical protein"). Esta ORF tem uma pequena similaridade com a entrada no NCBI AAC01717, do organismo *Amycolatopsis mediterranei,* cuja anotação corresponde a uma "aminodehydroquinate synthase – RifG". Porém este resultado não é suficientemente relevante pelas baixas identidades e simimilaridade produzidas no alinhamento. A Figura 62 mostra o melhor resultado de BLASTP da seqüência CV3270. Não há domínio algum identificado, e apenas 77 dos 351 aminoácidos foram alinhados; destes, apenas 30 são idênticos, e 38 são positivos, com 7 espaçamentos (*gaps*).

```
>gi|2792321|gb|AAC01717.1|   RifG [Amycolatopsis mediterranei]
          Length = 351
"score" = 41.6 bits (96), Expect = 0.012
Identities = 30/77 (38%), Positives = 38/77 (49%), Gaps = 7/77 (9%)
Query: 86  RTALGEQLCERPLDDETGPFAELFLPRDVLRRLGARHIGRRVVLGREADGWRYQRPGKGP 14
           RT +  +L ER  D   GP     LP +V+RRLGAR   R VV+      W    PG G
Sbjct: 2   RTTIPVRLAERSYDVLVGPGVRAALP-EVVRRLGAR---RAVVVSARPADWV---PGTGV 54

Query: 146 STLYLDAASGTPLRMVT 162
           TL L A  G P + ++
Sbjct: 55 ETLLLQARDGEPTKRLS 71
```

Figura 62: O melhor resultado do BLASTP da ORF CV3270.

Uma forte evidência de que a CV3270 faz parte do operon da biossíntese de violaceína, o *vioABCD*, é que os nucleotídeos localizados na região de 3.558.962 até a 3.558.981 do genoma da *C. violaceum*, parecem

formar um grampo (*hairpin*) e ter uma cauda poli-U, que possivelmente está relacionada à transcrição do operon da biossíntese de violaceína. A posição deste grampo está localizada na região terminal da seqüência de bases da CV3270 e transcrito na fita complementar, como no operon *vioABCD*. O diagrama deste grampo pode ser observado na Figura 63. Um terminador não foi localizado na região terminal do *vioD.*

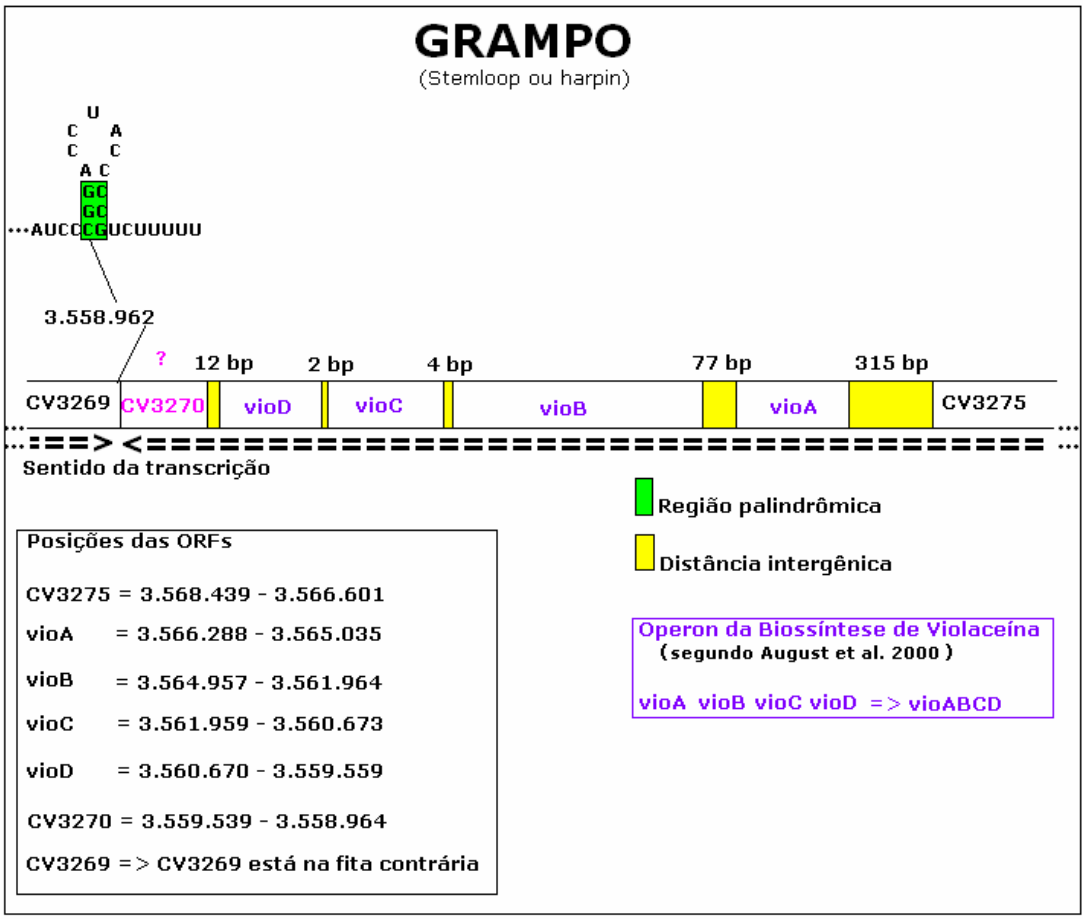


Figura 63: Grampo formado próximo à ORF CV3270, e estrutura do operon *vio*, da biossíntese de violaceína na *C. violaceum*, linhagem ATCC12472.

CAPÍTULO V

Conclusões e Sugestões

5.1 Conclusões

A base de dados Via/Genoma da *Chromobacterium violaceum* foi criada com sucesso. Várias consultas, alterações, inclusões e exclusões foram feitas para testar a estabilidade, flexibilidade e robustez do software Pathway Tools, e os resultados foram satisfatórios.

Foram encontrados alguns problemas com relação a funções bem específicas como, por exemplo, a impossibilidade de remover o status "hipotética" de uma reação incluída manualmente em uma nova via, ou erro ao adicionar uma via metabólica no mapa Overview. Todos os problemas encontrados foram relatados ao grupo do SRI e corrigidos sempre num prazo máximo de 48 horas.

Apesar do software Pathway Tools não ser específico para curar anotações genômicas, ao inferir as vias metabólicas da *Chromobacterium violaceum* a partir do genoma anotado, ele se mostrou uma boa ferramenta de depuração destas anotações. Também podem ser usadas outras informações que podem ser fontes de indícios de erros de anotação, como o relatório de pareamento entre os produtos, que compara o nome das enzimas anotadas com a base de dados das enzimas conhecidas já existentes no software.

A re-anotação genômica é uma ação que tem sido feita por muitos grupos de pesquisas com o objetivo de encontrar as funções de genes e proteínas, utilizando novos métodos para a anotação e bases de dados mais atualizadas.

Neste trabalho, foram analisadas sessenta e uma vias metabólicas. Trinta e nove ORFs foram alteradas na base CvioCyc. Isto representa aproximadamente 24,3% de erro numa amostra de 160 ORFs analisadas. Este resultado está dentro da faixa de erros comumente encontrada na literatura, que varia de 8% a 25%. Vários erros se encaixam nos erros de

anotação mais comumente encontrados na literatura, como ORFs falso-positivas, falso-negativas, erros de digitação, e falta de padronização nos nomes de enzimas e de genes. Algumas alterações, por causa das atualizações nas bases de dados referenciais, também foram detectadas como, por exemplo, as entradas de novas enzimas na base de dados BRENDA, ou a atribuição de um EC_number até então não definido para uma dada enzima.

Na análise de anotação das vias metabólicas foi utilizada principalmente a relação entre a quantidade de enzimas encontradas em relação à quantidade total de enzimas necessárias para completar a via em questão. Quando esta verificação constata que há muito poucas enzimas presentes em relação ao número total das mesmas, atribui-se a isto possíveis erros de anotação.

Uma direção de fundamental importância para melhorar a anotação automática, e ser seguida como referência, é a adoção de uma padronização para anotação dos produtos de genes. Foram verificadas muitas ocorrências da não detecção de um produto de expressão gênica, mesmo este estando no genoma, porém anotado com uma observação no final ou no meio do nome, por exemplo, em "beta subunit" ou "alpha sub-unit" ou "firA protein" para finalizar uma descrição de produto. Algumas informações que são colocadas no nome poderiam ficar no campo de comentários. Outras que devem ficar, como as subunidades, deveriam ter padrões mais rígidos. Uma das formas de evitar esses erros poderia ser a rejeição por parte do NCBI das informações não padrões para os campos mais relevantes, como, nome de gene e descrição do produto do gene. Observa-se que, para nomes de enzimas, já existe o comitê "INTERNATIONAL UNION OF BIOCHEMISTRY AND MOLECULAR BIOLOGY" (<http://www.chem.qmul.ac.uk/iubmb/>), o qual define padrões para nomes de enzimas e seus respectivos ECnumbers. Este comitê dá acesso livre a consultas e deveria ser usado por todos os grupos de pesquisas ligados à anotação genômica.

Com relação ao novo operon para a biossíntese de violaceína proposto, verifica-se que, no trabalho de AUGUST e colaboradoes (2000), há uma ORF, que é chamada de ORF2, na mesma posição da CV3270.

Todavia, não há determinação do tamanho desta ORF2; por isto, não há como se fazer uma relação direta entre a CV3270 e a ORF2, no que diz respeito aos experimentos com transposição executados pelos autores. Isto sugere que é provável a existência de mais um gene no operon *vioABCD* por dois motivos, o primeiro por causa da distância de apenas 12 pares de bases entre o vioD e a ORF CV3270, o segundo foi a estrutura em grampo que pôde ser encontrada à jusante desta ORF, o que indica o término da transcrição após a ORF CV3270.

5.2 Sugestões para futuros trabalhos

A função da ORF CV3270 deverá ser melhor estudada, assim como a sua relação com a biossíntese de violaceína, pois ela pode revelar alguma característica, tanto catalítica quanto regulatória, da biossíntese de violaceína que não tenha sido ainda esclarecida. Experimentos com transposições para definir sua função poderão ser executados futuramente, para analisar a proposta de alteração no operon hoje conhecido.

Sugere-se explorar a função do PathwayTools para a análise da expressão gênica. O uso de dados simulados pode ser uma alternativa para o desenho de novos experimentos *in vitro,* que tragam esclarecimentos sobre as vias de interesse. Novos experimentos com genes relacionados com a biossíntese de violaceína deverão ser feitos e os resultados quantitativos dos compostos introduzidos no Pathway Tools para a análise de expressão gênica em todas as outras vias da *C. violaceum*, completando assim a análise e re-anotação de todo o seu genoma.

As vias metabólicas criadas pelo Pathway Tools deveriam ser analisadas integralmente, e pesquisas freqüentes na Internet devem ser feitas para que o CvioCyc seja continuamente atualizado.

Deverá ser instalada a atualização do Pathway Tools, para que o CvioCyc tenha sua base atualizada a partir da nova versão do MetaCyc, versão 9.1. O novo MetaCyc contempla novas vias metabólicas e é continuamente curado.

A migração da base de dados CvioCyc para um servidor Windows poderá ser testada para avaliar o desempenho do software no que diz respeito à estabilidade, flexibilidade e robustez perante a nova plataforma.

CAPÍTULO VI

Referências Bibliográficas

AKOCHY, P. M.; BERNARD, D.; ROY, P. H.; LAPOINTE, J. Direct glutaminyl-tRNA biosynthesis and indirect asparaginyl-tRNA biosynthesis in Pseudomonas aeruginosa PAO1. **J. Bacteriol**. 2004 Feb;186(3):767-76.

ALBERTS B.; JOHNSON A.; LEWIS J.; RAFF M.; ROBERTS K.; ALTERP. **Molecular Biology of the Cell**: Taylor e Francis Group Ltd, 2005.

ALTSCHUL, S. F.; GISH, W.; MILLER, W.; MYERS, E. W.; LIPMAN, D. J. Basic local alignment search tool**. J Mol Biol**. 1990 Oct 5;215(3):403-10.
Disponível em: <http://www.ncbi.nlm.nih.gov/BLAST/>. Acesso em:18 abr 2005.

ALTSCHUL, S. F.; MADDEN, T. L.; SCHAFFER, A. A.; ZHANG, J.; ZHANG, Z.; MILLER, W.; LIPMAN, D.J. Gapped BLASTP and PSI-BLAST: a new generation of protein database search programs. **Nucleic Acids Res**. 1997 Sep 1;25(17):3389-402.
Disponível em: <http://www.ebi.ac.uk/blast/>. Acesso em:18 abr 2005.
Disponível em: <http://www.ncbi.nlm.nih.gov/BLAST/>. Acesso em:18 abr 2005.

AMPLICON Express Disponível em: <http://www.genomex.com/dna_sequencing/sequencing_index.php>. Acesso em: 05 mar 2005.

ANDRADE, M. A.; BROWN, N. P.; LEROY, C.; HOERSCH, S.; DE DARUVAR, A.; REICH, C.; FRANCHINI, A.; TAMAMES, J.; VALENCIA, A.; OUZOUNIS, C.; SANDER, C. Automated genome sequence analysis and annotation. **Bioinformatics**. 1999 May;15(5):391-412.

ANDRIGHETTI-FROHNER, C. R.; ANTONIO, R. V.; CRECZYNSKI-PASA, T. B.; BARARDI, C. R.; SIMOES, C. M. Cytotoxicity and potential antiviral evaluation of violacein produced by *Chromobacterium violaceum*. **Mem Inst Oswaldo Cruz.** 2003 Sep;98(6):843-8. Epub 2003 Oct 29.

ANTONIO, R. V.; CRECZYNSKI-PASA, T. B. Genetic analysis of violacein biosynthesis by *Chromobacterium violaceum*. **Genet Mol Res**. 2004 Mar 31;3(1):85-91.

AUGUST, P. R.; GROSSMAN, T. H.; MINOR, C.; DRAPER, M. P.; MACNEIL, I. A.; PEMBERTON, J. M.; CALL, K. M.; HOL, D.; OSBURNE, M. S. Sequence analysis and functional characterization of the violacein biosynthetic pathway from *Chromobacterium violaceum*. **J Mol Microbiol Biotechnol**. 2000 Oct;2(4):513-9.

BAIROCH, A. *The ENZYME database in 2000* Nucleic Acids Res. 2000 Jan 1;28(1):304-5.
Disponível em:<http://www.expasy.ch/enzyme/>. Acesso em:20 ago 2005.
BAIROCH A,; APWEILER, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000.
Nucleic Acids Res. 2000 Jan 1;28(1):45-8

BARRETT, A. J. Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). Enzyme Nomenclature. Recommendations 1992. Supplement 4: corrections and additions (1997). **Eur. J. Biochem***.* 1997, 250; 1-6.

BERMAN, H. M.; WESTBROOK, J.; FENG, Z.; GILLILAND, G.; BHAT, T.N.; WEISSIG H.; SHINDYALOV I.N.; BOURNE P. E. The Protein Data Bank. **Nucleic Acids Research**, 28 pp. 235-242 2000.
Disponível em: <http://www.rcsb.org/pdb/>). Acesso em:15 abr. 2005

BETTS, H. J.; CHAUDHURI, R. R.; PALLEN, M. J. An analysis of type-III secretion gene clusters in *Chromobacterium violaceum*. **Trends Microbiol**. 2004 Nov;12(11):476-82

BIONET, Rede Paranaense de Bioinformática. Disponível em: http://www.bionet.pucpr.br/cursos.php acesso em 10 jun 2005.

BLAMIRE, J. Disponível em:
<http://www.brooklyn.cuny.edu/bc/ahp/BioInfo/TT/TscriptD.html>. Acesso em 14 ago 2005.

BOECKMANN, B.; BAIROCH, A.; APWEILER, R.; BLATTER, M. C.; ESTREICHER, A.; GASTEIGER, E.; MARTIN, M. J.; MICHOUD, K.; O'DONOVAN, C.; PHAN, I.; PILBOUT, S.; SCHNEIDER, M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. **Nucleic Acids Res**. 2003 Jan 1;31(1):365-70.

BORK, P.; BAIROCH, A. Go hunting in sequence databases but watch out for the traps. **Trends Genet**. 1996 Oct;12(10):425-7.

BRAZILIAN NATIONAL GENOME PROJECT CONSORTIUM The complete genome sequence of *Chromobacterium violaceum* reveals remarkable and exploitable bacterial adaptability. **Proc Natl Acad Sci** U S A. 2003 Sep 30;100(20):11660-5. Epub 2003 Sep 18.

BRENNER, S. E. Errors in genome annotation. **Trends Genet**. 1999 15, 132-133.

BRITO, C. F. A.; CARVALHO, C. M. B.; SANTOS, F. R.; GAZZINELLI, R. T.; OLIVEIRA, S. C.; AZEVEDO, V.; TEIXEIRA, S. M. R. *Chromobacterium violaceum* genome:molecular mechanisms associated with pathogenicity. **Genet Mol Res.** 2004 Mar 31;3(1):148-61.

BROMBERG, N.; DURAN, N. Violacein biotransformation by basidiomycetes and bacteria. **Lett Appl Microbiol**. 2001 Oct;33(4):316-9.

BROWN, T. A. Genomes.Reino Unido: **BIOS Scientific Publishers Ltd**, 2002 Disponível em: : <http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=genomes.figgrp.645>. Acesso em: 19 jun. 2005
Disponível em: <http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=genomes.figgrp.6211>. Acesso em: 19 jun. 2005

BRUCE, A.; ALEXANDER, J.; JULIAN, L.; MARTIN, R.; KEITH, R.; PETER, W. Disponível em: <http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=mboc4.figgrp.598>. Acesso em: 19 jun. 2005
Disponível em: <http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=mboc4.figgrp.1075>. Acesso: em 19 jun. 2005

CAMPBELL, S. C.; OLSON, G. J.; CLARK, T. R.; MCFETERS, G. (2001) Biogenic production of cyanide and its application to gold recovery. **J Ind Microbiol Biotechnol**. 2001 Mar;26(3):134-9.

COOPER, G. M. **The Cell A Molecular Approach, Second Edition** – Boston: EditionSinauer Associates, Inc 2000.

CURNOW, A. W.; TUMBULA, D. L.; PELASCHIER, J. T.; MIN, B.; SOLL, D. Glutamyl-tRNA(Gln) amidotransferase in Deinococcus radiodurans may be confined to asparagine biosynthesis. **Proc Natl Acad Sci U S A**. 1998 Oct 27;95(22):12838-43.

CRECZYNSKI-PASA, T. B.e ANTÔNIO, R. V. Energetic metabolism of *Chromobacterium violaceum*. **Genet. Mol. Res**. 3 (1): 162-166 2004.

DARASELIA N.; DERNOVOY D.; TIAN Y.; BORODOVSKY M.; TATUSOV R.; TATUSOVA T. Reannotation of Shewanella oneidensis genome. **OMICS**. 2003 Summer;7(2):171-5.

DELCHER, A. L.; HARMON, D.; KASIF, S.; WHITE, O.; SALZBERG, S. L. Improved microbial gene identification with GLIMMER. **Nucleic Acids Res**. 1999 Dec 1;27(23):4636-41.
Disponível em: <http://www.tigr.org/software/glimmer/>. Acesso em: 20 abr 2005.

DEVOS, D.; VALENCIA, A. Practical limits of function prediction. **Proteins**. 2000 Oct 1;41(1):98-107.

DEVOS, D.; VALENCIA, A. Intrinsic errors in genome annotation. **Trends Genet**. 2001 Aug;17(8):429-31.

DURAN, N. e MENCK, C. F. M. *Chromobacterium violaceum*: a review of pharmacological and industrial perspective. Crit. Rev. **Microbiol**. 27: 201-222 2001.

EWING, B.; GREEN, P.; HILLIER, L., WENDL, M. Basecalling of automated sequencer traces using phred. I. Accuracy assessment. **Genome Research** 8:175-185 1998.
Disponível em: <http://www.phrap.org/consed/consed.html#howToGet>. Acesso em:22 abr. 2005.

FORSYTH, W. G.; HAYWARD, A. C.; ROBERTS, J. B. Occurrence of poly-beta-hydroxybutyric acid in aerobic gram-negative bacteria. **Nature**. 1958 Sep 20;182(4638):800-1.

GALPERIN, M. Y.; KOONIN, E. V. Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement, and operon disruption. **Silico Biol**, 1998.

GIBAS, C.; JÁMBECK P. **Desenvolvendo Bioinformática: ferramentas de software para aplicações em biologia** / Cynthia Gibas e Per Jambeck; tradução Milarepa Ltda - Rio de Janeiro: Campus, 2001.

GRIFFITHS, A. J. F.; MILLER, J. H.; SUZUKI, D. T.; LEWONTIN, R. C.; WILLIAM, M. G. An Introduction to Genetic Analysis: **W. H. Freeman and Company**, 2000.

GREEN, M. L.; KARP, P. D. A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. **BMC Bioinformatics**. 2004 Jun 09;5(1):76.

HUNGRIA, M.; NICOLAS, M. F.; GUIMARAES, C. T.; JARDIM, S. N.; GOMES, E. A.; VASCONCELOS, A. T. Tolerance to stress and environmental adaptability of *Chromobacterium violaceum*. **Genet Mol Res**. 2004 Mar 31;3(1):102-16.

ILIOPOULOS I.; TSOKA S.; ANDRADE M. A.; ENRIGHT A. J.; CARROLL M.; POULLET P.; PROMPONAS V.; LIAKOPOULOS T.; PALAIOS G.; PASQUIER C.; HAMODRAKAS S.; TAMAMES J.; YAGNIK A. T.; TRAMONTANO A.; DEVOS D.; BLASCHKE C.; VALENCIA A.; BRETT D.; MARTIN D.; LEROY C.; RIGOUTSOS I.; SANDER C.; OUZOUNIS C. A. Evaluation of annotation strategies using an entire genome sequence. **Bioinformatics**. 2003 Apr 12;19(6):717-26.

JACOB, F.; MONOD, J. Genetic regulatory mechanisms in the synthesis of proteins. **J Mol Biol**. 1961 Jun;3:318-56.

KANEHISA, M A database for post-genome analysis. **Trends Genet**. 13, 375-376 1997.

KANEHISA, M. Post-genome informatics. **USA: Oxford University Press**:2000.

KANEHISA, M; GOTO, S KEGG: kyoto encyclopedia of genes and genomes, **Nucleic Acids Res**. 2000 Jan 1;28(1):27-30. Disponível em: <http://www.genome.ad.jp/kegg/kegg.html>. Acesso em: 21 maio 2005.

KARP, P. D.; KRUMMENACKER, M.; PALEY, S.; WAGG, J. Integrated pathway-genome databases and their role in drug discovery. **Trends Biotechnol**. 1999 Jul;17(7):275-81.

KARP, P. D.; RILEY, M.; PALEY, S M; PELLEGRINI-TOOLE, A.; KRUMMENACKER, M. EcoCyc: encyclopedia of Escherichia coli genes and metabolism**. Nucleic Acids Res**. 1999 Jan 1;27(1):55-8.

KARP, P.D.; RILEY, M.; SAIER, M.; PAULSEN, I.T.; COLLADO-VIDES, J.; PALEY, S.M.; PELLEGRINI-TOOLE, A.; BONAVIDES, C.; GAMA-CASTRO, S. The EcoCyc Database. **Nucleic Acids Res**. 2002 Jan 1;30(1):56-8

KARP, P. D.; RILEY, M.; PALEY, S. M.; PELLEGRINI-TOOL,E A. The MetaCyc Database. **Nucleic Acids Res**. 2002 Jan 1;30(1):59-61.

KARP, P. D.; PALEY, S.; ROMERO, P. The Pathway Tools software. **Bioinformatics**. 2002;18 Suppl 1:S225-32.

KARP, P.D., RILEY, M.; PALEY, S.M.; PELLEGRINI-TOOLE, A. The MetaCyc Database. **Nucleic Acids Res**. 2002 Jan 1;30(1):59-61.

KESELER, I. M.; COLLADO-VIDES, J.; GAMA-CASTRO, S.; INGRAHAM, J.; PALEY, S.; PAULSEN, I. T.; PERALTA-GIL, M.; KARP, P.D. EcoCyc: a comprehensive database resource for Escherichia coli. **Nucleic Acids Res**. 2005 Jan 1;33(Database issue):D334-7.

KOONIN, E. V.; GALPERIN, M. Y. Sequence-Evolution-Function: Computational Approaches in Comparative Genomics. Kluwer **Academic Publishers**. Massachusetts-USA: 2003. Disponível em: <http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=sef.section.168>. Acesso em:21 maio 2005.

KRIEGER, C. J.; ZHANG, P.; MUELLER, L. A.; WANG, A.; PALEY, S.; ARNAUD, M.; PICK, J.; RHEE, S. Y.; KARP, P. D. MetaCyc: A multiorganism database of metabolic pathways and enzymes**, Nucleic Acids Res**., 32:D438-D442 2004.
Disponível em: <http://metacyc.org/>. Acesso em:20 ago 2005.

KRUMMENACKER, M.; PALEY, S.; MUELLER, L.; YAN. T.; KARP, P. D.
Querying and computing with BioCyc databases. **Bioinformatics**. 2005 Aug 15;21(16):3454-5. Epub 2005 Jun 16.
Disponível em: http://biocyc.org/. Acesso em:22 ago 2005.

LANGE B. M.; GHASSEMIAN M. Comprehensive post-genomic data analysis approaches integrating biochemical pathway maps. **Phytochemistry.** 2005 Feb;66(4):413-51.

LEE, S.; YOOK, J.; MIN, S. H.; KOO, H. A Werner syndrome protein homolog affects *C. elegans* development, growth rate, life span and sensitivity to DNA damage by acting at a DNA damage checkpoint. **Development**. 2004 Jun;131(11):2565-75. Epub 2004 Apr 28.

LEON, L. L.; MIRANDA, C. C.; DE SOUZA, A. O.; Durán, N. Antileishmanial activity of the violacein extracted from Chromobacterium violaceum**. Journal of Antimicrobial Chemotherapy** (2001) 48, 449-450.

LEWIN, B. **Genes VII**. Porto Alegre: 2001.

LIPMAN, D. J.; PEARSON, W.R. Rapid and sensitive protein similarity searches. **Science**. 1985 Mar 22;227(4693):1435-41.
Disponível em: <http://www.ebi.ac.uk/fasta/> e <http://fasta.genome.jp/>. Acesso em: 19 abr. 2005.

LODISH, H.; BERK, A.; ZIPURSKY, L.; MATSUDAIRA, P.; BALTIMORE, D.; DARNELL, J**. Molecular Cell BiologyFourth** Edition W. H. Freeman and Company, 2000.

MADIGAN, M. T.; MARTINHO, J. M.; PARKEY, J. **Biology of Microorganisms**. Prentice-Hall, Inc. USA: 2000.

MARCHLER-BAUER, A.; BRYANT, S. H. *CD-Search: protein domain annotations on the fly.*, **Nucleic Acids Res**. 32:W327-331 2004.

MATHEWS, C. K.; VAN HOLDE, K. E.; AHERN, K. G. **Biochemistry**. Addison Wesley Longman, Inc. San Francisco-CA: 1999.

MATZ, C.; DEINES, P.; BOENIGK, J.; ARNDT, H.; EBERL, L.; KJELLEBERG, S.; JURGENS, K. Impact of violacein-producing bacteria on survival and feeding of bacterivorous nanoflagellates. **Appl Environ Microbiol**. 2004 Mar;70(3):1593-9

MAYER, G. Microbiology and Imunnology on-line **University of South Carolina - School of Medicine** The Board of Trustees of the University of South Carolina, 2005.

MELO, P.S.; MARIA, S. S.; VIDAL, B. C.; HAUN, M.; DURAN, N. Violacein cytotoxicity and induction of apoptosis in V79 cells. In Vitro **Cell Dev Biol Anim**. 2000 Sep;36(8):539-43.

MOMEN, A. Z.; HOSHINO, T. Biosynthesis of violacein: intact incorporation of the tryptophan molecule on the oxindole side, with intramolecular rearrangement of the indole ring on the 5-hydroxyindole side. Biosci **Biotechnol Biochem**. 2000 Mar;64(3):539-49.

MUELLER, L.A.; ZHANG, P.; RHEE, S.Y. AraCyc: a biochemical pathway database for Arabidopsis. **Plant Physiol**. 2003 Jun;132(2):453-60.

Disponível em: <http://www.arabidopsis.org/tools/aracyc/>. Acesso em: 21 ago 2005.

NCBI - **NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION** A science primer molecular modeling: a method for unraveling protein structure and function.
Disponível em:
<http://www.ncbi.nih.gov/About/primer/molecularmod.html#molecular>.
Acesso em: 22 maio 2005.

NCBI - **NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION** The NCBI Handbook Part 3. Querying and Linking the Data Created:October 9,2002 Updated: August 13, 2003. Disponível em: <http://www.ncbi.nlm.nih.gov/
entrez/query.fcgi?cmd=Searchedb=booksedoptcmdl=GenBookHLeterm=QUERYING+AND+LINKING+THE+DATA+AND+handbook%5Bbook%5D+AND+154791%5Buid%5Derid=handbook.part.587>. Acesso em 28 mai 2005.

Ocelot Computer Services Inc., 2002. Disponível em: <http://www.ocelot.ca/>. Acesso em 10 set 2005.

OUZOUNIS, C. A.; KARP, P. D. The past, present and future of genome-wide re-annotation. **Genome Biol**. 2002;3(2):COMMENT2001. Epub 2002 Jan 31.

OVERBEEK, R.; LARSEN, N.; PUSCH, G. D.; D'SOUZA, M.; SELKOV, E. JR.; KYRPIDES, N.; FONSTEIN, M.; MALTSEV, N.; SELKOV, E. WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction**. Nucleic Acids Res**. 2000 Jan 1;28(1):123-5.

PALEY, S. M.; KARP, P. D. Evaluation of computational metabolic-pathway predictions for Helicobacter pylori. **Bioinformatics**. 2002 May;18(5):715-24.

PAN D.; SUN N.; CHEUNG K. H; GUAN Z.; MA L.; HOLFORD M.; DENG X.; ZHAO H. PathMAPA: a tool for displaying gene expression and performing statistical tests on metabolic pathways at multiple levels for Arabidopsis. **BMC Bioinformatics**. 2003 Nov 7;4(1):56.

PANGEASYSTEMS, **Pathway Tools User's Guide Version 7.5**, Copyright SRI international, 1(2002), p. 100.

RILEY, M. Functions of the gene products of Escherichia coli. **Microbiol Rev**. 1993 Dec;57(4):862-952.

RILEY, M. e SPACE, D. B. Genes and proteins of Escherichia coli (GenProtEc) **Nucleic Acids Res**. 1996 January 1; 24(1).

ROGOZIN, I. B.; MAKAROVA, K. S.; NATALE, D. A.; SPIRIDONOV, A. N.; TATUSOV, R. L.; WOLF, Y. I.; YIN, J.; KOONIN, E. V. Congruent evolution of

different classes of non-coding DNA in prokaryotic genomes **Nucleic Acids Res**. 2002 Oct 1;30(19):4264-71.

ROMERO, P. R. e KARP, P. D. Using Functional and Organizational Information to Improve Genome-Wide Computational Prediction of Transcription Units on Pathway-Genome. **Bioinformatics**. 2004 Mar 22;20(5):709-17. Epub 2004 Jan 29.

ROSKOSKI, R. Jr. e BRAZDA, F. G. **Biochemistry**. W. B. Saundres Company 1996.

SALGADO, H.; MORENO-HAGELSIEB G.; SMITH, T.F.; COLLADO-VIDES, J. Operons in Escherichia coli: Genomic analyses and predictions. **Proc Natl Acad Sci** U S A. 2000 Jun 6;97(12):6652-7.

SALGADO, H.; GAMA-CASTRO. S.; MARTINEZ-ANTONIO, A.; DIAZ-PEREDO, E.; SANCHEZ-SOLANO, F.; PERALTA-GIL, M.; GARCIA-ALONSO, D.; JIMENEZ-JACINTO, V.; SANTOS-ZAVALETA, A.; BONAVIDES-MARTINEZ, C.; COLLADO-VIDES, J. RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in Escherichia coli K-12. **Nucleic Acids Res**. 2004 Jan 1;32(Database issue):D303-6.

SANGER Institute. Disponível em: <http://www.sanger.ac.uk/Info/Intro/sanger.shtml>. Acesso em: 10 mar. 2005.

SARAIVA, V.S.; MARSHALL, J.C.; COOLS-LARTIGUE, J.; BURNIER, M. N. Jr. Cytotoxic effects of violacein in human uveal melanoma cell lines **Melanoma Res**. 2004 Oct;14(5):421-4.

SCHOMBURG, I; CHANG, A.; EBELING, C.; GREMSE, M.; HELDT, C.; HUHN, G.; SCHOMBURG, D. BRENDA, the enzyme database: updates and major new developments. **Nucleic Acids Res**. 2004 Jan 1;32(Database issue):D431-3.
Disponível em: <http://www.brenda.uni-koeln.de/>. Acesso em: 20 maio 2005.

SELKOV, E.Jr.; GRECHKIN, Y.; MIKHAILOVA, N.; SELKOV E. MPW: the Metabolic Pathways Database. **Nucleic Acids Res**. 1998 Jan 1;26(1):43-5. Disponível em: <http://www.empproject.com/>. Acesso em: 14 ago 2005.

SILVA, R.; ARARIPE J. R.; RONDINELLI, E.; ÜRMÉNYI, T. P. Gene expression in *Chromobacterium violaceum.* **Genet. Mol. Res**. 3 (1): 64-75, 2004.

SRI International. Escherichia coli DNA-Segments. Disponível em: <http://biocyc.org/ECOLI/new-image?object=DNA-Segments>. Acesso em: 15 mar 2005).

STOCSITS, R. R.; HOFACKER, I. L.; FRIED, C.; STADLER, P. F. Multiple sequence alignments of partially coding nucleic acid sequences. **BMC Bioinformatics**. 2005 Jun 28;6:160.

Disponível em: <http://www.ebi.ac.uk/clustalw/>. Acesso em: 22 ago 2005.

SUTTON, G. G.; WHITE, O.; ADAMS, M. D.; KERLAVAGE, A. R. TIGR Assembler: A New Tool for Assembling Large Shotgun Sequencing Projects. **Genome Science e Technology** 1: 9-19, 1995.
Disponível em: <http://www.tigr.org/software/assembler/>. Acesso em: 22 abr. 2005.

TATUSOV, R. L.; NATALE, D.A.; GARKAVTSEV, I.V.; TATUSOVA, T.A.; SHANKAVARAM, U. T.; RAO, B. S.; KIRYUTIN, B.; GALPERIN, M. Y.; FEDOROVA, N. D.; KOONIN, E. V. The COG database: new developments in phylogenetic classification of proteins from complete genomes**. Nucleic Acids Res**. 2001 Jan 1;29(1):22-8.

TURCHIN, A.; KOHANE, I. S. Gene Homology resources on the World Wide Web Physiological Genomics 11: 165-177, 2002.

UMBIOLOGY **University of Miami – Departament of Biology** – Disponível em: <http://fig.cox.miami.edu/~cmallery/150/gene/dideoxy.htm>. Acesso em: 08 mar 2005.

VALENTIN-HANSEN, P. Uridine-cytidine kinase from Escherichia coli. **Methods Enzymol**. 1978;51:308-14.

VOET, D.; VOET, J. G.; PRATT, C. W. trad:.Arthur Germano Fett Neto **Fundamentos de Bioquímica**. Porto Alegre: ArtMed Editora SA., 2000.

WITTIG, U.; DE BEUCKELAER, A. Analysis and comparison of metabolic pathway databases. **Brief Bioinform**. 2001 May;2(2):126-42.

WATSON, J.D.; CRICK, F.H. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. **Nature** 1953

WU, C. H.; YEH, L. S.; HUANG, H.; ARMINSKI, L.; CASTRO-ALVEAR, J.; CHEN, Y.; HU, Z.; KOURTESIS, P.; LEDLEY, R.S.; SUZEK, B.E.; VINAYAKA, C.R.; ZHANG, J.; BARKER, W.C. (2003) The Protein Information Resource.**Nucleic Acids Res**. 2003 Jan 1;31(1):345-7.

ZHANG P.; FOERSTER H.; TISSIER C. P.; MUELLER L.; PALEY S.; KARP P. D.; RHEE S. Y.; MetaCyc and AraCyc. Metabolic Pathway Databases for Plant Research. **Plant Physiol**. 2005 May;138(1):27-37.

ZHANG Z.; SCHAFFER, A. A.; MILLER, W.; MADDEN, T. L.; LIPMAN, D.J.; KOONIN, E. V.; ALTSCHUL, S. F. Protein sequence similarity searches using patterns as seeds. **Nucleic Acids Res**. 1998 Sep 1;26(17):3986-90.
Disponível em: <http://www.ncbi.nlm.nih.gov/BLAST/>. Acesso em: 18 abr 2005.

YEH, I.; HANEKAMP, T.; TSOKA, S.; KARP, P. D.; ALTMAN, R. B. Computational analysis of Plasmodium falciparum metabolism: organizing

genomic information to facilitate drug discovery. **Genome Res**. 2004 May;14(5):917-24. Epub 2004 Apr 12.

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo