Universidade de São Paulo
Instituto de Matemática e Estatística
Programa de Pós-Graduação em Bioinformática

Fatores moleculares da determinação do sexo e casta e evolução de famílias multigênicas na abelha eusocial *Apis mellifera*

Alexandre dos Santos Cristino

Tese apresentada ao Programa de Pós-Graduação em Bioinformática do Instituto de Matemática e Estatística da Universidade de São Paulo para obtenção do título de *Doutor* em *Bioinformática*.

São Paulo
2007

Universidade de São Paulo
Instituto de Matemática e Estatística
Programa de Pós-Graduação em Bioinformática

Fatores moleculares da determinação do sexo e casta e evolução de famílias multigênicas na abelha eusocial *Apis mellifera*

Alexandre dos Santos Cristino

Tese apresentada ao Programa de Pós-Graduação em Bioinformática do Instituto de Matemática e Estatística da Universidade de São Paulo para obtenção do título de *Doutor* em *Bioinformática*.

Orientador: Profa. Dra. Zilá Luz Paulino Simões

São Paulo
2007

Dedicatória

À minha linda esposa e amiga Mariana.

A toda minha Família.

Agradecimentos

À Profa. Dra. Zilá Luz Paulino Simões pela orientação e apoio constantes durante a execução deste trabalho.

Ao Prof. Dr. Luciano da Fontoura Costa pela co-orientação e apoio prestados para a condução e interpretações das análises computacionais.

Ao Dr. Charles Claudianos pela co-orientação nos estudos e experimentos de evolução molecular desenvolvidos na Australian National University, Canberra, Austrália.

Ao Dr. John Oakeshott e Dra. Robyn Russell por todo o suporte técnico oferecido para a realização dos experimentos de evolução molecular conduzidos no CSIRO, Canberra, Austrália.

Ao Dr. Denis Anderson pelo fornecimento de abelhas *Apis* coletadas na Ásia.

Ao Prof. Dr. Marco Antônio Del Lama pelo fornecimento de amostras da abelha *Eulaema nigrita*.

Ao Prof. Dr. Klaus Hartfelder pela oportunidade de colaboração durante o projeto de anotação dos genes de casta sob sua liderança.

Aos colegas Profa. Dra. Márcia M. G. Bitondi, Francis de Morais Franco Nunes e Carlos Henrique Lobo pelo excelente trabalho durante as anotações dos genes de casta.

À Dra. Adriana Mendes do Nascimento pela valiosa ajuda na identificação do gene *doublesex* de determinação do sexo.

Ao técnico Luiz R. Aguiar pelo trabalho de manejo com as colônias de abelhas em Ribeirão Preto, São Paulo, Brasil.

À Michelle Williams e Sunita Biswas pela constante ajuda durante os experimentos de evolução molecular.

Ao amigo Christian "Kiko" Reis por incontáveis durante vários momentos deste trabalho.

À Patrícia Cristina Martorelli, da secretaria de Pós-graduação em Bioinformática do IME-USP, pelos auxílios administrativos prestados.

À CAPES pelo financiamento da bolsa de doutorado e da bolsa de estágio de doutorado no exterior.

Conteúdo

Lista de Figuras

Lista de Tabelas

Resumo

A bioinformática se torna cada vez mais importante nas análises e gerenciamento de projetos em biotecnologia e nas ciências biológicas e da saúde. Esta nova área do conhecimento aplica princípios de tecnologia e ciência da informação para transformar a enorme e complexa quantidade de dados biológicos em informações úteis. Os fatores moleculares que determinam o sexo e as castas de *Apis mellifera*, e a expansão de famílias gênicas observadas no genoma deste inseto social são as questões que nos motivam neste estudo. A partir de seqüências genéticas disponíveis na rede e de informações na literatura, objetivamos identificar genes e processos biológicos que participam da determinação do sexo e casta e famílias multigênicas relacionadas à evolução da eusocialidade nestas abelhas. Genes de determinação e diferenciação do sexo em *Drosophila melanogaster* foram identificados em *A. mellifera* e revelaram fatores de transcrição (*dsx* e *ix*) conservados em metazoários. Redes regulatórias putativas controladas por estes fatores estão principalmente relacionadas aos processos metabólicos celulares, à transcrição e ao desenvolvimento de órgãos. A determinação de castas é um modelo para estudos de plasticidade fenotípica. Diferenças nutricionais no desenvolvimento de larvas fêmeas determinam os fenótipos de rainhas e operárias que se desenvolvem pela expressão diferencial de genes dos processos metabólicos e do desenvolvimento. Elementos regulatórios putativos foram descobertos por análises computacionais dos promotores dos genes de casta. O número e a organização destes elementos são diferentes em rainhas e operárias. As implicações da determinação do sexo e casta no sistema eusocial em *A. mellifera* ainda são pouco conhecidas. Expansões de algumas famílias multigênicas podem estar relacionadas com a evolução do sistema social nestas abelhas. Fragmentos de 4 genes nucleares (*ache2*, *or83b*, *lw-rh* e *mrjp*) de 7 espécies de abelhas com corbícula (*A. mellifera*, *A. cerana*, *A. dorsata*, *A. florea*, *M. quadrifasciata*, *B. terrestris*) foram usados para o cálculo de taxas de mutações silenciosas. O sistema eusocial nas abelhas com corbícula parece ter evoluído a partir de um ancestral comum e coincide com a radiação das angiospermas. As duplicações gênicas de maior interesse codificam proteínas receptoras de olfato (OR), proteínas de degradação de xenobiontes e feromônios (P450) e proteínas da geléia real (MRJP). As radiações destes genes parecem ter ocorrido principalmente nas abelhas altamente eusociais (tribo Apini). Embora, as funções destes novos genes ainda sejam pouco conhecidas, algumas delas têm sido caracterizadas como específicas de sexo e casta. Os resultados apresentados neste estudo foram obtidos, principalmente, por análises computacionais que direcionaram o delineamento de hipóteses biológicas e validações experimentais para identificação de genes e processos biológicos envolvidos na determinação do sexo e casta e na evolução da eusocialidade.

Abstract

Bioinformatics has become a very important tool in analysing and managing projects in biotechnology and in biological sciences and health care. This new area applies principles of Technology and Information Science in order to transform the complex and enormous quantity of biological data into useful information. The molecular factors that determine sex and caste in *Apis mellifera* and the expansion of multigenic families observed in the genome of this social insect are the questions that motivate our studies. Considering genetic sequences available on the internet and information from the literature, the aim of the present study was to identify genes and biological processes that take place in sex and caste determination and also multigenic families related to the evolution of eusociality in honey bees. Sex determination and differentiation genes in *Drosophila melanogaster* were identified in *A. mellifera* and showed conserved transcription factors (*dsx* and *ix*) in metazoa. Putative regulatory networks controlled by these factors are mainly related to cellular metabolic processes, to transcription and to organs development. Caste determination is a model to study phenotypic plasticity. Nutritional differences during the development of female larvae determine the phenotype of queens and workers that are formed through the differential expression of genes related to metabolic and developmental processes. Putative regulatory elements were discovered by computational analysis of the promoters of caste genes. The number and the organization of these cis elements are different in queens and workers. The implications of caste and sex determination in the eusocial system in *A. mellifera* are still not very known. Expansions of some multigenic families can be related to the evolution of the social system in these bees. Fragments of 4 nuclear genes (*ache2*, *or83b*, *lw-rh* and *mrjp*) from 7 species of corbiculate bees (*A. mellifera*, *A. cerana*, *A. dorsata*, *A. florea*, *M. quadrifasciata*, *B. terrestris* were used for the calculation of silent mutation rates. The eusocial system in corbiculate bees seems to have evolved from a common ancestor and coincides with the radiation of the angiosperms. The gene duplications of highest interest encode olfactory receptor (OR) proteins, xenobiontes and pheromones degradation (P450) proteins and royal jelly (MRJP) proteins. The radiations of these genes seem to have occurred mainly in highly eusocial bees. Even though the function of these novel genes are still poorly understood, some of them have been characterized as sex and caste-specific. The results presented in this study were obtained mainly through hypothesis-driven computational analysis and experimental validations for the identification of genes and biological processes most envolved in sex and caste determination and in the evolution of eusociality in honey bees.

1 Introdução

1.1 Contexto histórico do estudo

Atualmente, mais de 1 bilhão de letras do código genético representam genes e genomas parciais e completos de mais de 165.000 organismos. Esta quantidade massiva de dados biológicos está armazenada na base de dados GenBank (National Center for Biotechnology Information - NCBI) que conta com a colaboração de outras 2 grandes bases, EMBL (European Molecular Biology Laboratory) e DDBJ (DNA Data Bank of Japan). Obviamente, o uso de computadores se torna indispensável no gerenciamento e análise de dados transformando-os em informação útil para a sociedade. A bioinformática[1] surgiu principalmente da necessidade do desenvolvimento de um conjunto de ferramentas prático que permita o gerenciamento e análise dos dados destas seqüências genéticas.

Outras bases de dados (ex: Gene Ontology, KEGG, etc) que descrevem propriedades funcionais e estruturais dos produtos gênicos objetivam estabelecer um vocabulário controlado de atributos biológicos e bioquímicos já bem definidos pelas ciências biológicas. Tais relacionamentos são estabelecidos a partir de observações experimentais de fenômenos biológicos através da análise de expressão gênica, manipulação genética de genes em células e organismos e análises funcionais e estruturais. Toda esta rede de informação deve ser analisada de maneira global para fornecer informações biológicas relevantes que podem ser direta ou indiretamente aplicadas na agricultura, medicina e em outras áreas de interesse comercial e econômico.

[1] Bioinformática é uma área interdisciplinar envolvendo biologia, ciência da computação, matemática e estatística para a análise de dados de seqüências genéticas, conteúdo genômico e predições funcionais e estruturais de macromoléculas. A nova área se sobrepõe a outras áreas de pesquisa, tais como informática e biologia computacional, que estão tradicionalmente envolvidas com o desenvolvimento de tecnologias para o gerenciamento de informação e o desenvolvimento de algoritmos eficientes que solucionem problemas computacionais (ex: alinhamento múltiplo) (MOUNT, 2004).

O acesso a programas de distribuição livre juntamente com a expansão da rede de *internet* formam as bases para a evolução da bioinformática. Seria difícil imaginar um modelo mais eficiente para a manutenção semântica dos conceitos e métodos desenvolvidos por pesquisadores das diferentes áreas (biólogos, matemáticos, cientistas da computação, engenheiros, entre outros) e partes do mundo. O uso, a modificação e a redistribuição de vários programas de bioinformática são completamente livres e permitem que métodos eficientes na análise de dados sejam melhorados e aplicados sem nenhuma restrição.

O presente estudo é conduzido com a colaboração de 3 grupos de pesquisa: Laboratório de Biologia do Desenvolvimento de Abelhas - FFCLRP/USP, Grupo de Visão Cibernética - IFSC/USP e Group of Visual Science - RSBS/ANU. Estes grupos trabalham conceitos e métodos que auxiliam a compreensão de processos biológicos envolvidos na formação de padrões fenotípicos e unem esforços para estudar os mecanismos moleculares envolvidos na formação de fenótipos alternativos em abelhas que são determinados geneticamente (sexo) e pelo ambiente (casta).

1.2 Motivação

A motivação que nos orienta neste estudo visa criar elementos para a compreensão de um problema teórico em biologia: os mecanismos genéticos envolvidos na determinação e desenvolvimento de sexo e casta em *Apis mellifera*. Embora estas abelhas e outros insetos da mesma ordem (Hymenoptera) tenham importância em vários aspectos da sociedade humana (ex: indústria de alimentos, produção agrícola e saúde) pouco se conhece sobre os mecanismos genético-moleculares envolvidos em questões biológicas importantes como a determinação do sexo e de casta e a evolução da eusocialidade nestes insetos.

O genoma da abelha *A. mellifera* foi publicado na revista Nature (The Honeybee Genome Sequencing Consortium, 2006). Este fato possibilita avanços importantes para o estudo da biologia básica que há muito tempo utiliza a abelha melífera como um modelo de estudos biológicos. Entretanto, o uso dos métodos tradicionais da genética (ex: mutagênese e transgenia) é impraticável (ainda que possível) com estes insetos, dado sua complexidade de ciclo de vida e manutenção de linhagens mutantes. Neste contexto, a bioinformática torna-se essencial no

desenvolvimento de ferramentas e predições que apontem candidatos a genes envolvidos em diferentes aspectos do desenvolvimento sexual (ex: expressão diferencial de genes, formação de tecidos e órgãos, casta, comportamento, etc).

A bioinformática catalisa a transformação dos dados em informações hipotéticas (ou putativas) úteis na escolha dos marcadores de interesse bem como no delineamento experimental hipótese-dirigido. A determinação do sexo pela partenogênese, a determinação de castas por diferenças nutricionais e hormonais, o sistema de comunicação entre indivíduos de uma colônia, a capacidade de memória e aprendizagem, entre outros fatores, fazem destas abelhas um sistema modelo único para estudos de processos biológicos que expliquem a formação de padrões fenotípicos (sexo, casta, órgãos, comportamento) através da integração da genética, ambiente e evolução.

1.3 Fundamentos de biologia

A diversidade no desenvolvimento das formas nos animais está de alguma maneira codificada no genoma. Entretanto, ter apenas o DNA (ácido desoxiribonucléico) como informação não garante o entendimento dos mecanismos pelos quais a forma é adquirida durante o desenvolvimento e na evolução (DAVIDSON, 2001). Todos os organismos se desenvolvem e alguns mecanismos primordiais do desenvolvimento já podem ser observados nos seres unicelulares. Estes seres eucariontes possuem formas simples de diferenciação celular e reprodução sexuada (GILBERT, 2003), mas os padrões de desenvolvimento se tornam realmente interessantes e complexos nos organismos multicelulares. Uma extraordinária variedade de tipos celulares (ex: células do osso, células musculares, células nervosas, entre muitas outras) se desenvolvem, no momento e local específico do organismo, a partir de uma única célula-ovo (ALBERTS et al., 2002).

O DNA de uma única célula contém a informação necessária para a construção de um organismo com centenas de tipos celulares que desempenham funções especializadas para a manutenção viável do sistema biológico. As diferenças fenotípicas destas células são determinadas pela transcrição seletiva da informação armazenada no DNA em resposta a sinais ambientais. Tais variações externas são percebidas por receptores protéicos da membrana celular

e disparam cascatas reguladoras específicas para finalmente controlar a expressão de genes formando o novo repertório de proteínas e RNA reguladores dando origem a um determinado fenótipo (ex: tecidos, órgãos, sexo, casta, comportamento, etc)(ALBERTS et al., 2002; GILBERT, 2003).

A evolução da complexidade morfológica e comportamental dos organismos multicelulares não é explicada apenas pelo aumento no número de genes no genoma (ex: *Caenorhabditis elegans* com $\approx$ 20.000 genes; *Drosophila* com $\approx$ 14.000 genes; *Homo sapiens* com $\approx$ 30.000 genes). A diferença no número de genes dos vertebrados é apenas o dobro dos invertebrados e, mesmo assim, este aumento se origina principalmente da duplicação dos genes já existentes no genoma. Desta maneira, a complexidade dos seres vivos surge progressivamente de mecanismos mais elaborados de regulação da expressão gênica(DAVIDSON, 2001; LEVINE; TJIAN, 2003).

1.3.1 A anatomia do gene

O DNA é a informação hereditária de todos os organismos celulares. Tradicionalmente, o gene é definido com um segmento de DNA que codifica uma proteína ou uma molécula de RNA funcional. Os genes de organismos procariontes e eucariontes são estruturalmente diferentes em dois aspectos, (1) os genes eucariontes ficam empacotados em um complexo de interações entre DNA e proteínas (histonas) chamado cromatina e (2) os genes eucariontes não são necessariamente colineares com seus respectivos produtos protéicos (LEWIN, 2000) (Figura 1.1). Genes podem ser classificados em estruturais (ex: enzimas, proteínas do citoesqueleto, de membrana, tRNA, rRNA, etc) ou reguladores (ex: miRNA, fatores de transcrição, etc).

A síntese de uma proteína envolve um processo de decodificação pelo qual a informação genética transcrita em mRNA é traduzida em aminoácidos. Existem 20 aminoácidos que podem ser usados na tradução do mRNA. A tradução é o reconhecimento seqüencial de trincas de nucleotídeos (códon) e cada códon é traduzido em um aminoácido segundo um conjunto de regras chamado código genético (Tabela 1.3.1) (LI; GRAUR, 1991).

A arquitetura típica de um gene eucarionte que codifica proteína(LEWIN, 2000; DAVIDSON, 2001; ALBERTS et al., 2002; GILBERT, 2003) está representado na Figura 1.2 e consiste em:

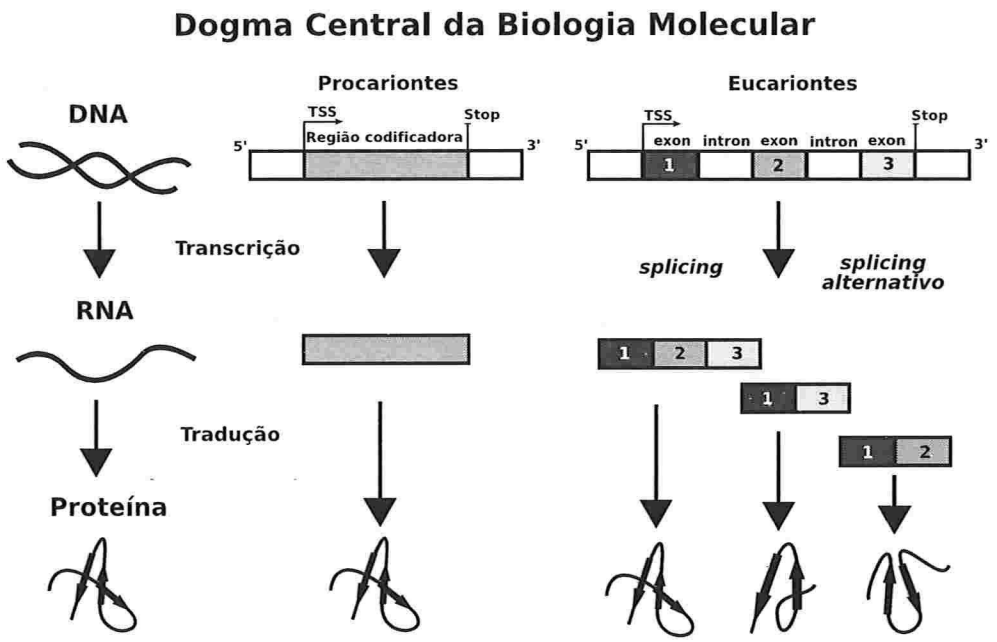


Figura 1.1: O dogma central de biologia molecular considera que o fluxo da informação genética ocorre em duas etapas principais, (1) transcrição do DNA em RNA e (2) tradução do RNA em proteína. Os eucariontes apresentam maior complexidade no fluxo de informação possibilitando a produção de proteínas combinando diferentes domínios funcionais codificados a partir de um único gene.

Região promotora é responsável pela ativação da transcrição gênica, que requer a presença de fatores de transcrição específicos e a RNA polimerase (RNApol) para iniciar a transcrição. Nos eucariontes, esta região reguladora pode ser dividida em seqüências promotoras e *enhancers*. A seqüência promotora (TATAbox) é o local onde se ligam os fatores de transcrição basais (ex: TFIIs[2], TBP[3], TAFs[4] e RNApol). Os *enhancers* são seqüências *cis* que, mesmo à grande distância do promotor, controlam a eficiência e a taxa de transcrição do gene. Eles podem funcionar tanto à montante (*upstream*) como à jusante (*downstream*) da unidade transcricional do gene.

Sítio de inicialização de transcrição (TSS) este sítio é conhecido como *cap* e é acrescentado pela ação sucessiva de 3 enzimas na extremidade 5' do RNA mensageiro (mRNA) que está sendo transcrito. Este sinal *cap* é utilizado pela célula para diferenciar os mRNAs de

[2]TFII abreviado do inglês *transcription factor for RNA polymerase II*
[3]TBP abreviado do inglês *TATA-binding protein*
[4]TAF abreviado do inglês *TBP-associated factors*

Tabela 1.1: O código genético é determinado por trincas de nucleotídeos (códons) que correspondem a aminoácidos. O código é degenerado onde cada aminoácido pode ser decodificado por um ou mais códons. Em geral, as substituições na terceira posição dos códons podem ser silenciosas, isto é, mutações nesta posição nem sempre implicam em alterações no aminoácido codificado.

	2a Base				
1a Base	**T**	**C**	**A**	**G**	**3a Base**
T	TTT Phe (F)	TCT Ser (S)	TAT Tyr (Y)	TGT Cys (C)	T
	TTC Phe (F)	TCC Ser (S)	TAC Tyr (Y)	TGC Cys (C)	C
	TTA Leu (L)	TCA Ser (S)	TAA STOP	TGA STOP	A
	TTG Leu (L)	TCG Ser (S)	TAG STOP	TGG Trp (W)	G
C	CTT Leu (L)	CCT Pro (P)	CAT His (H)	CGT Arg (R)	T
	CTC Leu (L)	CCC Pro (P)	CAC His (H)	CGC Arg (R)	C
	CTA Leu (L)	CCA Pro (P)	CAA Gln (Q)	CGA Arg (R)	A
	CTG Leu (L)	CCG Pro (P)	CAG Gln (Q)	CGG Arg (R)	G
A	ATT Ile (I)	ACT Thr (T)	AAT Asn (N)	AGT Ser (S)	T
	ATC Ile (I)	ACC Thr (T)	AAC Asn (N)	AGC Ser (S)	C
	ATA Ile (I)	ACA Thr (T)	AAA Lys (K)	AGA Arg (R)	A
	ATG Met (M)	ACG Thr (T)	AAG Lys (K)	AGG Arg (R)	G
G	GTT Val (V)	GCT Ala (A)	GAT Asp (D)	GGT Gly (G)	T
	GTC Val (V)	GCC Ala (A)	GAC Asp (D)	GGC Gly (G)	C
	GTA Val (V)	GCA Ala (A)	GAA Glu (E)	GGA Gly (G)	A
	GTG Val (V)	GCG Ala (A)	GAG Glu (E)	GGG Gly (G)	G

outros tipos de RNAs e serem propriamente processados e exportados para o citoplasma.

Sítio de inicialização de tradução (Met) este sítio é o primeiro códon[5] (AUG, codifica uma metionina) da proteína e está localizado a dezenas de pares de base (50-200pb) à jusante do TSS.

Região não traduzida 5' (5' UTR) é a seqüência entre AUG e TSS que determina a eficiência que a tradução é iniciada. Algumas 5' UTR[6] contêm seqüências que se ligam a proteínas repressoras de tradução (ex: proteínas que se ligam a íons e/ou hormônios).

Éxon seqüência de nucleotídeos que contêm a informação codificada (ou traduzida) em aminoácidos.

[5]Códon é uma combinação de 3 nucleotídeos que especificam um aminoácido seguindo uma determinada regra de tradução.

[6]UTR abreviado do inglês *untranslated region*

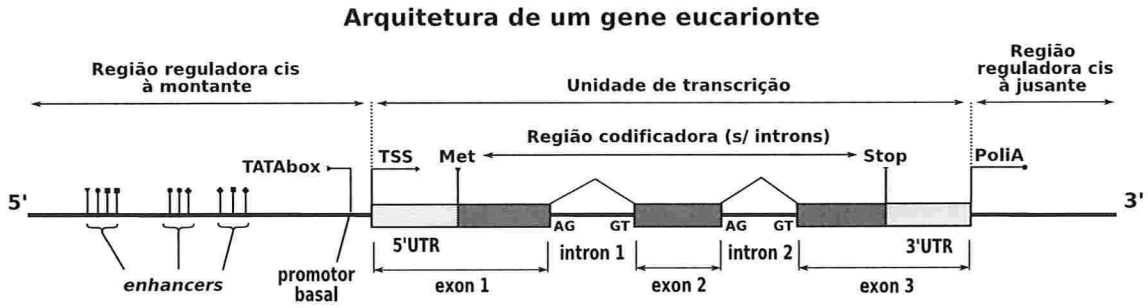


Figura 1.2: Diagrama ilustrando a arquitetura típica de um gene de eucarionte. A unidade de transcrição é flanqueada por regiões intergênicas reguladoras à montante e à jusante da unidade transcricional. Uma série de padrões conservados sinalizam a maquinaria genética responsável pela iniciação de transcrição e posterior edição e tradução do mRNA.

Íntron seqüência nucleotídica não codificante que fica entre os éxons. Esta região é removida do pré-mRNA para se transformar em mRNA maduro pronto para deixar o núcleo e ser traduzido.

Códon de término de tradução (Stop) é o códon de sinalização (em geral, UAA) para o fim da proteína. O complexo ribossomal se desmonta (ou dissocia) ao ler este códon e a proteína é liberada.

Sítio de poliadenilação (PoliA) este sítio (AAUAAA) sinaliza o que será a extremidade 3' do mRNA. Assim que a RNApol passa pelo sítio, duas enzimas reconhecem este sinal na molécula de mRNA, clivando e inserindo uma cauda com 200-300 resíduos de adenina (poliA). A cauda poliA confere estabilidade ao mRNA, permite que o mRNA seja exportado do núcleo para o citoplasma e que o mRNA seja traduzido em proteína.

Região não traduzida 3' (3' UTR) esta região está localizada entre o *stop* códon e a poliA. Esta seqüência possui sinais que direcionam a localização do mRNA na célula. Recentemente, descobriu-se que uma nova classe de RNA (miRNA) está relacionada com o controle negativo pós-transcricional através de sítios específicos na região 3' UTR que levam a molécula de mRNA para a maquinaria de degradação(LAI, 2002).

Elementos *cis* são seqüências de DNA nas proximidades do gene que não são convertidas em nenhuma outra forma de molécula mas funcionam no controle específico de proteínas e RNA reguladores. Algumas das entidades já definidas acima podem ser consideradas

elementos *cis*, por exemplo, TATA box, motivos de ligação de fatores de transcrição, poliA, motivos que regulam o mecanismo de *splicing*, entre outros.

1.3.2 Controle da expressão gênica

O fluxo de informação genética em um célula de bactéria ou humana percorre o mesmo caminho, isto é, a informação contida no DNA é transcrita em RNA, que, por sua vez, é traduzida em proteína. Este princípio é conhecido como *dogma central* da biologia molecular (Figura 1.1). Ainda que células procariontes (bactérias) e eucariontes (célula humana) compartilhem princípios fundamentais no fluxo de informação genética, somente as células eucariontes executam uma série de processamentos no RNA antes que ele deixe o núcleo. Tais passos intermediários incluem o mecanismo de *splicing* do RNA (remoção dos íntrons). Este mecanismo de edição do RNA pode produzir formas alternativas de mRNA pela combinação diferencial dos éxons (*splicing* alternativo; Figura 1.1). Um único gene eucarionte pode ser codificado em uma variedade de isoformas que assumem funções diferentes na célula (LATCHMAN, 1998).

Os dois aspectos que diferenciam genes eucariontes dos procariontes, cromatina e estrutura éxon-íntron, representam uma mudança no paradigma de controle de expressão gênica. O modelo clássico de controle em bactérias é, em geral negativo, isto é, uma proteína repressora é necessária para inibir a ação da RNApol na transcrição do gene (Figura 1.3, A). Em eucariontes, o estado de cromatina garante que os genes fiquem inativos e a ativação da expressão gênica necessita de alguns fatores, como os de transcrição, que desempenham papel fundamental no desempacotamento de regiões da cromatina, dando acesso a outros fatores de transcrição que se ligam a determinadas seqüências na região promotora permitindo que a RNApolII se ligue ao TATAbox e inicie a transcrição do gene.

Assim que a RNApolII inicia a transcrição, a extremidade 5' é modificada pela adição do *cap* e, ao final da transcrição, a extremidade 3' é reconhecida (sinal poliA), clivada e mais de 200 resíduos de adenina (cauda poliA) são adicionados a ela. O pré-mRNA contendo *cap*, éxons e íntrons e poliA é agora processado pelo mecanismo de *splicing* pelo qual íntrons são removidos do RNA. O mRNA funcional é transportado do núcleo para o local adequado no citoplasma onde será traduzido em proteína (Figura 1.3, A) (LEWIN, 2000). A expressão gênica específica em

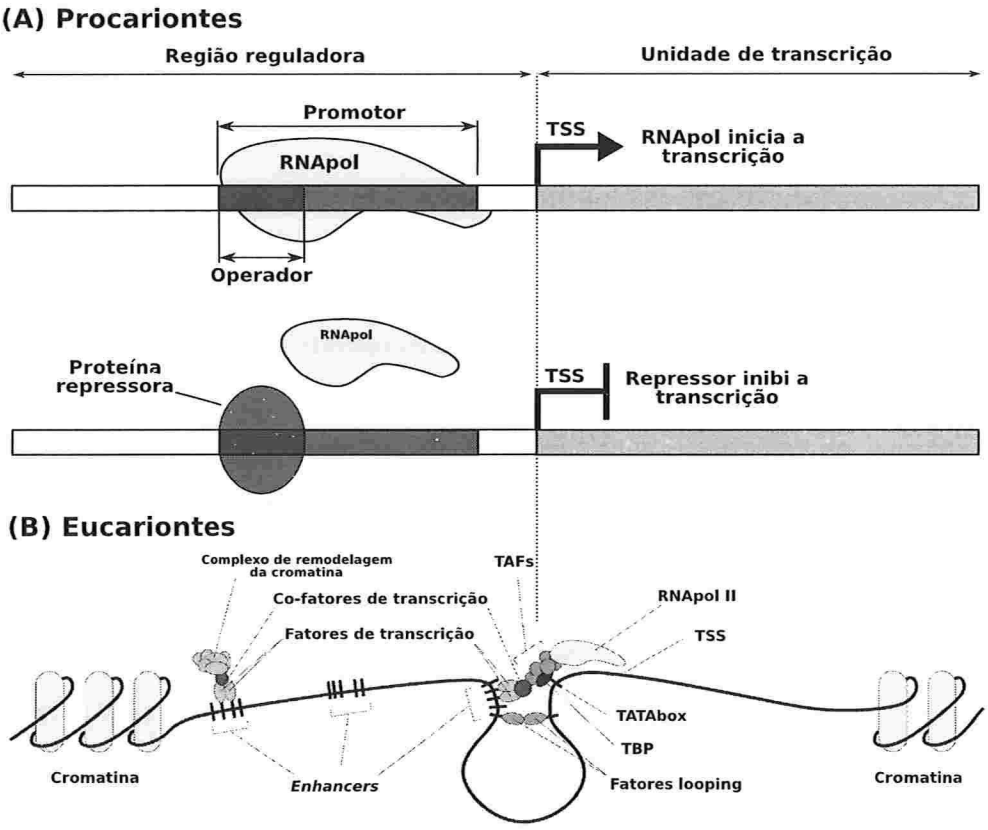


Figura 1.3: Esquema ilustrativo do controle de transcrição gênica em procariontes e eucariontes. (A) Em procariontes, o controle é geralmente negativo sendo necessário a presença de um repressor para inibir a ligação de RNApol que transcreve a informação do DNA em RNA. (B) Em eucariontes, o DNA está normalmente compactado em uma estrutura denominada cromatina portanto a transcrição gênica está reprimida por *default*. Para que a transcrição gênica seja ativada é necessária a presença de vários fatores e cofatores de transcrição que primeiramente remodelam a cromatina permitindo o acesso de fatores que se ligam de maneira complexa a região promotora e estabilizam a RNApol para só então iniciar a transcrição do gene (modificado de (WRAY et al., 2003)).

determinados tecidos poderia surgir pelo controle, em qualquer nível ou estágio, da produção de um mRNA funcional. No entanto, o controle primário da expressão gênica em qualquer célula ocorre, principalmente, na inicialização da transcrição quando fatores de transcrição específicos se ligam aos promotores e *enhancers* criando condições favoráveis para a atividade da RNApol (LATCHMAN, 1998).

Embora o controle primário da expressão gênica esteja relacionado ao processo de transcrição, existem situações onde as alterações na taxa de síntese de uma proteína ocorrem sem uma

alteração na taxa de transcrição do gene correspondente. Em alguns organismos inferiores, a regulação pós-transcricional constitui a forma predominante de regulação gênica. A regulação pós-transcricional pode ocorrer em muitos estágios entre o transcrito primário[7] e a tradução do RNAm processado. Os diferentes mecanismos de regulação pós-transcricional do RNA podem ocorrer durante o *splicing*, o transporte, a estabilidade e a tradução.

Os genes eucarióticos são intercalados por seqüências codificadoras (éxons) e não codificadoras (íntrons) de proteínas. As regiões intrônicas são removidas do transcrito primário pelo mecanismo de *splicing*. Diferentes tecidos e células podem ou não realizar *splicing* de um determinado transcrito de RNA, produzindo ou não uma proteína correspondente.

O *splicing* alternativo possibilita a tradução de diferentes proteínas a partir de uma mesma seqüência de DNA. Este mecanismo ocorre em genes envolvidos em uma variedade de processos celulares que vão desde genes reguladores do desenvolvimento embrionário ou de determinação do sexo (como em *D. melanogaster*, por exemplo) até aqueles envolvidos na contração muscular ou função neuronal em mamíferos.

Os mecanismos de *splicing* alternativo podem ser divididos em três grupos:

- Situações onde o final 5' de transcritos diferencialmente processados é diferente;
- Situações onde o final 3' de transcritos diferencialmente processados é diferente;
- Situações onde ambos trechos terminais 5' e 3' de transcritos diferencialmente processados são idênticos.

Nas situações onde o final 5' é diferente, o *splicing* alternativo surge a partir de diferenças no sítio de inicialização transcricional (sítio promotor). Neste caso, a variação no *splicing* é secundária à seleção dos diferentes promotores em diferentes tecidos ou condições. Nas situações onde o final 3' é diferente, o *splicing* alternativo ocorre após a produção do transcrito primário que em seguida sofre uma clivagem e poliadenilação (sítio poliA) em diferentes posições dentro do transcrito primário em diferentes tecidos ou condições. Nas situações onde as extremidades 5' e 3' dos transcritos são idênticas, o *splicing* alternativo pode ser explicado pela existência de fatores de transcrição que regulam este mecanismo em diferentes tecidos

[7]Molécula de RNAm transcrita diretamente do gene sem nenhum processamento, isto é, com íntrons e éxons

ou condições, já que não há, neste caso, o uso diferencial de promotores ou sítios de poliA (LATCHMAN, 1998; LEWIN, 2000; ALBERTS et al., 2002).

1.3.3 A teoria neutra da evolução molecular

A teoria neutra da evolução molecular (ou neutralismo) considera que a maioria das modificações moleculares na evolução ocorrem pela fixação randômica de mutações neutras ou quase neutras (KIMURA, 1968; KING; JUKES, 1969). A neutralidade, nesta hipótese, significa que o destino dos alelos[8] é determinado principalmente pela deriva genética[9].

O processo de duplicação do DNA garante que a informação genética seja copiada com muita precisão. Entretanto, eventuais erros (mutações) ocorrem. Quando estas mutações acontecem nas células germinativas[10], elas são propagadas para seus descendentes, possibilitando modificações genéticas que alteram o *valor adaptativo* (*fitness*) do indivíduo ou da população. Por outro lado, a maioria das mutações é neutra e não causa nenhum tipo de alteração que possa sofrer pressão de seleção. A mutação pode ser causada por substituição de um nucleotídeo por outro, por deleção de um ou mais nucleotídeos na seqüência de DNA, por inserção de um ou mais nucleotídeos no DNA e por inversão da polaridade de uma seqüência de DNA (LI; GRAUR, 1991; FUTUYMA, 1998).

Para estudar a dinâmica de substituição de nucleotídeos entre duas seqüências é necessário assumir as probabilidades de substituição de um nucleotídeo por outro. Existem vários modelos para calcular o número de substituições entre seqüências (LI; WU; LUO, 1985; KIMURA, 1980; JUKES; CANTOR, 1969). Os modelos mais usados são: um-parâmetro[11] (JUKES; CANTOR, 1969) e dois-parâmetros[12] (KIMURA, 1980).

O processo básico na evolução do DNA é a substituição de nucleotídeos ao longo do tempo. As substituições que ocorrem nas regiões codificadoras das proteínas podem ser caracterizadas como sinônimas (silenciosas) ou não-sinônimas (alterações de aminoácido). A

[8] variações ou polimorfismos de um gene

[9] é por definição um processo estocástico pelo qual os alelos são transmitidos de uma geração para outra dado que apenas uma fração de todas as possíveis combinações alélicas compõe o zigoto que se torna maduro

[10] linhagem de células responsáveis pela transmissão hereditária do material genético

[11] o modelo assume que qualquer substituição ocorra randomicamente entre os 4 tipos de nucleotídeos.

[12] o modelo considera a ocorrência de um viés na direção das substituições. Geralmente, as transições (trocas entre A e G e entre C e T) são mais freqüentes do que as transversões (todas as outras trocas).

taxa de substituição sinônima é maior do que a não-sinônima e é similar para genes diferentes (MIYATA; YASUNAGA; NISHIDA, 1980). O código genético é degenerado, isto é, um aminoácido pode ser codificado por mais de um códon (Figura 1.3.1). A degeneração do código ocorre, principalmente, na terceira posição do códon, onde 70% das substituições são silenciosas e na primeira posição, onde 5% das substituições são sinônimas ou silenciosas, enquanto todas as substituições na segunda posição são não-sinônimas (NEI; GOJOBORI, 1986).

O número de substituições sinônimas e não-sinônimas pode ser estimado a partir do modelo de Nei-Gojobori (NEI; GOJOBORI, 1986) que é uma simplificação do modelo proposto por Miyata *et al.* (MIYATA; YASUNAGA; NISHIDA, 1980). O número de sítios sinônimos (s_d) e não-sinônimos (n_d) é computado para cada códon, considerando as propriedades de mudança destes (Figura 1.3.1). Para uma seqüência de DNA com r códons, o número total de sítios sinônimos e não-sinônimos é dado por $S_d = \sum_{j=1}^{r} s_{dj}$ e $N_d = \sum_{j=1}^{r} n_{dj}$, onde s_{dj} e n_{dj} é o valor de s_d e n_d para o códon j, respectivamente. Assim, a proporção de diferenças sinônimas e não-sinônimas pode ser calculada por $p_S = \frac{S_d}{S}$ e $p_N = \frac{N_d}{N}$, onde S e N são o número médio de sítios sinônimos e não-sinônimos, respectivamente (NEI; GOJOBORI, 1986). Quando o tempo de divergência entre as seqüências é grande (¿ 100Ma), é necessário considerar a possibilidade de ocorrência de múltiplas substituições em um mesmo sítio. A fórmula mais utilizada para esta correção é dada por $K = -\frac{3}{4}\ln(1 - \frac{4}{3}p)$, onde p é p_S ou p_N (JUKES; CANTOR, 1969).

O modelo apresentado acima é muito útil para estudos em evolução molecular e, a partir deles, pode-se estimar taxas evolutivas e reconstruir a história evolutiva dos organismos. A taxa de mutação em uma dada seqüência é aproximadamente constante no tempo em todas as linhagens e, sobre esta observação, é definido o conceito de relógio molecular (ZUCKERKANDL; PAULING, 1965). O relógio pode ser calibrado a partir do tempo absoluto de divergência das espécies, estimado pelos registros fósseis (MIYATA; YASUNAGA; NISHIDA, 1980).

1.3.4 Sistemas genéticos de reprodução

Há uma grande diversidade de padrões reprodutivos nos organismos. Eles podem ser sexuados ou assexuados, haplóides, diplóides ou poliplóides, ter sexos separados ou não, muitos ou poucos cromossomos e ter ou não cromossomos sexuais. Todas estas diferenças alteram os

padrões de variação genética de indivíduos e populações (BULL, 1983; FUTUYMA, 1998).

A reprodução sexuada ocorre em quase todos os organismos eucariontes e envolve a produção de dois gametas haplóides (n) através da divisão reducional (ou meiose), que se unem para formar um zigoto diplóide (2n). Embora o custo genético seja maior na reprodução sexuada (meiose parental contribui apenas com 50% do conteúdo genético de sua prole), a recombinação possibilita a remoção de mutações letais bem como a criação de novas combinações de genes e elementos genômicos, oferecendo uma maior flexibilidade para adaptação das populações em mudanças ambientais (BARTON; CHARLESWORTH, 1998).

A existência de tipos sexuais implica na existência de mecanismos e processos biológicos envolvidos na (1) razão sexual (proporção de machos), (2) alocação sexual (alguns organismos são hermafroditas, outros possuem sexos completamente separados) e (3) determinação do sexo. Todas estas 3 questões têm influência na adaptação dos mecanismos moleculares da determinação do sexo.

Mecanismos de determinação do sexo

A determinação do sexo atraiu interesse desde o momento da descoberta dos cromossomos sexuais em insetos (MCCLUNG, 1902; WILSON, 1905). Estas descobertas foram cruciais para a síntese da genética mendeliana com a citogenética e subseqüente elucidação dos fenômenos cromossômicos envolvidos na determinação do sexo em *D. melanogaster* (MORGAN, 1910; BRIDGES, 1916).

Os sistemas de determinação do sexo podem ser divididos em 2 amplas categorias: (1) sistema ambiental de determinação do sexo (ESD[13]), onde o sinal inicializador está nas condições ambientais e (2) sistema genético de determinação do sexo (GSD[14]), onde o sinal inicial é dado por combinações de elementos genéticos (BULL, 1983).

O sistema ESD pode ser exemplificado por quase todos os répteis (exceto algumas tartarugas e a maioria das cobras que possuem padrões de cromossomos sexuais similares às aves, ZZ/ZW) em que a temperatura durante um período específico na embriogênese determina o destino sexual do ovo (FERGUSON; JOANEN, 1982; HARVEY; SLATKIN, 1982). O

[13]ESD abreviado do inglês *Environment Sex Determination*
[14]GSD abreviado do inglês *Genetic Sex Determination*

sistema GSD ocorre principalmente em invertebrados (insetos e nematódeos) e vertebrados (aves e mamíferos), apresentando grande diversidade de padrões envolvendo combinações de cromossomos sexuais (ex: XX/XY em mamíferos e ZZ/ZW em aves) ou não (ex: haplodiploidia em muitos invertebrados) (BULL, 1983). Em mamíferos, as fêmeas são o sexo homogamético (XX) enquanto machos, o hetorogamético (XY) (LYON, 1992). Entretanto, em aves este padrão é invertido sendo o macho homogamético (ZZ) e a fêmea, heterogamética (ZW) (SMITH; SINCLAIR, 2004).

Nos invertebrados, as diferenças nos padrões de determinação do sexo são grandes e podem envolver cromossomos sexuais e a razão entre o cromossomo X e os autossomos (por exemplo, em insetos dípteros tal como *D. melanogaster* e em vermes nematódeos tal como *Caenorhabditis elegans*), (CLINE; MEYER, 1996) bem como um ou alguns genes do mecanismo complementar do sexo, observado em insetos himenópteros (por exemplos, abelhas, vespas e formigas) (WHITING, 1933, 1943). Os himenópteros não possuem cromossomos sexuais e determinam o sexo pela haplodiploidia, onde um óvulo haplóide não fecundado se desenvolve em macho e um ovo diplóide fecundado dá origem a uma fêmea (WHITE, 1973).

As diferenças nos sinais inicializadores (ex: cromossomos sexuais, razão X:A, loco sexual e temperatura) da determinação do sexo contrastam com uma considerável homologia entre alguns genes (ex: *DMRT1* em vertebrados, *MAB3* em nematódeos e *dsx* em insetos) das cascatas reguladoras da determinação das gônadas e outras características fenotípicas entre os metazoários (MARíN; BAKER, 1998; RAYMOND et al., 1998; GRAHAM; PENN; SCHEDL, 2002; CREWS, 2003) (Figura 1.4. A evolução das cascatas gênicas envolvidas na determinação do sexo é mais rápida no topo da cascata (sinais iniciadores), enquanto nos genes da base destas vias de determinação do sexo, a evolução é mais lenta e os genes são mais conservados em estrutura e função (WILKINS, 1995).

A haplodiploidia (ou arrenotoquia, um tipo de partenogênese, Figura 1.5) é observada em aproximadamente 20% dos metazoários e na maioria dos insetos da ordem Hymenoptera. O sistema de determinação do sexo destes organismos têm propriedades exclusivamente genéticas, sem nenhuma influência de fatores ambientais (BULL, 1983). A hipótese mais aceita para explicar a determinação do sexo em Hymenoptera é a determinação complementar do sexo (CSD, sigla em inglês) com um loco sexual e vários alelos (WHITING, 1933, 1943) e CSD com vários locos e vários alelos (SNELL, 1935; CROZIER, 1971).

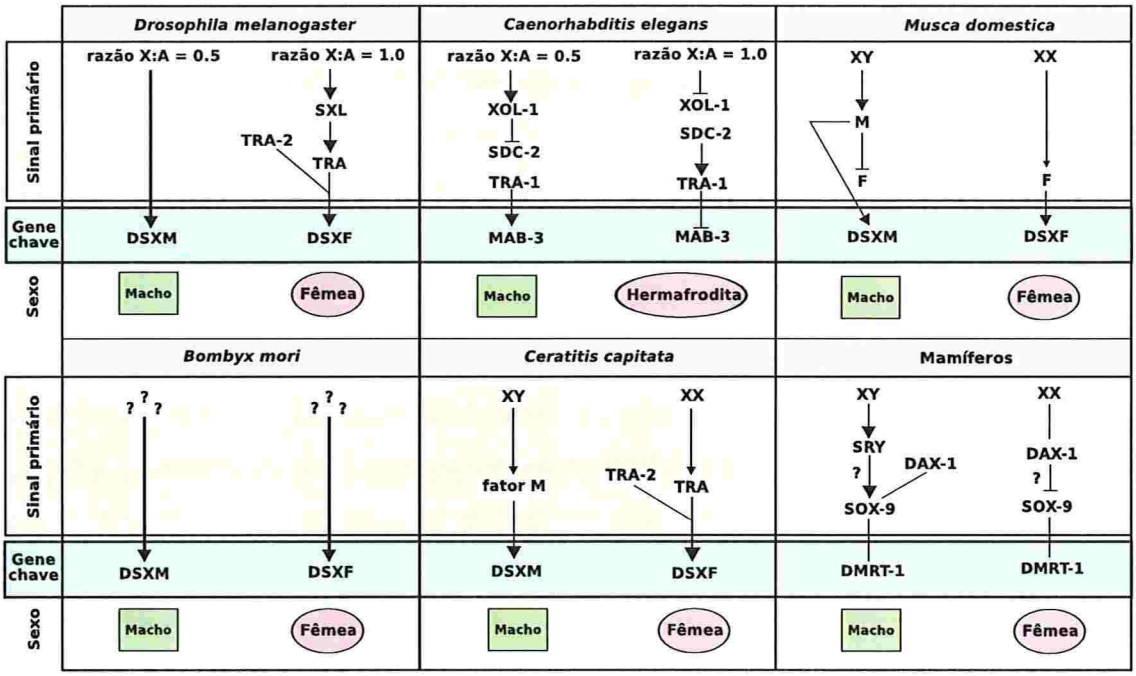


Figura 1.4: Comparação das vias genéticas de determinação do sexo em metazoários. As diferenças ocorrem, principalmente, no sinal primário da cascata. Os genes chave (ou terminais) são bem conservados entre os metazoários e codificam fatores de transcrição que regulam, de forma alternativa (ou dualista), a transcrição de genes de diferenciação sexual.

Experimentos de genética clássica demonstraram a existência de machos diplóides em abelhas a partir de acasalamentos endogâmicos. Estes machos diplóides são retirados da célula de cria pelas operárias da colônia assim que os ovos eclodem e o favo parece ter falhas na postura da rainha (MACKENSEN, 1951) . Recentemente, Beye e col (BEYE et al., 2003) demonstraram experimentalmente que a hipótese de (WHITING, 1943) (CSD com um loco multi-alélico) é a mais aceita para explicar a determinação do sexo em *Apis mellifera*. O gene *complementary sex determiner* (*csd*) encontrado em abelha codifica uma proteína de expressão constitutiva (SR-RB proteína) capaz de regular o "splicing" do mRNA de *dsx* formando uma proteína DSX (nenhum dado concreto sobre o gene

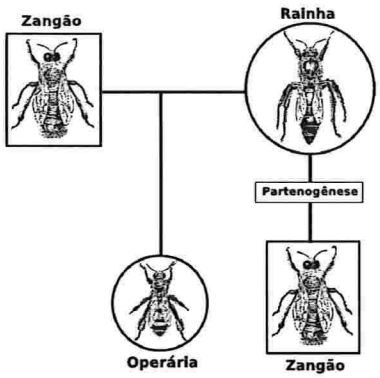


Figura 1.5: Sistema de determinação do sexo por haplodiploidia (ou arrenotoquia) que é um tipo de partenogênese. Os machos haplóides se desenvolvem a partir de um ovo não fecundado enquanto as fêmeas diplóides se desenvolvem de ovos fecundados.

dsx foi publicado em abelhas) específica de fêmea nos heterozigotos. Por outro lado, o homozigoto ou hemizigoto transcreve somente um tipo de molécula CSD, direcionando a via para a produção de uma proteína DSX específica de macho, por *default* (Figura 1.6, A). Os embriões de fêmeas heterozigotas para o loco *csd* tratadas com csd dsmRNA formaram testículo (Figura 1.6, B).

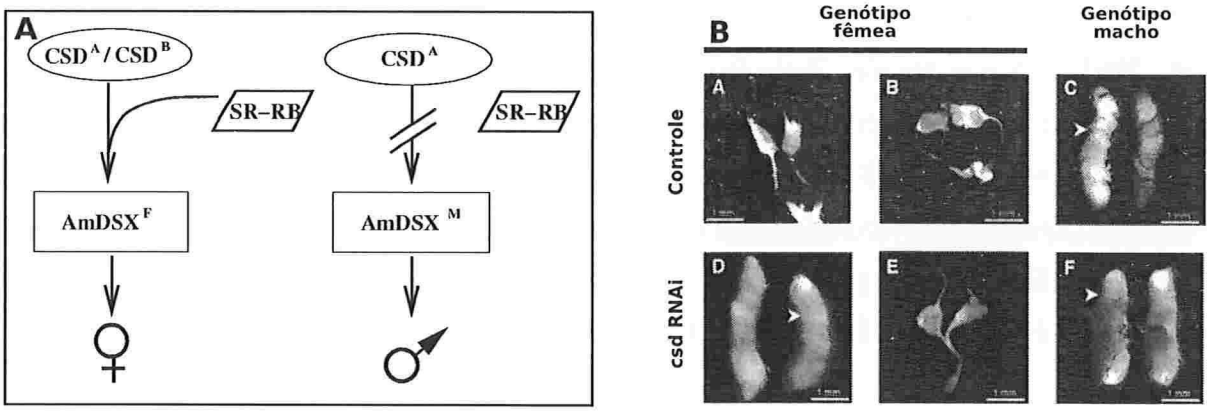


Figura 1.6: Determinação complementar do sexo em *A. mellifera*. (A) Modelo complementar de determinação do sexo em *A. mellifera* sugere que o gene *csd* (*complementary sex determiner*) precisa estar em heterozigose para codificar proteínas SR (ricas em serina e arginina) diferentes. Esta complementação promove o *splicing* de fêmea no mRNA do gene *dsx*. Os homozigotos (2n) ou hemizigotos (n) para o gene *csd* não necessitam de CSD para que o *splicing* de macho ocorra (*default*). (B) Experimento de RNAi com o gene *csd* demonstrou o que os produtos CSD de alelos diferentes são essenciais para o desenvolvimento dos ovários mas não para diferenciação dos testículos (BEYE et al., 2003).

A base da via genética de determinação do sexo apresenta características bem conservadas entre espécies distintas (Figura 1.4 e 1.6). O gene *doublesex* (*dsx*) está na base da cascata de determinação do sexo e parece ser o principal fator na regulação de vias específicas de cada sexo. Entretanto, o sinal primário responsável por regular a expressão sexo-específica do DSX apresenta grande variação entre as diferentes espécies, corroborando a idéia de uma evolução base-topo das vias metabólicas (WILKINS, 1995). A proteína DSX é um fator de transcrição e atua na região promotora dos genes que possuem o sítio de ligação específico. Em *Drosophila*, o fator DSX regula genes responsáveis pela formação de órgãos sexuais (KEISMAN; BAKER, 2001; KEISMAN; CHRISTIANSEN; BAKER, 2001), pelo comportamento de corte do macho (DAUWALDER et al., 2002), pela produção de proteínas do ovo (YP, do inglês yolk protein) (AN; WENSINK, 1995b) e pelo padrão de pigmentação em

cada sexo (CHRISTIANSEN et al., 2002).

A função mais conhecida do fator de transcrição DSX é na regulação da expressão das YP-1 em *Drosophila* (*yp-1*), mas, recentemente, também tem sido descrito como um fator importante na expressão de outros genes (ex: hexamerinas e vitelogeninas) em outros insetos, tal como no gene *hexamerina-1.2* (*hex-1.2*) do mosquito *Ochlerotatus atropalpus* (ZAKHARKIN et al., 2001; JINWAL et al., 2006) e nos genes *hex*, *Bmvg* (*vitelogenina*) de *B. mori* (SUZUKI et al., 2003). Outras funções deste fator de transcrição estão relacionadas com a formação da genitália de machos e fêmeas de *Drosophila*, que se desenvolvem a partir dos três últimos segmentos abdominais (8,9 e 10). Nestes segmentos, diferentes populações de células têm destinos específicos de acordo com a isoforma de DSX. O DSXM inibe o gene *wingless* (*wg*), enquanto o DSXF inibe o gene *decapentaplegic* (*dpp*), determinando o crescimento de diferentes regiões do disco genital (KEISMAN; BAKER, 2001; KEISMAN; CHRISTIANSEN; BAKER, 2001).

A expressão sexo e tecido-específica do gene *yp-1* é regulada, principalmente, por 3 fatores de transcrição (AEF-1, BZIP-1, DSX) que competem pela ligação a sítios sobrepostos localizados na região reguladora à montante de *yp-1*, denominada FBE (do inglês "Fat Body Enhancer"). O DSX se liga ao seu respectivo sítio (dsxa) e em cooperação com o BZIP-1 e outro fator ainda não conhecido (REF-1), ativa a expressão de YP-1. No corpo gorduroso (CG)[15], o DSX específico de fêmea (DSX-F) é mais abundante que AEF-1 ocorrendo a expressão de YP-1. Além disso, AEF-1 parece não ter nenhuma influência repressora neste órgão (Figura 1.7, A). Por outro lado, nos ovários (Ov), o DSX-F não é expresso e a abundância do repressor AEF-1 é muito maior, inibindo a expressão de YP-1 pela sua ligação ao sítio aef1 no FBE (Figura 1.7, B). Neste órgão reprodutor feminino, o DSX é irrelevante para a transcrição ovariana de *yp-1* (principalmente nas fases do desenvolvimento que precedem a oogênese), sendo BZIP-1 suficiente para expressar YP-1. Nos machos, o DSX-M possui uma região C-terminal maior que inibe a expressão de YP-1 pela oclusão do sítio de ligação de BZIP-1. (Figura 1.7, A) (AN; WENSINK, 1995b, 1995a).

[15]Este é principal órgão sintetizador que mantém, na hemolinfa, a homeostase de proteínas, lipídios e carboidratos desempenhando papel essencial no metabolismo, desenvolvimento e reprodução.

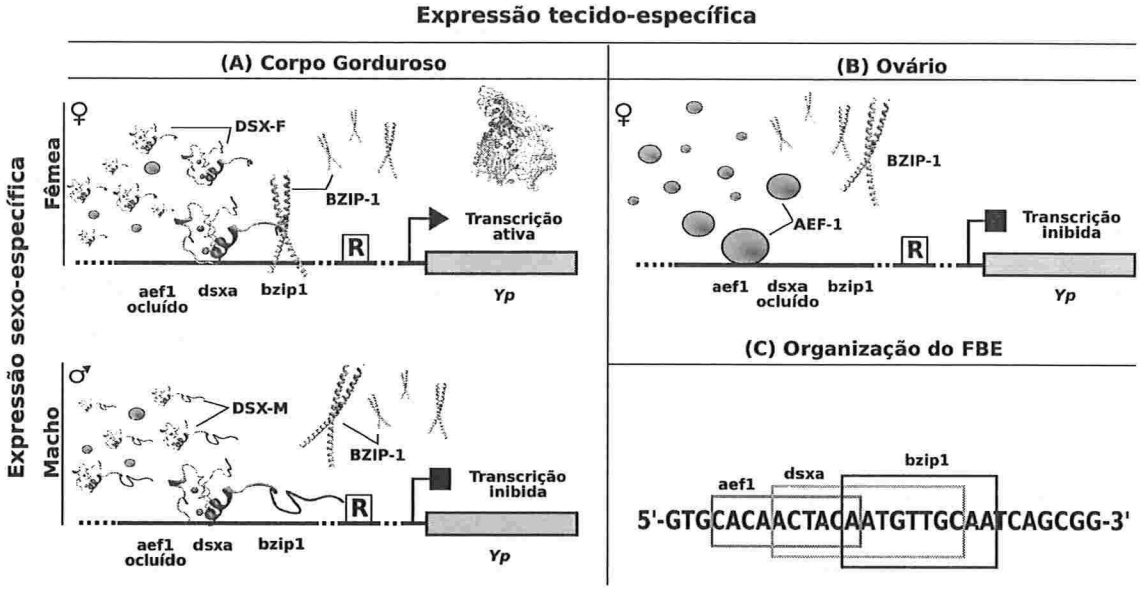


Figura 1.7: Modelo de expressão tecido e sexo-específica do gene *yp-1* de *D. melanogaster* proposto por An e Wensink (1995). (A) No corpo gorduroso, os fatores DSX-F e BZIP-1 formam um complexo sinérgico para, juntamente com um fator desconhecido R, ativar a transcrição de *yp-1* em fêmeas. Nos machos, o DSX-M bloqueia o acesso de BZIP-1 ao sítio bzip1 reprimindo a transcrição. (B) Em ovários, DSX-F está em concentrações muito baixas ou mesmo ausente enquanto o repressor AEF-1 está em maior abundância e compete pela ligação ao sítio aef1 que impede a ligação do fator principal BZIP-1, reprimindo a transcrição de *yp-1* neste tecido. (C) Os sítios de ligação dos três fatores (DSX, AEF-1 e BZIP-1) são sobrepostos resultando na competição ou cooperação entre estes fatores que se combinam para ativa ou inibir a transcrição de *yp-1* em diferentes sexos e tecidos.

1.3.5 Fatores ambientais e a plasticidade fenotípica

A capacidade de um indivíduo expressar fenótipos distintos sob certas circunstâncias ambientais é definida como *plasticidade fenotípica*. Dois tipos de plasticidade fenotípica podem ser observados: (1) polifenismo[16] e (2) norma de reação[17]. Muitos insetos apresentam polifenismos que podem ser expressos a partir de variações ambientais tais como sazonalidade, densidade populacional e diversidade nutricional (WEST-EBERHARD, 2003). Um dos exemplos de polifenismo mais intrigantes na biologia é a formação das castas em alguns himenópteros sociais, em particular, nas abelhas *A. mellifera*.

[16]padrões fenotípicos descontínuos determinados pela expressão gênica diferencial a partir de um mesmo genótipo em resposta ao ambiente

[17]as variações fenotípicas se expressam em um intervalo contínuo determinado por um genótipo em resposta a variações ambientais

Nestas abelhas altamente sociais, apenas uma rainha se desenvolve na colônia e se especializa em tarefas reprodutivas, enquanto um número muito maior de operárias (algumas dezenas de milhares) facultativamente estéreis[18] cuidam da manutenção da colônia, desempenhando uma grande diversidade de tarefas (nutrizes, faxineiras, guardas e forrageiras) (MICHENER, 1974; WILSON, 1975). As duas trajetórias do desenvolvimento de castas (rainhas ou operárias) das fêmeas de *A. mellifera* são determinadas nas fases pós-embrionárias (entre os períodos larvais L3 e L4) por diferenças na qualidade e quantidade de alimento larval. Larvas fêmeas que são alimentadas continuamente com uma mistura de secreções glandulares (geléia real) se tornam rainhas, enquanto larvas fêmeas destinadas a operárias se alimentam de uma mistura de geléia real, pólen e nectar a partir de uma determinada fase larval (L4) (HAYDAK, 1943).

As diferenças nutricionais experimentadas pelas fêmeas de abelha acionam uma resposta endócrina que se expressa como um elevado título de hormônio juvenil (HJ) em larvas de rainhas quando comparadas a larvas de operárias (HARTFELDER; ENGELS, 1998). As diferenças morfológicas, fisiológicas e comportamentais das castas são resultado de uma expressão diferencial de genes que responde de maneira direita ou indireta às variações nutricionais e hormonais. Muitos dos genes diferencialmente expressos encontrados estão relacionados a taxas metabólicas e respostas celulares a hormônios (EVANS; WHEELER, 1999, 2001).

A determinação de sexo e casta em *A. mellifera* está intimamente relacionada com a evolução da estrutura altamente social destes insetos. No entanto, outros genes desempenham funções importantes após a definição de sexo no início da embriogênese e de casta logo após a embriogênese. Nos insetos, genes que codificam proteínas de estocagem, tais como vitelogeninas (Vg) e hexamerinas (Hex), são importantes em uma série de processos biológicos que vão da formação da cutícula à reprodução. Em geral, o controle da expressão destes genes sofre forte influência de concentrações de hormônios, da nutrição, do fotoperíodo e da sexualidade (HARSHMAN; JAMES, 1998).

A Vg é uma glico-lipoproteína precursora da vitelina que compõe o vitelo do ovo e é sintetizada no corpo gorduroso, liberada na hemolinfa e finalmente internalizada nos oócitos, onde será utilizada como fonte de energia para o desenvolvimento do embrião (WYATT, 1999). Em *A. mellifera*, a Vg não é somente sintetizada pela rainha, mas também pelas operárias,

[18] na ausência da rainha algumas operárias são capazes de reativar o seu ovário e produzir ovos não fertilizados que se desenvolverão em zangões

funcionalmente estéreis, e zangões. Isto sugere que esta proteína participa de outras funções além da maturação dos oócitos e suprimento de energia para a embriogênese (PIULACHS et al., 2003; AMDAM et al., 2003; GUIDUGLI et al., 2005b).

O perfil de expressão do gene da vitelogenina e hexamerinas de *A. mellifera* (*Amvg* e *Amhex*, respectivamente) é diferente em rainhas, operárias e zangões, constituindo bons sistemas genéticos para o estudo do controle de expressão gênica nos sexo, casta e tecidos. O *Amvg* é expresso em ovários de rainhas mas não em ovários de operárias. Em operárias, o HJ tem função de inibir a expressão de *Amvg* (GUIDUGLI et al., 2005b). As operárias jovens expressam Vg enquanto desempenham tarefas dentro do ninho (ex: nutrizes e faxineiras) e o título de HJ é baixo. Quando passam a realizar tarefas de forrageira (coleta de pólen e nectar) os níveis de HJ sobem e a Vg decresce. O HJ controla negativamente a síntese de Vg em abelhas e isto tem implicações importantes na evolução e manutenção da socialidade destes insetos, já que as trajetórias de desenvolvimento das castas são definidas, primariamente, por alimentação diferencial e conseqüente variação no título de HJ (GUIDUGLI et al., 2005a) Experimentos de silenciamento gênico comprovaram esta regulação, ou seja, quando operárias tiveram o gene codificador de Vg (*Amvg*) silenciado, os níveis de HJ aumentaram (AMDAM et al., 2003).

As hexamerinas também desempenham papel fisiológico importante nas diferentes castas e sexos e respondem a concentrações de HJ e ecdisteróides (Ecd). Os genes de hexamerina em *A. mellifera* são expressos em diferentes períodos do desenvolvimento de um indivíduo bem como em sexos, castas e tecidos (BITONDI et al., 2006; CUNHA et al., 2005).

1.4 Fundamentos de bioinformática

Os avanços na biologia molecular e tecnologia da informação foram notáveis a partir da segunda metade do século 20. Após a identificação da estrutura do DNA (1953), iniciam-se inúmeros estudos sobre estas seqüências, que vão de questões teóricas em biologia e física à aplicações na medicina, agricultura e nanotecnologia. Entretanto, a primeira base de dados de seqüências genéticas foi estabelecida apenas em 1960 por Margaret Dayhoff e seus colaboradores da National Biomedical Research Foundation (NBRF) em Washington, D.C., que seqüenciaram famílias de proteínas mais comuns em várias espécies de organismos (ex: citocromos) e

organizaram estas proteínas em famílias de acordo com o grau de similaridade entre elas. Após os anos 70, as seqüências de DNA passaram a ser coletadas pela base de dados GenBank que logo estabeleceu uma colaboração internacional com outras 2 bases importantes, EMBL e DDBJ.

O surgimento das bases com milhares de seqüências genéticas de vários tipos (ex: DNA, RNA, proteína) e organismos demanda uma eficiente estratégia de organização e desenvolvimento de algoritmos computacionais capazes de analisar, comparar e fazer predições a respeito de características físicas e biológicas destas seqüências digitais. Especialistas em diferentes áreas do conhecimento juntaram esforços para estabelecer uma série de tipos de base de dados (funcional e estrutural) e de ferramentas computacionais de domínio público e distribuição livre que permitiram a padronização das informações biológicas a serem depositadas nas bases de dados mundiais bem como o uso e desenvolvimento aberto e livre dos algoritmos computacionais essenciais para o estabelecimento da bioinformática. Os conceitos e estratégias utilizados para o desenvolvimento da bioinformática neste estudo estão brevemente descritos nas próximas seções e têm como principal objetivo apresentar, principalmente aos biólogos, noções sobre os métodos e cálculos utilizados aqui.

1.4.1 Coletando e armazenando seqüências genéticas

Atualmente, a coleta de seqüências de DNA, RNA ou proteína é realizada de maneira automatizada e sem muitas dificuldades técnicas em um laboratório de biologia molecular. Restrições técnicas destes métodos moleculares estão, principalmente, no comprimento do fragmento seqüenciado, que não supera 700 bases. Desta maneira, o seqüenciamento completo de seqüências expressas ou genômicas deve ser feito em vários trechos separadamente e programas de computador desempenham funções essenciais que vão da análise dos arquivos gerados pelo seqüenciador até a reunião dos fragmentos que se sobrepõe a trechos maiores, denominados *contigs*.

As seqüências geradas em laboratórios de biologia devem ser devidamente formatadas em arquivos contendo a seqüência e informações biológicas e técnicas relacionadas a ela. A maneira mais adequada de organizar o crescente número de seqüências é armazená-las em bases de dados que permitam a recuperação e análise de informações de interesse. O esquema da base de dados e a tecnologia utilizada na sua implementação (relacional, orientada a objeto, objeto-relacional)

dependem da natureza e quantidade de dados. As bases de dados de seqüências de domínio público mais utilizadas entre os interessados neste assunto fornecem informações sobre o formato necessário para que a seqüência seja facilmente submetida e posteriormente compartilhada pela comunidade científica internacional.

O formato GenBank contém várias informações relacionadas à seqüência (Figura 1.8). As marcações em letras maiúsculas na primeira coluna a esquerda são definidas pela base GenBank e contêm informações do número de acesso na base (ACCESSION), a definição da seqüência (DEFINITION), os autores (AUTHORS), qualificadores (FEATURES), seqüência codificadora (CDS), etc (para maiores detalhes das marcações definidas ver http://www.ebi.ac.uk/embl/).

```
LOCUS       NM_001043640            680 bp    mRNA    linear   INV 18-NOV-2006
DEFINITION  Bombyx mori intersex (Ix), mRNA.
ACCESSION   NM_001043640 NM_001043641 NM_001043642
VERSION     NM_001043640.1  GI:112982868
KEYWORDS    .
SOURCE      Bombyx mori (domestic silkworm)
  ORGANISM  Bombyx mori
            Eukaryota; Metazoa; Arthropoda; Hexapoda; Insecta; Pterygota;
            Neoptera; Endopterygota; Lepidoptera; Glossata; Ditrysia;
            Bombycoidea; Bombycidae; Bombyx.
REFERENCE   1  (bases 1 to 680)
  AUTHORS   Siegal,M.L. and Baker,B.S.
  TITLE     Functional conservation and divergence of intersex, a gene required
            for female differentiation in Drosophila melanogaster
  JOURNAL   Dev. Genes Evol. 215 (1), 1-12 (2005)
   PUBMED   15645316
COMMENT     PROVISIONAL REFSEQ: This record has not yet been subject to final
            NCBI review. The reference sequence was derived from AY648340.1.
            On or before Sep 6, 2006 this sequence version replaced
            gi:112982866, gi:112982883.
FEATURES             Location/Qualifiers
     source          1..680
                     /organism="Bombyx mori"
                     /mol_type="mRNA"
                     /db_xref="taxon:7091"
     gene            1..680
                     /gene="Ix"
                     /note="intersex"
                     /db_xref="GeneID:692650"
     CDS             29..607
                     /gene="Ix"
                     /codon_start=1
                     /product="intersex"
                     /protein_id="NP_001037105.1"
                     /db_xref="GI:112982869"
                     /db_xref="GeneID:692650"
                     /translation="MNHNQRNNHVPMNQVAGAPNVAMQMPVPGPIMQQQSPQQMQPAP
                     VPQQTQQDKMDNISKVKSLMGSLRESIPMTLKSAAQILEQNENADSNTQKGMDNPVPR
                     FEKNLKEFFSICDQMELHLRTATTCIQQAQSAAHYLPLSVIASRLDSGPTTQETTLSY
                     PQYLKTVGLQISYAKDIHDTLVAAAQNISPPE"
ORIGIN
        1 aaatagtttt agtccaaaaa ttttaaatat gaatcacaat caaatgaata tgcacgtacc
       61 aatgaaccaa gttgotggag ctccaaatgt agccatgcaa atgcctgtac caggacccat
      121 aatgaaacag caatctcctc aacaaatgca acctgcaacca gtgcctcaac aaacacaaca
      181 agataaaatg gacaatatat ctaaagtgaa gtctttaatg ggatccttac gcgaatctat
      241 accgatgaca cttaagtctg cagcccaaat attacaccaa aatcacaatg cagactccaa
      301 tacacaaaaa ggtatggata atcctgttcc tagatttgaa aaaaacttag aagagttttt
      361 ctcaatctgt gatcaaatgg aactacattt gcgaacagca acaacgtgta tccagcaago
      421 gcaatcagcc gcccattatc tgcggctctc ggtgataacc tcaagactcg actctggacc
      481 taccacgcag gaaacgacat taagctatcc gcaatatttg aagacagttg gcctccaaat
      541 ctcatacgcg aaagatattc acgacacatt ggtcgcggct gctcagaata tatcaccgcc
      601 cgaatgaaaa tgaacottat ccacataagt ttaaagaata aagaataaat tattatttgt
      661 aatttataaa aaaaaaaaaa
//
```

Figura 1.8: Arquivo texto contendo uma seqüência genética no formato GenBank.

Outro formato muito usado é o FASTA e a grande maioria dos programas de análise de seqüências recebe este formato como entrada. A primeira linha é iniciada com o marcador "$>$" seguido por informações relevantes e, a partir da segunda linha, está a seqüência, como pode ser visto na Figura 1.9.

```
>gi|112982868|ref|NM_001043640.1| Bombyx mori intersex (Ix), mRNA
AAATAGTTTTAGTCCAAAAATTTTAAATATGAATCACAATCAAATGAATATGCACGTACCAATGAACCAA
GTTGCTGGAGCTCCAAATGTAGCCATGCAAATGCCTGTACCAGGACCCATAATGCAACAGCAATCTCCTC
AACAAATGCAACCTGCACCAGTGCCTCAACAAACACAACAAGATAAAATGGACAATATATCTAAAGTGAA
GTCTTTAATGGGATCCTTACGCGAATCTATACCGATGACACTTAAGTCTGCAGCCCAAATATTACACCAA
AATCACAATGCAGACTCCAATACACAAAAAGGTATGGATAATCCTGTTCCTAGATTTGAAAAAAACTTAG
AAGAGTTTTTCTCAATCTGTGATCAAATGGAACTACATTTGCGAACAGCAACAACGTGTATCCAGCAAGC
GCAATCAGCCGCCCATTATCTGCCGCTCTCGGTGATAGCCTCAAGACTCGACTCTGGACCTACCACGCAG
GAAACGACATTAAGCTATCCGCAATATTTGAAGACAGTTGGCCTCCAAATCTCATACGCGAAAGATATTC
ACGACACATTGGTCGCGGCTGCTCAGAATATATCACCGCCCGAATGAAAATGAACCTTATCCACATAAGT
TTAAAGAATAAAGAATAAATTATTATTTGTAATTTATAAAAAAAAAAAAA
```

Figura 1.9: Arquivo texto contendo uma seqüência genética em formato FASTA.

Uma série de outros formatos também é freqüentemente utilizadas (ex: EMBL, GCG, PIR, entre muitos outros) e facilmente convertidos de um para outro com programas já disponíveis ou linguagens de *script* (Perl, Python).

1.4.2 Alinhamento de seqüências

Um dos métodos mais importantes já desenvolvidos em bioinformática é o alinhamento de seqüências que é utilizado na comparação entre duas (alinhamento par-a-par) ou mais

(alinhamento múltiplo de seqüências) seqüências. Tal comparação pode revelar informações relevantes sobre a função, estrutura e evolução das seqüências. Os algoritmos de alinhamento são geralmente aplicados em comparações entre seqüências de DNA, RNA ou proteínas. As aplicações dos algoritmos de alinhamentos variam de acordo com o estudo comparativo de interesse e são usadas para estabelecer relações evolutivas entre genes e espécies, na busca por seqüências similares em bases de dados de seqüências já conhecidas e por padrões que representem domínios funcionais em proteínas ou sítios de ligação de fatores de transcrição em seqüência de DNA.

O alinhamento entre duas seqüências deve considerar os eventos evolutivos que possam gerar diferenças entre seqüências homólogas[19]. A seqüência de genes homólogos pode apresentar maior ou menor similaridade dependendo do tempo de divergência e pressão de seleção que, por sua vez, influenciam na ocorrência de eventos tais como substituições, inserções e/ou deleções de resíduos (nucleotídeos ou aminoácidos). O alinhamento pode ser representado em duas (alinhamento par-a-par) ou mais (alinhamento múltiplo) linhas, onde resíduos similares ou idênticos são colocados na mesma coluna, enquanto inserções e deleções acrescentam um ou mais espaços, otimizando o alinhamento para que o maior número possível de colunas contenha resíduos similares[20] ou idênticos.

Os principais métodos computacionais utilizados para alinhar seqüências genéticas são: (1) matriz de pontos, (2) programação dinâmica e (3) dicionário de palavras ou k-tupla (MOUNT, 2004). A análise de matriz de pontos é eficiente para revelar graficamente a presença de *indels* (inserção/deleção) ou repetições diretas ou invertidas que não são facilmente percebidas pelos outros métodos (GIBBS; MCINTYRE, 1970). Os algoritmos de programação dinâmica foram implementados em dois tipos de análise de seqüências, (1) alinhamento global (NEEDLEMAN; WUNSCH, 1970) e (2) local (SMITH; WATERMAN, 1981). O algoritmo de Needleman-Wunsch produz alinhamento global, isto é, seqüências inteiras são alinhadas do início ao fim e, geralmente, é usado quando as seqüências possuem alta similaridade e mesmo comprimento. O algoritmo de Smith-Waterman produz alinhamentos locais evidenciando somente regiões de alta similaridade, que podem ser bem menores que o comprimento das seqüências originalmente comparadas. Este último algoritmo é muito útil na identificação de

[19] seqüências que descendem de uma seqüência ancestral comum

[20] resíduos de propriedades químicas semelhantes

regiões e domínios conservados. O método de dicionário de palavras ou k-tupla alinha seqüências muito rapidamente, pela busca inicial de seqüências curtas e a posterior junção destas seqüências por alinhamento, usando programação dinâmica. O programa BLAST (ALTSCHUL et al., 1990) é dos mais usados para busca de seqüências similares em bases de dados com milhares de outras seqüências e utiliza o método de palavra, que otimiza as buscas em bases de dados muito grandes.

Estabelecer um alinhamento de seqüências coerente com alguns possíveis eventos evolutivos e a avaliação do grau de similaridade ou identidade entre estas seqüências alinhadas implica na quantificação de semelhanças e diferenças medidas por métodos baseados em pontuação (*score*). Para o cálculo destas pontuações são definidos alguns parâmetros (ex: *match*, *mismatch*, *gap*) que refletem os eventos de mutação importantes na evolução das seqüências genéticas. Um exemplo ilustrativo de uma matriz de alinhamento entre duas seqüências hipotéticas (ACCGTGA e ACGTGT) considerando que os valores dos parâmetros *match*, *mismatch* e *gap* sejam +3, -1 e -5, respectivamente, está representado na Figura 1.10. Diferentes valores destes parâmetros podem alterar os resultados de alinhamentos ótimos entre as seqüências e, portanto, é importante que se tenha uma compreensão mínima de como a matriz de pontuações é calculada.

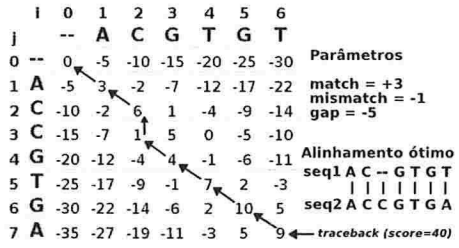


Figura 1.10: Esquema do alinhamento local de seqüências por algoritmo de programação dinâmica.

A formulação matemática do algoritmo de alinhamento de i resíduos da seqüência 1 contra j resíduos da seqüência 2 é dada pela seguinte equação:

$$S(i,j) = max \begin{cases} S(i-1,j-1) + c(i,j) \\ S(i-1,j) + c(i,-) \\ S(i,j-1) + c(-,j) \end{cases} \quad (1.1)$$

onde $c(i,j)$ é a pontuação para o alinhamento do resíduo i e j que, neste exemplo, assumem os valores $+3$ para *match* e -1 para *mismatch*, e $c(i,-)$ ou $c(-,j)$ são as penalidades de alinhamento da seqüência 1 ou 2, com um *gap*, respectivamente, e assumem o valor -5. A execução deste algoritmo gera os valores observados na matriz bidimensional da Figura 1.10 e o valor de cada célula é o valor máximo dado por um dos três eventos possíveis, $S(i-1,j-1)+c(i,j)$, $S(i-1,j)+c(i,-)$ ou $S(i,j-1)+c(-,j)$. O melhor alinhamento entre as seqüências é então determinado seguindo-se a trajetória de

pontuações da matriz de volta ao início (*traceback*) que gerou a mais alta pontuação possível. A diferença entre os alinhamentos local e global está no tipo de pontuação que é mantida na matriz de pontuação. No alinhamento global (Needleman-Wunsch) todos as pontuações são mantidas, não importando se os valores são negativos ou positivos. No alinhamento local (Smith-Waterman), somente os valores positivos são mantidos, enquanto os valores negativos se tornam zero. Tal modificação permite que o algoritmo alinhe regiões com as maiores pontuações, independente do comprimento total das seqüências alinhadas.

O alinhamento de seqüências de proteína utiliza o mesmo algoritmo de programação dinâmica, entretanto, os valores de substituição (*match* e *mismatch*) entre aminoácidos são determinados por matrizes de substituição de aminoácidos que representam as freqüências de substituição de um aminoácido por outro. Os dois tipos mais usados de matrizes de substituição são PAM (*Percent Accepted Mutation*) e BLOSUM (*Blocks Substitution Matrix*). As matrizes PAM (DAYHOFF, 1979) são construídas a partir do alinhamento global de seqüências de proteínas de espécies evolutivamente próximas. A PAM1 é uma matriz calculada a partir de comparações de seqüências com não mais de 1% de divergência. Outras matrizes PAM(x) (ex: PAM100, PAM250) são calculadas a partir da PAM1 com uma quantidade x de divergência evolutiva. A BLOSUM (HENIKOFF; HENIKOFF, 1992) é derivada de observações das freqüências de substituições em blocos de alinhamentos locais em proteínas de espécies aparentadas. As diversas BLOSUM(x) são calculadas a partir de comparações entre seqüências de proteínas que não sejam menos que x% divergentes (ex: BLOSUM62 é construída a partir de proteínas com 62% de identidade). As matrizes BLOSUM com maiores valores e matrizes PAM com menores valores são, em geral, usadas em comparações de seqüências menos divergentes (BLOSUM80 e PAM120 são equivalentes), enquanto BLOSUM com menores valores e PAM com maiores valores são freqüentemente usadas para comparar seqüências mais divergentes (BLOSUM45 e PAM250 são equivalentes) (http://www.ebi.ac.uk/help/matrix_frame.html).

O alinhamento de seqüências pode ser aplicado a uma variedade de estudos em biologia. Algumas das aplicações mais comuns são (1) em buscas por similaridades entre seqüências desconhecidas e outras já descritas em bases de dados para inferências funcionais e/ou estruturais, (2) na identificação de regiões conservadas ou não, para desenhar oligonucleotídeos (iniciadores ou *primers*[21]) que sejam muito ou pouco específicos dependendo do caso em estudo

[21] *primers* são seqüências curtas de nucleotídeos que são utilizados como moléculas iniciadoras de uma PCR

(ex: filogenia, expressão gênica, entre muitos outros), (3) em análises de substituições sinônimas e não-sinônimas que devem quantificar a substituição de nucleotídeos em cada um dos 3 sítios do códon. Muitas outras aplicações poderiam ser listadas. No entanto, alguns erros que resultam do alinhamento de seqüências de DNA podem gerar artefatos, principalmente, nas análises (2) e (3). Uma estratégia eficiente na solução destes erros é a construção do alinhamento de DNA a partir da tradução reversa de um alinhamento de proteínas, garantindo que os *gaps* nas seqüências de DNA estejam alinhados com as fronteiras dos códons (Figura 1.11).

1.4.3 Predições filogenéticas

Por uma série de razões, pode ser interessante saber as relações evolutivas de um grupo de seqüências. A partir desta informação, pode-se propor relações evolutivas de grupos de espécies ou mesmo a história evolutiva de famílias gênicas. Tradicionalmente, os relacionamentos evolutivos entre organismos ou membros de famílias gênicas são representados por árvores evolutivas. As árvores podem ter raiz (*rooted*) ou não (*unrooted*) e isto altera não só a topologia, mas a quantidade de árvores possíveis, sendo muito maior quando uma raiz é estabelecida. Os nós representam unidades taxonômicas (que podem ser espécies, populações, indivíduos ou genes). As ramificações (ou bifurcações) das árvores representam os eventos de especiação ou duplicação gênica, enquanto o comprimento destas ramificações representa o número de substituições que ocorreram nos ramos (LI; GRAUR, 1991).

(A) Alinhamento comum

```
ATG CT- --G ATA GGG
ATG CTC AAG ATA GGG
ATG CTC AA- --A GGG
```

(B) Alinhamento reverso

```
ATG CTG --- ATA GGG
ATG CTC AAG ATA GGG
ATG CTC AAA --- GGG

 M   L   K   I   G
```

Figura 1.11: O alinhamento comum de seqüências de nucleotídeos pode conter erros. (A) Alinhamento comum mostrando inserções incorretas de *gaps*. (B) Alinhamento reverso utiliza a informação do alinhamento de aminoácidos como ponto de partida para correto alinhamento de nucleotídeos.

Os métodos mais utilizados na reconstrução de árvores filogenéticas são: (1) máxima parcimônia, (2) distância, (3) máxima verossimilhança. Embora estes métodos tenham mecanismos diferentes para calcular os relacionamentos entre as unidades taxonômicas (ou OTU[22]), eles

(*Polymerase Chain Reaction*) que juntamente com a enzima Taq DNA polimerase e uma solução tampão com os 4 tipos de nucleotídeos amplifica fragmentos genéticos que contenham regiões similares a estes iniciadores.

[22] abreviado do inglês *operational taxonomic unit*

objetivam identificar a topologia mais congruente a partir dos dados observados (LI; GRAUR, 1991; MOUNT, 2004; GIBSON; MUSE, 2004). Todos os imétodos calculam suas medidas a partir de um alinhamento múltiplo de seqüências (msa) e, portanto, este alinhamento deve conter o mínimo de erros possíveis.

Máxima parcimônia. Este método faz predições das melhores topologias de árvores, minimizando o número de passos necessários para gerar as variações observadas em cada sítio, nas seqüências a partir de uma seqüência ancestral comum. O método é mais adequado para um número pequeno de seqüências com muita similaridade. Além disso é importante que as taxas de substituição sejam homogêneas em todas as ramificações. Árvores com ou sem raiz podem ser geradas por este método.

Distância genética. Este método se baseia em distância genética medida entre pares de seqüências, assumindo-se modelos probabilísticos de evolução de seqüências. Os modelos de substituição de nucleotídeos (ex: modelo de Jukes-Cantor) e aminoácidos (ex: matrizes PAM, JTT[23], entre outras) são usados no cálculo de distância evolutiva para todos os pares de seqüências e a reconstrução filogenética pode, então, ser feita por alguns métodos diferentes (ex: UPGMA[24] e NJ[25]) (LI; GRAUR, 1991).

Máxima verossimilhança. O mecanismo de reconstrução filogenética deste método é semelhante ao de parcimônia, onde a análise é feita em cada sítio (cada coluna do msa). Entretanto, as taxas de substituições entre resíduos não são equivalentes, mas seguem modelos probabilísticos (assim como nos métodos de distância) que descrevem as modificações em seqüências no tempo. Os métodos de verossimilhança tendem a ser muito eficientes na reconstrução filogenética mas, requisitam muito processamento computacional e apresentam limitações no número de seqüências analisadas (não mais do que 50 seqüências) (MOUNT, 2004; GIBSON; MUSE, 2004).

[23]modelo Jones-Taylor-Thornton.
[24]abreviado do inglês *unweighted pair group method with arithmetic mean.*
[25]abreviado do inglês *neighbor-joining* (SAITOU; NEI, 1987).

1.4.4 Predições de padrões conservados em seqüências

Existe uma série de ferramentas computacionais que modelam estatisticamente as características dos genes, domínios protéicos e motivos de ligação de fatores de transcrição. A construção de bons modelos destas entidades é útil na localização de seqüências homólogas ainda desconhecidas, com alta probabilidade de atenderem ao perfil do modelo. Alguns métodos, tais como modelo de Markov oculto (HMM[26]), redes neurais (NN) e matrix de pontuação de posição específica (PSSM[27]) são amplamente usados e vários programas de bioinformática já possuem estes métodos (ou derivações deles) implementados.

Predição de genes e domínios de proteínas

A predição de genes ou domínios de proteínas é, em geral, realizada por modelos HMM ou NN que são devidamente treinados a partir de padrões já conhecidos de seqüências de genes ou proteínas. Um considerável número de programas de predição de genes ou domínios de proteínas está disponível na *internet* (ex: GenScan (BURGE; KARLIN, 1997), GlimmerM (MAJOROS et al., 2003), HMMER (EDDY, 1998) entre muitos outros). O desenvolvimento de métodos eficientes de predições de genes e padrões conservados em proteínas é uma importante área da bioinformática; No entanto, nenhum detalhe dos métodos é descrito neste manuscrito. Descrições matemáticas e computacionais detalhadas sobre o funcionamento destes modelos podem ser encontradas em vários artigos científicos (KROGH et al., 1994; EDDY, 1998; SONNHAMMER et al., 1998; HUGHEY; KROGH, 1996; WINTERS-HILT, 2006) e livros (DURBIN et al., 1998; MOUNT, 2004; GIBSON; MUSE, 2004).

Predição de motivos em seqüências reguladoras

Muitas outras regiões funcionais importantes podem ser descobertas por busca de padrões em seqüências além dos próprios genes e motivos (ou domínios) conservados de proteínas. Alguns exemplos destes motivos são de grande interesse e incluem ilhotas CpG[28], sítios de *splicing* que

[26] abreviado do inglês *Hidden Markov Model*.

[27] abreviado do inglês *position-specific scoring matrix*.

[28] regiões com um grande número de citosinas e guaninas adjacentes umas às outras, ligadas por pontes de fosfodiéster. A metilação dos sítios CpGs nas regiões promotoras de um gene pode inibir a expressão do gene.

delimitam as fronteiras íntron/éxon e sítios de ligação de proteínas reguladoras (ou fatores de transcrição). A expressão diferencial de genes em resposta a variações ambientais e sinalizações para o desenvolvimento de fenótipos depende da ação destes fatores de transcrição (TF) e a identificação dos motivos regulatórios, aos quais estes TFs se ligam (sítios de ligação do DNA), pode fornecer informações cruciais na compreensão dos mecanismos de regulação transcricional (MACISAAC; FRAENKEL, 2006).

A predição computacional de sítios de ligação de TFs em genomas de organismos eucariontes é, muitas vezes, uma tarefa desafiadora, mas estratégias computacionais bem elaboradas tem possibilitado grandes avanços na descoberta de motivos regulatórios em seqüências de DNA (STORMO, 2000; BULYK, 2003; WASSERMAN; SANDELIN, 2004; TOMPA et al., 2005; MACISAAC; FRAENKEL, 2006). Os motivos podem ser representados por vários tipos de modelos. As duas maneiras mais utilizadas na representação destes motivos são a (1) seqüência consenso e (2) matriz de pontuação específica de posição (PSSM, também conhecida como PWM[29]). A seqüência consenso é uma maneira simplificada de representar ambigüidade no motivo e é de fácil visualização, mas não contém as informações quantitativas que caracterizam as variações em cada sítio de um motivo (Figura 1.12).

A PWM é um modelo mais adequado para representar os motivos ou sítios de ligação de TF. A partir de um alinhamento de todos os sítios de ligação de TF conhecidos, o número total de observações de cada nucleotídeo é registrado para cada posição, produzindo uma matriz de freqüência por posição (PFM, Figura 1.12, A-B). Uma PWM é, então, calculada a partir de uma transformação logarítmica da PFM normalizada, assumindo-se que o nucleotídeo em uma dada posição apresenta uma distribuição independente dos nucleotídeos em outras posições (Figura 1.12, C-D). Uma pontuação (*score*) para o alinhamento entre qualquer seqüência de DNA e a matriz pode ser, então, calculada diretamente pela somatória dos valores que correspondem a cada nucleotídeo observado em cada posição (STORMO, 2000; WASSERMAN; SANDELIN, 2004). Os motivos representados desta maneira podem ser visualizados por seqüências de logos que consistem em uma pilha de letras (neste caso os 4 nucleotídeos, A, T, C e G) ordenadas, onde a altura das letras indica o conteúdo de informação em cada posição (SCHNEIDER; STEPHENS, 1990; MACISAAC; FRAENKEL, 2006) (Figura 1.12, E).

[29]abreviado do inglês *Position Weight Matrix*.

O desenvolvimento e as aplicações de algoritmos computacionais para análise e predição de sítios de ligação do DNA podem ser divididos em duas abordagens: (1) dada uma coleção de sítios de ligação conhecidos, o motivo pode ser representado a partir dos sítios e usado na busca por sítios de ligação adicionais e novas seqüências, e (2) dada uma coleção de um conjunto de seqüências que contenham sítios de ligação, ainda desconhecidos, para um ou mais fatores de transcrição, é possível descobrir a localização destes sítios em cada seqüência (STORMO, 2000).

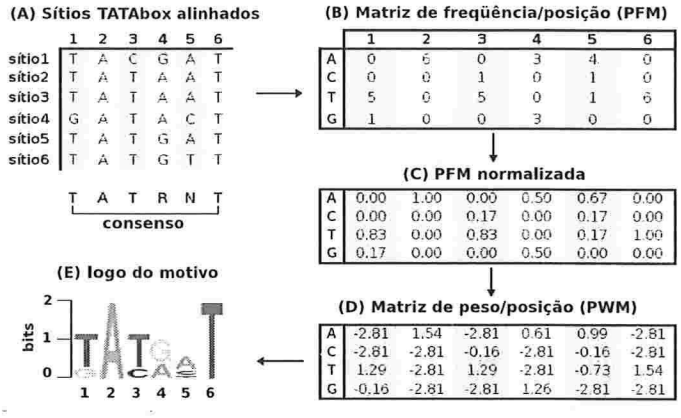


Figura 1.12: Representação de um motivo regulatório em uma matriz de contagem de nucleotídeos por posição. (A) Alinhamento de sítios de ligação de RNApolII. (B) Matriz de freqüência de nucleotídeos por posição (PFM) calculada a partir do alinhamento dos sítios TATA. (C) Normalização da matriz freqüência/posição para calcular as probabilidades de ocorrência de cada nucleotídeo em cada posição. (D) Matriz de peso/posição (PWM) calculada a partir da transformação logarítmica da PFM. (E) Logo do motivo plotado a partir do conteúdo de informação em cada posição.

TRANSFAC (WINGENDER et al., 2000) e JASPAR (VLIEGHE et al., 2006) são bases de dados de sítios de ligação de fatores de transcrição definidos experimentalmente em vários organismos que podem ser usadas na abordagem (1) mencionada acima, revelando quais genes possuem sítios similares aos sítios de ligação de TF já conhecidos. Entretanto, uma grande proporção destes sítios preditos pode ser ou falso-positivo, ou não funcional (sítio espúrio). Para contornar estes problemas técnicos, pode-se observar algumas características complementares, tais como regiões conservadas entre genes homólogos pelo método de *footprinting* filogenético (BLANCHETTE; TOMPA, 2002; FRAZER et al., 2003; WANG; STORMO, 2003) e o agrupamento de sítios dentro de módulos cis-regulatórios (CRM) (DAVIDSON, 2001). O método de *footprinting* filogenético é muito eficiente na identificação de regiões conservadas entre regiões promotoras de genes homólogos que provavelmente compartilham sítios regulatórios conservados (WASSERMAN; SANDELIN, 2004; MACISAAC; FRAENKEL, 2006). No entanto, é importante que genes ortólogos[30] e

[30] genes homólogos de diferentes espécies que se originaram de um gene ancestral comum.

parálogos [31] sejam corretamente identificados e isto nem sempre é possível.

O problema de descoberta de motivos regulatórios é geralmente formulado a partir da hipótese de que genes co-regulados são provavelmente regulados por um ou alguns TFs e que, portanto, devem compartilhar sítios regulatórios conservados aos quais estes TFs se ligam (MACISAAC; FRAENKEL, 2006). As técnicas mais freqüentemente utilizadas para esta busca envolvem o desenvolvimento de modelos probabilísticos a partir dos dados de seqüência observados e otimização para encontrar motivos comuns em todas as seqüências analisadas. Os dois métodos mais usados para estas buscas são: (1) *expectation-maximization* (EM) (LAWRENCE; REILLY, 1990) e (2) *Gibbs sampling* (LAWRENCE et al., 1993). A descrição dos detalhes destes métodos exige conceitos estatísticos que estão além do escopo deste texto.

O sucesso nas buscas por motivos regulatórios desconhecidos é bem maior se o conjunto de seqüências analisadas for de alta confiabilidade, isto é, se a maioria dos genes considerados no conjunto de entrada forem realmente co-regulados. Assim, muitas seqüências ou seqüências de comprimento muito grandes aumentam o ruído na análise e sinais (motivos) com pouco conteúdo de informação, muitas vezes, não são localizados pelos algoritmos. Os genes co-regulados podem ser identificados a partir do agrupamento de genes que compartilham categorias funcionais ou são co-expressos sob determinadas condições ambientais ou estágios do desenvolvimento (ROTH et al., 1998; HUGHES et al., 2000) bem como experimentos de imunoprecipitação de cromatina (ChIP) (LIU; BRUTLAG; LIU, 2002; MUKHERJEE et al., 2004; HARBISON et al., 2004). A partir de um conjunto de seqüências que potencialmente compartilham motivos regulatórios conservados, é possível identificar estes elementos regulatórios por vários programas já devidamente validados (ex: AlignAce, MDScan, MEME) e o uso comparativo de mais de um destes programas é altamente recomendado (HARBISON et al., 2004; TOMPA et al., 2005).

Uma vez que os motivos são identificados, é necessário avaliar se estes padrões não são simplesmente sítios espúrios que ocorram ao acaso no genoma. Esta avaliação dos motivos descobertos deve ser feita por uma métrica de pontuação que permita que os motivos sejam comparados e ordenados independente de sua descoberta. Em geral, métricas, tais como (1) *group-specificity score* (Church score) (HUGHES et al., 2000), (2) *hypergeometric enrichment* (BARASH; BEJERANO; FRIEDMAN, 2001) e (3) *area under the curve for a receiver operator*

[31] genes homólogos que se originaram a partir de eventos de duplicação gênica e ocupam dois locos em um genoma.

characteristic plot (ROC-AUC) (CLARKE; GRANEK, 2003) avaliam o viés na ocorrência de um motivo no conjunto de entrada em relação ou conjunto total (*background*). Deste modo, com a descoberta de motivos regulatórios, é possível dar início a uma série de outras buscas no genoma. A interpretação do papel biológico destes elementos possibilita comparações entre redes regulatórias associadas aos mecanismos de controle de expressão gênica e aos padrões fenotípicos.

1.4.5 Genômica funcional e comparativa

Uma das principais aplicações da bioinformática é na análise de genomas recém-seqüenciados. Em geral, estas análises envolvem comparações e predições realizadas a partir de uma grande quantidade de métodos (alguns deles descritos nas seções anteriores) e critérios que devem ser aplicados de forma coerente com as propriedades biológicas dos organismos usados nestas comparações. O ponto de partida na análise de um novo genoma (já devidamente seqüenciado e montado) é a identificação e mapeamento dos genes. Só então, é possível dar início a estudos que envolvam a identificação de famílias gênicas, organização espacial dos genes no genoma (ex: sintenia[32]), estrutura éxon/íntron, regiões reguladoras, entre muitas outras.

A diversidade na evolução de genes e famílias gênicas é o resultado de uma complexa relação entre os mecanismos genéticos, processos evolutivos e pressão de seleção. Alguns genes ou famílias gênicas podem estar muito conservados em número, estrutura e função mesmo entre espécies muito distantes na evolução, mas, também, podem ter evoluído especificamente em um ou alguns poucos organismos. Talvez, ainda mais intrigante e informativo sejam os padrões evolutivos que moldam as regiões intergênicas nos genomas. Tais regiões são particularmente importantes no controle inicial que determina o fluxo de informação (DNA $\rightarrow$ RNA $\rightarrow$ proteína) necessário para construir o cenário molecular adequado que caracterize cada estágio do desenvolvimento e/ou tecidos e órgãos. A formação dos padrões fenotípicos é determinada por redes genéticas que integram todas estas entidades biológicas (genes e seus produtos, e elementos regulatórios) às condições ambientais. No entanto, estes fenômenos genéticos de formação de padrões ainda são muito pouco compreendidos.

A genômica comparativa é uma comparação entre dois ou mais genomas, onde se utilizam

[32] é a colinearidade na ordem dos genes em duas espécies.

várias características, tais como radiações de famílias gênicas, localizações relativa de genes, disposição da ordem de genes, arquitetura do genes, polimorfismos genéticos, seqüências repetitivas, entre outras características. O estudo da evolução de novos genes e a diversificação funcional em famílias multigênicas acrescentam informações importantes para a compreensão da evolução da diversidade de organismos (OAKESHOTT et al., 1999; LONG et al., 2003). Algumas questões importantes sobre como a resistência a inseticidas em mosquitos (*A. gambiea*) evolui estão intimamente relacionadas a expansões em membros das três grandes famílias de enzimas, citocromo P450s, glutationa transferases (GST) e carboxilesterases (COE), que ocorreram como resposta a mudanças ambientais (RANSON et al., 2002). Por outro lado, em abelhas *A. mellifera*, estes genes de resistência a inseticidas estão numericamente reduzidos, o que é coerente com a alta sensibilidade destes insetos a inseticidas, enquanto alguns clados das P450s e dos genes da família dos receptores de olfato sofreram radiações que possivelmente estão correlacionadas com a evolução de processos hormonais e quimiosenssoriais envolvidos na organização da eusocialidade destes insetos (CLAUDIANOS et al., 2006; The Honeybee Genome Sequencing Consortium, 2006). Contudo, só é possível conduzir este tipo de estudo com a correta identificação dos genes ortólogos ou parálogos entre os genomas comparados e, para isto, uma criteriosa anotação do genoma seguida de análises filogenéticas (ver seção 1.4.3) são obrigatórias.

Anotação de genomas

A crescente quantidade de genomas já seqüenciados e devidamente anotados acelera consideravelmente o processo de anotação[33] de novos genomas. Os genes são inicialmente preditos da seqüência de um novo genoma e podem conter artefatos. Estas predições são, em boa parte, validadas experimentalmente por análises que confirmem a presença de transcritos (ex: seqüenciamento de cDNAs[34], ESTs[35] e *tilling array*). Entretanto, algumas questões, tais como a presença de pseudogenes e *splicing* alternativo, podem não ser resolvidas de maneira satisfatória, exigindo ensaios experimentais mais específicos e detalhados sobre o gene.

A anotação do genoma se inicia, de fato, quando as seqüências de aminoácidos derivadas

[33] é o processo de marcação de seqüências genômicas com informações funcionais (MOUNT, 2004).
[34] abreviado do inglês *complementary DNA*, é uma molécula sintetizada a partir de um mRNA
[35] abreviado do inglês *Expressed Sequence Tags*.

dos genes preditos são alinhadas contra bases de dados de proteínas de várias espécies para busca de seqüências similares. Estas comparações podem fornecer informações sobre a estrutura e função da seqüência do gene que está sendo anotado. Porém, nem todos os genes preditos podem ser comparados a proteínas de outras espécies, por falta de similaridade. A busca por domínios protéicos (padrões conservados de seqüências de aminoácidos) conservados pode ajudar na identificação de assinaturas de famílias de proteínas ou de características bioquímicas e/ou estruturais. Pfam (BATEMAN et al., 2004) é uma das bases de dados mais utilizadas para estas buscas de domínios e famílias de proteínas e contém uma grande coleção de alinhamentos múltiplos de seqüências e modelos HMM.

Em geral, os estudos de função de um gene são desenvolvidos por vários grupos de pesquisa que, muitas vezes, objetivam responder questões biológicas muito diversificadas e distintas envolvendo a participação do respectivo gene. Desta maneira, estabelecer um vocabulário padronizado para a descrição funcional dos genes, independente do tipo de experimento ou espécie, possibilita a construção de bases de dados com uma uniformidade semântica importante para os estudos comparativos e sistemáticos. O consórcio Gene Ontology (GO, http://www.geneontology.org/) (ASHBURNER et al., 2000) representa uma iniciativa de estabelecer esta uniformidade de termos e descrições biológicas que são modeladas em uma rede de estrutura hierárquica representada por um grafo (Figura 1.13). O GO classifica a função de proteínas em três categorias gerais ou ontológicas (Função Molecular, Processo Biológico e Componente Celular) que se organizam em definições mais específicas, por exemplo, processo biológico (nível 2) → processo de desenvolvimento (nível 3) → desenvolvimento multicelular do organismo (nível 4) → determinação somática do sexo (nível 5). Entretanto, a estrutura hierárquica do GO não corresponde a uma árvore, mas a uma rede onde um termo específico pode estar relacionado com outros vários termos de outros processos biológicos, funções moleculares ou componentes celulares (Figura 1.13).

Esta transferência direta do conhecimento da função de um gene de uma espécie para outra facilita a anotação do genoma e a formulação de hipótese de estudo e, muitas vezes, representa o único ponto de partida para anotações funcionais em larga escala. A evolução de novas funções em famílias multigênicas ilustra bem as dificuldades de atribuição de funções por similaridade de seqüências como no caso de genes da família das carboxil/colinesterases (CCE), que sofrem expansões e reduções ocupando um amplo espectro de nichos funcionais nos metazoários.

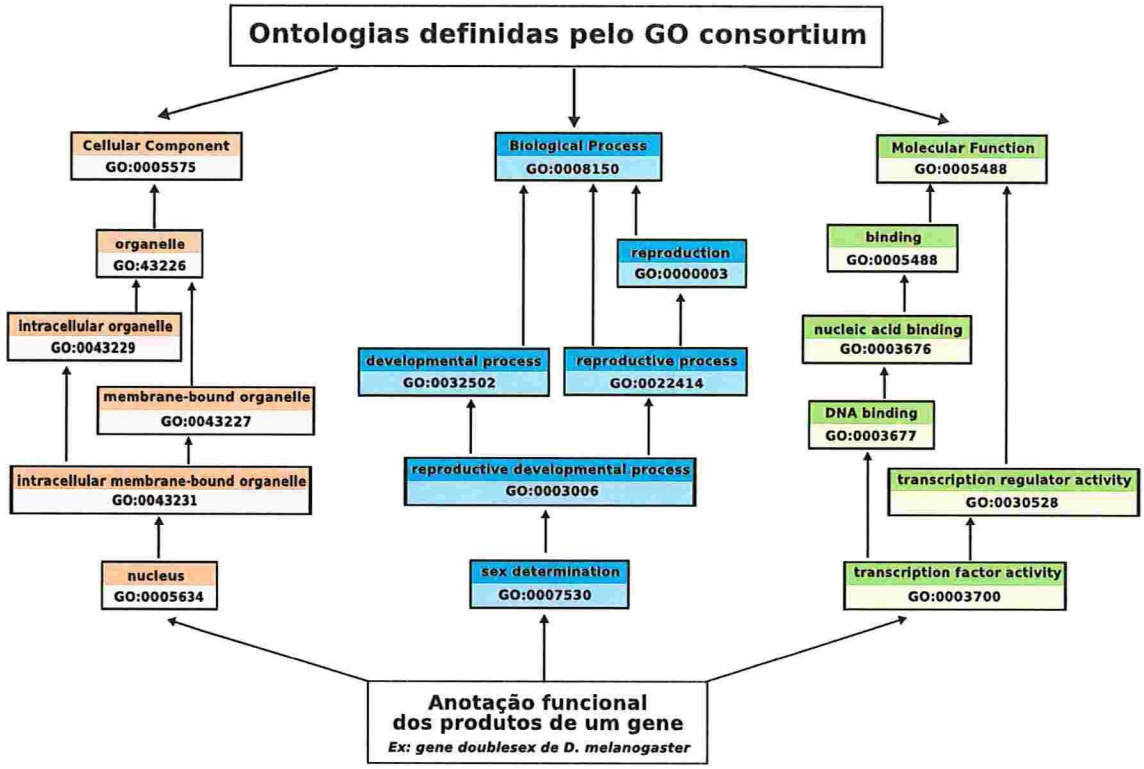


Figura 1.13: Rede de anotação funcional de acordo com os termos definidos pelo GO. O exemplo descreve algumas anotações funcionais do gene *dsx* de *D. melanogaster*. A anotação possui uma estrutura conceitual hierárquica que vai da definição mais geral (nível 2) à mais especializada (o exemplo vai até o nível 5, mas pode chegar a descrições mais específicas).

Algumas CCEs muito similares em seqüência e estrutura podem assumir funções completamente diferentes pela substituição de alguns poucos aminoácidos que, por exemplo, resultam na perda de função catalítica e adaptação em vias de transdução de sinal (OAKESHOTT et al., 1999).

Análise de expressão gênica

Análises de expressão gênica podem revelar quais genes são expressos em determinados estágios do desenvolvimento, em tecidos e órgãos específicos, em resposta a sinais ambientais, em diferentes tipos de doenças, etc. Há uma considerável diversidade de métodos experimentais para este fim (ex: construção de bibliotecas subtrativas, *differential display*, *microarray*, SAGE[36] e RT-PCR[37] em tempo real). Os experimentos de expressão gênica em larga escala (ex:

[36] abreviado do inglês *Serial Analysis of Gene Expression*.
[37] abreviado do inglês *Reverse Transcriptase PCR*.

microarray) são geralmente utilizados no levantamento global dos genes que possam estar relacionados com processos biológicos ou vias bioquímicas específicas. Entre estes genes co-expressos podem estar muitos grupos de genes que estão sob o mesmo controle transcricional e, portanto, são co-regulados. Embora, a grande maioria dos experimentos de microarray atualmente identifique os genes que estão diferencialmente expressos nas situações comparadas, eles não garantem que todos aqueles genes co-expressos são necessariamente co-regulados e, então, uma busca por padrões conservados nas regiões reguladoras destes genes exige uma abordagem mais complexa do que simplesmente o alinhamento de motivos já conhecidos contra os promotores destes genes (muitos motivos espúrios são identificados desta maneira).

Após este levantamento global dos genes expressos em determinados fenótipos, alguns deles devem ser eleitos para validação experimental mais precisa e criteriosa. A criação de organismos mutantes ou experimentos de RNAi (RNA de interferência) combinados com análises de RT-PCR em tempo real (PFAFFL, 2001; BUSTIN, 2002) podem ser conduzidos em laboratórios de biologia molecular para uma validação da função daqueles genes evidenciados inicialmente pelas análise do perfil de expressão gênica. Desta maneira, as anotações funcionais dos genes por similaridade de seqüência podem ser parcialmente validadas por estes experimentos. Entretanto, uma validação completa da função gênica envolve testes bioquímicos e funcionais das proteínas codificadas por estes genes, sendo, muitas vezes, complexa, necessitando de especialistas e laboratórios equipados para estas metodologias em estudos de proteínas.

1.4.6 Redes complexas em biologia

A crescente quantidade de informações disponíveis sobre genomas, transcriptomas, proteomas e metabolomas de vários organismos representa o alicerce para a reconstrução do conhecimento biológico de maneira menos reducionista e conseqüentemente mais complexa. Embora a biologia molecular tenha contribuído muito (e certamente continuará contribuindo) para a criação de um detalhado inventário sobre genes, proteínas e metabólitos em vários organismos, o mesmo não é suficiente para entender a complexidade do funcionamento de uma célula ou de um organismo multicelular (OLTVAI; BARABASI, 2002). Os sistemas biológicos são sistemas complexos constituídos primariamente de interações moleculares que ocorrem com algum grau interno de ordem (BARABASI; OLTVAI, 2004).

O estudo de redes biológicas utiliza conceitos da teoria dos grafos (CHARTRAND; LESNIAK, 1996) e redes complexas (ALBERT; BARABÁSI, 2002) que permite uma poderosa visualização gráfica das inter-relações das entidades biológicas (ex: genes, proteínas, RNA reguladores, elementos regulatórios) e possibilita uma quantificação objetiva dos sistemas biológicos. Tal abordagem permite uma melhor compreensão das características e propriedades relevantes para a formação dos padrões biológicos. As redes são representadas com grafos contituídos por nós (ou vértices) e ligações (ou arestas) entre estes nós (Figura 1.14). A rede pode ser direcionada ou não e pode representar ligações entre proteínas, células ou organismos (Figura 1.14A) ou mesmo reações químicas ou padrões regulatórios dos genes (Figura 1.14B).

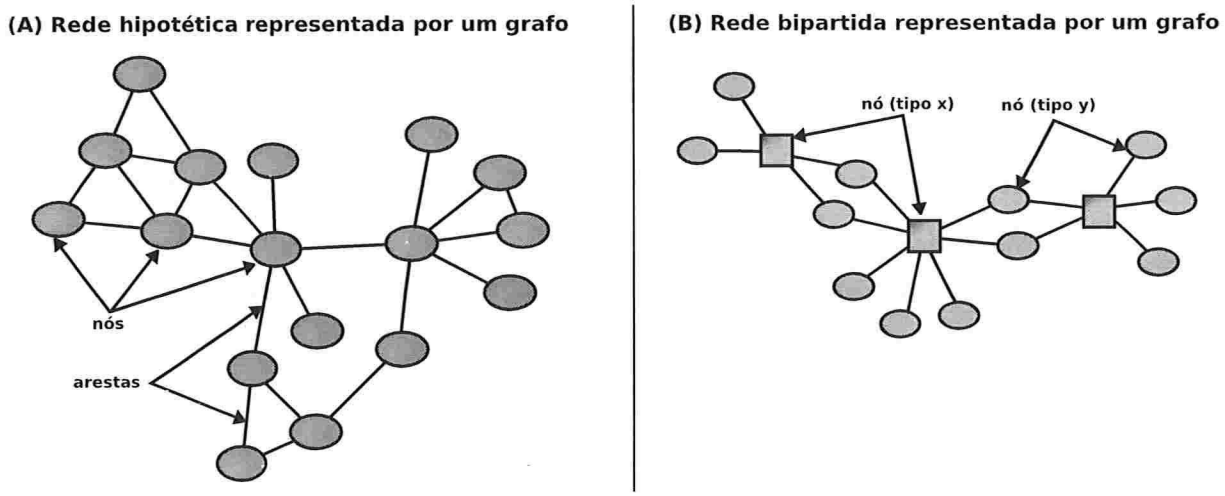


Figura 1.14: Os grafos são estruturas muito úteis na representação de redes complexas. (A) Um grafo pode representar uma rede de interações entre entidades, onde os nós representam as entidades e as arestas, as interações. (B) Uma rede bipartida representa as interações entre duas entidades diferentes.

As aplicações dos conceitos e métodos de redes complexas na biologia cresce em ritmo acelerado e já são amplamente utilizados em estudos de cadeias alimentares, redes celulares e metabólicas e redes de regulação gênica (LEE et al., 2002; NEWMAN, 2003; COSTA, 2004; PROULX; PROMISLOW; PHILLIPS, 2005; COSTA; SPORNS, 2006; COSTA; ROCHA, 2006; COSTA, 2006)

1.5 Objetivos

- identificar, anotar e caracterizar os principais genes e elementos regulatórios cis putativos envolvidos na determinação do sexo em *A. mellifera*, por meio de ferramentas de bioinformática e experimentos de PCR.
- entender a dinâmica temporal da expressão dos genes identificados pelo item anterior em embriões e tecidos de cada sexo pela técnica de RT-PCR em tempo real.
- identificar o perfil funcional dos genes alvo potencialmente regulados pelos principais fatores de transcrição envolvidos na via de determinação do sexo.
- anotar os genes envolvidos na determinação de casta a partir de dados experimentais já disponíveis na literatura.
- descobrir elementos regulatórios cis putativos nos genes de determinação de casta que possam revelar as redes regulatórias responsáveis pela ativação destes genes.
- estimar as taxas de evolução em 7 espécies de abelhas corbiculadas a partir do cálculo de substituições sinônimas em 4 genes de famílias diferentes.
- utilizar as taxas de substituição sinônima com relógio molecular para identificar os possíveis eventos de duplicação gênica correlacionados com a evolução da eusocialidade de *A. mellifera*.

2 *Material e Métodos*

2.1 Métodos computacionais

2.1.1 Sistema operacional, programas e linguagem de programação

O trabalho foi desenvolvido em um IBM eServer 110, intel x86 PentiumIII , 1,2Gb de memória RAM, 3 discos scsi de 36Gb em RAID-5, sistema operacional Linux distribuição Ubuntu (LTS - http://www.ubuntu.com). Uma considerável quantidade de programas foi utilizada para as análises. Todos os programas são de distribuição livre e a maioria com o código aberto. As informações sobre os principais programas utilizados estão listadas na Tabela 2.1.

Python é uma linguagem de programação interpretada, interativa, orientada ao objeto e extensível (FOUNDATION, 2007). Python é uma linguagem muito flexível e multiplataforma (Linux/Unix, Windows, Mac OS X, entre outros sistemas) e, além disso, é facilmente adaptada ao módulos já escritos em Python ou em linguagens compiladas como C, C++ e Fortran. Existem vários módulos escritos para integrar diferentes aplicações de bioinformática e estatística que aceleram o desenvolvimento de *pipelines* para analisar grandes quantidades de dados biológicos.

Diversos *scripts* foram escritos em Python para integrar vários programas e bases de dados. Algumas das aplicações desenvolvidas em Python e Zope estão disponíveis no servidor do Laboratório de Biologia do Desenvolvimento de Abelhas (http://zulu.fmrp.usp.br/beelab).

2.1.2 Bases de dados de seqüências genéticas e de propriedades funcionais

A grande maioria das bases de dados utilizadas neste estudo foi parcial ou integralmente baixada em nosso servidor a partir de diversas localidades. As principais bases de dados usadas

Tabela 2.1: Principais programas de computador utilizados

Programa	Utilidade	Endereço Web	Referência
BLAST	Alinhamento local	http://www.ncbi.nlm.nih.gov/BLAST	(ALTSCHUL et al., 1990)
CLUSTALW	Alinhamento múltiplo	ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalW/	(THOMPSON; HIGGINS; GIBSON, 1994)
T-COFFEE	Alinhamento múltiplo	http://www.tcoffee.org/Projects_home_page/... ...t_coffee_home_page.html	(NOTREDAME; HIGGINS; HERINGA, 2000)
HMMER	Cria padrões HMM a partir de alinhamentos múltiplos	http://hmmer.janelia.org/	(EDDY, 1998)
DNAsp	Análise de polimorfismo de nucleotídeos	http://www.ub.es/dnasp/	(ROZAS; ROZAS, 1999)
RevTrans	Alinhamento múltiplo reverso	http://www.cbs.dtu.dk/services/RevTrans/	(WERNERSSON; PEDERSEN, 2003)
PHYLIP	Inferência filogenética	http://evolution.genetics.washington.edu/... ...phylip/getme.html	(FELSENSTEIN, 1989)
TREE-PUZZLE	Análise de máxima verossimilhança	http://www.tree-puzzle.de/	(SCHMIDT et al., 2002)
MRBAYES	Inferência bayesiana de filogenia	http://mrbayes.csit.fsu.edu/	(HUELSENBECK; RONQUIST, 2001)
MEME	Localiza motivos de seqüência por método estatístico	http://meme.sdsc.edu/meme/intro.html	(BAILEY; ELKAN, 1994)
MDSCAN	Descobre motivos conservados em seqüências de DNA	http://ai.stanford.edu/ xsliu/MDscan/	(LIU; BRUTLAG; LIU, 2002)
ALIGNACE	Descobre motivos conservados em seqüências de DNA	http://atlas.med.harvard.edu/	(ROTH et al., 1998)
PHYLOCON	Identifica motivos de DNA por consenso filogenético	http://ural.wustl.edu/ twang/PhyloCon/	(WANG; STORMO, 2003)
TAMO	Módulo Python para análise de motivos em seqüência de DNA	http://fraenkel.mit.edu/TAMO/	(GORDON et al., 2005)
BIOPYTHON	Módulos Python para biologia molecular computacional	http://biopython.org/wiki/Biopython	ver site
WEBLOGO	Gera de logo de seqüência	http://weblogo.berkeley.edu/	(CROOKS et al., 2004)
FATIGO	Ferramenta web para associação de termos Gene Ontology	http://fatigo.bioinfo.cipf.es/	(AL-SHAHROUR; DIAZ-URIARTE; DOPAZO, 2004)
PHRED	Nomeia nucleotídeos a partir de arquivo de cromotograma	http://www.phrap.org/	(EWING et al., 1998)
PHRAP	Reuni seqüências de DNA	http://www.phrap.org/	ver site
CAP3	Reuni seqüências de DNA	http://genome.cs.mtu.edu/cap/cap3.html	(HUANG; MADAN, 1999)
EGENE	Gera pipelines para análise automatizada de seqüências	http://www.lbm.fmvz.usp.br/egene/	(DURHAM et al., 2005)
ARTEMIS	Ferramenta para visualização e anotação de seqüências de DNA	http://www.sanger.ac.uk/Software/Artemis/	(RUTHERFORD et al., 2000)
R	Programa de análise estatística e geração de gráficos	http://www.r-project.org/	ver site
ZOPE	Servidor de aplicações web	http://www.zope.org/	ver site

estão na Tabela 2.2.

As duas bases de dados de regiões promotoras à montante foram geradas por *scripts* Python desenhados para analisar toda a anotação do genoma de *A. mellifera* e *D. melanogaster* contida em arquivos no formato GFF (http://www.sanger.ac.uk/Software/formats/GFF/). A análise de todo o mapeamento dos genes nos genomas destas duas espécie foi estruturada em um dicionário onde cada chave é o número de identificação do gene e os valores são as coordenadas de posição destes genes no genoma. Assim, as regiões intergênicas de 1000pb à montante foram extraídas com base na coordenada genômica da extremidade 5' de cada gene e, em seguida, alinhadas contra as seqüências de proteínas de OfficialSet_pep para excluir quaisquer regiões codificadoras de genes adjacentes. As duas bases de região promotora à montante do gene contêm 10123 (Amel-PromDB) e 14380 (Dmel-PromDB) seqüências de *A. mellifera* e *D. melanogaster*, respectivamente. A região promotora do gene da *vitelogenina* de *Bombyx mori* (*Bmvg*; GenBank: NP_001037309) foi extraída manualmente a partir da anotação manual deste gene no genoma (SilkDB: scaffold002333).

Tabela 2.2: Principais bases de dados usadas

Base de dados	Endereço Web	Referência
GenBank	http://www.ncbi.nlm.nih.gov/	ver site
FlyBase	http://flybase.bio.indiana.edu/	(FlyBase Consortium, 2003)
Pfam	http://www.sanger.ac.uk/Software/Pfam/	(SONNHAMMER et al., 1998)
TRANSFAC	http://www.gene-regulation.com/	(WINGENDER et al., 2000)
Gene Ontology	http://www.geneontology.org/	(ASHBURNER et al., 2000)
Amel	http://www.hgsc.bcm.tmc.edu/projects/honeybee/	(The Honeybee Genome Sequencing Consortium, 2006)
OfficialSet_pep	ftp://ftp.beegenome.hgsc.bcm.tmc.edu	(The Honeybee Genome Sequencing Consortium, 2006)
OfficialSet_cds	ftp://ftp.beegenome.hgsc.bcm.tmc.edu	(The Honeybee Genome Sequencing Consortium, 2006)

2.1.3 Construção do *pipeline* para descoberta de motivos regulatórios super-representados em conjuntos de seqüências

Um *pipeline* escrito em Python foi construído para integrar outros módulos Python úteis (ex: TAMO, módulo para análise de motivos em seqüências de DNA) e vários programas já bem estabelecidos para descoberta de motivos conservados em conjuntos de seqüências, tais como AlignAce (ROTH et al., 1998), MDScan (LIU; BRUTLAG; LIU, 2002) e MEME (BAILEY; ELKAN, 1994)). Os valores de parâmetros de cada programa não foram alterados, com exceção do conteúdo GC (background), calculado para toda a base de dados de regiões promotoras do genoma de *A. mellifera* (Amel-PromDB, 25% de conteúdo GC). As bases de regiões promotoras dos genes (Amel-PromDB e Dmel-PromDB) são utilizadas pelo *pipeline* FindMotif como principal repositório para extrair as seqüências listadas nos arquivos de entrada.

Pontuações que quantificam a especificidade dos motivos descobertos em um dado grupo de seqüências foram calculadas por meio de várias métricas. As métricas usadas foram: MAP (*maximum a priori log likelihood*) (ROTH et al., 1998), Church (*group specificity score*) (HUGHES et al., 2000), ROC AUC (*area under the curve for a receiver-operator characteristic plot*) e MNCP (*mean normalized conditional probability*) (CLARKE; GRANEK, 2003). Para reduzir o número de motivos espúrios ou falso-positivos, os valores das principais métricas utilizadas como filtro para seleção dos motivos mais representativos são MAP $\geq$ 5.0, ROC AUC $\geq$ 0.7, Church $\leq$ 1e-5.

As métricas usadas para pontuação dos motivos descobertos já foram publicadas em outros estudos (ROTH et al., 1998; HUGHES et al., 2000; CLARKE; GRANEK, 2003). Usamos as pontuações calculadas por estas métricas para os testes estatísticos contra grupos controle, que são as pontuações resultantes de seqüências selecionadas randomicamente. O teste estatístico não paramétrico Kolmogorov-Smirnov é aplicado para testar se a distribuição dos valores de cada métrica é igual entre os motivos encontrados em grupos de genes escolhidos randomicamente e

grupos de genes que compartilham alguma propriedade biológica em comum.

Os conjuntos de motivos significativamente diferentes dos motivos descobertos de grupos randômicos são filtrados pelos valores das pontuações, considerando apenas os motivos mais relevantes para outras análises. Os motivos representados por PWMs que atenderam às condições do filtro são alinhados contra todas as seqüências descritas na base TRANSFAC que representam sítios de ligação de fatores de transcrição de *D. melanogater*. Em geral, somente os motivos com pelo menos 80% de similaridade com sítios contidos na base TRANSFAC foram considerados, mas este valor pode ser alterado para buscas menos conservadoras.

O *pipeline* FindMotif gera três arquivos de saída: (1) um arquivo contendo todas as seqüências em formato fasta referentes aos genes listados no arquivo de entrada; (2) um arquivo com todos os motivos que passaram no filtro em formato específico do módulo TAMO; e (3) um arquivo contendo uma tabela com a seqüência consenso dos motivos, pontuações e sítio similar encontrado no TRANSFAC. Um esquema ilustrando o fluxo de trabalho do *pipeline* FindMotif está apresentado na Figura 2.1.

2.1.4 Análise dos resultados de seqüenciamento

Os resultados gerados pelo seqüenciamento dos fragmentos de cDNA de *Amdsx* e de DNA genômico dos três genes (*ache2*, *or83b*, *mrjp*) em sete espécies de abelhas foram analisados e reunidos pelo programa Egene (DURHAM et al., 2005). Egene é um sistema de geração de *pipelines* para análise automatizada de seqüências, que integra diferentes programas, tais como PHRED, PHRAP, CAP3 e BLAST (ver referências na Tabela 2.1). O diagrama de fluxo de trabalho dos módulos do *pipeline* usado neste estudo está ilustrado na Figura 2.2.

2.1.5 Análise computacional de genes conservados na determinação do sexo entre *D. melanogaster* e *A. mellifera*

Seleção e anotação dos genes de determinação do sexo

Uma lista de 29 genes anotados para determinação do sexo (GO:0007530) e diferenciação do sexo (GO:0007548) em *D. melanogaster* foi recuperada da base Gene Ontology (GO), excluindo-se as seqüências redundantes (UniProt e isoformas). As seqüências de proteínas destes 29

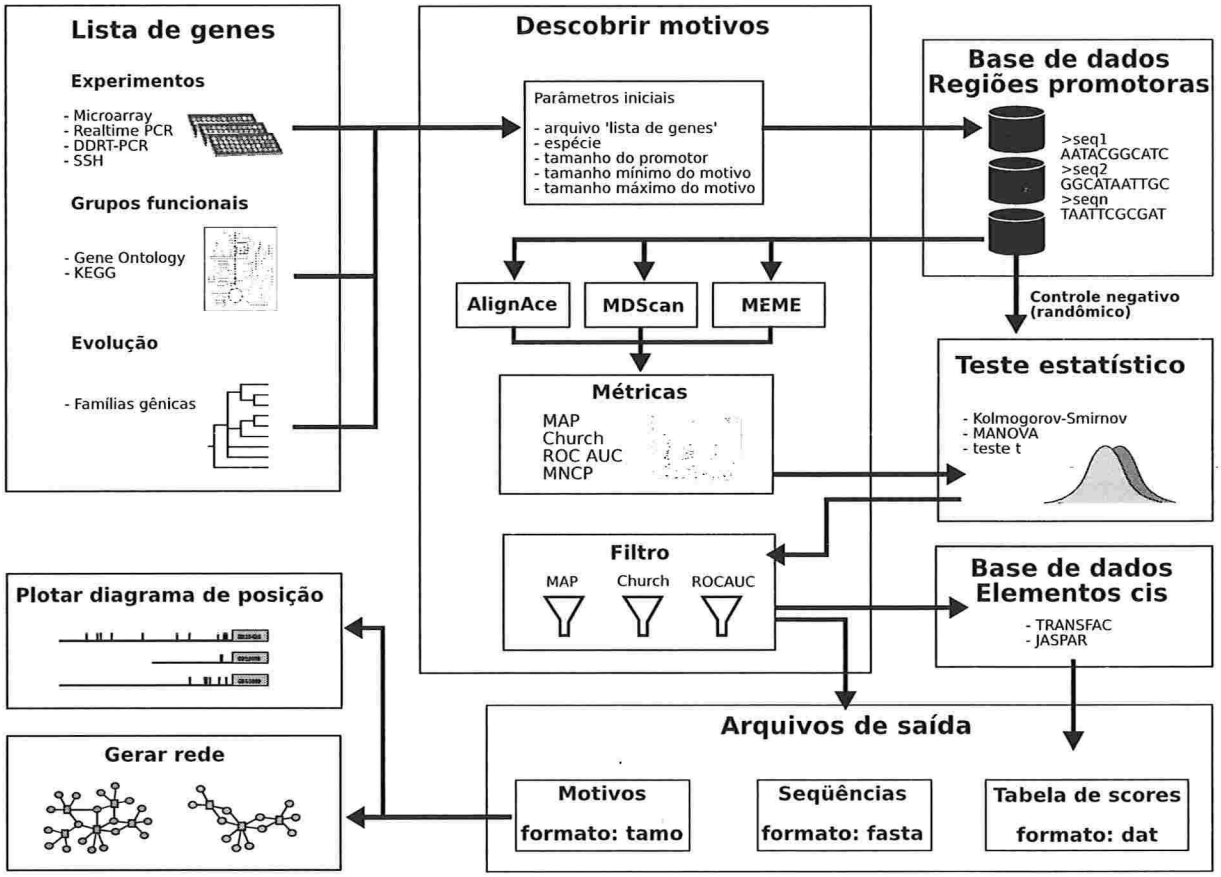


Figura 2.1: Programa para descoberta de motivos conservados nas regiões promotoras dos genes

genes foram alinhadas contra a base de seqüências de proteínas de *A. mellifera* (OfficialSet_pep) usando BLASTP. Neste caso, somente os alinhamentos muito conservados foram considerados ($E \leq$ 1e-18 e no mímino 40% de identidade). As seqüências do gene *dsx* de outros insetos foram extraídas do GenBank: *Bactrocera oleae* (CAD67987), *Anastrepha obliqua* (AAY25167), *B. mori* (BAB13472), *D. melanogaster* (AAF54169), *Megaselia scalaris* (AAK38832), *Musca domestica* (AAR23813), *Anopheles gambiae* (AAX48939), *Ceratitis capitata* (AAN63597). Todas as seqüências muito similares ao *dsx* encontradas nos dois conjuntos de seqüências de proteínas de *D. melanogaster* e *A. mellifera* foram identificadas por um método baseado na reciprocidade de melhor alinhamento, incluindo todos os genes muito conservados de um mesma família gênica ($E \leq$ 1e-18) (CHERVITZ et al., 1998).

Alinhamentos múltiplos das seqüências de proteínas dos ortólogos de *dsx* nos insetos e dos genes da mesma família de *dsx* em *D. melanogaster* e *A. mellifera* foram gerados pelo

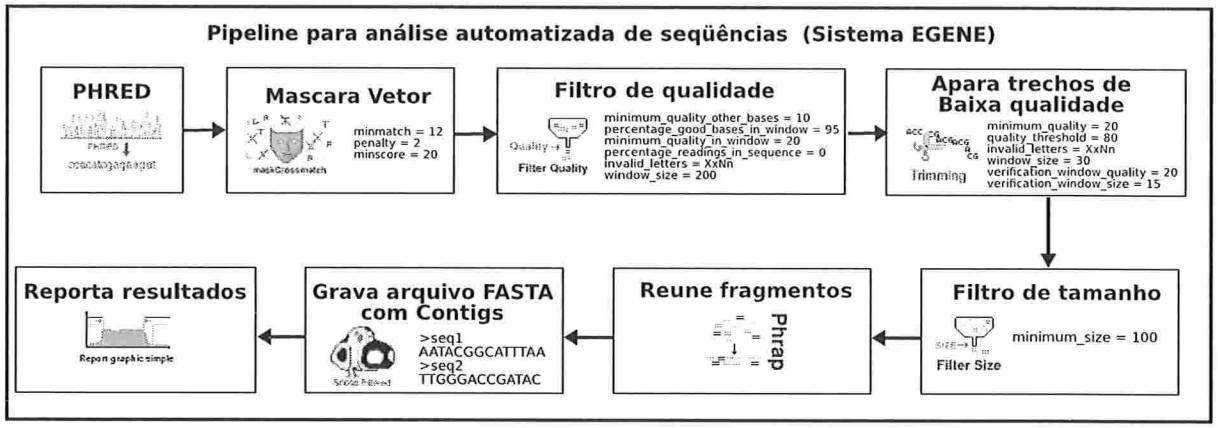


Figura 2.2: Esquema do programa Egene para o processamento dos resultados de seqüenciamento (DURHAM et al., 2005).

CLUSTALW (THOMPSON; HIGGINS; GIBSON, 1994) com os valores *default* de parâmetros (*gap opening* = 10; *gap extension penalty* = 0,1). A filogenia dos genes da família *dsx* foi inferida pelo PHYLIP v3.64 (FELSENSTEIN, 1989). SEQBOOT foi usado para gerar 1000 replicações de *bootstrap*. PROTDIST gerou matrizes de distância a partir da matriz de substituição PAM (DAYHOFF, 1979). As árvores de *neighbor-joining* (SAITOU; NEI, 1987) foram geradas pelo NEIGHBOR e CONSENSE gerou uma árvore consenso.

Os domínios de proteína conservados na família do gene *dsx* foram procurados na base de dados Pfam (SONNHAMMER et al., 1998). O programa HMMER foi usado para localizar dois domínios bem conservados (E $\leq$ 1e-10) entre estes genes, *DM DNA-binding domain* (PfamID: PF00751) e *DMRTA motif* (PfamID: PF03474).

O cDNA de vitelogenina de *A. mellifera* (*Amvg*) foi clonado e seqüenciado em estudo anterior (PIULACHS et al., 2003). A seqüência do gene *Amvg* foi recuperada do GenBank (NM_001011578) e alinhada (BLASTN) com todas as seqüências de nucleotídeos das proteínas preditas do genoma de abelha (OfficialSet_cds). Uma única seqüência foi identificada na base OfficialSet_cds (GB13999-RA). O gene *Amvg* é, portanto, identificado como GB13999. O gene *Bmvg* foi identificado pelo alinhamento da seqüência de proteína de *Amvg* contra a base nr do GenBank (NP_001037309; E $\leq$ 9e-118). A seqüência do gene *yp-1* de *D. melanogaster* foi identificada experimentalmente (HUNG; WENSINK, 1981) e está disponível no Flybase (CG2985). Os genes de vitelogenina de *B. mori* e *A. mellifera* são homólogos e bem conservados enquanto o *yp-1* é um análogo das VGs e não teve a mesma origem evolutiva.

Análise das regiões promotoras dos genes de determinação do sexo

As análises das regiões promotoras de genes envolvidos na determinação do sexo foram conduzidas por duas abordagens diferentes: (1) localizar os sítios de ligação para os três fatores de transcrição (DSX, BZIP-1, AEF-1) conhecidos como reguladores importantes na expressão sexo e tecido-específica em insetos; (2) descobrir motivos conservados entre genes co-regulados e ortólogos entre *A. mellifera* e *D. melanogaster*. Na primeira abordagem, os motivos DSX, BZIP-1 e AEF-1 correspondentes a cada fator de transcrição são representados por PWMs geradas a partir do alinhamento dos respectivos sítios de ligação descritos na literatura (FALB; MANIATIS, 1992; AN; WENSINK, 1995b). Na segunda abordagem, assumimos a hipótese de que os dois genes terminais da via de determinação do sexo (*dsx* e *ix*) são co-regulados em *Drosophila* e *Apis* e um motivo conservado entre estes ortólogos potencialmente co-regulados foi descoberto pelo programa PhyloCon (WANG; STORMO, 2003) e apresenta uma similaridade de pelo menos 70% com o sítios GAGA descrito no TRANSFAC (WINGENDER et al., 2000). Este motivo foi denominado GAGA-DIX.

Todas as análises relativas aos motivos foram desenvolvidas por programas escritos em Python, utilizando principalmente o módulo TAMO (GORDON et al., 2005) como base para a construção das PWMs e alinhamentos destas matrizes contra seqüências das regiões promotoras dos genes de interesse. Em geral, os sítios com pelo menos 60% de similaridade aos motivos descritos pelas PWMs são considerados neste estudo, entretanto, alguns sítios com até 50% de similaridade também foram incluídos nas análises. A distribuição de três motivos (GAGA-DIX, DSX, BZIP-1) nas regiões promotoras de todos os genes preditos no genoma de *A. mellifera* e *D. melanogaster* foi calculada a partir de alinhamentos com similaridade de pelo menos 60%.

O viés de posição destes três motivos foi investigado a partir de comparações da distribuição de ocorrência de sítios dos motivos canônicos (GAGA-DIX, DSX, BZIP-1) e dos motivos embaralhados (*shuffled*) com uma distribuição uniforme. Os motivos embaralhados derivaram de cada um dos motivos canônicos a partir do embaralhamento aleatório da informação de cada sítio (coluna da PWM). Foram gerados 5 motivos embaralhados para cada motivo canônico e as ocorrências destes 5 motivos são as bases de dados utilizadas para os testes estatísticos contra uma distribuição uniforme. Os testes de Kolmogorov-Smirnov foram conduzidos sob a hipótese nula de que motivos espúrios e sem função biológica apresentam uma distribuição

aproximadamente uniforme ao longo das seqüências genômicas. A distribuição dos motivos canônicos e embaralhados foi comparada com uma distribuição uniforme teórica.

Os potenciais genes-alvo do fator GAGA foram identificados pelo alinhamento do motivo GAGA-DIX contra 1000pb das regiões promotoras de todos os genes preditos (Amel-PromDB). Os genes potencialmente regulados de maneira sexo-específica foram identificados pela ocorrência de dois motivos, DSX e BZIP-1, sobrepostos em pelo menos um sítio, localizados até 600pb das regiões promotoras dos genes de abelha. As duas listas de genes-alvo (GAGA-DIX e DSX/BZIP-1) foram usadas para determinar o perfil funcional dos genes regulados pelo fator de transcrição GAGA e pelo complexo DSX/BZIP-1, através da atribuição de anotação funcional com base nos termos GO para os ortólogos recíprocos em *D. melanogaster*. As anotações funcionais foram recuperadas pelo programa FatiGO (AL-SHAHROUR; DIAZ-URIARTE; DOPAZO, 2004).

2.1.6 Análise computacional de genes relacionados à determinação de casta em *A. mellifera*

Seleção e anotação de ESTs diferencialmente expressos no desenvolvimento de castas em *A. mellifera*

As análises partiram de 164 ESTs recuperadas do GenBank (BG101532–BG101697) geradas por experimentos de SSH (EVANS; WHEELER, 1999). Estas ESTs foram posteriormente validadas por análises de macroarranjos (EVANS; WHEELER, 2001), que identificaram três principais grupos de genes: (1) genes super-expressos em larva jovem; (2) genes super-expressos no último instar larval de rainha; e (3) genes super-expressos no último instar larval de operária. O grupo 1 de ESTs foi excluído da análise porque se refere às diferenças entre larva no início do desenvolvimento (casta ainda indefinida) e larva no final do desenvolvimento (casta já definida). O grupo 2 de ESTs de rainha (82) foi expandido pela inclusão de uma seqüência de cDNA (AY601642) gerada por DDRT-PCR (CORONA; ESTRADA; ZURITA, 1999). Ao grupo 3 de ESTs de operária (40) foram incluídas mais 7 ESTs (BG149167–BG149173) geradas por DDRT-PCR em desenvolvimento de ovários (HEPPERLE; HARTFELDER, 2001).

As seqüências de EST foram alinhadas contra o genoma de abelha (Amel3.0) e os conjuntos de genes preditos (OfficialSet_cds) usando o programa BLASTN. As seqüências genômicas

(*scaffolds*) e de genes preditos identificadas pelo BLAST foram recuperadas para anotação manual. Sete ESTs de rainha não alinharam com nenhuma região do genoma seqüenciado e foram retiradas da análise. As ESTs de castas diferentes que representavam o mesmo gene também foram excluidas da análise. Algumas ESTs alinharam na região de um mesmo gene e foram agrupadas com representantes deste. Para as ESTs não identificadas nas seqüências de genes preditos (OfficialSet_cds) buscou-se por seqüências codificadoras nas adjacências da região genômica onde a EST foi mapeada, principalmente à montante (5') da EST, já que estas bibliotecas apresentam um viés de ocorrência na região 3'-UTR que não está contida nas seqüências de genes preditos. Este procedimento resultou na identificação de um conjunto de 51 genes diferencialmente expressos em larvas de rainha (34 genes) e operária (17 genes).

As seqüências de proteínas do OfficialSet foram alinhadas contra o genoma (Amel3.0) usando TBLASTN para anotação das coordenadas destas proteínas no genoma. As ESTs também foram usadas para correções dos genes preditos quando necessário e, subseqüentemente, todas as seqüências de proteínas anotadas foram alinhadas com BLASTP contra as bases não redundantes (nr) do GenBank e proteínas anotadas do Flybase. A anotação manual de cada gene foi conduzida com o programa Artemis v7.0. O arquivo final de anotação incluindo todos genes foi gerado no formato GFF (http://www.sanger.ac.uk/Software/formats/GFF/) por um *script* escrito em Python.

Todos os 51 genes anotados em *A. mellifera* foram submetidos à análise de melhor alinhamento recíproco com a base de proteínas anotadas de *D. melanogaster*, identificando os genes ortólogos entre estas duas espécies. A anotação funcional destes genes foi feita a partir dos nomes dos ortólogos em *Drosophila* e suas atribuições a termos GO de processos biológicos e funções moleculares. O programa FatiGO disponível na web foi utilizado para recuperar as anotações funcionais (em nível 3) atribuídas aos genes de *Drosophila*. Além disso, as definições de dois novos termos GO na ontologia Processo Biológico foram coordenadas juntamente com o consórcio GO (ASHBURNER et al., 2000). São eles: desenvolvimento de casta e polifenismo (GO:0048651 e GO:0048650, respectivamente). Os domínios de proteínas conservados foram identificados pela busca de padrões similares a todos os modelos HMM contidos na base Pfam usando o programa HMMER (E-value $\leq$ 1e-10).

Descoberta de motivos conservados nas regiões promotoras de genes diferencialmente expressos em castas

Para a descoberta de motivos conservados nas regiões promotoras dos dois conjuntos de genes relacionados com as castas, foram selecionados dois subconjuntos com base em dois critérios: (1) genes com os mais altos níveis de expressão diferencial nas duas castas (EVANS; WHEELER, 2001) e (2) genes com a região 5' conservada em seus ortólogos de *D. melanogaster*. Desta maneira, foram selecionados 6 genes (top6) de rainha (GB13072, GB11628, GB19380, GB14798, GB16047 e GB18242) e de operária (GB10869, GB12371, GB12239, GB10428, GB19006, GB14758). As duas listas de genes foram submetidas independentemente ao *pipeline* FindMotif (ver seção 2.1), resultando na descoberta de 2 motivos super-representados na região promotora dos genes de rainha e 12 motivos super-representados na região promotora dos genes de operária.

2.1.7 Análise computacional dos genes *mrjp*, *ache2*, *or83b* e *lw-rh* em abelhas

Aquisição de seqüências disponíveis em bases de dados públicas

A maioria das seqüências dos genes usados nas análises de taxas evolutivas foram identificadas e seqüenciadas neste estudos (ver seção 2). Algumas seqüencias de MRJP das espécies de Apini foram recuperadas do GenBank e estão apresentadas na Tabela 2.3.

Tabela 2.3: Seqüência das MRJP recuperadas do GenBank. np, não publicado

Espécie	Gene	GenBank	Referência
A.cerana	*mrjp1*	AY279539	(SU et al., 2005)
A.cerana	*mrjp2*	AY392758	(SU et al., 2005)
A.cerana	*mrjp3*	AY663105	(ALBERTOVA et al., 2005)
A.cerana	*mrjp4*	AY532368	np
A.cerana	*mrjp5*	AY392757	(SU et al., 2005)
A.cerana	*mrjp6*	AY732221	(ALBERTOVA et al., 2005)
A.cerana	*mrjp7*	AY862496	(SU et al., 2005)
A.dorsata	*mrjp3*	AY663106	(ALBERTOVA et al., 2005)
A.florea	*mrjp3*	AY663107	(ALBERTOVA et al., 2005)

O gene *long-wavelength rhodopsin* (*lw-rh*) das 6 espécies de abelhas estudas foi recuperado do GenBank e está apresentado na Tabela 2.4.

Tabela 2.4: Seqüência de *lw-rh* das 6 espécies de abelhas recuperadas do GenBank. np, não publicado

Espécie	GenBank	Referência
A.cerana	AB091502	np
A.dorsata	AF091733	(MARDULYN; CAMERON, 1999)
A.florea	AB091500	np
Melipona sp.	AF344607	(ASCHER; DANFORTH; JI, 2001)
B.terrestris	AF091722	(MARDULYN; CAMERON, 1999)
E.nigrita	AJ581732	(MICHEL-SALZAT; CAMERON; OLIVEIRA, 2004)

Alinhamento das seqüências de nucleotídeos, cálculo de taxas evolutivas e tempo de divergência

As seqüências de nucleotídeos foram alinhadas usando o alinhamento de aminoácidos como referência para manter as posições dos sítios de códons alinhadas. O programa clustalw foi usado para o alinhamento múltiplo das seqüencias de aminoácidos e o alinhamento das seqüências de nucleotídeos foi conduzido com o programa T-coffee usando o alinhamento de aminoácidos.

Os cálculos de substituição sinônima (K_s) e não sinônima (K_n) foram conduzidos pelo programa DNAsp. As árvores filogenéticas reconstruídas pelos métodos de *neighbor-joining* usando o modelo de substituição de nucleotídeo *maximum composite likelihood* e máxima parcimônia foram feitas com o programa MEGA3.1, e as árvores de máxima verossimilhança foram feitas com o Tree-puzzle, usando o modelo de substituição de Tamura-Nei (TAMURA; NEI, 1993). *E. nigrita* foi considerada como grupo externo. Os testes de neutralidade e de taxa relativa de mutações foram realizados com o programa DNAsp e MEGA3.1, respectivamente.

O tempo de divergência entre as espécies de abelhas foi estimado por uma função linear aproximada por regressão linear a partir do K_s médio estimado pela comparação dos três genes (*ache2*, *or83b* e *lw-rh*) de *A. mellifera* e *Melipona* e do registro fóssil de um meliponini (ENGEL, 2000, 2001). O tempo de divergência entre as espécies foi, então, calculado a partir desta equação usando K_s como variável de entrada. A taxa de evolução absoluta foi calculada como já descrito na literatura (KIMURA, 1980).

As árvores filogenéticas de tempo de divergência das espécies de abelhas e das duplicações dos parálogos de *A. mellifera* foram primeiramente reconstruídas por *neighbor-joining* (1000 replicações), considerando somente o terceiro sítio do códon, usando o modelo de substituição

de nucleotídeo de Nei e Gojobori (1986) com correção de Jukes e Cantor (1969). Estas árvores foram, então, linearizadas pelo algoritmo implementado no programa MEGA, usando a taxa evolutiva absoluta V_s e *E. nigrita* como grupo externo.

2.1.8 Análises estatísticas

As análises estatísticas de Kolmogorov-Smirnov, chi-quadrado, regressão linear e MANOVA foram realizadas durante estes estudos, utilizando principalmente o programa R.

2.2 Material Biológico

2.2.1 Coleta de embriões e tecidos de macho e fêmea para análise de genes de determinação do sexo em *A. mellifera*

Os embriões e tecidos de abelhas adultas foram coletados a partir de duas colônias (num. 52 e 37) de abelhas africanizadas (*A. mellifera*) instaladas no apiário da Faculdade de Medicina de Ribeirão Preto (USP, Ribeirão Preto). A coleta dos embriões foi realizada em sete fases (0, 12, 24, 36, 48, 60 e 72hs) do desenvolvimento embrionário de macho e fêmea de *A. mellifera*. Cada fase do período embrionário foi representada por uma amostra de 100 embriões oriundos de duas colônias. Dois zangões recém-nascidos e quatro rainhas recém-nascidas forneceram as amostras de tecidos. A carcaça abdominal dos zangões foi separada do aparelho reprodutor e sistema digestivo e é composta do tegumento e corpo gorduroso. Os ovários e corpo gorduroso de cada rainha foram cuidadosamente retirados da cavidade abdominal com auxílio de Estereomicroscópio (Leica) e instrumentos adequados para dissecção. Os embriões e tecidos foram armazenados a -70oC em criotubos de 2ml contendo Trizol (Gibco) para posterior extração de RNA total.

2.2.2 Coleta de espécies de abelhas com corbícula para estudos de taxas evolutivas

As abelhas fêmeas de 7 espécies representantes das 4 tribos da família Apidae e subfamília Apinae foram coletadas em diferentes localidades como descreve a Tabela 2.5. As amostras

foram preservadas em etanol 70% a -20^{o}C até o momento da extração de DNA genômico de três abelhas de cada espécie.

Tabela 2.5: Espécies de abelhas utilizadas no estudo de taxas evolutivas. (a) amostras fornecidas por Dr. Denis Anderson, CSIRO-Entomology, Austrália. (b) amostras fornecidas por Profa. Dra. Zilá Luz Paulino Simões. (c) amostras fornecidas por Prof. Dr. Marco Antônio Del Lama, UFSCar, Brasil.

Espécie	Tribo	Modo de vida	Local
Apis mellifera	Apini	Eusocial	Canberra, Austrália
Apis cerana (a)	Apini	Eusocial	Malinau, Bornéo
Apis dorsata (a)	Apini	Eusocial	Filipinas
Apis florea (a)	Apini	Eusocial	Banglore, Índia
Melipona quadrifasciata (b)	Meliponini	Eusocial	São Paulo, Brasil
Bombus terrestris	Bombini	Semi-social	Tasmânia, Austrália
Eulaema nigrita (c)	Euglossini	Solitário	São Paulo, Brasil

2.3 Manipulação de ácidos nucléicos e técnicas de biologia molecular

2.3.1 Extração de RNA total

Amostras de embriões (0-72hs) e tecidos (carcaça de zangões recém-nascidos e ovários e corpo gorduroso de rainhas recém-nascidas) foram guardadas em Trizol (Gibco) a -70^{o}C. O RNA total de cada amostra foi extraído utilizando o protocolo de Trizol, de acordo com as especificações dos fabricantes. Para inibir a ação de RNAses e evitar a degradação do RNA, as amostras foram suspensas em água tratada com dietilpirocarbonato (DEPC) 0,1% (v/v) e, em seguida, incubadas em presença de 3 unidades de DNAse I por 30 min a 37^{o}C para eliminar uma possível contaminação com DNA. A concentração do RNA foi estimada com um espectrofotômetro a um comprimento de onda de 260 nm.

2.3.2 Extração de DNA genômico

O DNA genômico de cada uma das sete espécies de abelhas (ver Tabela 2.5) foi extraído a partir do tórax de três representantes de cada espécie. A extração de DNA foi feita com *DNA*

Purification Kit (QIAGEN), seguindo o protocolo especificado pelo fabricante. Para os ensaios de amplificação por PCR, foram usadas soluções de DNA de 0,1μg/μl.

2.3.3 Construção dos estoques de cDNA por transcrição reversa seguida por PCR (RT-PCR)

Moléculas de cDNA foram sintetizadas a partir de 0,5–5μg de RNA total de cada amostra de embriões e tecidos. O protocolo é dividido em 2 etapas:

Tratamento do RNA com DNAse Para eliminar eventual contaminação com DNA, uma solução contendo o RNA total e DNAse I é preparada a partir de 0,5–5μg de RNA em água com DEPC e DNAse I diluída 10x. A solução é incubada a 37^{o}C por 40 min. Em seguida, a DNAase I é inativada a 70^{o}C por 15 min e finalmente mantida em gelo.

Transcrição reversa (RT) A reação de transcrição reversa sintetiza uma fita de DNA complementar (cDNA) a partir de uma fita de RNA. Os estoques ou bibliotecas de cDNA construídos a partir das moléculas de RNA mensageiro representadas em uma amostra são muito mais estáveis para manipulação em experimentos de amplificação por PCR. Para esta reação foi usado o kit *SuperScript II Reverse Transcriptase* (Invitrogen).

A reação é realizada em 3 etapas: (1) 0,5–5μg de RNA total tratado com DNAase I, 500μg/μl Oligo(dt)$_{12-18}$ e 10mM de cada dNTP, incubando a 65^{o}C por 5 min e, em seguida, colocado em gelo; (2) acrescentar a mistura *First Strand Buffer 5x*, 100mM de DTT e 40 unidades/μl de *RNAse Out*, incubando a 42^{o}C por 2 min e, em seguida, colocar em gelo; (3) finalmente, acrescentar *SuperScript II Reverse Transcriptase* e deixar a 42^{o}C por 50 min, a 70^{o}C por 15 min e, então, pode ser armazenada a -20^{o}C para posterior utilização em PCR.

2.3.4 Reação em cadeia da polimerase (PCR)

As amplificações de fragmentos de cDNA e DNA genômico foram realizadas de acordo com protocolos já publicados (MULLIS; FALOONA, 1987), mas com algumas adaptações.

PCR de fragmentos de cDNA em amostras de embriões e tecidos

As reações ocorreram em volume final de 25μl contendo 10μl de *Master mix* (Eppendorf), 1μl de cada iniciador (10ρmoles/μl), 1-2μl da solução de cDNA molde e água Milli-Q completando o volume final. Os iniciadores utilizados aos pares (*forward* e *reverse*) estão relacionadas na Tabela 2.6. A mistura de reação foi submetida a PCR na seguinte condição: 40 ciclos, desnaturação (95^{o}C por 30 s), *annealing*[1] (60^{o}C por 2 min) e polimerização (72^{o}C por 30 s). Finalizados os 40 ciclos, foi feita uma etapa de extensão final a 72^{o}C por 5 minutos.

PCR de fragmentos de DNA genômico em amostras de espécies de abelhas

As reações ocorreram em volume final de 50μl contendo 5μl de tampão 10x concentrado (200nM Tris-HCl pH 8.0, 0,1 mM EDTA, 1mM DTT, 50% glicerol), 5μl de $MgCl_2$ (50mM), 2μl de dNTP mix (10mM), 1μl de tampão Taq DNA polimerase (5unidades/μl), 1μl de cada iniciador (10ρmoles/μl), 1μl de solução de DNA genômico como molde e 34μl de água Milli-Q. Os iniciadores utilizados aos pares (*forward* e *reverse*) estão relacionadas na Tabela 2.7. As misturas de reação contendo combinações de iniciadores diferentes e como DNA molde de espécies diferentes foram submetidas a PCRs com diferentes condições de temperatura na etapa de *annealing*: 40 ciclos, denaturação (96^{o}C por 1 min), *annealing* (55–65^{o}C por 1 min) e polimerização (72^{o}C 2 min). Após os 40 ciclos, uma etapa final de extensão dos fragmentos foi feita a 72^{o}C por 5 minutos.

Eletroforese em gel de agarose para fracionamento de moléculas de DNA

O produto da PCR foi carregado com tampão *DNA loading* (Ficoll 400 15%, azul de bromofenol 0,25% e xileno cianol 0,25%) em gel de agarose 1% (1g de agarose/100ml de tampão TBE 1x) contendo 0,5μg/ml de Brometo de etídio. A eletroforese ocorreu submersa em tampão TBE 1x na condição de 100V. O fracionamento das moléculas de DNA foi visualizado com luz ultravioleta (UV_{260nm}).

[1] A capacidade de combinar ácidos nucléicos complementares por um processo de aquecimento e resfriamento.

2.3.5 Iniciadores para RT-PCR e PCR em tempo real

Os iniciadores desenhados para a amplificação específica dos fragmentos dos genes estudados estão listados nas Tabelas 2.6 e 2.7.

Tabela 2.6: Iniciadores (ou *primers*) desenhados para RT-PCR e PCR em tempo real dos genes de determinação do sexo. Tm, temperatura de dissociação de uma dupla fita de DNA; F, *forward*; R, *reverse*.

Iniciador	Orientação	Seqüência do iniciador (5'→3')	Tm (^oC)
AMDSX-F1	F	TGCGAGAAGTGTAAGATCAC	60
AMDSX-R2	R	GTGCTCCAATAGAATTTCCAC	60
AMDSX-R3	R	GCACGACTAGGTTGGGACAT	60
AMDSX-F4	F	ATTTCTTCGGTCCCTCAACC	60
AMDSX-R5	R	GCGGCAATTTTCTTGTAGGA	60
AMIX-F1	F	GGTGGTATTGATCAGCCAGA	60
AMIX-R2	R	TGGCATGACAGGTACTGGAA	60
AMVG-F1	F	GCAGAATACATGGACGGTGT	60
AMVG-R2	R	GAACAGTCTTCGGAAGCTTG	60
RP49-F1	F	CGTCATATGTTGCCAACTGGT	60
RP49-R2	R	TTGAGCACGTTCAACAATGG	60

2.3.6 Clonagem

Preparo de meios de cultura

Meio LB sólido para placas O meio de cultura LB sólido é utilizado para selecionar bactérias que possuem o plasmídeo com fragmento de DNA inserido. O meio LB sólido é feito a partir de 10g/l de Triptona, 5g/l de extrato de levedura, 10g/l de NaCl e 15g/l de *Botton agar* misturado em água destilada e a mistura é esterilizada em auto-clave. Em uma câmara de fluxo ou capela, acrescentar 100mg/ml de Ampicilina e XGAL/IPTG (50mg/ml e 200mg/ml, respectivamente) com a mistura ainda líquida a aproximadamente 40^oC (mas não superior). Cada placa de petri deve receber um fina camada com uma distribuição homogênea ($\approx$ 50ml). Após a solidificação do meio, as placas devem ser guardadas em geladeira preferencialmente protegidas de luz e devem ser usadas em um prazo médio de 1 mês para evitar a degradação da Ampilicina e/ou XGAL/IPTG.

Tabela 2.7: Iniciadores (ou *primers*) desenhados para PCR dos genes usados nas medidas de taxas evolutivas. Tm, temperatura de dissociação de uma dupla fita de DNA; F, *forward*; R, *reverse*. I - inosina; R - A/G; Y - T/C; W - A/T; S - G/C; M - A/C; K - G/T.

Iniciador	Orientação	Seqüência do iniciador (5'→3')	Tm (^{o}C)
MRJP-F1	F	AYIAATGGAAIYWYITIGATTAYRAYTTYG	51
MRJP-F3	F	GGAAATACTTCGATTATAATTTYG	48
MRJP-R2	R	GAAGWCAWTCGWTGRAARGAATYRTCIG	57
MRJP9-E3-R1	R	AGATCAAAGACGACAATCTGCGAAG	56
ACHE2-F3	F	GGTCGTGGAGACGACCAG	55
ACHE2-F5	F	AACACCAACATATCCGAGGACTG	55
ACHE2-R4	R	CCTCGTTCTCGTTGTTACCG	54
ACHE2-R6	R	CTCGTTGTTACCGATCAGTATCTC	56
ACHE2-R6A	R	CTCGTTGTTACCGATSAGWATCTC	56
OR83B-F3	F	GTCCACCTGGTCCTGATACTGAT	57
OR83B-F5	F	ATACCAAGGTTGATGGTTCGTTC	53
OR83B-F7	F	TATGGTGTAGCTTTGCTGCTACATATG	57
OR83B-R2	R	TGCTTGTGICICTCIACCCAGTACTTG	59
OR83B-R4	R	AACGATATGCTTGTGTCTCTCTACC	56
OR83B-R4A	R	CKMACSAYRTGCTTGTGYCKCTCKACC	63
OR83B-R8	R	CACCATGAAGTAGGTAACCATAGC	56

Meio para crescimento de bactérias transformadas O LB SOC é usado como meio para o crescimento de células, garantindo uma máxima eficiência de transformação. Este meio é composto por 0,5% de extrato de levedura, 2% de triptona, 10mM de NaCl, 2,5mM de KCl, 10mM $MgCl_2$, 20mM $MgSO_4$ e 20nM de glucose.

Meio LB líquido para crescimento de clones O meio de cultura LB líquido é usado para o crescimento dos clones de colônias transformadas contendo o fragmento de DNA. O LB líquido segue as mesmas proporções do LB sólido mas sem agar, sendo composto por 1% de triptona, 0,5% de extrato de levedura, 10mM de NaCl e Ampilicina 1μl/ml de LB.

Purificação de fragmento de DNA a partir de gel de agarose

Fragmentos de DNA foram recuperados do gel de agarose 1% com auxílio do *QIAquick Gel Extraction Kit* (QIAGEN). A banda de interesse é recortada do gel e, em seguida, solubilizada com solução tampão. A solução contendo DNA e gel é aplicada em uma coluna de filtração que retém o DNA; Após algumas lavagens, o DNA purificado é eluído e está pronto para ser usado na reação de ligação.

Reação de ligação

Uma alíquota do DNA recuperado do gel de agarose (ver seção 2.3.6) ou do bloco de gel (agarose *low melting*) derretido a 65^oC contendo o fragmento de DNA foi alternativamente utilizada para as reações de ligação com 25 ng de vetor *pGEM-T easy vector*. A reação de ligação ocorreu *overnight* a 10^oC, em solução de ligação contendo 5μl de tampão (2x) indicado pelo fabricante, 1μl de vetor, 1μl de enzima *T4 DNA ligase* fornecida com o kit e 3μl do DNA recuperado. O volume final da reação foi 10μl.

Vetor para clonagem

Para subclonagem de fragmentos de DNA amplificados por PCR usamos o vetor de alto número de cópias *pGEM-T Easy Vector* (Promega). Este vetor é preparado pela clivagem com *EcoR V*, adicionando uma timidina em cada uma das extremidades 3' do vetor linearizado. Estas timidinas 3' protrusas no sítio de inserção aumentam muito a eficiência de ligação de produtos de PCR no plasmídio porque evitam a recircularização do vetor e fornecem maior compatibilidade com produtos de PCR gerados por alguns tipos de DNA polimerases (Taqs) que freqüentemente inserem uma deoxiadenosina na ponta 3' de fragmentos amplificados. O vetor é composto por 3015 pares de bases, contendo os promotores de RNA polimerase T7 e SP6 flanqueando o sítio para clonagens múltiplas que fica dentro da região codificadora do α-peptídeo da enzima β-galactosidase. A inativação insercional do α-peptídeo por clonagem permite a identificação dos clones recombinantes por cor (ex: cor branca = positivo; cor azul = negativo).

Transformação de bactéria quimio-competentes

As células de bactérias quimio-competentes (*Escherichia coli* linhagem JM109) são mantidas em freezer -80^oC e devem ser descongeladas em gelo alguns minutos antes (5 min) do início do processo de transformação. Com as células descongeladas, acrescenta-se 10μl da reação de ligação mantendo-as em gelo por mais 30 minutos. O choque térmico deve ser feito a 42^oC por 45-50 segundos e, em seguida, os tubos voltam para o gelo por mais 5 minutos. O SOC é então acrescentado (0,8-1ml) e as bactérias são incubadas a 37^oC por 1-2 horas agitando a 240rpm. Centrifuga-se os tubos a 3400rpm por 5 minutos e retira-se o excesso de SOC (500-700μl),

suspende-se cuidadosamente as bactérias no meio e espalha-se o volume nas placas de petri com meio LB sólido. As placas são incubadas a 37^oC *overnight*.

As colônias brancas (com fragmentos de DNA inseridos no plasmídeo) são selecionadas e transferidas para tubos com 5ml de LB + Ampicilina (5μl/5ml de LB). Os clones são incubados a 37^oC *overnight*.

Extração de plasmídeos das bactérias (MiniPrep)

Os plasmídeos contendo insertos de DNA foram extraídos dos clones de coloração branca crescidos em meio LB + AMP, e purificados com auxílio do kit *QIAprep Spin Miniprep*. As soluções contendo apenas os plasmídeos foram armazenas a -20^oC.

Digestão de plasmídeo com enzima de restrição

Os plasmídeos extraídos das bactérias foram submetidos à digestão com enzima de restrição que permite verificar se o inserto tem o tamanho esperado. A reação de digestão foi feita em volume final de 28μl contendo 3μl de DNA plasmidial, 0,5μl da enzima NotI, 2,5μl de *10x buffer O* e 22μl água Milli-Q. A solução foi incubada a 37^oC por 1 hora. O resultado da digestão foi verificado em gel de agarose 1%.

2.3.7 Seqüenciamento de DNA

Para determinação da seqüência de resíduos de nucleotídeos em DNA foi utilizado um protocolo adaptado a partir do original descrito por Sanger e col. (SANGER; NICKLEN; COULSON, 1977). As reações de seqüenciamento foram preparadas utilizando o reagente *BigDye Terminator Cycle Sequencing Ready Reaction v3.1* (Applied Biosystems) numa reação com volume final de 10 μl contendo 2 μl de *BigDye*, 1 μl de iniciador (M13-F ou M13-R) a 10 pmoles/μl e 1μl de DNA plasmidial (200–500ng/μl), 2μl de tampão 2,5x e 3μl de água Milli-Q. As reações foram incubadas por 1 minuto a 95°C e submetidas a 35 ciclos com denaturação a 95°C por 10 segundos, anelamento a 50°C por 5 segundos, e extensão a 60°C por 4 minutos em termociclador (GeneAmp PCR System 9700).

Após a reação de seqüenciamento e precipitação do DNA, foram adicionados, a cada amostra, 30 μl de água Milli-Q, 5 μl de acetato de potássio 3M, pH 5,2 e 100 μl de etanol absoluto. A mistura foi feita por inversão seguida de incubação por 15 minutos à temperatura ambiente e de centrifugação a 12000 x*g* por 30 minutos, a 4°C. O sobrenadante foi descartado e seu resquício retirado por inversão do tubo sobre papel absorvente. Foram adicionados 200 μl de etanol 70% e após incubação por 2 minutos à temperatura ambiente procedeu-se nova centrifugação a 15000 x*g* por 6 minutos, a 4°C. O sobrenadante foi descartado e a lavagem com etanol 70% foi repetida. Em seguida, centrifuga-se o tubo a 720 x*g* por 3 minutos, o excesso de etanol é deixado evaporar por 30 minutos, mantendo-se o tubo aberto sobre a bancada. O DNA precipitado foi solubilizado em tampão formamida *hidi* (Applied Biosystems) e enviado ao seqüenciador automático *ABI Prism 310* (Applied Biosystems) produzindo um arquivo de cromotograma para cada reação de seqüenciamento.

2.3.8 Quantificação relativa por PCR em tempo real

Com o objetivo de obter uma quantificação relativa mais precisa dos transcritos dos genes *Amdsx*, *Amix* e *Amvg*, o ensaio de PCR em tempo real foi otimizado. O controle interno ou constitutivo escolhido como referência foi o gene *rp49* (GenBank: AF441189), por apresentar o menor coeficiente de variação entre amostras de diferentes estágios de desenvolvimento e tecidos, quando comparado com outros controles normalmente usados (*actina*, *elf2-*α e *tbp*; dados ainda não publicados por Lourenço et al). Os iniciadores foram desenhados para que os amplicons tenham $\approx$ 150pb e a eficiência de amplificação próxima de 2. As combinações de iniciadores foram as seguintes: *Amdsx*, AMDSX-P4 e AMDSX-R5; *Amix*, AMIX-F1 e AMIX-R2; *Amvg*, AMVG-F1 e AMVG-R2; *rp49*, RP49-F1 e RP49-R2. As seqüências dos iniciadores estão listadas na Tabela 2.6.

A PCR quantitativa foi conduzida em tampão de amplificação *SYBRGreen PCR Master Mix* (Applied Biosystems) e as amplificações foram detectadas com o equipamento *Real Time PCR Applied Biosystems 7500*. A reação de amplificação para cada estágio embrionário (0-72hs) e tecido (MARC, FCG, FOV) ocorreu em 1μl de cDNA, 1μl de cada iniciador, 10μl de *SYBRGreen PCR Master Mix* (Applied Biosystems) e 8 μl de água Milli-Q e submetidas à reação de amplificação com a seguinte configuração: 40 ciclos, denaturação (95^{o}C por 15s), *Anneling*

(60^oC por 60s). O modelo usado para o cálculo de quantificação relativa foi o da aproximação por $2^{-\triangle\triangle Ct}$ (PFAFFL, 2001).

3 *Resultados e Discussão*

3.1 Análise comparativa dos genes conservados de determinação do sexo entre *A. mellifera* e *D. melanogaster*

Uma comparação dos genes mais conservados da via de determinação do sexo entre a mosca-de-fruta, *D. melanogaster*, e a abelha-de-mel (*A. mellifera*), foi realizada a partir de ferramentas de bioinformática desenvolvidas para análises de seqüências genéticas. As seqüências de proteínas já descritas em mosca-de-fruta como participantes da cascata de determinação do sexo foram recuperadas da base de dados Gene Ontology e posteriormente alinhadas contra um base de dados de seqüências protéicas preditas a partir do genoma de *A. mellifera*. O gene *dsx* é considerado um dos mais conservados da via de determinação do sexo em insetos. Um fragmento de cDNA do *dsx* de *A. mellifera* (*Amdsx*) específico de macho foi experimentalmente identificado, clonado e seqüenciado. O gene *ix* de *A. mellifera* (*Amix*) também foi identificado por PCR de um fragmento de cDNA em embriões de ambos os sexos. O perfil de expressão dos genes *Amdsx*, *Amix* e *Amvg* foi investigado em embriões de 0 a 72h de machos e fêmeas bem como a expressão em dois tecidos de rainhas recém-nascidas (corpo gorduroso e ovário) e carcaça abdominal de macho (corpo gorduroso + tegumento).

3.1.1 Identificação computacional e anotação de genes conservados da via de determinação do sexo em abelhas

A identificação dos 13 genes mais conservados em *A. mellifera* envolvidos no processo de determinação do sexo (GO:0007530) e diferenciação do sexo (GO:0007548) em *D. melanogaster* está relacionada na tabela 3.1. A comparação se iniciou com 29 seqüências protéicas de *D. melanogaster* anotadas na base GO como participantes destes processos biológicos e somente

as seqüências com pelo menos 40% de identidade e *e-value* $< 1e-18$ foram consideradas na análise. Entre os 13 genes mais conservados, existem 3 deles que são altamente conservados e participam de ambos os processos (determinação e diferenciação do sexo), são eles: *sexlethal* (*Drosophila sxl* apresenta 68% de identidade com GB13127-PA de *Apis*), *intersex* (*ix* de mosca-da-fruta apresenta 43% de identidade com GB19364-PA de abelha) e *doublesex* (*dsx* de mosca-da-fruta apresenta 51% de identidade com GB18426-PA de abelha) (Tabela 3.1). Dentre estes 13 genes estão elementos importantes para o completo desenvolvimento e manutenção do dimorfismo sexual, mas nenhum deles é tão bem estudado e conhecido como o gene *dsx*.

O gene *dsx* codifica um fator de transcrição Zn-finger altamente conservado que apresenta o motivo DM de ligação ao DNA. Este domínio apresenta um padrão novo de cisteínas e histidinas que possibilitam ligar ao sulco menor do DNA (ERDMAN; BURTIS, 1993; ZHU et al., 2000). O domínio DM foi assim nomeado pela sua ocorrência em DSX (*D. melanogaster*) e MAB-3 (*C. elegans*) (RAYMOND et al., 1998). O motivo DM é conservado na determinação do sexo dos metazoários. Em humanos, os genes contendo estes motivos estão envolvidos na diferenciação dos testículos. A deleção de *dmrt1* e *dmrt2* está relacionada com a reversão sexual XY, mesmo que o gene determinante de macho, *sex-determining region Y* (*sry*), esteja íntegro (SINCLAIR et al., 1990). Em peixes e mamíferos, um gene DM, *terra*, tem função na formação de padrão do mesoderme em ambos os sexos (MENG et al., 1999; VOLFF; ZARKOWER; SCHARTL, 2003). Os genes homólogos ao *dsx* podem ser os genes mais antigos na cascata de determinação do sexo nos metazoários de acordo com a hipótese "base-topo" (*bottom-up*) da evolução destas cascatas (WILKINS, 1995).

Os 13 genes reportados na Tabela 3.1 estão envolvidos em 3 diferentes níveis de regulação molecular: (1) *splicing* alternativo, (2) transcrição gênica e (3) transdução de sinal. Os 6 fatores de transcrição (TFs) são: *dsx* (TF tipo Zn-finger com um domínio de ligação DM), *ix* (TF com domínio de ligação à proteína), *fruitless* (*fru*, TF tipo Zn-finger com domínio de ligação à proteína), *deadpan* (*dpn*, TF tipo bHLH), *dissatisfaction* (*dsf*, TF tipo Zn-finger com domínio receptor de esteróides), *runt* (*run*, TF com atividade de RNApolII), *bric-a-brac* (*bab1*, TF tipo helix-turn-helix) e *sex combs reduced* (*scr*, TF com atividade RNApolII). Três genes participam do mecanismo de *splicing*, são eles: *sexlethal* (*sxl*), *transformer2* (*tra2*) e *female lethal* (*fl*). Dois genes participam em vias de transdução de sinal, são eles: *hopscotch* (*hop*) e *protein kinase 61C* (*pk61C*). A nomenclatura dos genes e anotação funcional foram retiradas das bases FlyBase e

Tabela 3.1: Genes conservados envolvidos em dois processos biológicos (determinação do sexo, GO:0007530; diferenciação do sexo, GO:0007548) em *D. melanogaster* e *A. mellifera* (só estão relacionadas as seqüências com pelo menos 40% de identidade e E-value $< 1e-18$). ***doublesex* é um dos mais importantes genes conhecidos como reguladores do desenvolvimento do fenótipo sexual em metazoários. *genes envolvidos nos dois processos.

Determinação do sexo (GO:0007530)					Diferenciação do sexo (GO:0007548)				
Nome	**FB-ID**	**Amel-ID**	**%Id**	**Evalue**	**Nome**	**FB-ID**	**Amel-ID**	**%Id**	**Evalue**
dpn	CG8704	GB12076-PA	45.73	4e-46	babl	CG9097	GB13762-PA	62.81	2e-62
dsf	CG9019	GB14217-PA	46.91	4e-58			GB15064-PA	41.72	4e-54
		GB14217-PA	71.79	6e-47			GB16756-PA	52.73	2e-42
		GB10077-PA	43.00	7e-42			GB17640-PA	50.66	3e-40
		GB10077-PA	64.52	1e-29			GB14194-PA	53.85	9e-39
		GB20053-PA	74.74	1e-37			GB18625-PA	58.12	1e-37
		GB17775-PA	53.64	6e-28			GB19033-PA	45.95	3e-37
		GB16648-PA	47.06	9e-21			GB10633-PA	57.26	9e-36
		GB18358-PA	54.32	2e-21			GB16366-PA	56.90	3e-35
dsx**	CG11094	GB15791-PA	44.70	4e-20			GB15346-PA	56.41	1e-34
		GB18426-PA	51.06	1e-19			GB14649-PA	52.07	1e-33
fl	CG6315	GB13927-PA	55.64	3e-77			GB12094-PA	54.87	2e-33
fru	CG14307	GB17617-PA	81.20	7e-52			GB17617-PA	43.54	5e-31
		GB15064-PA	51.13	2e-37			GB18588-PA	45.22	6e-30
		GB11420-PA	57.98	2e-36			GB14319-PA	45.22	2e-25
		GB18625-PA	41.33	2e-36			GB11337-PA	40.60	3e-25
		GB14194-PA	41.62	1e-33			GB14696-PA	42.37	7e-22
		GB14649-PA	43.04	1e-32	**dsx****	CG11094	GB15791-PA	44.70	4e-20
		GB16756-PA	53.85	1e-31			GB18426-PA	51.06	1e-19
		GB12094-PA	50.43	2e-31	ix*	CG13201	GB19364-PA	43.23	1e-36
		GB18588-PA	46.21	1e-30	pk61C	CG1210	GB15780-PA	51.41	6e-161
		GB10633-PA	52.21	1e-30	scr	CG1030	GB13491-PA	47.36	1e-85
		GB14070-PA	47.33	3e-30			GB13409-PA	69.41	5e-30
		GB14243-PA	49.57	2e-28			GB13409-PA	82.54	5e-26
		GB19568-PA	50.00	9e-28			GB18813-PA	72.62	8e-30
		GB11400-PA	42.45	9e-28			GB19738-PA	67.47	3e-24
		GB19033-PA	43.23	1e-27			GB11524-PA	84.48	4e-23
		GB16366-PA	41.03	1e-27			GB18792-PA	41.50	5e-19
		GB17640-PA	46.49	7e-27	sxl*	CG18350	GB13127-PA	68.18	4e-70
		GB11337-PA	48.70	4e-26			GB18785-PA	46.67	2e-45
		GB17083-PA	44.35	8e-26					
		GB14319-PA	44.72	1e-22					
hop	CG1594	GB16422-PA	45.42	9e-59					
		GB17556-PA	41.09	5e-19					
ix*	CG13201	GB19364-PA	43.23	1e-36					
run	CG1849	GB11654-PA	40.25	5e-64					
		GB16431-PA	46.69	4e-50					
		GB15836-PA	50.00	7e-50					
sxl*	CG18350	GB13127-PA	68.18	4e-70					
		GB18785-PA	46.67	2e-45					
tra2	CG10128	GB11130-PA	45.81	3e-47					

GO.

Com exceção do *ix*, todos os outros genes apresentaram similaridade com mais de um gene no genoma de abelha. Estes genes homólogos são membros das mesmas famílias gênicas e podem compartilhar um ou mais domínios conservados. Quatro genes (*fru*, *dsf*, *babl* e *scr*) possuem um considerável número de parálogos membros de famílias de TF que são comuns

em *A. mellifera* e *D. melanogaster*. Dois genes distintos em *Apis* são reportados nesta primeira análise como homólogos de *dsx*, e uma comparação entre todos os possíveis homólogos dos genomas dos dois insetos (mosca e abelha) é necessária para a predição do melhor candidato a ortólogo do gene *dsx*.

A cascata de determinação do sexo evoluiu da base para topo (WILKINS, 1995; MARíN; BAKER, 1998; SCHüTT; NöTHIGER, 2000) e de acordo com esta hipótese de evolução base-topo, os genes *dsx* e *ix* estão bem conservados e são considerados os genes terminais nesta via regulatória. A função destes dois genes tem sido reportada como conservada não só entre os insetos mas entre os metazoários (RAYMOND et al., 1998; SUZUKI et al., 2003; HEDIGER et al., 2004; SIEGAL; BAKER, 2005). O produto do gene *ix* (IX) pode formar um complexo com DSX-F que juntos teriam um efeito funcional similar ao do DSX-M sozinho, inibindo ou ativando genes-alvo de maneira específica em fêmeas (SIEGAL; BAKER, 2005). Um exemplo bem conhecido desta interação é a ligação do complexo DSX-F/IX à região reguladora dos genes *yolk-protein1* (*yp1*) e *yolk-protein2* (*yp2*) de *D. melanogaster* controlando a transcrição destes genes expressos especificamente em fêmeas (COSCHIGANO; WENSINK, 1993; GARRETT-ENGELE et al., 2002).

3.1.2 Relações filogenéticas entre as proteínas com domínios DM em *A. mellifera* e *D. melanogaster*

Um total de 8 seqüências de proteínas homólogas ao *dsx* foi encontrado em *A. mellifera* e *D. melanogaster* considerando somente aquelas seqüências com alto grau de conservação (E-value $< 1e-18$). As relações filogenéticas entre as 8 seqüência de proteínas alinhadas codificadas a partir de genes homólogos de *dsx* revelaram 4 grupos de parálogos, em concordância com estudos anteriores (OTTOLENGHI et al., 2002) (Figura 3.1, A). Todos os 8 genes comparados codificam proteínas com domínio DM (Pfam ID: PF00751; Figura 3.1, B, quadrado preto) na região N-terminal e apenas dois grupos de parálogos apresentam um outro domínio conservado na região C-terminal conhecido como domínio DMRTA (Pfam ID: PF03474, Figura 3.1, B, quadrado cinza). Existem três fortes evidências que levam a crer que GB18426-PA e *Dmdsx* são ortólogos e não GB15791-PA e *Dmdsx*. Em primeiro lugar, a seqüência de amino ácidos codificada pelo GB18426-PA e *Dmdsx* apresentam 51% de identidade, enquanto GB15791-PA

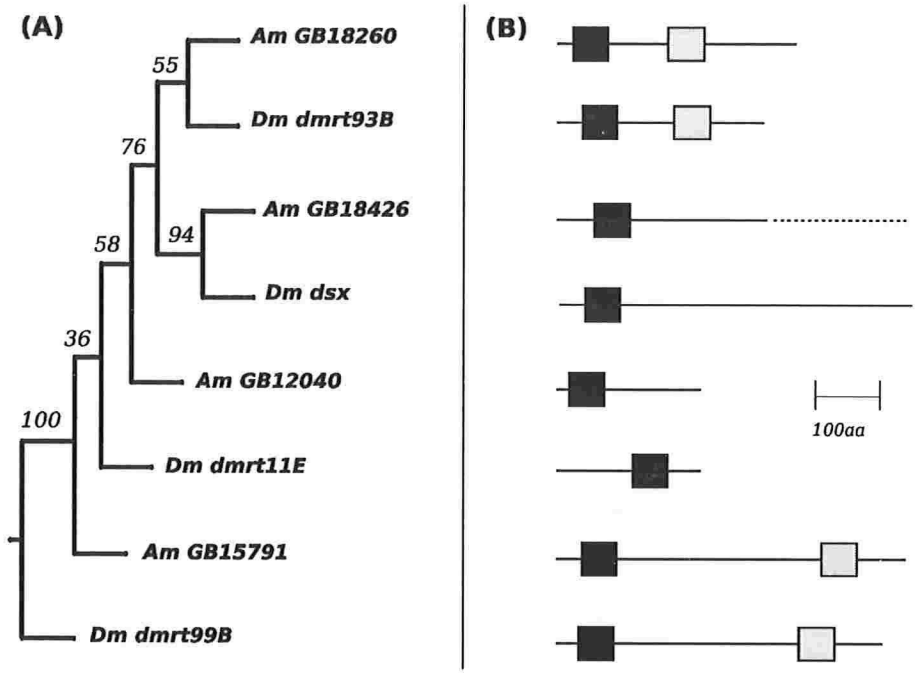


Figura 3.1: (A) Árvore enraizada no ponto médio construída pelo método de *neighbor-joining* mostrando as relação filogenéticas de todos os homólogos de *dsx* de *A. mellifera* (Am) e *D. melanogaster* (Dm). Os valores de *bootstrap* são apresentados em cada nó. (B) Diagrama ilustrando o comprimento e posição dos domínios protéicos. O quadrado preto representa o domínio DM e o cinza o domínio DMRTA. *Dmdmrt99B* (GenBankID: AAF56919); *Dmdmrt11E* (GenBankID: AAF48261); *Dmdmrt93B* (GenBankID: AAF55843); *Dmdsx* (GenBankID: AAF54169); linha pontilhada em GB18426-PA representa a região C-terminal sexo-específica, ainda não identificada.

e *Dmdsx* são 45% idênticas. Segundo, GB18426-PA e *Dmdsx* alinham com 34% de identidade em uma outra região, desta maneira mostrando duas região altamente conservadas e que não está presente em GB15791-PA (dado não apresentado na Tabela 3.1 por não superar o E-value estabelecido como corte). Finalmente, somente o domínio DM foi encontrado em GB18426-PA e *Dmdsx* na região N-terminal, enquanto dois domínios (DM e DMRTA) estão presentes em GB15791-PA sendo que DMRTA não está presente em *Dmdsx* (Figura 3.1, B).

Os mais prováveis ortólogos, GB18426-PA e *Dmdsx*, estão localizados nos cromossomos 5 e 3R, respectivamente. GB18260-PA e *Dmdmrt93B* são 44% idênticos apresentando os dois domínios (DM e DMRTA) e estão localizados nos cromossomos 8 e 3R, respectivamente (Figura 3.1, B). Os genes ortólogos mais conservados, GB12040-PA e *Dmdmrt11E*, apresentam 78% de identidade e estão situados nos cromossomos 5 e X, respectivamente (Figura 3.1, B). O

último grupo, GB15791-PA and *Dmdmrt99B*, são os menos conservados com 39% de identidade compartilhando os domínios DM e DMRTA e localizados nos cromossomos 1 e 3R (Figura 3.1, B). Nenhuma sintenia parece existir entre estes genes DM, nestes insetos.

As seqüências de proteínas dos genes ortólogos de *dsx* em outros sete insetos também foram comparados neste estudo. No total, nove seqüências de DSX de insetos foram alinhadas: *Bactrocera oleae* (GenBankID: CAD67987), *Anastrepha obliqua* (GenBankID: AAY25167), *Bombyx mori* (GenBankID: BAB13472), *D. melanogaster* (GenBankID: AAF54169), *Megaselia scalaris* (GenBankID: AAK38832), *Musca domestica* (GenBankID: AAR23813), *Anopheles gambiae* (GenBankID: AAX48939), *Ceratitis capitata* (GenBankID: AAN63597) e *A. mellifera* (GB18426-PA). Duas regiões bem conservadas são evidenciadas no alinhamento múltiplo (Figura 3.2), e são funcionalmente importantes para as proteínas DSX. A região N-terminal conservada inclui um padrão distinto (Zn-finger intercalado) do domínio de ligação ao DNA (DBD) que se liga à cavidade menor do DNA (*minor groove*) (Figura 3.2, DBD sítio I e II) (ERDMAN; BURTIS, 1993; ZHU et al., 2000), e um domínio não sexo-específico responsável pela formação dímeros de DSX que então se ligam aos sítios regulatórios de genes alvo (Figura 3.2, OD1). A região C-terminal corresponde ao domínio de oligomerização 2 (OD2) que ocorre parcialmente nas DSX-F (DSX de fêmea) e DSX-M (DSX de macho) apresenta um trecho sexo-específico (Figura 3.2, OD2). O domínio OD2 possui uma seqüência comum em ambos os sexos mas também um trecho sexo-específico que é importante na oligomerização via OD2.

3.1.3 Identificação, clonagem e seqüenciamento do cDNA parcial de *dsx* em *A. mellifera*

Um fragmento de cDNA de 443 pares de bases (pb) do *Amdsx* foi isolado a partir de uma amostra de RNA extraído de embriões fêmeas usando um par de iniciadores (AMDSX-F1 e AMDSX-R2) desenhados nas duas regiões mais conservadas do gene *dsx*, DBD/OD1 e OD2 (Figura 3.3, A e D. Esta seqüência parcial de *Amdsx* representa a região comum nos dois sexos composta de três éxons (depositada no GenBankID: AY375535). Um fragmento específico de macho de 560pb foi obtido a partir do RNA extraído de larvas de macho usando a combinação dos iniciadores AMDSX-F1 e AMDSX-R3 (Figura 3.3, A e D). O iniciador AMDSX-P3 foi desenhado com base no éxon transcrito apenas em machos (Figura 3.3, A e D). A seqüência

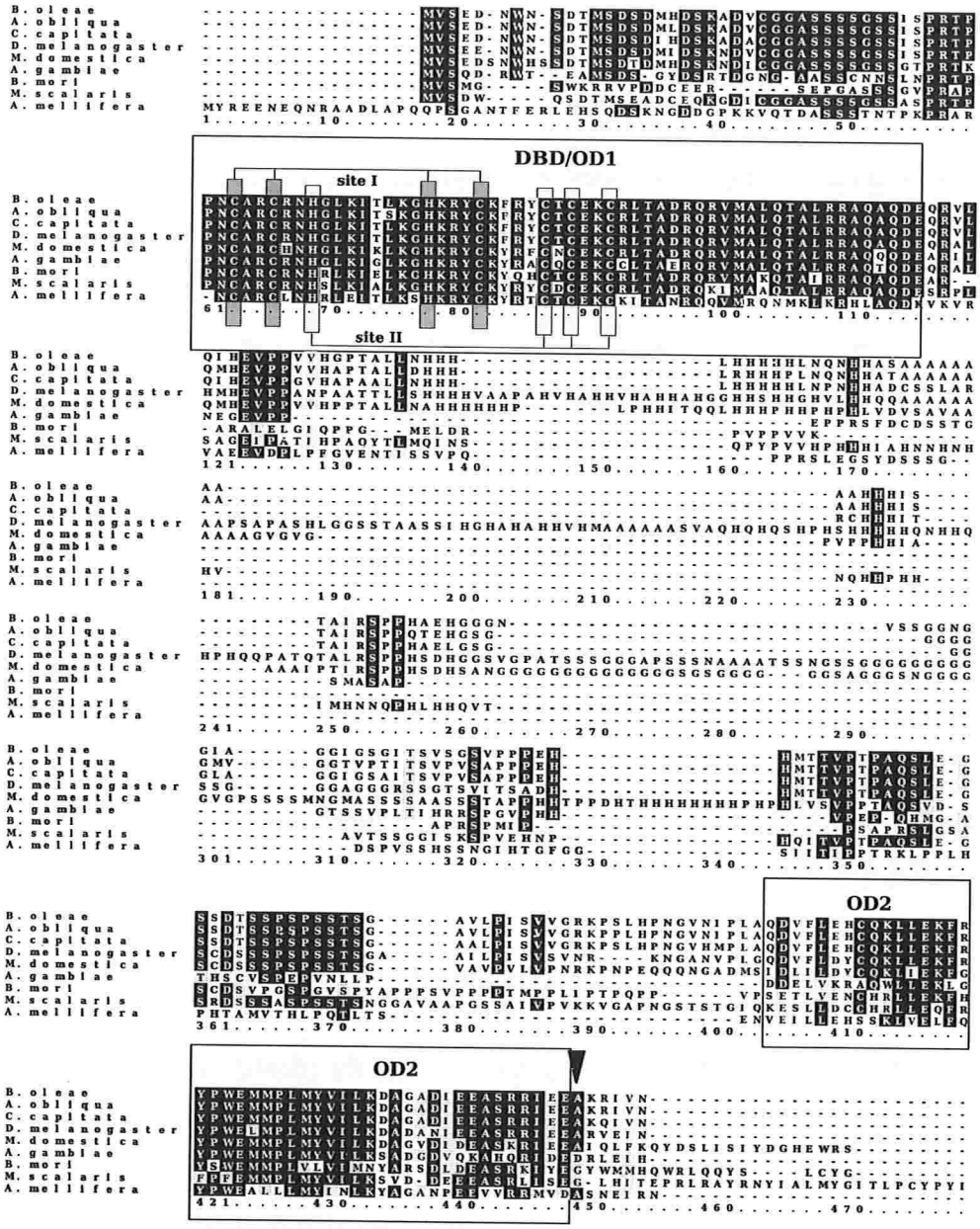


Figura 3.2: Alinhamento múltiplo das seqüências de proteínas DSX de macho de nove espécies de insetos, *B. oleae*, *A. obliqua*, *B. mori*, *D. melanogaster*, *M. scalaris*, *M. domestica*, *A. gambiae*, *C. capitata* e *A. mellifera*. A região N-terminal é bem conservada incluindo o domínio de ligação ao DNA (DBD com o sítio I em barras cinza escuro e o sítio II em barras cinza claro). Um domínio de oligomerização presente em seqüências de ambos os sexos (OD1). A região C-terminal conservada corresponde ao domínio de oligomerização 2 (OD2) que apresenta uma região comum em ambos os sexos e outra sexo-específico.

do *Amdsx* hipotético específica de macho foi alinhada contra todos os genes da família DM de *D. melanogaster* e apresentou maior similaridade com *Dmdsx*. Além disso, comparando esta seqüência parcial com os genes preditos em abelha, a maior similaridade encontrada foi com GB18426-PA (Figura 3.3, B), desta maneira confirmando que esta seqüência identificada foi transcrita a partir de *Amdsx*.

O gene *Amdsx* está localizado no *scaffold* do Group5.6 (equivalente ao cromossomo 5) e possui pelo menos três éxons comuns nos transcritos de ambos os sexos. O primeiro éxon contém o domínio DM (DBD/OD1) e no terceiro éxon está o domínio OD2 que possui tanto uma região comum nos sexos como sexo-específica (Figura 3.3, C). O quarto éxon foi obtido através de predição computacional e evidências experimentais indicam ser específico de macho. O padrão de *splicing* alternativo sexo-específico do *dsx* parece estar conservado entre os insetos mantendo até mesmo a presença de uma alanina (A) na primeiro posição do éxon 4, enquanto em fêmeas o resíduo mais comum na posição do éxon específico de fêmea é a glicina (G) (Figura 3.2, seta preta; somente o alinhamento das seqüências de machos é apresentado, o segmento sexo-específico ainda é desconhecido em abelha). Embora, outros autores já tenham mencionado uma possível existência de um ortólogo do *dsx* em *A. mellifera*, nenhuma evidência experimental foi apresentada até o momento (BEYE et al., 2003).

3.1.4 Identificação de um fragmento de cDNA de *ix* em *A. mellifera*

Um único fragmento de cDNA de 153 pb do *Amix* foi identificado por RT-PCR a partir de amostras de RNA de embriões de macho e fêmea de 24h, usando a combinação de um par de iniciadores (AMIX-F1 e AMIX-R2) desenhados nas regiões intermediárias do gene (Figura 3.4, A-B). A estrutura do *Amix* é bem simples com apenas um éxon com 570 pb e codifica uma proteína com 190 aminoácidos (Figura 3.4, B-C). Nenhuma outra duplicação no genoma de abelha foi encontrada. O gene *Amix* está localizado no *scaffold* do Group.Un735 mas o cromossomo em que se encontra o gene ainda não foi identificado (Figura 3.4, A).

O *Amix* é expresso constitutivamente nos dois sexos, não havendo evidências de isoformas específicas em macho e fêmea, nem mesmo um domínio de ligação ao DNA, atuando como cofator de transcrição juntamente com DSX-F, mas não com DSX-M (GARRETT-ENGELE et al., 2002). O complexo IX/DSX-F é necessário para garantir a diferenciação de fêmea em *D.*

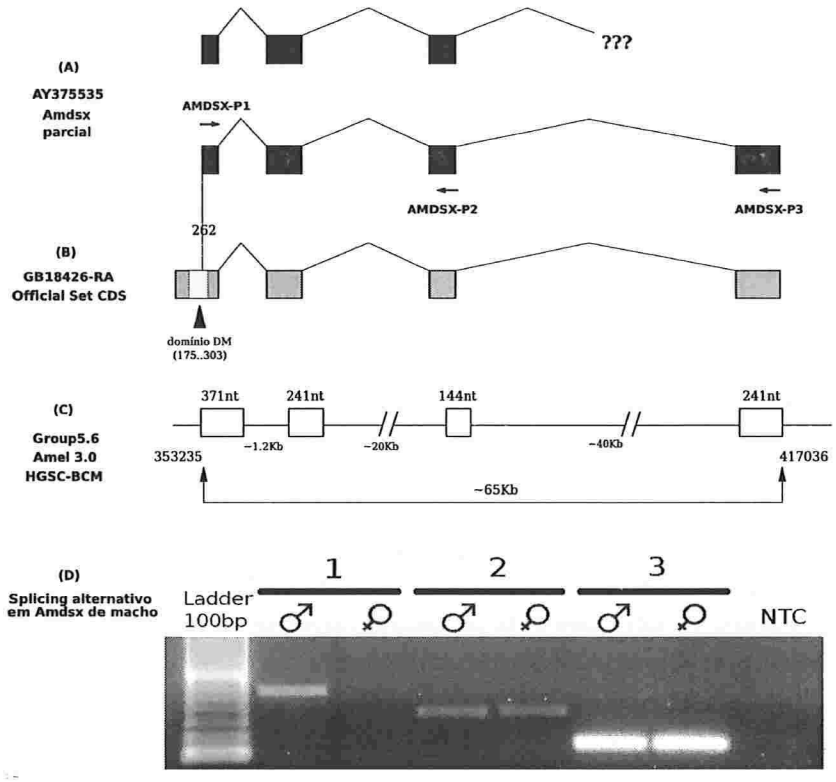


Figura 3.3: Diagrama ilustrando a estrutura do gene *Amdsx* e as evidências experimentais de *splicing* alternativo. (A) Fragmentos de cDNA de *Amdsx* comuns aos dois sexos foram isolados, clonados e seqüenciados a partir de amostras de embriões de abelha (GenBankID: AY375535). A presença dos éxons específicos de fêmeas ainda é desconhecida, entretanto, um fragmento de 560pb transcrito especificamente em macho é identificado como GB18426-RA, com base na seqüência de nucleotídeos predita (HGSC-BCM). (B) GB18426-RA apresenta o éxon 4 que está presente somente em transcritos de machos. (C) O *Amdsx* está situado no *scaffold* do Group5.6 (cromossomo 5) versão 3.0 do genoma de *A. mellifera* (HGSC-BCM). (D) Gel de agarose (1%) revelando fragmentos de cDNA amplificados por RT-PCR em amostras dos dois sexos por a partir de 3 combinações de iniciadores: (1) AMDSX-F1 e AMDSX-R3 revela um fragmento encontrado somente em macho; (2) AMDSX-F1 e AMDSX-R2 é o controle positivo que demonstra a presença de transcritos de *Amdsx* em ambos os sexos; (3) fragmentos de EF-alpha1 como controle positivo de um gene constitutivo. NTC = controle negativo; CDS = seqüência codificadora; HGSC-BCM = Human Genome Sequencing Center - Baylor College of Medicine.

melanogaster (SIEGAL; BAKER, 2005). O alto grau de conservação destes genes (*ix* e *dsx*) em insetos e as evidências de que estes genes também estão presentes em *A. mellifera* com várias propriedades em comum, nos possibilita investigar como estes genes estão se expressando durante o desenvolvimento de abelha e se eles estão desempenhando funções similares na

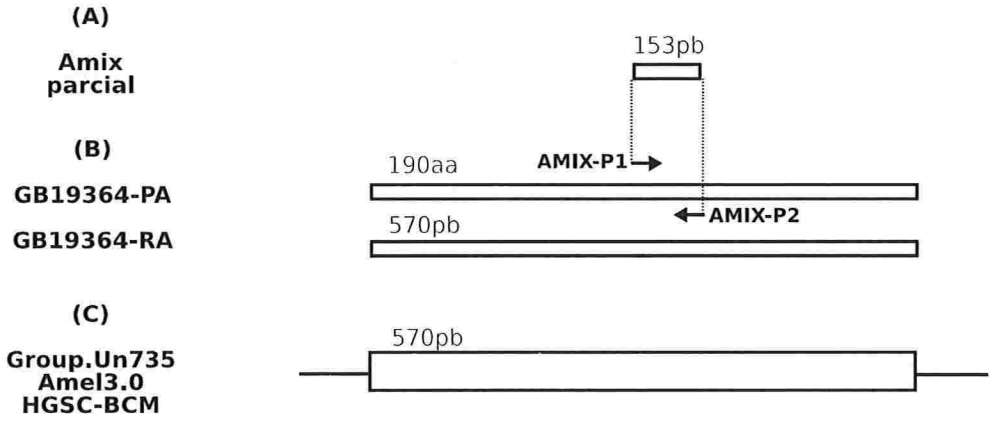


Figura 3.4: Diagrama ilustrando a estrutura do gene *Amix* e as evidências experimentais de transcrição em embriões 24h, de macho e fêmea. (A) Fragmento de 153pb de cDNA de *Amix* isolados por PCR em amostras de embriões 24h de macho e fêmea de abelha. (B) Diagrama das regiões codificadoras de *Amix* em aminoácidos (GB19364-PA, 190aa) e nucleotídeos (GB19364-RA, 570pb). (C) O *Amix* está situado no *scaffold* do Group.Un735 (cromossomo não identificado) versão 3.0 do genoma de *A. mellifera* (HGSC-BCM). NTC = controle negativo; CDS = seqüência codificadora; HGSC-BCM = Human Genome Sequencing Center - Baylor College of Medicine.

regulação de genes alvo, tal como na regulação do gene *yp-1* de *Drosophila* e *vg* de outros insetos (ex: *B. mori* e *M. domestica*) que embora não sejam homólogos parecem apresentar expressão tecido- e sexo-específicas regulada principalmente por DSX-F (SUZUKI et al., 2003; HEDIGER et al., 2004).

3.1.5 Quantificação relativa do perfil de transcrição de *vg*, *dsx* e *ix* em embriões e tecidos de *A. mellifera*

Estudar a transcrição de *Amvg*, *Amdsx* e *Amix* nas fases iniciais do desenvolvimento embrionário fornece informações úteis que podem ajudar a compreender algumas diferenças dos fenótipos sexuais, que são determinadas geneticamente ainda no início do desenvolvimento. Embora a Vg tenha, tradicionalmente, função mais bem conhecida durante a oogênese (WYATT, 1999; HARTFELDER; ENGELS, 1998; BARCHUK; BITONDI; SIMõES, 2002; PIULACHS et al., 2003), outras funções bem diversificadas têm sido atribuídas a Vg de *A. mellifera* (AmVg). A AmVg é expressa em ovários de rainhas mas não no de operárias e também é expressa durante o período larval em rainhas, operárias e zangões (GUIDUGLI et al., 2005b).

Outros estudos com AmVg indicam que ela também desempenha funções metabólicas tais como na síntese do alimento larval produzido pelas operárias, no transporte de zinco, no sistema imune, na senescência e até mesmo na regulação da dinâmica hormonal (AMDAM; OMHOLT, 2002; AMDAM et al., 2003, 2004; GUIDUGLI et al., 2005a). Assim, frente a tanta diversidade funcional pode-se esperar que AmVg tenha algum padrão de expressão atípico na fase embrionária que possa ser importante para o desenvolvimento destes insetos.

Considerando esta potencial diversidade funcional de AmVg, investigamos o perfil de transcrição de *Amvg*, *Amdsx* e *Amix* em fases embrionárias e tecidos de fêmea e macho de *A. mellifera*, por métodos de quantificação relativa a partir de análises da cinética de amplificação por PCR em tempo real (Figura 3.5). *Amvg* não é transcrito em nenhum momento do desenvolvimento embrionário em ambos os sexos e portanto não é possível estabelecer qualquer relação com os genes terminais de determinação de sexo (*Amdsx* e *Amix*) neste período. Entretanto, os fatores de transcrição (DSX e IX) codificados por estes genes terminais mostraram dinâmicas de transcrição bem distintas entre os sexos mas com semelhanças muito interessantes entre as dinâmicas de transcrição dos dois genes em cada sexo, principalmente em fêmeas (Figura 3.5, A).

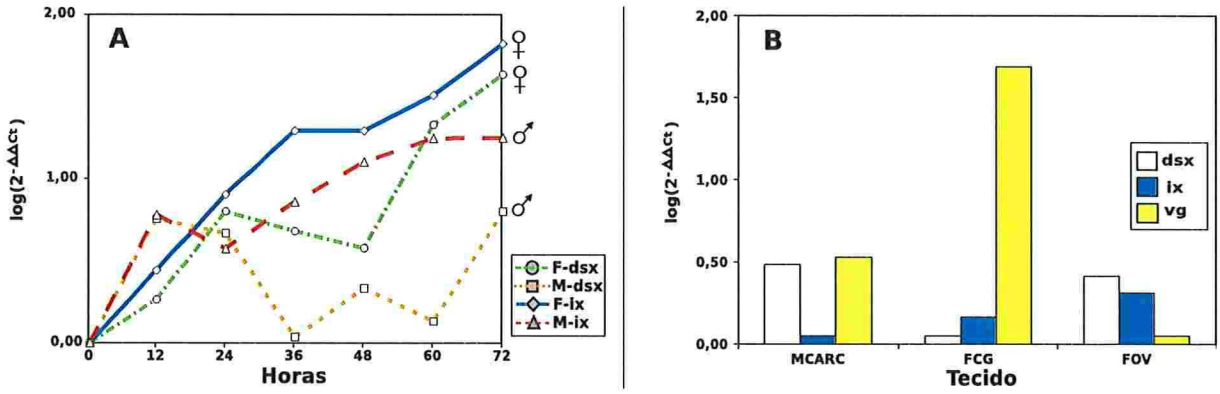


Figura 3.5: Perfil de transcrição de *Amvg*, *Amdsx* e *Amix* em embriões e diferentes tecidos de macho e fêmea de *A. mellifera*. (A) A transcrição de *Amdsx* e *Amix* em embriões de macho (M-dsx e M-ix) e fêmea (F-dsx e F-ix) são claramente diferentes tanto em qualidade como em quantidade. *Amvg* não é transcrito em nenhuma fase do desenvolvimento embrionário em ambos os sexos. (B) A transcrição de *Amvg*, *Amdsx* e *Amix* não caracteriza nenhuma evidência de que *Amvg* seja regulado por DSX e IX como ocorre em outros insetos. MCARC, carcaça abdominal (tegumento + corpo gorduroso) de zangões recém-nascidos; FCG, corpo gorduroso de rainhas recém-nascidas; FOV, ovários de rainhas recém-nascidas.

A transcrição de *Amdsx* e *Amix* é diferente entre macho e fêmea tanto em qualidade como em quantidade. Nas primeiras 24hs do desenvolvimento dos embriões de fêmea e macho, o *Amdsx* e *Amix* mostram uma expressão muito similar em cada sexo mas diferente entre os sexos (Figura 3.5, A). Após 24hs, as diferenças nos níveis de transcrição se acentuam principalmente em macho enquanto em fêmea, estes dois genes demonstram uma correlação maior ($R^2 = 0,96$ entre F-dsx e F-ix; Tabela 3.2) do que o mesmo gene em cada sexo ($R^2 = 0,44$ entre F-dsx e M-dsx; $R^2 = 0,33$ entre F-ix e M-ix; Tabela 3.2). Tal observação nos permite assumir a hipótese de que *Amdsx* e *Amix* são co-regulados em fêmeas. O perfil qualitativo da transcrição de *Amdsx* entre macho e fêmea é muito diferente e apresenta uma dinâmica oposta em boa parte do período de desenvolvimento do embrião, isto é, quando a transcrição de *Amdsx* aumenta em um sexo, diminui no outro (F-dsx e M-dsx; Figura 3.5, A).

Além disso, os níveis de transcritos de *Amdsx* e *Amix* em fêmeas são nitidamente maiores do que em machos principalmente nas fases finais da embriogênese, quando ocorre a organogênese (Figura 3.5, A). Em rainhas, a diferença , ao longo das 72h, pode chegar a 42 e 65 vezes em quantidade de transcritos de DSX e IX, respectivamente (Figura 3.5, A). Em zangões, esta diferença é bem menor chegando a 5 e 16 vezes, respectivamente (Figura 3.5, A). Estas diferenças na quantidade de transcritos destes dois genes em machos e fêmeas indicam uma possível ausência de um mecanismo de compensação de dose em *A. mellifera* em que os machos haplóides (n) transcrevem menos RNA que as fêmeas diplóides (2n).

Tabela 3.2: Correlação entre os perfis de transcrição de *Amdsx* e *Amix* durante o desenvolvimento embrionário de machos e fêmeas

Correlação	F-dsx	M-dsx	F-ix	M-ix
F-dsx	1			
M-dsx	0,44	1		
F-ix	0,96	0,33	1	
M-ix	0,78	0,15	0,85	1

Nas análises de tecidos, os perfis de transcrição de *Amvg*, *Amdsx* e *Amix* em carcaça de zangões recém-nascidos (MARC) e no corpo gorduroso e ovários de rainhas recém-nascidas (FCG e FOV, respectivamente) resultaram em algumas observações interessantes (Figura 3.5, B). O gene *Amvg* é transcrito em ambos os sexos e portanto reforça o fato de Vg não ser uma proteína sexo-específica como em alguns insetos, entretanto tal observação já havia sido

feita em estudos anteriores (PIULACHS et al., 2003; GUIDUGLI et al., 2005a, 2005b). Em rainhas recém-nascidas, a transcrição de *Amvg* é muito maior em CG do que em Ov ($\approx$ 48 vezes maior) em concordância com estudos anteriores (GUIDUGLI et al., 2005b). No entanto, a maior contribuição deste estudo não diz respeito ao perfil de transcrição de *Amvg*, mas da análise conjunta de *Amvg*, *Amdsx* e *Amix*. (Figura 3.5, B). Em *D. melanogaster*, a expressão de *yp-1* ocorre em grande quantidade no CG, como em abelhas, mas nenhuma expressão é detectada em Ov. O gene *Amdsx* apresentou um perfil oposto ao observado em *D. melanogaster*, onde *dsx* é transcrito em CG mas não em Ov. Em abelhas, *Amdsx* é levemente mais transcrito em Ov do que em CG de rainhas ($\approx 1,5$ vezes; Figura 3.5, B).

3.1.6 Análise computacional das regiões reguladoras dos genes envolvidos na determinação do sexo em *D. melanogaster* e *A. mellifera*

As análises das regiões reguladoras de genes envolvidos nos mesmos processos biológicos em espécies diferentes acrescentam informações importantes que ajudam na compreensão da história evolutiva destes genes bem como das redes gênicas de regulação. Estudar de maneira comparativa e integrada as regiões codificadoras e intergênicas adjacentes (regiões reguladoras cis) destes genes pode revelar propriedades e padrões funcionais que se complementam criando um cenário mais realista da evolução das redes regulatórias envolvidas no desenvolvimento de padrões fenotípicos homólogos, tal como os fenótipos sexuais nos metazoários.

De acordo com estudos anteriores sobre as funções biológicas dos ortólogos *dsx* em *D. melanogaster* e *B. mori*, os genes análogos *yp-1* e *Bmvg* (*vitelogenina* de *B. mori*) são transcritos de forma específica nos tecidos e sexos, com a participação de DSX (AN; WENSINK, 1995b; GARRETT-ENGELE et al., 2002; YANO et al., 1994b; SUZUKI et al., 2003). As hexamerinas de *B. mori* e *O. atropalpus* também apresentam expressão tecido- e sexo-específicas envolvendo a participação de DSX. (SUZUKI et al., 2003; JINWAL et al., 2006). A presença de um módulo regulatório semelhante ao FBE tem sido observada não só em *yp-1* de *D. melanogaster*, mas em *hex-1.2* de *O. atropalpus*. Em *Bmvg*, embora nenhum módulo tenha sido descrito, uma região 200pb à montante do gene possui sítio de ligação para DSX. Uma busca por sítios de ligação putativos de DSX, AEF-1 e BZIP-1 nas regiões promotoras (1000pb) à montante dos genes *yp-1*, *Bmvg* e *Amvg* revelou a ocorrência de sítios já descritos em estudos anteriores bem como sítios

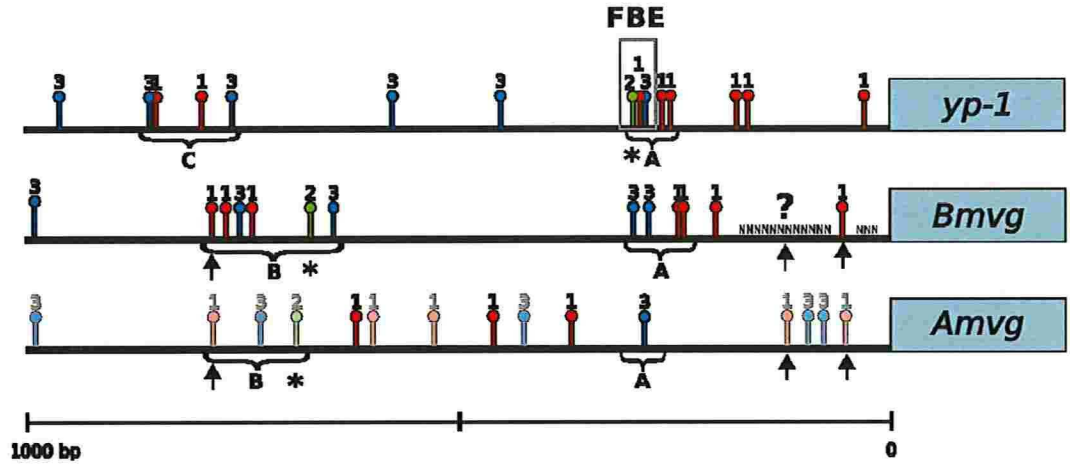


Figura 3.6: Diagrama da ocorrência de sítios regulatórios putativos nas regiões promotoras de *yp-1*, *Bmvg* e *Amvg*. Três módulos principais são considerados neste estudo, A, B e C. O módulo A está mais conservado entre os genes análogos *yp-1* e *Bmvg*, mas este último não apresentou o sítio 2 (aef1). Em *Amvg*, o módulo A é muito pouco conservado e só apresenta um sítio 3 (bzip1). O módulo B possui algumas características conservadas como a ocorrência dos sítios 1, 2 e 3 (dsx, aef1, e bzip1, respectivamente) em uma mesma região entre os genes homólogos *Bmvg* e *Amvg*, mas sem nenhuma colinearidade. O módulo C é principalmente caracterizado pela ocorrência de pelo menos 2 sítios dsx e 2 bzip1 sobrepostos ou muito próximos. Em *Amvg* os sítios representados em segundo plano só foram encontrados quando o limiar de similaridade foi diminuído para 50%, todos os outros têm similaridade de pelo menos 60%. Sítio 1, dsx; Sítio 2, aef1; Sítio 3, bzip1; (FBE) *Fat Body Enhancer*; (*) ocorrência do sítio aef1 para ligação do repressor AEF-1; (N) seqüência genômica não determinada; (?) sítio dsx caracterizado em *B. mori*; (↑) sítios dsx conservados em posição, entre *B. mori* e *A. mellifera*

putativos de DSX, AEF-1 e BZIP-1 (sítios 1 [dsx], 2 [aef1] e 3 [bzip1], respectivamente; Figura 3.6).

O gene *Bmvg* compartilha maior similaridade com *yp-1* dentro do módulo A considerando a ocorrência dos sítios dsx e bzip1, mas não possui algo semelhante com o FBE (Figura 3.6). Os homólogos *Amvg* e *Bmvg* possuem alguns sítios conservados próximos ao início da região codificadora e dentro do módulo A (bzip1; Figura 3.6) e B (dsx e aef1; Figura 3.6). É interessante notar que embora o módulo A de *yp-1* e *Bmvg* tenha uma semelhança nas múltiplas ocorrências de sítios dsx e bzip1, o sítio aef1 de *Bmvg* acontece no módulo B onde também ocorre este mesmo sítio em *Amvg*. A conservação mais evidente entre os três genes está no módulo A com o ocorrência do sítio bzip1 a aproximadamente 300pb do início da região codificadora (Figura 3.6).

A partir dos estudos do controle de transcrição de *yp-1* em *D. melanogaster*, sabe-se que os sítios que compõem o FBE (dsxa, aef1, bzip1) não são os únicos responsáveis por tal controle de transcrição sendo este dependente de contexto, isto é, depende de vários outros fatores (ex: títulos de ecdisteróides e hormônio juvenil) que determinam os níveis em que *yp-1* é transcrito em determinados órgãos, nos sexos e em estágios do desenvolvimento. O transplante de ovários para machos não inibe a expressão de YP-1 que continua nas células do folículo ovariano (BOWNES, 1994). Além disso, a regulação sexo-específica é diferente em CG e Ov. Entretanto, algumas propriedades deste sistema parecem bem resolvidas. O sítio dsxa é irrelevante em Ov, mas essencial para a ativação sexo-específica em CG. O sítio aef1 não é funcional em CG mas inibi a expressão em Ov, atuando de forma oposta ao dsxa. Embora DSX-F não seja transcrito em Ov de *Drosophila*, sua atividade é muito importante para a formação deste e deve atuar de forma indireta (AN; WENSINK, 1995a; BAKER; RIDGE, 1980). O sítio bzip1 é suficiente para promover um forte ativação de expressão ovariana quando a repressão em aef1 é removida (AN; WENSINK, 1995a).

No desenvolvimento embrionário de *A. mellifera*, o perfil de transcrição de *Amdsx* e *Amix* de cada sexo (de 0 à 24hs; Figura 3.5, A) indica haver uma forte correlação na dinâmica de transcrição destes dois genes em fêmeas, e portanto, nos permite assumir a hipótese de serem co-regulados neste período. A Análise das regiões promotoras dos genes *dsx* e *ix* de *D. melanogaster* e *A. mellifera* foi conduzida com o programa PhyloCon (WANG; STORMO, 2003), que leva em consideração a conservação entre genes ortólogos e co-regulação entre os genes dentro da espécie. Um motivo regulatório putativo bem conservado foi localizado na região promotora dos 2 grupos de ortólogos utilizados na análise. O motivo foi representando por uma matriz PWM (consenso: CrCyYsRGCT) e então alinhado contra todos os sítios de ligação de *D. melanogaster* descritas na base TRANSFAC (v4.0) apresentando uma similaridade de 70% com o sítio de ligação do fator GAGA codificado pelo gene *Trithorax-like* (*Trl*; Figura 3.7, sítio 4 [trl]).

O motivo encontrado foi denominado, GAGA-DIX (DIX = contração de DSX e IX). O ortólogo de *Trl* foi encontrado em abelha (GB14696) e codifica em sua região N-terminal um motivo de ligação ao DNA muito conservado (BTB/POZ; 76% de identidade ao *Trl* de *Drosophila*). Tal observação reforça a possibilidade de que este motivo regulatório de ligação do fator GAGA esteja conservado entre as duas espécies de insetos. O fator GAGA pertence a um grupo de genes (*Trithorax*) necessários para o controle de expressão de genes homeóticos durante

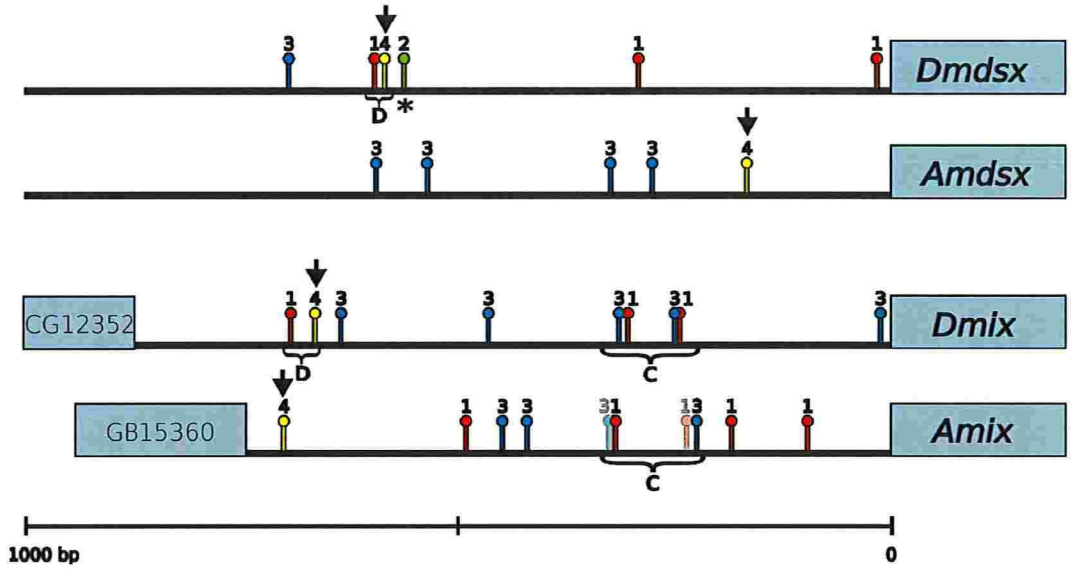


Figura 3.7: Diagrama da ocorrência de sítios putativos de ligação de DSX, AEF-1, BZIP-1 e GAGA nas regiões promotoras de *Dmdsx*, *Amdsx*, *Dmix* e *Amix*. Os sítios 1, 2, 3 e 4 representam dsx, aef1, bzip1 e gaga, respectivamente. O módulo C combina uma ocorrência redundante de dsx e bzip1, muito próximos ou sobrepostos um do outro. Este módulo está presente em uma região muito semelhante entre os ortólogos *Dmix* e *Amix*. A ocorrência de um sítio dsx e gaga adjacentes foi denominado módulo D. Sítios de ligação para o fator GAGA (sítio 4) foram encontrados nos 2 grupos de ortólogos (*dsx* e *ix*) de *D. melanogaster* e *A. mellifera*. Em *Amvg*, os sítios representados em segundo plano só foram encontrados quando o limiar de similaridade foi diminuído para 50%, todos os outros têm similaridade de pelo menos 60%. sítio 1, dsx; sítio 2, aef1; sítio 3, bzip1; sítio 4, gaga; (*) ocorrência do sítio aef1 para ligação do repressor AEF-1; (↓) sítios gaga encontrados pelo PhyloCon.

a fase embrionária (BEJARANO; BUSTURIA, 2004). A presença de sítios de ligação do fator GAGA na região reguladora à montante destes genes de desenvolvimento é observada entre os regulados por este fator (BHAT et al., 1996). Sua principal função é a remodelagem da cromatina permitindo o acesso de fatores de transcrição ao DNA, necessários para ativação da transcrição gênica; Além disso, outras funções globais incluem o controle da compensação de dose em machos de *D. melanogaster* e do ciclo celular mitótico durante a blastoderme sincicial (FARKAS et al., 1994; BHAT et al., 1996; GREENBERG; YANOWITZ; SCHEDL, 2004; BEJARANO; BUSTURIA, 2004).

Os padrões de ocorrências dos motivos nas regiões promotoras dos 2 grupos de ortólogos podem nos ajudar a explicar os perfis observados em *A. mellifera* e *D. melanogaster*. No *Dmdsx*, o sítio aef1 (Figura 3.7) pode ser o principal responsável pela inibição da expressão de DSX-F

em Ov, já que este órgão apresenta altos níveis do fator AEF-1. O módulo D é composto de um sítio dsx e gaga e de maneira interessante acontecem ao lado do repressor aef1 podendo controlar a expressão do próprio *Dmdsx* de maneira tecido e sexo-específicas (Figura 3.7). No *Amdsx*, não houve nenhuma ocorrência do sítio dsx ou aef1 e o sítio gaga fica localizado em uma região muito distinta do seu ortólogo *Dmdsx*. Por outro lado, a quantidade de sítios bzip1 é muito maior quando comparado com seu ortólogo da mosca-de-fruta (Figura 3.7).

Entre os ortólogos *Amix* e *Dmix*, observou-se uma maior similaridade entre suas regiões promotoras. O módulo C (também ocorre no *yp-1*; Figura 3.6) é composto de sítios dsx e bzip1 distribuídos com relativa semelhança, e dada a proximidade dos sítios dsx e bzip1, pode ser uma região importante no controle sexo-específico do *ix* (Figura 3.7). Esta hipótese é reforçada pelo fato de *ix* apresentar níveis de transcrição mais elevados em fêmeas do que machos (Figura 3.5, A); Além disso, este gene é mais expresso em Ov do que em outros órgãos. Este fenômeno parece ocorrer tanto em *A. mellifera* (Figura 3.5, B) como em *D. melanogaster* (GARRETT-ENGELE et al., 2002). O módulo D aparece em *Dmix* assim como em *Dmdsx*, mas não é observado em seu ortólogo *Amix* (Figura 3.7 e 3.6). No entanto, o sítio gaga está presente em uma região próxima entre os ortólogos (Figura 3.7).

Com base nos estudos de *D. melanogaster* e nos resultados dos padrões de transcrição de *Amdsx*, *Amix* e *Amvg* aqui apresentados, assumimos que a proximidade ou sobreposição dos sítios dsx e bzip1 é importante para o controle sexo-específico da transcrição gênica através da formação do complexo sinérgico DSX-F/BZIP-1/IX responsável por um controle de transcrição ativador em fêmeas mas repressor em machos já que nestes o DSX-M não parece formar complexo algum com BZIP-1 ou IX. A presença de elementos cis similares a estes motivos nas adjacências dos genes pode significar que estes estejam sob o controle destes fatores de transcrição. No entanto, muitos destes elementos cis encontrados nas regiões intergênicas são espúrios e ocorreram ao acaso sendo boa parte falso-positivo. Identificar quais ocorrências são realmente importantes é um desafio ainda sem uma solução computacional razoável. Apenas evidências experimentais podem garantir a importância biológica que um elemento cis desempenha e mesmo assim a complexidade de interações entre os fatores de transcrição e cofatores é dependente de contexto (ex: sexo, tecido e estágio do desenvolvimento) e determinam quão importante é um sítio ou módulo regulatório no controle da expressão de um gene. Entretanto, algumas análises computacionais e estatísticas simples podem nos ajudar a

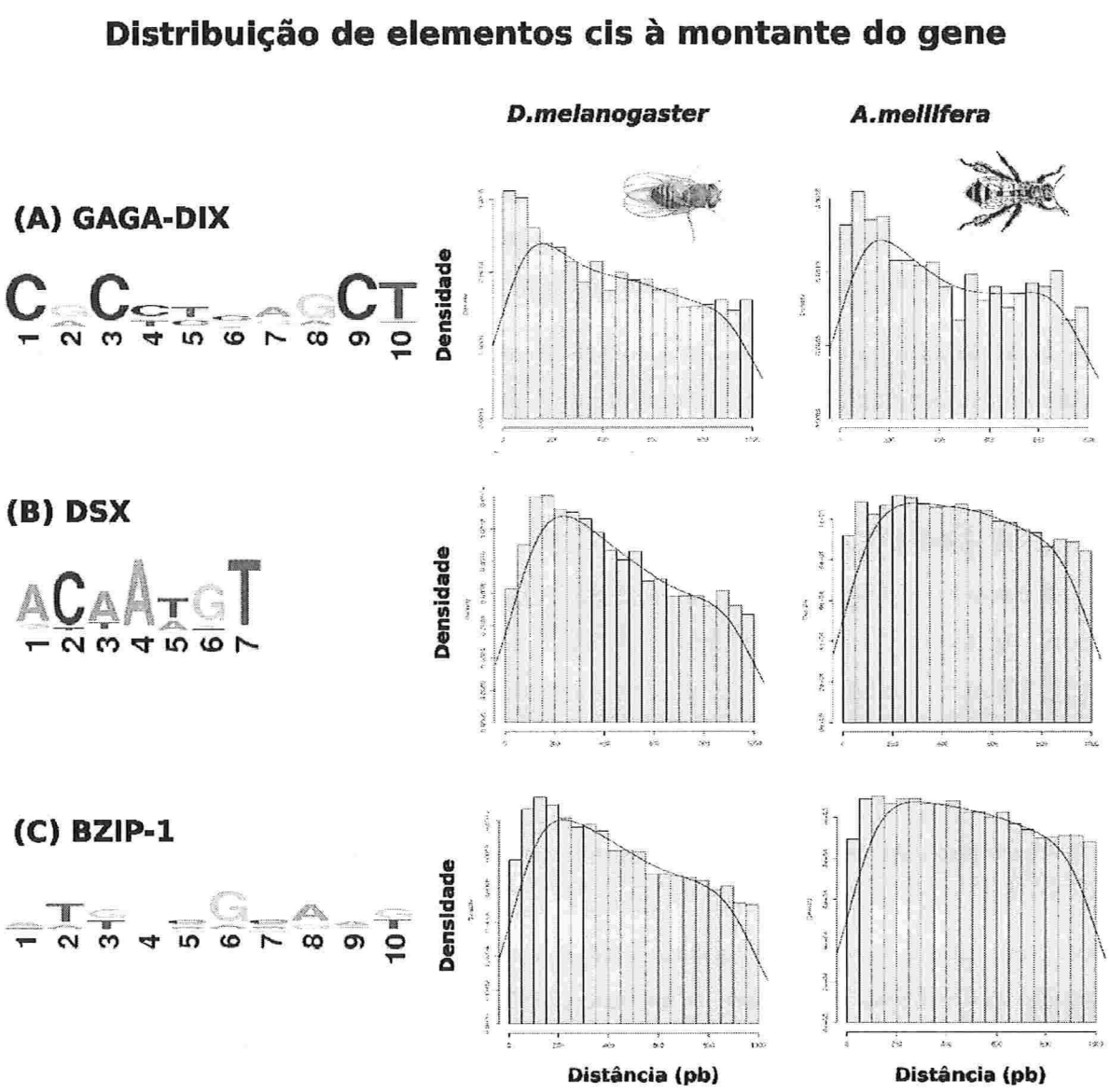


Figura 3.8: Distribuição dos motivos GAGA-DIX, DSX e BZIP-1 na região promotora de todos os genes de *A. mellifera* e *D. melanogaster*. (A) O motivo GAGA-DIX não apresenta uma distribuição uniforme e ocorre com maior freqüência até 500pb da região codificadora do gene. (B) O motivo DSX é mais freqüente entre 100 e 600pb em *D. melanogaster*. Em *A. mellifera*, a distribuição é mais uniforme, mas ainda assim ocorre com maior freqüência em direção ao início da região codificadora. (C) O motivo BZIP-1 apresenta uma distribuição muito semelhante ao DSX em cada uma das duas espécies de insetos.

avaliar se tais motivos têm algum tipo de específico de distribuição nas regiões promotoras e que não corresponda a uma distribuição uniforme esperada pelo acaso (Figura 3.8 e Tabela 3.3).

Tabela 3.3: Análise de Kolmogorov-Smirnov testando a significância estatística da distribuição dos motivos regulatórios canônicos (GAGA-DIX, DSX, BZIP-1) e seus respectivos motivos embaralhados (*shuffled*) contra uma distribuição uniforme.

D. melanogaster	Canônico vs Uniforme		*Shuffled* vs Uniforme	
Motivo	D	P-valor	D	P-valor
GAGA-DIX	0,128	1,5e-07	0,049	0,18
DSX	0,094	2,2e-16	0,012	0,87
BZIP-1	0,093	2,2e-16	0,026	0,06
A. mellifera	Canônico vs Uniforme		*Shuffled* vs Uniforme	
Motivo	D	P-valor	D	P-valor
GAGA-DIX	0,107	2,1e-05	0,043	0,31
DSX	0,032	0,0112	0,016	0,51
BZIP-1	0,053	1,6e-06	0,040	0,0006

A distribuição de ocorrência dos três motivos (GAGA-DIX, DSX, BZIP-1) considerados mais relevantes para controle de transcrição relacionados ao desenvolvimento dos fenótipos sexuais foi avaliada estatisticamente pelo teste de Kolmogorov-Smirnov (Tabela 3.3). Cada motivo foi alinhado com as regiões promotoras de todos os genes de *D. melanogaster* e *A. mellifera* resultando nas distribuições da Figura 3.8. O viés de ocorrência destes motivos na região promotora até 600pb foi significativo (Tabela 3.3). O motivo BZIP-1 foi o único que mesmo após o embaralhamento ainda mostrou viés de posição, entretanto isto pode estar relacionado com o baixo conteúdo de informação deste motivo. Nesta região mais próxima ao início de tradução do gene é onde se encontram boa parte dos elementos cis com maior influência na regulação da transcrição gênica (DAVIDSON, 2001). Estes resultados nos dão suporte para investigar quais genes possuem estes elementos cis em sua região promotora e quais processos biológicos estes fatores de transcrição podem regular.

Os motivos GAGA-DIX, DSX e BZIP-1 foram alinhados com todos os promotores dos genes de *D. melanogaster* e *A. mellifera*. Somente os genes que apresentaram pelo menos uma ocorrência do sítio gaga e dos sítios dsx e bzip1 sobrepostos em ambos os ortólogos destas espécies foram utilizados nas análises de perfil funcional. Um total de 355 genes apresentou sítios gaga e 215, os sítios sobrepostos dsx e bzip1. Estes foram identificados por seus IDs de acordo com a base de dados FlyBase e suas respectivas anotações funcionais recuperadas da base GO (consideramos as anotações no nível 5). Os histogramas das distribuições dos principais processos biológicos potencialmente regulados pelo fator GAGA e pelo complexo DSX/BZIP-1

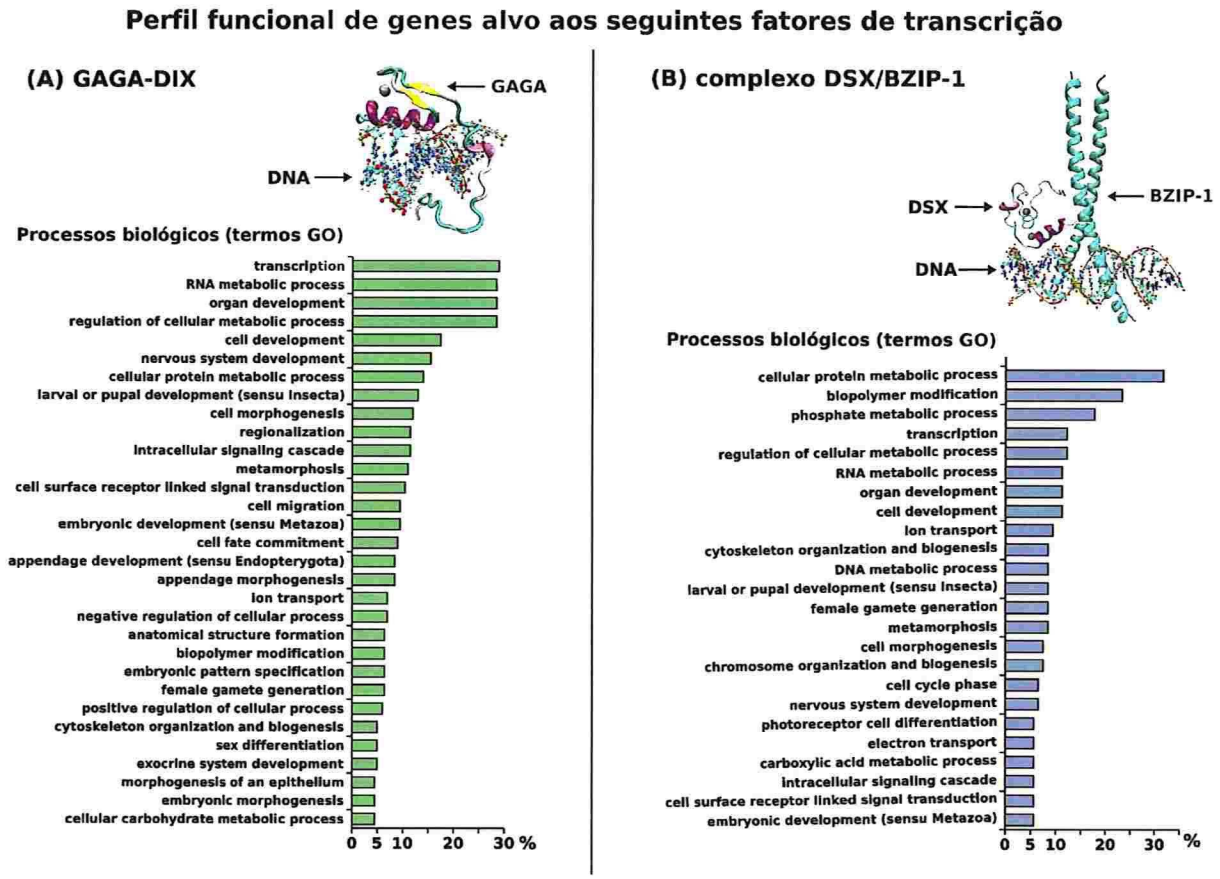


Figura 3.9: Histogramas dos processos biológicos mais representados entre os genes potencialmente regulados pelo fator GAGA e pelo complexo DSX/BZIP-1. (A) Processos biológicos predominantes entre os genes putativamente regulados pelo fator GAGA. (B) Processos biológicos predominantes entre os genes putativamente regulados pelo complexo DSX/BZIP-1.

estão apresentados na Figura 3.9 (A) e (B), respectivamente.

Dentre os genes putativamente regulados pelo fator GAGA, podemos destacar os seguintes processos biológicos predominantes: transcrição (28,9%), processo de metabolismo do RNA (28,4%), desenvolvimento de órgãos (28,4%), regulação do processo metabólico celular. Todos os outros 27 processos estão relacionados com o desenvolvimento dos padrões morfológicos de órgãos e estruturas do organismo bem como com os mecanismos celulares importantes na embriogênese e comunicação celular (Figura 3.9, A). Além disso, dois processos estão claramente relacionados com a determinação do sexo, geração de gameta feminino (6,4%) e diferenciação do sexo (5%). Dentre os genes potencialmente regulados pelo complexo DSX/BZIP-1 (209 com

anotações GO dos 215 inicialmente identificados), os processos biológicos mais representados são: processo metabólico celular de proteína (31,8%), modificação de biopolímeros (23,3%) e processo metabólico do fosfato (17,7%). Os outros 21 processos estão relacionados com embriogênese e organogênese além de mecanismos celulares importantes na formação e manutenção dos organismos multicelulares (Figura 3.9, B).

Um modelo simplificado do mecanismo genético de determinação do sexo em *A. mellifera* é apresentado na Figura 3.10. O sinal primário é o mais simples já descrito entre os metazoários sendo a complementaridade entre isoformas o que determina o mecanismo de *splicing* alternativo. Uma vez estabelecido este sinal, a produção de proteínas DSX sexo-específicas ocorre constitutivamente (dependendo da quantidade de mRNA), sendo que as cooperações e competições com outros fatores e cofatores de transcrição devem funcionar de maneira oposta em cada sexo, isto é, em fêmeas, um complexo com DSX-F regula positivamente os genes de diferenciação do fenótipo feminino e negativamente os genes masculinizantes, sendo o inverso na presença de DSX-M. O fator GAGA pode ser um dos fatores mais importantes na produção inicial de mRNA de *Amdsx* e *Amix* já no início da embriogênese. O sistema de controle mais bem conhecido envolve a formação do complexo DSX-F/IX/BZIP-1 específico de fêmea. O fator DSX-M possuí uma região C-terminal maior, com um domínio de oligomerização que deve funcionar de maneira similar ao complexo DSX-F/IX, regulando de forma alternativa os genes de diferenciação sexual.

3.1.7 Conclusões

Uma busca a partir de 29 genes anotados na base de dados GO como envolvidos na determinação e diferenciação do sexo de *D. melanogaster* apontaram para 13 genes altamente conservados em *A. mellifera*. Oito destes 13 genes codificam uma variedade de fatores de transcrição (*dsx*, *ix*, *fru*, *dpn*, *dsf*, *run*, *bab1* e *scr*; Tabela 3.1, pg.76). Três destes 13 genes conservados em abelha codificam proteínas que participam no sinal inicial da determinação do sexo (*sxl*, *tra2* e *fl*) e estão envolvidos nos mecanismos de *splicing* que determinam o destino sexual de *D. melanogaster*. Embora estes genes do início da cascata de determinação do sexo estejam conservados no genoma de *A. mellifera*, nenhuma evidência que permita inferir conservação funcional destes sinais iniciais foi encontrada. Ao contrário, estes sinais

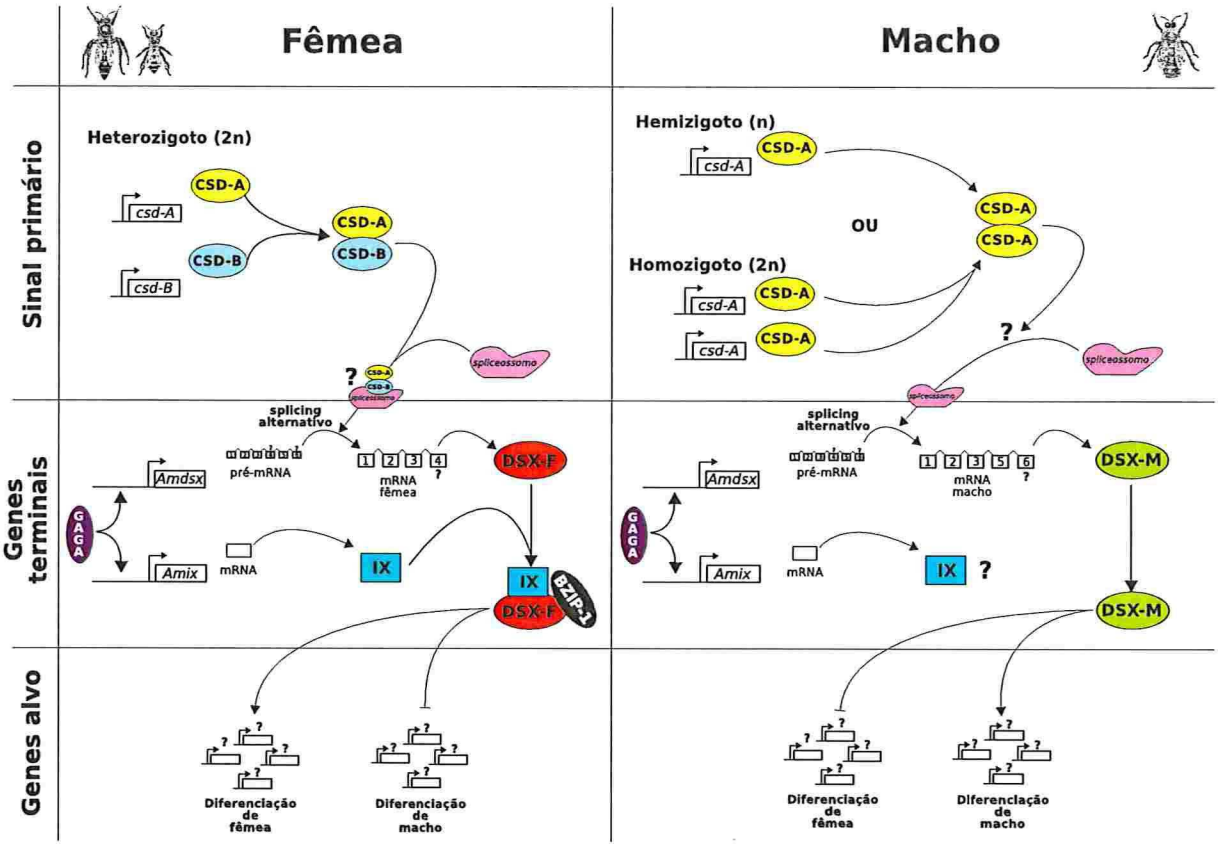


Figura 3.10: Modelo simplificado do mecanismo genético de determinação do sexo em *A. mellifera*

apresentam variações mesmo entre o gênero de moscas *Musca* (MEISE et al., 1998). Os outros dois genes conservados em abelhas estão relacionados com processos de comunicação celular através de vias de transdução de sinal (*hop* e *pk61C*). Estas vias de transdução de sinal são consideradas muito importantes na formação das estruturas genital e anal, por meio de uma regulação sexo-específica dos genes *decapentaplegic* (*dpp*), *hedgehog* (*hh*) e *wingless* (*wg*) (ESTRADA; CASARES; SáNCHEZ-HERRERO, 2003). Todos estes 13 genes são importantes para a determinação e desenvolvimento sexual em *D. melanogaster*, mas dois deles, *dsx* e *ix*, são considerados terminais na cascata gênica e altamente conservados em estrutura e função mesmo em espécies evolutivamente distantes (CLINE; MEYER, 1996; MARíN; BAKER, 1998; GRAHAM; PENN; SCHEDL, 2002; SIEGAL; BAKER, 2005).

Em abelhas, o sinal inicial não depende da razão X:A ou qualquer tipo de gene ligado a cromossomos sexuais. Na verdade, estes insetos não possuem cromossomos sexuais e

a determinação do sexo é estabelecida pelo produto de um único gene (ou loco) que se complementa. No mecanismo complementar de determinação do sexo (CSD) em *A. mellifera*, a proteína codificada a partir de um mesmo alelo não é funcional e então os machos hemizigotos (haplóides) ou homozigotos (diplóides) se desenvolvem (*default*). Por outro lado se o indivíduo é portador de dois alelos diferentes (heterozigoto), os produtos destes alelos se combinam e disparam o mecanismo de determinação de fêmea (WHITING, 1943; BEYE et al., 2003). O gene *complementary sex determiner* (*csd*) codifica uma proteína SR[1] que deve atuar no topo da cascata de determinação do sexo regulando a formação de transcritos de sexo-específicos de *Amdsx* de maneira similar ao TRA de *D. melanogaster* e *C. capitata* (BEYE et al., 2003; BEYE, 2004).

As cascatas de determinação do sexo evoluíram da base para o topo (WILKINS, 1995; MARíN; BAKER, 1998; SCHüTT; NöTHIGER, 2000) e de acordo com esta hipótese os genes, *dsx* e *ix*, estão altamente conservados na base da via de determinação do sexo. Estes dois genes também estão conservados em função nos metazoários (RAYMOND et al., 1998; SUZUKI et al., 2003; HEDIGER et al., 2004; SIEGAL; BAKER, 2005) e no presente estudo encontramos evidências de que estes genes terminais também estão conservados em *A. mellifera*. O produto de *ix* é essencial para que DSX-F desempenhe sua função ativadora ou repressora em genes alvo de maneira específica em fêmeas, enquanto DSX-M não depende de IX e atua de maneira contrária às fêmeas nos genes alvo (SIEGAL; BAKER, 2005). O complexo DSX-F/IX/BZIP-1 que se liga à região reguladora, FBE, de *yp-1* em *D. melanogaster*, ativa a transcrição deste gene em fêmeas. Em machos, as proteínas DSX-M não permite a ligação de BZIP-1 ao seu sítio bzip1 no FBE e inibe a transcrição de *yp-1* (COSCHIGANO; WENSINK, 1993; AN; WENSINK, 1995a; GARRETT-ENGELE et al., 2002).

Em *A. mellifera*, o controle de transcrição de *Amvg* compartilha pouca similaridade com o seu análogo, *yp-1*, em mosca-de-fruta ou mesmo com seu homólogo, *Bmvg*, no bicho-da-seda. A quantidade de transcritos no CG é muito maior que em Ov, garantindo a posição do CG como principal órgão sintetizador de Vg. Além disso, o *Amvg* não é um gene de expressão específica do sexo feminino sendo transcrito em níveis consideráveis em zangões. Uma análise computacional comparativa entre as regiões promotoras dos três genes codificadores de proteínas

[1]proteínas com resíduos Serina/Arginina envolvidas na regulação e seleção de sítios de splicing em mRNA eucariontes.

do ovo revelaram poucos elementos conservados. Entretanto, algumas características destas regiões reguladoras são particularmente interessantes: (1) em *yp-1* e *Bmvg*, observa-se alguns sítios dsx e bzip1 em redundância e sobrepostos em uma região que inclui o FBE, podendo ser responsável pelo controle de transcrição sexo-específico observado nestes genes; (2) em *Amvg*, não há nenhum sítio dsx nas proximidades da região equivalente ao FBE, mas apenas um sítio bzip1 que é o principal ativador da transcrição destes genes em Ov; (3) nos três genes, observamos a ocorrência do sítio aef1 que é um repressor de transcrição dependente da abundância do fator AEF-1 e que pode ser o principal responsável pelas diferenças nos níveis de transcrição destes genes em diferentes tecidos.

Considerando apenas os perfis qualitativos da transcrição de *Amvg*, *Amdsx* e *Amix* nos tecidos analisados neste estudo podem-se observar algumas semelhanças e diferenças nos genes equivalentes em *D. melanogaster*. As semelhanças são: (1) os genes das proteínas do ovo (*yp-1* e *Amvg*) são transcritos principalmente no CG, mas quando são transcritos nos ovários e células do folículo ovariano, respectivamente, não estão sob o controle da via de determinação do sexo; (2) *Amdsx* é regulado de maneira sexo-específica com a produção de transcritos alternativos por *splicing* alternativo; (3) os ortólogos *ix* são mais expressos em fêmeas e em Ov. As diferenças são: (1) *Amvg* é transcrito em Ov de rainhas e em machos enquanto *yp-1* só se expressa em fêmeas mas não em Ov; (2) *Amdsx* é mais transcrito em Ov do que em CG e *Dmdsx* é transcrito em CG mas não em Ov embora seja transcrito nas células do folículo ovariano, em estágios específicos de desenvolvimento.

No desenvolvimento dos embriões, não encontramos nenhuma evidência de que *Amvg* seja transcrita em machos e fêmeas e mesmo com toda a diversidade funcional que a AmVg apresenta é intrigante verificar esta ausência de transcrição nos embriões. Por outro lado, a transcrição de *Amdsx* e *Amix* durante a embriogênese revelou um padrão muito interessante que deve traduzir a complexidade das interações entre estes dois genes com seus genes alvo que determinam a dualidade no destino da formação de estruturas morfológicas e posteriormente fisiológicas e comportamentais. Durante todo o desenvolvimento embrionário a transcrição de *Amdsx* em machos e fêmeas é notavelmente antagônica, isto é, quando a transcrição aumenta em macho, diminui em fêmea. Este padrão se manifesta em todas as fases amostradas com exceção da última fase que caracteriza a organogênese. Tal alternância no perfil de transcrição de *Amdsx* fornece evidências de que os genes alvo controlados pelo fator DSX podem ser expressos em ondas que

seriam responsáveis pela determinação de padrões embrionários sexo-específicos. Além disso, *Amdsx* e *Amix* em ambos os sexos são co-expressos no início da embriogênese (até 24hs), mas a partir deste período a co-expressão parece continuar apenas em fêmeas.

Observando tal co-expressão assumimos a hipótese de co-regulação entre *Amdsx* e *Amix* e portanto devem compartilhar elementos regulatórios cis conservados em suas regiões promotoras. A correlação entre o perfil de transcrição de um mesmo gene em machos e fêmeas é menor ($R^2 = 0,44$ para F-dsx vs M-dsx; $R^2 = 0,85$ para F-ix vs M-ix) do que entre os dois genes em fêmeas ($R^2 = 0,96$ para F-dsx vs F-ix). O alto índice de correlação no perfil de transcrição de *Amdsx* e *Amix* em fêmeas está em concordância com necessidade da formação de um complexo DSX-F/IX essencial para o controle da expressão gênica específica de fêmeas. Um motivo muito similar ao sítio de ligação do fator GAGA foi descoberto nas regiões promotoras dos dois grupos de genes ortólogos (*dsx* e *ix*), denominado GAGA-DIX. O fator GAGA é codificado pelo gene *Trl* e é muito conservado no genoma de abelha. Este fator regula a transcrição gênica principalmente pela remodelagem da cromatina permitindo o acesso ao DNA de outros fatores de transcrição importantes para o controle transcricional de genes relacionados com o estabelecimento de padrões no desenvolvimento do embrião. Encontrar sítios gaga nas regiões promotoras destes genes terminais significa que a dualidade regulatória dos sexos se inicia já nas primeiras horas da embriogênese (até 24h) e que respostas (ou "feed-back") alternativas do sistema regulatório de transcrição determinam o destino sexual (dualista) deste sistema.

A presença de outros motivos nos promotores de *Amdsx* e *Amix*, tais como DSX, AEF-1 e BZIP-1, formam combinações que podem atuar como módulos regulatórios. A presença destes módulos pode fornecer idéias dos padrões de regulação aos quais estes genes estão submetidos. O *Dmdsx* apresenta uma região com sítios gaga, dsx e aef1, e pode estar submetido ao controle positivo do fator GAGA, mas de forma sexo (auto-regulação) e tecido-específicas (DSX não é transcrito em Ov onde o repressor AEF-1 é abundante). Em *Amdsx*, pouca similaridade é observada em seu promotor com relação ao seu ortólogo em *Drosophila*. No entanto, a ausência de sítios dsx indica não haver auto-regulação (pelo menos de forma direta) e a presença de vários sítios bzip1 pode ser uma das razões de *Amdsx* ser transcrito em Ov. Em *D. melanogaster*, BZIP-1 é o principal fator responsável pela transcrição de *yp-1*. Os três motivos considerados mais importante para o controle do desenvolvimento sexual (GAGA-DIX, DSX, BZIP-1) apresentaram distribuições de ocorrência nos promotores significativamente diferentes

da esperada pelo acaso, e portanto, devem estar sob pressão de seleção para manter o viés de posição (ocorrem preferencialmente até $\approx$ 600pb do início de tradução do gene). Os genes alvo potencialmente controlados por estes fatores participam de processos metabólicos, transcrição gênica, desenvolvimento embrionário e de estruturas anatômicas, bem como de mecanismos de comunicação celular, reprodução e comportamento.

Contudo, a principal contribuição deste estudo está na possibilidade da construção de um modelo regulatório hipotético envolvido na determinação genética do sexo em abelhas que pode ser utilizado como ponto de partida para validações experimentais orientadas por hipóteses. Ainda que boa parte do modelo seja teórico e necessite de validações experimentais mais detalhadas em biologia molecular e genética (ex: silenciamento gênico, estudo de interação proteína-proteína, ensaios de ligação proteína-DNA e expressão gênica), muito pouco se conhecia sobre as entidades genéticas e possíveis mecanismos moleculares envolvidos na determinação do sexo em *A. mellifera*.

3.2 Genômica funcional da determinação de casta

A determinação de castas em abelhas é um sistema modelo para o estudo de plasticidade do desenvolvimento. Uma diferença na composição alimentar de larvas de fêmea promove o desenvolvimento de fenótipos alternativos. Rainhas e operárias se desenvolvem pela expressão diferencial de genes a partir de um genótipo idêntico e representam um dos exemplos mais conhecidos de polifenismo em insetos. A diferença mais marcante entre as rainhas e operárias de *A. mellifera* está nas particularidades do sistema reprodutivo. As rainhas desenvolvem grandes ovários para a postura de centenas de ovos por dia, enquanto as operárias são estéreis com o ovário pouco desenvolvido. Os genes diferencialmente expressos em rainhas e operárias de *A. mellifera* foram identificados em diversos estudos com aplicações de técnicas experimentais bem distintas que, por sua vez, identificaram genes distintos (CORONA; ESTRADA; ZURITA, 1999; EVANS; WHEELER, 1999, 2001; HEPPERLE; HARTFELDER, 2001). No presente estudo, os genes identificados como diferencialmente expressos nas fases larvais onde as diferenças nutricionais ocorrem. As seqüências de cDNA e ESTs dos estudos anteriores foram anotadas no genoma de abelha e seus respectivos genes identificados.

O perfil funcional dos genes diferencialmente expressos em cada casta foi proposto segundo anotações dos genes ortólogos em *D. melanogaster*. Os genes que regulam o metabolismo estão mais expressos em rainhas, enquanto genes que regulam vias do desenvolvimento estão mais expressos em operárias. Genes envolvidos no processamento pós-transcricional e com atividade de hidrolases e oxidoredutases também estão muito representados entre os genes envolvidos na determinação de castas. As regiões promotoras dos genes diferencialmente expressos em castas foram analisadas e motivos putativos específicos de alguns grupos de genes foram identificados e se organizam de maneira distinta em cada casta.

3.2.1 Identificação e anotação de genes envolvidos na determinação de casta

Os genes diferencialmente expressos em larvas de rainha e operária estão listados nas Tabelas 3.4 e 3.5. Os genes diferencialmente expressos em rainhas (34 genes; Tabela 3.4) e operárias (17 genes; Tabela 3.5) foram identificados a partir de ESTs disponíveis no GenBank publicados em (EVANS; WHEELER, 1999, 2001). Os genes ortólogos em *Drosophila* foram identificados por similaridade de seqüência e os domínios de proteína conservados descritos na base Pfam (E-value $\leq$ 1e-15, não mostrado nas Tabelas 3.4 e 3.5) também foram localizados nos genes de abelha (Tabelas 3.4 e 3.5).

A maioria dos genes diferencialmente expressos em castas são representados por uma ou duas ESTs, exceto o gene da *hex70b* (GB10869 é homólogo de Lsp2-PA de *D. melanogaster*). Este gene foi representado por 10 ESTs, uma localizada na região 5' (primeiro éxon) e nove na região 3' (cinco ESTs no éxon 7 e região 3'-UTR, e outras quatro ESTs nos éxons 6 e 7). Experimentos de *macroarray* (EVANS; WHEELER, 2001) evidenciaram *hex70b* como super-expresso em operárias. As hexamerinas são uma importante classe de proteínas de estocagem que apresentam um interessante padrão de expressão relacionado ao sexo, à casta e à reprodução em muitos insetos sociais (JINWAL et al., 2006; MARTINEZ et al., 2000; HUNT; BUCK; WHEELER, 2003; ZHOU; OI; SCHARF, 2006; ZHOU et al., 2006). Um cDNA referente a *hex70b* de *A. mellifera* foi recentemente clonado e seqüenciado, e experimentos com manipulação de hormônio mostraram que a abundância de transcritos de *hex70b* no desenvolvimento larval é positivamente correlacionado com altos níveis de JH e ecdisteróides

Tabela 3.4: Lista de genes super-expressos em rainhas identificados e anotados a partir de dados de cDNA e ESTs publicados em estudos anteriores (CORONA; ESTRADA; ZURITA, 1999; EVANS; WHEELER, 1999, 2001). (⋆) Genes usados para descoberta de motivos regulatórios putativos.

Officialset	GenBank	Scaffold	Flybase	E-value	Nome(Pfam)	Descrição (Pfam)
GB18242⋆	BG101680	Group1.68	Hmgcr	0,0	HMG-CoA_red	Hidroximetilglutaril-coenzima A redutase
GB11628⋆	BG101591	Group13.10	CG4851	8e-90	–	–
GB19626	BG101692	Group15.24	Osi9	9e-15	DUF1676	Proteína de função desconhecida (DUF1676)
GB18446	BG101673	Group16.2	yps	1e-31	CSD	Domínio de ligação ao DNA Cold-shock
GB13051	BG101697	Group16.19	CG4662	1e-131	–	–
GB15755	BG101571	Group5.12	RpS11	3e-53	Ribosomal_S17	Proteína ribossomal S17
GB15754	BG101555	Group6.19	CG32405	5e-19	Chitin_bind_4	Proteína cuticular de insetos
GB14874	BG101573 BG101695	Group7.47	Sop2	1e-135	WD40	Domínio WD, repetição G-beta
GB11177	BG101636	Group9.22	eEF1delta	4e-48	EF1_GNE	Domínio de troca de guanina EF-1
GB15686	BG101696	Group1.33	–	–	–	–
GB10042	AF069739	GroupUn.1189	CG12413	e-145	GTP_EFTU	Domínio de ligação GTP do EF Tu
GB17825	BG101532	Group6.53	RpI135	0,0	RNA_pol_Rpb2_6	Domíno 6, RNA polimerase Rpb2
GB15817	BG101643 BG101641	GroupUn.85	–	–	–	–
GB10156	BG101691	Group5.24	CG14511	6e-55	–	–
GB14262	BG101592	GroupUn.33	ade5	1e-142	SAICAR_synt	sintetase SAICAR
GB13368	BG101559 BG101600	Group12.20	CG9914	1e-101	3HCDH_N	3-hidroxiacil-CoA dehidrogenase
GB11029	BG101543 BG101568	Group2.22	RpL18	3e-73	Ribosomal_L18e	Proteína ribossomal eucarionte L18
GB12441	BG101574 BG101597	Group9.17	RpL35A	3e-42	Ribosomal_L35Ae	Proteína ribossomal L35Ae
GB14798⋆	BG101635	Group10.38	Gapdh1	1e-141	Gp_dh_C	Gliceraldeído 3-fosfato desidrogenase, C-
GB14476	BG101634	Group15.19	mRpL35	2e-33	–	–
GB18769	BG101610 BG101614	Group3.18	oho23B	1e-33	Ribosomal_S21e	Proteína ribossomal S21e
GB12741	BG101583	Group7.16	Aldh	0,0	Aldedh	Desidrogenase de aldeído
GB13776	BG101617	Group1.12	mRpL3	1e-97	Ribosomal_L3	Proteína ribossomal L3
GB18969	BG101585	Group16.4	Hsp60	0,0	Cpn60_TCP1	Chaperonina TCP-1/cpn60
GB17423	BG101577	Group1.67	CG12130	1e-73	NHL	Repetição NHL
GB16047⋆	BG101595	Group11.41	loj	2e-73	EMP24_GP25L	Família emp24/gp25L/p24
GB11973	BG101579	Group16.9	Cyp4g15	0,0	p450	Citocromo P450
GB13072⋆	BG101582	Group2.23	RpS27	2e-31	Ribosomal_S27e	Proteína ribosomal S27
GB19801	BG101570	Group2.38	CG4747	4e-76	NAD_binding_2	Domínio de ligação NAD do 6-fosfogluconato
GB10002	BG101605	Group5.13	CG5639	1e-131	Thyroglobulin_1	Repetição tiroglobulina tipo-1
GB19380⋆	BG101596	Group12.5	CG6888	2e-86	AhpC-TSA	Família AhpC/TSA
GB19465	BG101594	GroupUn.179	RpL12	5e-74	Ribosomal_L11_N	Proteína ribossomal L11, N-terminal
GB12811	BG101564	Group13.9	CG2150	1e-12	–	–
GB11303	BG101576	Group2.33	lid	0,0	JmjC	Domínio JmjC

(CUNHA et al., 2005). No cupim, *Reticulitermes flavipes*, a razão entre Hex1 e Hex2 controla a disponibilidade de JH para diferenciação casta-específica de tecidos (ZHOU et al., 2006). Estas observações permitem reflexões sobre possíveis interações entre as diferentes subunidades de Hexamerinas de *A. mellifera* desempenhando funções regulatórias de *feedback* no controle dos títulos de JH em castas de abelhas (ZHOU et al., 2006).

O perfil funcional dos genes de rainhas e operárias foi analisado com auxílio do programa FatiGO a partir do nome dos genes ortólogos em *D. melanogaster* (coluna FlyBase; Tabelas 3.4 e 3.5). Dos 34 genes de rainhas, apenas 2 não possuem similaridade com nenhum gene de *Drosophila*. Em operárias, apenas 1 dos 17 genes não possui homólogo em *Drosophila*. Os processos biológicos predominantes (considerando o nível 3 de termos GO) entre todos os genes diferencialmente expressos em castas (total 48 genes) são: processo fisiológico celular (95%; Figura 3.11, A) e metabolismo (90%; Figura 3.11, A). As diferenças nos processos biológicos

Tabela 3.5: Lista de genes super-expressos em operárias identificados e anotados a partir de dados de cDNA e ESTs publicados em estudos anteriores (EVANS; WHEELER, 1999, 2001). (⋆) Genes usados para descoberta de motivos regulatórios putativos.

Officialset	GenBank	Scaffold	Flybase	E-value	Nome(Pfam)	Descrição (Pfam)
GB17508	BG101588	Group11.37	CG9769	1e-108	Mov34	Família Mov34/MPN/PAD-1
GB10428⋆	BG101584	Group2.30	CG30413	1e-05	–	–
GB19006⋆	BG101586	Group4.7	CG3884	2e-30	–	–
GB11132	BG101562	Group5.33	RpS5a	1e-99	Ribosomal_S7	Proteína ribossomal S7p/S5e
GB17499	BG101550	Group7.37	sesB	1e-141	Mito_carr	Proteína mitocondrial carreadora
GB11059	BG101556	Group8.20	RfaBp	1e-139	Vitellogenin_N	Região amino terminal de Lipoproteína
GB12239⋆	BG101547	Group9.15	zeelin1	4e-34	–	–
GB18643	BG101603	GroupUn.472	Dhc64C	0,0	Dynein_heavy	Cadeia pesada de dineína
GB12230	BG101662	Group10.38	RpLP1	8e-29	Ribosomal_60s	Proteína ribossomal 60s
GB10469	BG149167	Group2.31	klg	2e-42	I-set	Domínio imunoglobulina I-set
GB12371⋆	BG101639	Group2.38	Mgstl	5e-25	–	–
	BG101674					
	BG101686					
GB14758⋆	BG101609	Group7.33	Hsp83	0,0	HSP90	Proteína Hsp90
GB14634	BG101609	Group9.22	Hexo2	6e-55	Glyco_hydro_20	Glicosil hidrolase, família 20
GB10869⋆	BG101646	GroupUn.53	Lsp2-PA	7e-97	Hemocyanin_C	Domínio tipo ig, Hemocianina
	BG101569					
	BG101537					
	BG101541					
	BG101558					
	BG101632					
	BG101539					
	BG101633					
	BG101552					
GB14791	BG101690	GroupUn.4897	blw	0,0	ATP-synt_ab	Família alfa/beta ATP sintase
GB13410	BG149169	GroupUn.416	xmas-2	1e-109	SAC3_GANP	Família p25 SAC3/GANP/Nin1/mts3/eIF-3
GB10536	BG101629	Group15.11	–	–	–	–

dos quais os genes de rainhas e operárias podem participar estão principalmente na diferenciação celular (28,5% dos genes em operárias e nenhum em rainhas; Figura 3.12, A) e no metabolismo (96% dos genes em rainhas e 78,5% em operárias; Figura 3.12, A).

Com relação às funções moleculares, a grande maioria dos termos estão relacionados com a tradução de mRNA (ligação a ácido nucléico, 38%; estrutura do ribossomo, 24%; ligação a proteína, 12%; ligação a nucleotídeo, 12%; atividade de fator de tradução e ligação, 7%; Figura 3.11, B). Outras funções moleculares importantes também foram observadas, tais como, atividade oxidoredutase (19%; Figura 3.11, B) e hidrolase (16,5%; Figura 3.11, B). Para estes dois termos notamos diferenças interessantes potencialmente relacionadas à formação de castas, com atividade hidrolase estando muito representada entre os transcritos de operária, e a atividade oxidoredutase representada somente por genes super-expressos em rainha (Figura 3.11, B). Ainda que estas atribuições funcionais sejam baseadas em evidências obtidas de ortólogos de *D. melanogaster*, boa parte dos genes de casta possuem domínios funcionais conservados em seqüência de proteína e, portanto, reforçam as nossas observações sobre as tendências funcionais destes 51 genes de casta.

Em termos gerais, as diferentes representações de vias funcionais em cada casta, oxidoredutase e hidrolase, podem ser reflexo da mudança na dieta que a operária recebe durante o

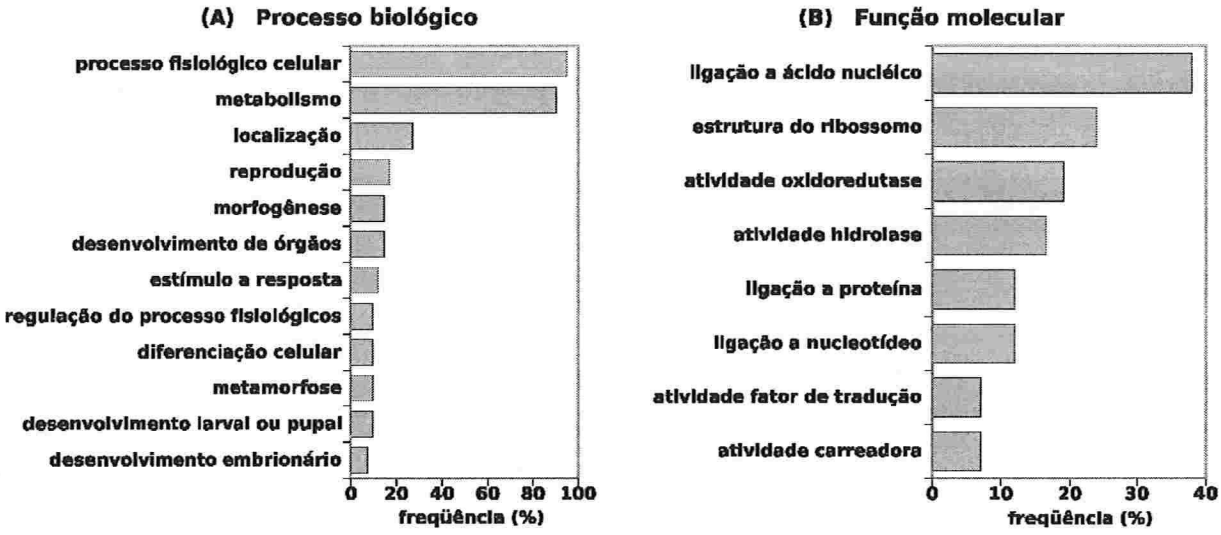


Figura 3.11: Perfil funcional de todo o conjunto de genes diferencialmente expressos em castas de *A. mellifera*. (A) Os Processos biológicos predominantes entre estes genes mostram grande participação do metabolismo e fisiologia celular na determinação de casta em *A. mellifera*. (B) As funções moleculares predominantes neste conjunto de genes mostram a importância de genes envolvidos nos mecanismos de controle pós-transcricional e atividades hidrolase e oxidoredutase. As propriedades funcionais dos genes de abelha foram inferidas através da anotação funcional dos ortólogos de *D. melanogaster* no GO (nível 3).

quarto e quinto ínstares larvais. Esta alternativa significa uma mudança de uma dieta rica em lipídios e proteínas para outra rica em carboidratos (HAYDAK, 1970), e esta alternância vem acompanhada de um aumento na expressão de genes que codificam hidrolases. Alternâncias similares nos padrões de expressão gênica foram recentemente observadas para *D. melanogaster* em experimentos onde larvas foram trocadas de meios alimentares (dieta à base de milho para banana) resultando em 55 genes diferencialmente expressos. Entre eles estão 5 genes com atividade desidrogenase/oxidoredutase. Estes padrões semelhantes resultantes das alterações de dieta podem ser indicativos de redes regulatórias semelhantes. De grande interesse para a biologia é compreender como estas respostas alternativas das vias regulatórias evoluem para gerar diferentes fenótipos, tais como as castas. Entretanto, as plataformas experimentais para os estudos funcionais em abelhas começaram a ser desenvolvidas apenas recentemente (ex: RNAi, cultura de células, etc) com a disponibilidade do genoma e das ferramentas de bioinformática.

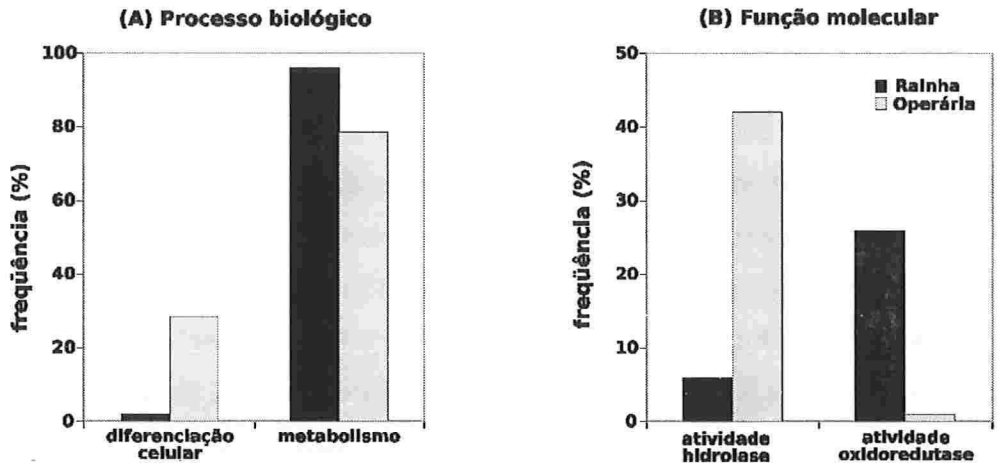


Figura 3.12: Perfil funcional dos genes diferencialmente expressos em rainhas e operárias de *A. mellifera*. (A) Os genes que regulam o processo de diferenciação celular estão super-expressos em maior proporção em operárias enquanto os genes de metabolismo são estão mais representados em rainhas. (B) Genes com atividade hidrolase são predominantes em operárias e ao contrário, genes com atividade oxidoredutase estão mais expressos em rainhas. As propriedades funcionais dos genes de abelha foram inferidas através da anotação funcional dos ortólogos de *D. melanogaster* no GO (nível 3).

3.2.2 Elementos regulatórios cis putativos nos genes de casta

Os genes relacionados ao desenvolvimento de casta estão entre os primeiros genes de *A. mellifera* para os quais dados de expressão validados experimentalmente foram gerados (CORONA; ESTRADA; ZURITA, 1999; EVANS; WHEELER, 1999, 2001; HEPPERLE; HARTFELDER, 2001; GUIDUGLI; HEPPERLE; HARTFELDER, 2004). Certamente, estes 51 genes representam apenas uma parcela de todos os genes realmente envolvidos no desenvolvimento de casta. No entanto, eles são os genes com diferenças de expressão mais proeminentes detectados por experimentos (SSH[2] e DDRT-PCR[3]) que evidenciam as maiores diferenças entre as duas situações comparadas, rainha e operária. Os 51 genes evidenciados não pertencem a um única família gênica mas a várias famílias com diferentes funções. Desta maneira, podemos assumir a hipótese de que os padrões de expressão dos genes em larvas de rainhas e de operárias podem estar associados à ocorrência de motivos regulatórios cis específicos nas regiões promotoras destes genes.

[2]Suppression Subtractive Hybridization
[3]Differential Display RT-PCR

O *pipeline* escrito em Python usa três algoritmos diferentes para a descoberta de motivos conservados: AlignACE (ROTH et al., 1998), MEME (BAILEY; ELKAN, 1994) e MDscan (LIU; BRUTLAG; LIU, 2002); que são integrados pelo módulo Python TAMO (GORDON et al., 2005). Este *pipeline* analisa as regiões promotoras de um conjunto de genes de interesse e retorna uma lista de motivos super-representados neste conjunto, comparando a ocorrência destes motivos nas regiões promotoras do conjunto inicial a todos os genes do genoma (para detalhes ver seção 2.1.3 e Figura 2.1). Dois conjuntos de genes foram selecionados pelos maiores níveis de expressão em cada casta (top6 de rainha e top6 de operária), de acordo com evidências experimentais (EVANS; WHEELER, 2001) e também um conjunto controle (controle negativo) foi selecionado randomicamente do conjunto total de regiões promotoras (Amel-PromDB). Para cada motivo descoberto, foram calculadas quatro métricas diferentes: MAP (ROTH et al., 1998), especificidade de grupo (HUGHES et al., 2000) e ROC AUC e MNCP (CLARKE; GRANEK, 2003). Um primeiro filtro foi usado para eliminar aqueles motivos mais espúrios (MAP $\geq$ 5; ROC AUC $\geq$ 0,7). Esta filtragem inicial resultou em 46 motivos de um total inicial de 123 encontrados no conjunto de promotores de rainha, e em 71 de um total de 261 motivos encontrados no conjunto de operárias.

Tabela 3.6: Teste de Kolmogorov-Smirnov para a análise da distribuição dos valores de ROC AUC, MNCP e Church calculados para cada motivo encontrado nos genes super-expressos em larvas de rainha e operária. Os grupos ou conjuntos de motivos de rainha e operária foram comparados com o conjunto de motivos encontrados em conjuntos de regiões promotoras selecionadas randomicamente da base Amel-PromDB.

Grupos	**ROC AUC**	**MNCP**	**Church**
random vs. rainha+operária	P<0.01	P>0.1	P>0.1
random vs. rainha	P<0.001	P>0.1	P>0.1
random vs. operária	P>0.1	P>0.1	P<0.025
rainha vs. operária	P<0.005	P>0.1	P<0.001

Dois testes estatísticos, um paramétrico (MANOVA; P<0.0001; Wilks=0.73; F=7.36) e outro não paramétrico (Teste de Kolmogorov-Smirnov; Tabela 3.6), foram conduzidos a partir dos valores de ROC AUC, MNCP e Church mostrando que a distribuição da pontuação (principalmente de ROC AUC e Church) destes dois conjuntos de motivos (rainhas e operárias) é significativamente diferente dos motivos encontrados no controle negativo. O ROC AUC tem sido descrito em outros estudos como uma métrica eficiente na eliminação de motivos falso-positivos (CLARKE; GRANEK, 2003; MACISAAC; FRAENKEL, 2006).

Para selecionar apenas os motivos mais específicos encontrados nos genes de cada casta, usamos a pontuação especificidade de grupo ou Church score (HUGHES et al., 2000) para identificar os mais prováveis motivos envolvidos nas vias regulatórias de determinação do desenvolvimento de rainha (2 motivos com Church score $\leq$ 1e-5; Figura 3.13) ou operária (12 motivos com Church score $\leq$ 1e-7; Figura 3.13). Considerando que os métodos experimentais utilizados (SSH e DDRT-PCR) identificam somente uma sub-população dos genes importantes para a determinação de castas, entendemos que os motivos descobertos neste estudo também representam um cenário parcial da rede regulatória de transcrição envolvida no desenvolvimento de casta. Estes motivos podem ser usados para buscas em larga escala de potenciais genes candidatos que também participem das redes regulatórias de desenvolvimento de castas e que poderão ser submetidos a validações experimentais em trabalhos futuros.

Cada motivo encontrado nas regiões promotoras dos genes de rainha (46 motivos) e operária (71 motivos) foi alinhado contra todas as seqüências de elementos regulatórios cis de *D. melanogaster* contidas na base de dados TRANSFAC v4.0 (WINGENDER et al., 2000). Somente alinhamentos com pelo menos 80% de identidade com motivos encontrados aqui foram considerados. Nenhum dos motivos mais específicos de cada casta mostra similaridade com alguma seqüência reguladora de *Drosophila* contida no TRANSFAC. No entanto, alguns motivos de ocorrência ubíqüa apresentaram considerável similaridade com sítios de ligação de fatores de transcrição, tais como *Antennapedia*, *Ultrabithorax*, *zerknüllt*, *even skipped*, *trithorax-like*, *tailless*, *paired*, *fushi tarazu*, *Adh-1*.

Quando plotamos as posições dos 2 motivos de rainha e os 12 motivos de operária nas regiões promotoras de todos os 51 genes diferencialmente expressos em cada casta, observamos um padrão muito diferente na distribuição dos motivos de rainha e operária (Figura 3.14). Os 2 motivos de rainha foram localizados em 14 do 34 genes super-expressos em rainha e em apenas um gene super-expresso em operária (GB10869; Figura 3.14, A). É interessante observar que este gene é a *hex70b* e possui uma única ocorrência do motivo Q2 específico de rainha (Figuras 3.13 e 3.14, A). Este motivo pode significar que um fator de transcrição (ainda desconhecido) comum nas castas atue de maneira oposta no controle de transcrição de alguns genes de rainha e operária, por exemplo, o *hex70b* que é super-expresso em larvas de operárias mas reprimido em larvas de rainhas.

Os 12 motivos de operária foram localizados em 9 dos 16 genes super-expressos em operária

(A) Rainha

Motivo	MAP	Church	ROC_AUC	MNCP	Logo
Q1	5.50	1.37e-07	0.82	208.38	
Q2	5.27	2.57e-05	0.95	19.20	

(B) Operária

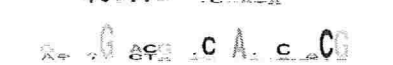
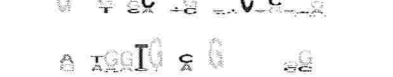
Motivo	MAP	Church	ROC_AUC	MNCP	Logo
W1	12.79	5.32e-10	0.74	113.04	
W2	11.39	5.32e-10	0.74	97.89	
W3	8.93	5.32e-10	0.83	237.48	
W4	7.04	5.32e-10	0.83	227.58	
W5	9.31	5.32e-10	0.90	118.10	
W6	9.86	5.32e-10	0.74	334.61	
W7	15.86	5.59e-10	0.74	206.11	
W8	8.55	1.37e-07	0.82	77.90	
W9	7.44	1.37e-07	0.90	35.16	
W10	13.83	1.37e-07	0.74	169.79	
W11	11.09	1.37e-07	0.83	184.30	
W12	10.59	1.37e-07	0.74	68.93	

Figura 3.13: Motivos regulatórios putativos descobertos nas regiões promotoras dos genes diferencialmente expressos em castas. (A) Motivos descobertos nas regiões promotoras de genes super-expressos em rainha. (B) Motivos descobertos nas regiões promotoras de genes super-expressos em operárias.

e em 2 genes dos 34 do conjunto de rainha (GB12811 e GB19626; (Figura 3.14, B). O gene GB12811 (super-expresso em rainha) possui um sítio do motivo W5 específico de operária. Este tem um ortólogo conservado em *D. melanogaster* que codifica uma proteína rica em glicina e, embora tenha sido considerado importante no desenvolvimento, pouco se sabe sobre sua função (FLAVELL; DYSON; ISH-HOROWICZ, 1987). Outro gene também super-expresso em rainha que apresenta um motivo específico de operária (W8; Figura 3.13, B) é o GB19626. Seu ortólogo em *Drosophila* é o *Osiris 9* (CG15592) que não tem nenhuma anotação funcional no GO mas é um gene que está localizado em um QTL relacionado à resistência à inanição e

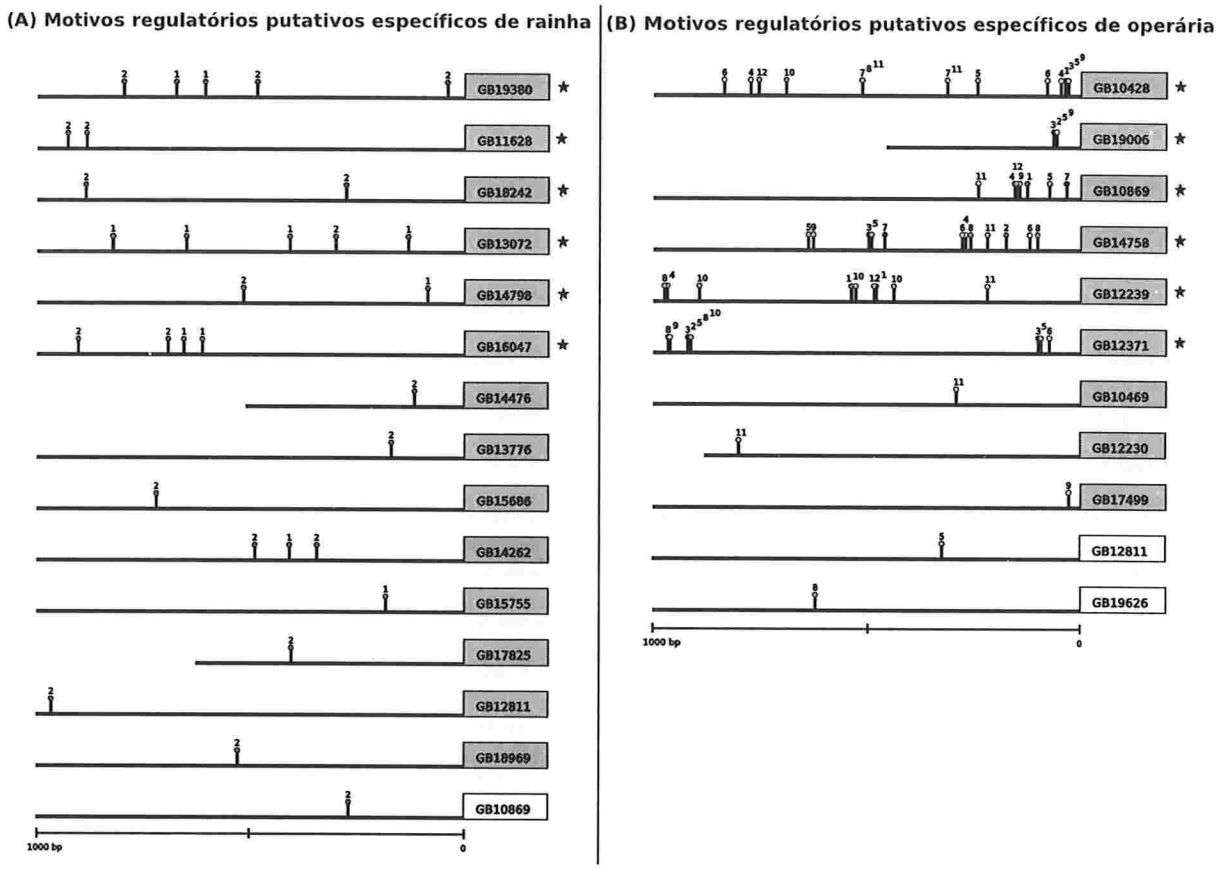


Figura 3.14: Diagrama de ocorrência dos motivos regulatórios putativos descobertos nos genes diferencialmente expressos em castas. (A) Os motivos descobertos em genes super-expressos em rainhas ocorrem em 14 de 34 genes de rainha e apenas um gene de operária, GB10869 (*hex70b*). (B) Os motivos descobertos em genes super-expressos em operárias ocorrem em 9 de 16 genes de operária e em 2 genes de rainha, GB12811 e GB19626. ($\star$) marca os top6 usados na descoberta dos motivos.

tem sua transcrição reprimida em experimentos de estresse por inanição em *D. melanogaster* (HARBISON et al., 2004, 2005). Assim como no exemplo do *hex70b*, a ocorrência destes motivos específicos de operária nos 2 genes super-expressos em rainha indica algum tipo de controle negativo destes genes durante o desenvolvimento da larva de operária. Além disso, um destes genes (GB19626) apresentou um padrão de expressão similar com seu ortólogo da mosca-de-fruta (*Osiris 9*, reprimido pela inanição) ao ser reprimido nas operárias que sofrem uma mudança drástica na dieta alimentar entre o quarto e quinto ínstares larvais.

Outro aspecto interessante dos motivos regulatórios putativos está na quantidade de motivos encontrados entre os genes de operária (12; Figura 3.13, B) em relação ao motivos de rainha (2;

Figura 3.13, A). Além do número superior de elementos conservados nas regiões promotoras dos genes de operária, os mesmos apresentam um maior agrupamento, com algumas exceções, em regiões próximas ao início de tradução dos genes. Tal observação está em concordância com o fato de que os elementos regulatórios cis com forte efeito regulatório estão próximos ao início do sítio de tradução (DAVIDSON, 2001).

Como o desenvolvimento de casta é altamente dependente de títulos de JH e ecdisteróides na hemolinfa, procuramos por sítios de ligação de receptores nucleares de hormônio nas regiões promotoras dos genes diferencialmente expressos em castas. Os elementos regulatórios envolvidos na resposta ao JH ainda não são muito bem entendidos. Assim, qualquer predição neste sentido estaria comprometida (WHEELER; NIJHOUT, 2003). Entretanto, elementos de resposta a ecdisona (EcRE) têm sido identificados e sabe-se que o complexo EcR/USP se liga a elementos regulatórios (sítios) com repetições diretas ou palindrômicas (invertidas) (RIDDIHOUGH; PELHAM, 1987; ANTONIEWSKI et al., 1995; PERERA et al., 2005). A representação canônica do motivo EcR/USP foi derivada da construção de uma PWM considerando toda a informação experimental atual disponível sobre os sítios de ligação do complexo EcR/USP (consenso: rGkTCAaTGamcy) (PERERA et al., 2005). Nenhum sítio similar (pelo menos 70%) foi encontrado em nenhuma das 51 regiões promotoras (1000pb) dos genes diferencialmente expressos em castas. No entanto, isto não determina que estes genes não possam responder às mudanças nos níveis de JH e/ou ecdisteróides, mas que eles possam responder indireta e tardiamente (LI; WHITE, 2003; SULLIVAN; THUMMEL, 2003) ou pela formação de complexos de receptores nucleares ainda pouco conhecidos.

3.2.3 Conclusões

As análises do perfil funcional e das regiões reguladoras de genes de determinação de casta foram conduzidas a partir de dados experimentais obtidos durante o desenvolvimento de larvas de *A. mellifera*. A anotação funcional destes genes, com base nas definições do GO referentes aos seus respectivos ortólogos de *Drosophila*, revelou os processos biológicos e funções moleculares mais relevantes para o desenvolvimento de castas. Os genes reguladores do metabolismo estão entre os mais importantes na determinação das castas, enquanto os genes codificadores de oxidoredutases estão mais expressos em rainhas e os genes de hidrolases mais expressos em

operárias.

O genoma de abelhas não só possibilita uma precisa anotação de genes com base em dados de cDNA e ESTs diferencialmente transcritos em castas, como também a exploração de regiões intergênicas na busca por elementos regulatórios relacionados com o controle de redes regulatórias de processos biológicos envolvidos no desenvolvimento de castas. No presente estudo, descobrimos 14 motivos conservados nas regiões promotoras (até 1000pb) dos genes de casta, sendo que 12 são motivos específicos de genes super-expressos em operária. Os sítios de ocorrência destes motivos apresentam uma distribuição agrupada e geralmente estão mais próximos (até ≈600pb) do sítio de início de tradução. O padrão de organização destes motivos nos promotores é de difícil interpretação e métodos de análise combinatória devem ser aplicados em estudos futuros para a identificação de possíveis módulos regulatórios putativos.

O polifenismo de casta nos insetos sociais é considerado um importante estudo de caso para o aparecimento de novidades nos processos evolutivos moleculares (WEST-EBERHARD, 2003). Embora o presente estudo não tenha a pretensão de discutir os detalhes dos mecanismos da plasticidade de desenvolvimento, outros dois estudos também apontam para modificações em propriedades funcionais e regulatórias . Estas modificações são responsáveis pela supressão do desenvolvimento de asas em formigas (ABOUHEIF; WRAY, 2002), bem como em alterações temporais nos padrões de expressão gênica durante o desenvolvimento pós-embrionário de formigas e abelhas (*B. terrestris*) (GOODISMAN et al., 2005; PEREBOOM et al., 2005). Tais mudanças regulatórias podem envolver elementos cis por meio de mutações nos sítios de ligação de fatores de transcrição que alteram a afinidade de ligação e conseqüentemente os padrões de resposta sob limiares específicos de concentração de fatores de transcrição (WRAY et al., 2003) ou de hormônios morfogenéticos (HARTFELDER; EMLEN, 2005). As modificações ou evolução das propriedades funcionais de genes podem ser exemplificadas pela importante participação das vitelogenina e hexamerinas. Estas proteínas são tradicionalmente reconhecidas nos outros insetos como proteínas de estocagem específicas de fêmeas e importantes na vitelogênese e oogênese. No entanto, em abelhas, elas evoluíram para funções também relacionadas com a determinação de castas e divisão de trabalho via novos rearranjos regulatórios com JH (AMDAM et al., 2004; GUIDUGLI et al., 2005a; ZHOU; OI; SCHARF, 2006; ZHOU et al., 2006).

A estratégia usada neste estudo combina informações experimentais de expressão gênica e

predições computacionais de motivos regulatórios conservados em grupos específicos de genes. Tal abordagem representa uma importante transição das descritivas em larga escala sem uma hipótese a priori para abordagens hipótese-dirigidas, podendo ajudar a explicar os padrões de expressão gênica dependentes de contexto na biologia das abelhas. Estas buscas podem servir como uma eficiente plataforma para uma análise integrada de experimentos em larga escala bem como para o delineamento de novos experimentos pela identificação de entidades importante nos fenômenos biológicos em abelhas e até mesmo outros metazoários. Além disso, este estudo exemplifica como programas e algoritmos de bioinformática já existentes podem ser integrados em uma ferramenta útil para o estudo de padrões de expressão gênica e de redes regulatórias em *A. mellifera*.

3.3 Evolução de famílias multigênicas em abelhas solitárias e sociais

A história evolutiva da diversificação das espécies de abelhas tem início no Período Cretáceo (≈120Ma[4]) juntamente com o período de maior expansão das angiospermas (vegetais que desenvolvem flores). É provável que as abelhas e angiospermas tenham co-evoluído de forma que a expansão do grupo das abelhas facilitou a expansão das angiospermas e vice-versa. Ainda hoje, a importância das abelhas na polinização é vital para a manutenção dos ecossistemas e da economia agrícola (DANFORTH, 2007). Além disso, as abelhas sociais, em particular as abelhas com corbícula[5], são úteis como modelo para estudos sobre a evolução da estrutura social e de modificações genéticas, morfológicas e comportamentais envolvidas neste modo de vida.

Mais de 16.000 espécies de abelhas já foram descritas mas apenas 6% são eusociais[6] (MICHENER, 2000). A eusocialidade em abelhas teve múltiplas origens sendo uma delas nas abelhas com corbícula (Hymenoptera: Apoidea: Apidae: Apinae). Este grupo de abelhas é subdividido em 4 tribos (Euglossini, Bombini, Meliponini e Apini) que apresentam modos de vida solitário (Euglossini), semi-social (Bombini) e eusocial (Meliponini e Apini). A estrutura

[4]Ma, Milhões de anos

[5]Corbícula, estrutura em forma de cesto que se desenvolve na tíbia da perna traseira de fêmeas para o transporte de pólen.

[6]Um sistema eusocial é caracterizado pela divisão reprodutiva de tarefas, sobreposição de geração e cooperação no cuidado da prole.

social neste grupo parece ter sua origem em um ancentral comum de Bombini, Meliponini e Apini (DANFORTH, 2007). A abelha *A. mellifera* pertence a tribo Apini que é um grupo monofilético, onde todas as espécies pertencem ao mesmo gênero *Apis*.

O seqüenciamento do genoma de *A. mellifera* adiciona uma nova dimensão na genômica comparativa e abre caminho para a compreensão das bases genéticas que garantem a realização das propriedades de um sistema eusocial, por exemplo, polifenismo de casta, polietismo de idade e divisão reprodutiva de tarefas (DANFORTH, 2007). Algumas observações sobre a evolução das famílias gênicas em *A. mellifera* são particularmente interessantes, tal como as diferenças no número de genes de algumas famílias multigênicas com possíveis efeitos no modo de vida destas abelhas. Algumas famílias multigênicas apresentam significativa redução no número de genes (ex: genes de detoxificação, proteínas cuticulares, receptores gustatórios, componentes do sistema imunológico), enquanto em outras, o número de genes é expandido (ex: proteínas da geléia real, receptores de olfato e algumas expansões de grupos das P450 e CCE associados a hormônios e processos quimiossensores) quando comparadas com o genoma de outros insetos (*Drosophila* e *Anopheles*) (CLAUDIANOS et al., 2006; The Honeybee Genome Sequencing Consortium, 2006).

No presente estudo, a taxa de substituição sinônima é utilizada como calibrador de um relógio molecular em abelhas. Este relógio permite identificar alguns eventos de duplicação gênica mais provavelmente correlacionados com a evolução das espécies de abelhas corbiculadas juntamente com aparecimento de propriedades importantes para a evolução do sistema eusocial nestes insetos. Estas aproximações foram feitas a partir de fragmentos dos genes *ache2* (*acetylcholinesterase 2*, *or83b* (*olfactory receptor 83b*), *lw-rh* (*long-wavelength rhodopsin*) e *mrjp* (*major royal jelly protein*) de sete espécies de abelhas com corbícula de modo de vida solitário à eusocial (*A. mellifera*, *A. cerana*, *A. dorsata*, *A. florea*, *M. quadrifasciata*, *B. terrestris* e *E. nigrita*). Os resultados apresentados a seguir suportam a hipótese de uma origem comum da eusocialidade nas abelhas com corbícula e apontam para os principais eventos de duplicação gênica ocorridos ao longo da história evolutiva destas abelhas.

3.3.1 Seleção dos genes utilizados para medidas de taxas evolutivas

Os genes usados no cálculo de taxas evolutivas entre sete espécies de abelhas foram selecionados principalmente pelos seguintes critérios: (1) os genes são de famílias gênicas diferentes e estas famílias apresentam diferenças em número comparado com outros insetos; (2) os genes estão localizados em cromossomos diferentes (locos não ligados) e codificam proteínas com funções distintas; (3) alguns genes são conservados em outras espécies de insetos e apresentam ortólogos evidentes (ortólogos 1:1:1 em *D. melanogaster*, *A. gambiae* e *A. mellifera*), enquanto outros genes são de uma mesma família que ocorre especificamente em abelhas. Os genes selecionados para este estudo estão relacionados na Tabela 3.7 que também descreve algumas características da organização genômica e funcional destes genes.

Tabela 3.7: Genes selecionados para os cálculos de taxas evolutivas. Grupo = Cromossomo. cds = seqüência codificadora

Nome	OfficiaSet	Grupo	cds(nt)	#éxons	GenBank	Pfam	E-value
Genes ortólogos 1:1:1							
ache2	GB14873	8	1821	5	NP_001035320	COesterase	1.3e-227
						Abhydrolase_3	1.8e-08
or83b	GB19990	1	1317	7	–	7tm_6	2.8e-19
lw-rh	GB19657	15	1134	5	AF091732	7tm_1	2.2e-67
Família gênica específica de abelha							
mrjp1	GB14888	11	1430	6	AF000633	MRJP	2.4e-169
mrjp2	GB16246	11	1544	6	AF000632	MRJP	6.8e-167
mrjp3	GB16459	11	1830	6	Z26318	MRJP	3.9e-168
mrjp4	GB11768	11	1612	6	Z26319	MRJP	7e-145
mrjp5	GB10622	11	1966	6	AF004842	MRJP	8.7e-138
mrjp6	GB13789	11	1529	6	AY313893	MRJP	1.1e-151
mrjp7	GB11022	11	1427	6	BK001420	MRJP	1.3e-149
mrjp8	GB14639	11	1329	6	AY398690	MRJP	3.3e-118
mrjp9	GB16324	11	1793	6	DQ000307	MRJP	6e-122

Os 12 genes foram anotados no genoma de *A. mellifera* a partir das predições (OfficialSet) e de fragmentos de cDNA depositados no GenBank, exceto *or83b* que até o presente momento não possui nenhuma seqüência expressa disponível nesta base e foi anotado a partir da seqüência predita. A organização estrutural dos genes e a localização de seus respectivos domínios funcionais estão representados no diagrama da Figura 3.15.

As sete espécies de abelhas selecionadas para este estudo pertencem a quatro tribos de Apinae, sendo quatro espécies (*A. mellifera*, *A. cerana*, *A. dorsata*, *A. florea*) da tribo Apini, uma

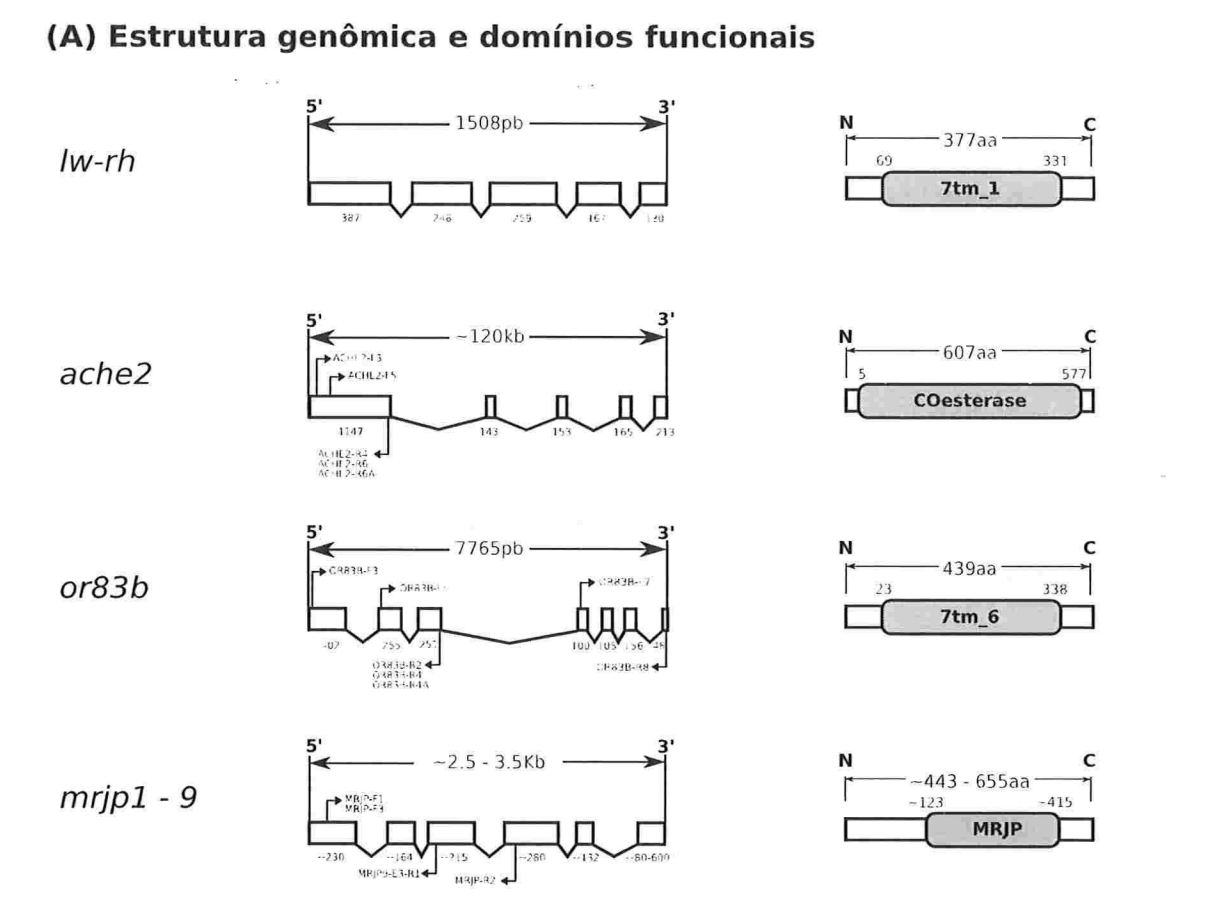


Figura 3.15: Diagrama da estrutura genômica dos genes *ache2*, *or83b*, *lw-rh* e *mrjp*s. (A) Arquitetura dos genes selecionados como calibradores do relógio molecular em abelhas. Todas as regiões em que os iniciadores foram desenhados estão representadas, bem como as regiões de seqüência codificadora utilizadas para a análise comparativa entre as sete espécies de abelhas. (B) Agrupamento dos genes da MRJPs no cromossomo 11. Estes genes codificam proteínas da geléia real e pertencem a uma família específica de abelhas que evoluiu a partir da *yellow-e3* (DRAPEAU et al., 2006a).

espécie de Meliponini (*M. quadrifasciata*), uma espécie de Bombini (*B. terrestris*) e uma espécie de Euglossini (*E. nigrita*). Estas abelhas possuem duas características muitos interessantes que

permitem estudos sobre a evolução de genes a partir de taxas evolutivas: (1) estas abelhas não divergem mais do que 100 Ma (ASCHER; DANFORTH; JI, 2001; MICHEL-SALZAT; CAMERON; OLIVEIRA, 2004; DANFORTH et al., 2006), o que possibilita usar substituições silenciosas sem o problema de saturação causado por múltiplas substituições e (2) são abelhas monofiléticas que possuem diferentes níveis de organização social que vão do modo de vida solitário ao modo vida eusocial. Deste modo, estudos sobre evolução molecular nestas espécies da subfamília Apinae podem ajudar a compreender melhor os eventos de duplicação gênica e a evolução de fenótipos relevantes para o surgimento da eusocialidade nestes insetos.

3.3.2 Identificação, clonagem e seqüenciamento de fragmentos da região genômica dos genes *mrjp*, *ache2* e *or83b* de sete espécies de abelhas

Um total de 30 contigs não redundantes são resultados da identificação, clonagem e seqüenciamento de fragmentos de DNA genômico derivados de regiões dos genes *mrjp*, *ache2* e *or83b* nas sete espécies de abelhas estudadas (Tabela 3.8 e 3.9). Os fragmentos foram amplificados por PCR a partir da combinação de vários iniciadores desenhados em regiões conservadas dos genes (Figura 3.15 e Tabela 2.7 e 3.9). Destas 30 seqüências não redundantes, 24 são trechos de DNA genômico inéditos que foram anotados para identificação das regiões codificadoras usadas para análises evolutivas.

Tabela 3.8: Resumo dos resultados do seqüenciamento de fragmentos de DNA genônimo de 3 genes em 7 espécies de abelhas.

Total de seqüências	353
Filtrado por qualidade	12
Reads positivos	100
Contigs	28
Singlets	4
Não redundantes	30

Todas as seqüências de *A. mellifera* adquiridas neste estudo só serviram como controle para acertar as condições experimentais e não foram usadas para as análises posteriores. As seqüências utilizadas para esta espécie foram extraídas das anotação oficiais do genoma de abelha (The Honeybee Genome Sequencing Consortium, 2006). Em todas as espécies, exceto em *A. florea*, existem dois contigs para o gene *or83b* que alinham com duas regiões diferentes (região

5' e 3'). Apenas os fragmentos da região 3' foram identificados e seqüenciados em todas as sete espécie e, por isso, somente este trecho foi usado para comparações de *or83b* nas espécies de abelhas (or83b-3'; Tabela 3.9).

Tabela 3.9: Resultados do seqüenciamento de fragmentos de DNA genômico de *mrjp*, *ache2* e *or83b* em sete espécies de abelhas. (*) marca os contigs inéditos seqüenciados neste estudo.

Gene	GB	Contigs	Reads	Tamanho(pb)	Clones	Iniciadores
Apis mellifera						
mrjp3	GB16459	AM_Contig3	2	1428	AM4-MRJP-2205-A1-18	MRJP-F1/MRJP-R2
mrjp7	GB11022	AM_Contig2	2	1369	AM4-MRJP-2205-A1-10	MRJP-F1/MRJP-R2
mrjp9	GB16324	AM_Singlet5	1	750	AM4-MRJP-0506-A3-1	MRJP-F3/MRJP-R2
		AM_Singlet6	1	706	AM4-MRJP-0506-A3-1	MRJP-F3/MRJP-R2
ache2	GB14873	AM_Contig4	4	1068	AM4-ACHE2-1005-A2-2	ACHE2-F3/ACHE2-R4
					AM4-ACHE2-0306-A-2	ACHE2-F3/ACHE2-R4
or83b-5'	GB19990	AM_Contig1	2	792	AM4-OR83B-1506-A1-1	OR83B-F3/OR83B-R4
					AM4-OR83B-1506-A1-2	OR83B-F3/OR83B-R4
Apis cerana						
mrjp5	GB10622	AC_Contig1	2	1261	AC208-MRJP-0406-A1-2	MRJP-F1/MRJP-R2
mrjp9*	GB16324	AC_Contig3	3	2042	AC208-MRJP-0506-A3-1	MRJP-F3/MRJP-R2
					AC208-MRJP-0506-A3-1	MRJP-F3/MRJP9-E3-R1
ache2*	GB14873	AC_Contig2	2	1067	AC208-ACHE2-0207-A3-1	ACHE2-F3/ACHE2-R6
or83b-5'*	GB19990	AC_Contig4	3	1388	AC208-OR83B-1506-A1-1	OR83B-F3/OR83B-R4
					AC208-OR83B-1506-A1-2	OR83B-F3/OR83B-R4
or83b-3'*	GB19990	AC_Contig5	4	910	AC208-OR83B-0907-A2-1	OR83B-F7/OR83B-R8
					AC208-OR83B-0907-A2-2	OR83B-F7/OR83B-R8
Apis dorsata						
mrjp1*	GB14888	AD_Contig1	9	1486	AD2-MRJP-0406-A1-1	MRJP-F1/MRJP-R2
					AD2-MRJP-0406-A1-4	MRJP-F1/MRJP-R2
					AD2-MRJP-0406-A1-6	MRJP-F1/MRJP-R2
					AD2-MRJP-0406-A1-7	MRJP-F1/MRJP-R2
					AD2-MRJP-0406-A1-8	MRJP-F1/MRJP-R2
mrjp5*	GB10622	AD_Contig6	2	1423	AD2-MRJP-0406-A1-11	MRJP-F1/MRJP-R2
ache2*	GB14873	AD_Contig5	2	1071	AD2-ACHE2-0306-B-1	ACHE2-F3/ACHE2-R4
or83b-5'*	GB19990	AD_Contig7	3	1409	AD2-OR83B-1506-A1-1	OR83B-F3/OR83B-R4
					AD2-OR83B-1506-A1-2	OR83B-F3/OR83B-R4
or83b-3'*	GB19990	AD_Contig2	4	1122	AD2-OR83B-0907-A2-1	OR83B-F7/OR83B-R8
					AD2-OR83B-0907-A2-2	OR83B-F7/OR83B-R8
Apis florea						
mrjp5*	GB10622	AF_Contig3	2	1385	AF8-MRJP-0406-A1-1	MRJP-F1/MRJP-R2
mrjp6*	GB13789	AF_Contig2	11	1442	AF8-MRJP-0406-A1-3	MRJP-F1/MRJP-R2
					AF8-MRJP-0406-A1-4	MRJP-F1/MRJP-R2
					AF8-MRJP-0406-A1-7	MRJP-F1/MRJP-R2
					AF8-MRJP-0406-A1-12	MRJP-F1/MRJP-R2
					AF8-MRJP-0406-A1-8	MRJP-F1/MRJP-R2
					AF8-MRJP-0406-A1-9	MRJP-F1/MRJP-R2
					AF8-MRJP-0406-A1-10	MRJP-F1/MRJP-R2
mrjp9*	GB16324	AF_Contig5	3	1874	AF8-MRJP-0506-A3-1	MRJP-F3/MRJP-R2
					AF8-MRJP-0506-A3-1	MRJP-F3/MRJP9-E3-R1
ache2*	GB14873	AF_Contig1	10	1060	AF8-ACHE2-1707-A6-2	ACHE2-F5/ACHE2-R6A
					AF8-ACHE2-1707-A6-3	ACHE2-F5/ACHE2-R6A
					AF8-ACHE2-1707-A1-3	ACHE2-F3/ACHE2-R4
					AF8-ACHE2-1707-A4-2	ACHE2-F5/ACHE2-R6
					AF8-ACHE2-1707-A6-1	ACHE2-F5/ACHE2-R6A
or83b-3'*	GB19990	AF_Contig4	2	1126	AF8-OR83B-0807-A2-2	OR83B-F7/OR83B-R8
Melipona quadrifasciata						
ache2*	GB14873	MQ_Contig3	2	1059	MQ1-ACHE2-0306-A-1	ACHE2-F3/ACHE2-R4
or83b-5'*	GB19990	MQ_Contig1	4	1136	MQ1-OR83B-0807-A1-1	OR83B-F5/OR83B-R4A
					MQ1-OR83B-0807-A1-2	OR83B-F5/OR83B-R4A
or83b-3'*	GB19990	MQ_Contig2	4	975	MQ1-OR83B-0807-A2-1	OR83B-F7/OR83B-R8
					MQ1-OR83B-0807-A2-2	OR83B-F7/OR83B-R8
Bombus terrestris						
ache2*	GB14873	BT_Contig2	2	755	BT5-ACHE2-0207-A3-4	ACHE2-F3/ACHE2-R6
or83b-5'*	GB19990	BT_Singlet3	1	709	BT5-OR83B-1506-A2-2	OR83B-F3/OR83B-R2
		BT_Singlet4	1	572	BT5-OR83B-1506-A2-2	OR83B-F3/OR83B-R2
or83b-3'*	GB19990	BT_Contig1	4	1515	BT5-OR83B-0907-A2-1	OR83B-F7/OR83B-R8
					BT5-OR83B-0907-A2-2	OR83B-F7/OR83B-R8
Eulaema nigrita						
ache2*	GB14873	EN_Contig1	2	810	EN3-ACHE2-0207-B4-1	ACHE2-F5/ACHE2-R6
or83b-5'*	GB19990	EN_Contig2	2	807	EN3-OR83B-0807-A1-2	OR83B-F5/OR83B-R4A
or83b-3'*	GB19990	EN_Contig3	4	1352	EN3-OR83B-0807-A2-1	OR83B-F7/OR83B-R8
					EN3-OR83B-0807-A2-2	OR83B-F7/OR83B-R8

Alguns genes da família das MRJPs foram identificados somente nas abelhas da tribo

Apini e nenhum gene desta família foi amplificado com sucesso em Meliponini, Bombini ou Euglossini. É possível que tal ausência de fragmentos amplificados ocorra por causa da ineficiência dos iniciadores na amplificação dos genes das MRJPs. Outra hipótese seria a de que estes genes tiveram sua origem após a especiação de um ancestral da tribo Apini e por isso nenhum ortólogo foi encontrado nas outras tribos de abelhas. Além disso, alguns genes *mrjp* não foram identificados nem mesmo em Apini. Por exemplo, *mrjp9* foi identificado em Apini exceto em *A. dorsata* (três amostras de diferentes regiões da Ásia foram usadas sem nenhum fragmento amplificado). Outros *mrjp*s foram identificados em apenas uma (*mrjp8*), duas (*mrjp2*, *mrjp4*, *mrjp6* e *mrjp7*) ou três espécies de Apini (*mrjp1* e *mjrp9*). Os únicos genes que estão representados nas quatro espécies são *mrjp3* e *mrjp5*.

A seguir, análises filogenéticas e medidas de substituição de nucleotídeos serão utilizadas para ajudar na compreensão da evolução das famílias multigênicas em abelhas corbiculadas. Mesmo utilizando espécies muito próximas evolutivamente, nem sempre é possível identificar claramente quem são genes ortólogos ou parálogos por causa de padrões evolutivos complexos, tais como, evolução em concerto e nascimento-e-morte (NEI; ROONEY, 2005). Por esta razão, nosso estudo incluiu genes de diferentes famílias com ortologia 1:1 bem resolvida (*ache2*, *or83b* e *lw-rh*) e genes de uma mesma família que sofreram uma recente radiação (*mrjp*s).

3.3.3 Filogenia e composição de nucleotídeos dos genes *ache2*, *or83b*, *lw-rh* e *mrjp* em abelhas Apinae

Os dados usados para as análises evolutivas foram, em grande parte, seqüenciados neste estudo (Tabela 3.9) mas também foram incluídas outras seqüências já disponíveis em bases de dados públicas (ex: genes da família yellow/MRJP da vespa *N. vitripennis* e *lw-rh* das sete espécies de abelhas; ver seção 2.1.7 pg 62). As regiões codificadoras foram determinadas a partir da anotação dos fragmentos de DNA genômico com base na similaridade com os genes *mrjp*, *ache2*, *or83b* e *lw-rh* de *A. mellifera*. Somente as regiões comuns dos genes *ache2*, *or83b* e *lw-rh* das sete espécies de abelhas foram usadas para as inferências filogenéticas e cálculos de taxas evolutivas. Além disso, todos os alinhamentos de seqüências de nucleotídeos (nt) foram resolvidos pelo alinhamento reverso a partir das seqüências de aminoácidos (aa) evitando, assim, os erros de inserção e deleção resultantes dos algoritmos de alinhamento.

O alinhamento dos 29 fragmentos de 537 pb dos genes da família MRJP/Yellow de abelhas da tribo Apini e da vespa *N. vitripennis* mostra que os genes da proteína da geléia real tiveram sua origem muito recentemente em abelhas e os genes *Nvrjpl* de vespa não são ortólogos dos genes *mrjp* mas descendem de um ancestral comum derivado de um gene *yellow* (Figura 3.16). Os genes *mrjp9* e *mrjp1* foram encontrados em três das quatro espécies de Apini estudadas. Nem todos os genes *mrjp* foram encontrados nas quatro espécies de Apini, apenas *mrjp3* e *mrjp5*. Entretanto, é difícil identificar, por estes resultados, se esta ausência de ortólogos é resultados de problemas de amostragem ou do processo de nascimento-e-morte das duplicações gênicas (NEI; ROONEY, 2005).

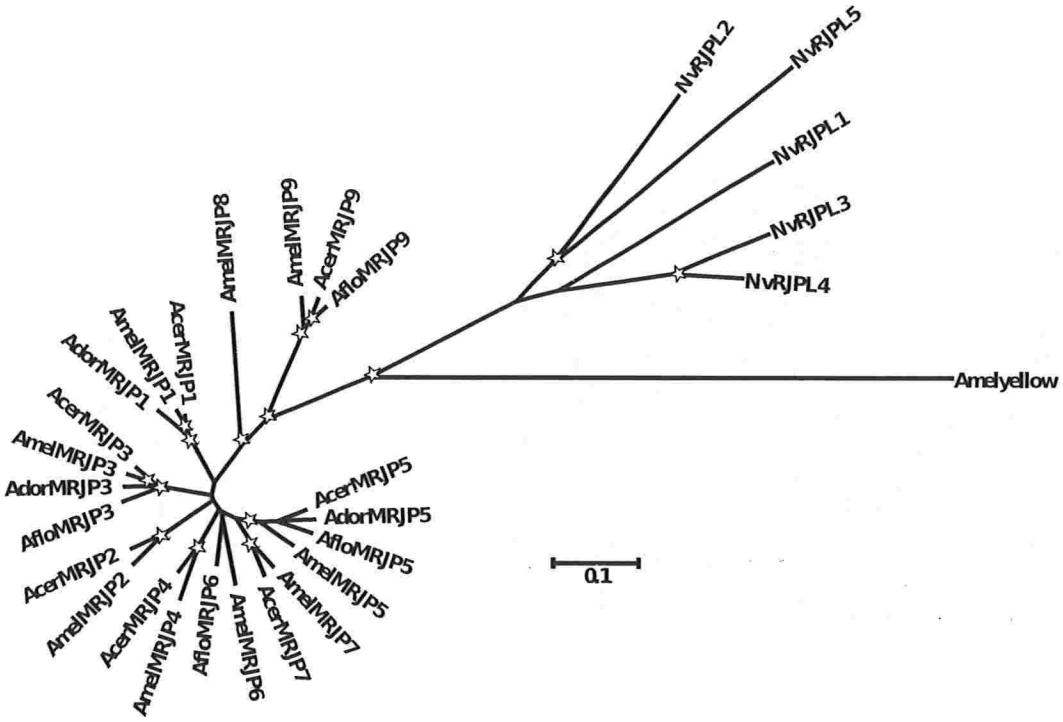


Figura 3.16: Filogenia dos genes da família das MRJP em abelhas e vespa. (⋆) suporte estatístico pelos métodos de MP e ML.

O alinhamento dos 7 fragmentos de 630 pb da região codificadora do gene *ache2* revelou um alto grau de conservação em seqüência de nucleotídeo e obviamente aminoácidos (80-93% de identidade em nt e 88-95% de identidade em aa; Figura 3.17, B). A composição de nucleotídeos em cada posição do códon mostrou diferença significativa no terceiro sítio entre as sete seqüências analisadas (χ^2 = 36,7, df = 18, P<0.01; Figura 3.17, A e B). A composição do primeiro e segundo sítios é homogênea nas sete seqüências de *ache2* o que significa que

estão variando de forma semelhante ao longo da evolução. A filogenia do gene *ache2* agrupou as espécies da tribo Apini (*A. mellifera*, *A. cerana*, *A. dorsata* e *A. florea*), enquanto *M. quadrifasciata*, *B. terrestris* e *E. nigrita* formaram um outro grupo com maior distância genética entre os tipos (Figura 3.17, C). As tribos Bombini e Euglossini teriam uma descendência comum e Meliponini estaria mais próxima de Apini.

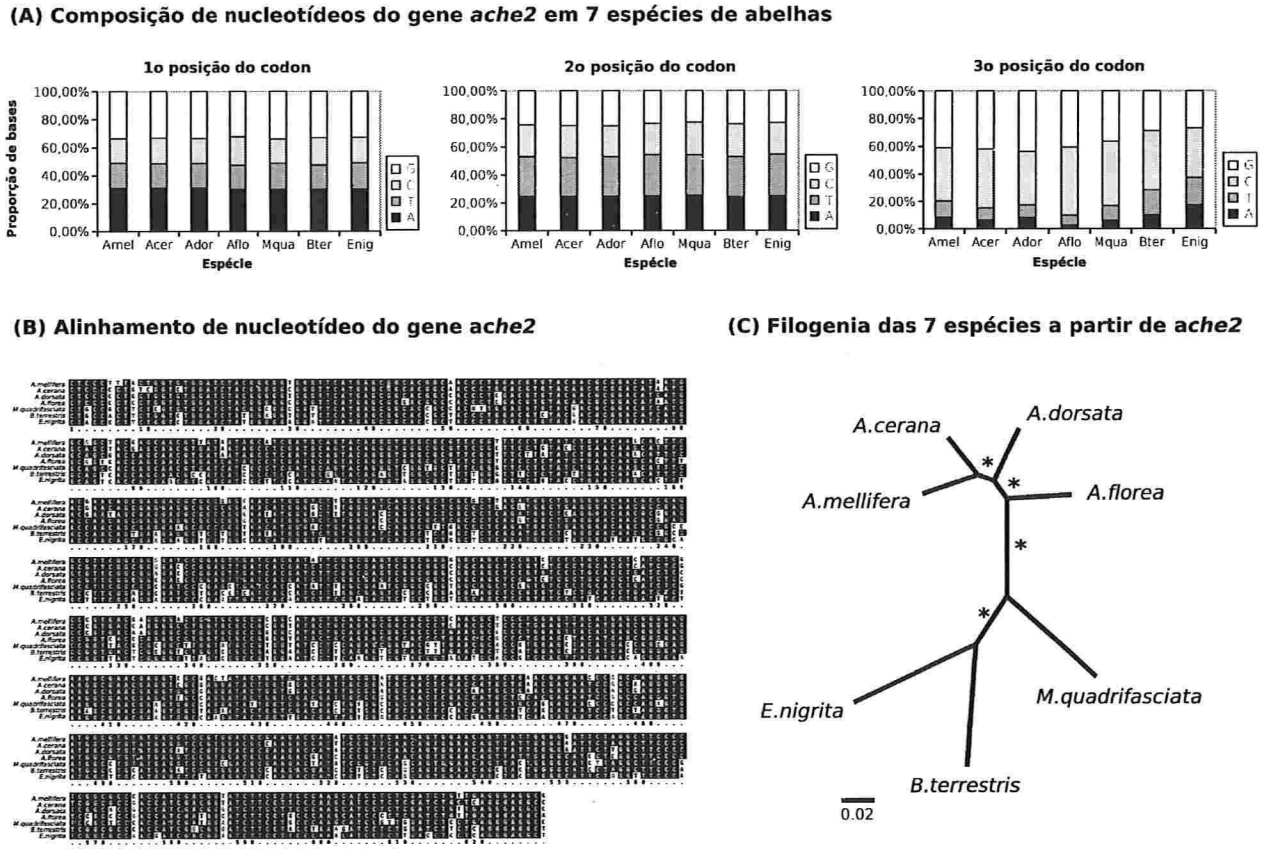


Figura 3.17: Análise das seqüências de DNA do gene *ache2* em 7 espécies de abelhas. (A) Composição de nucleotídeos em cada posição do códon. (B) Alinhamento múltiplo reverso das seqüências de DNA das 7 espécies de abelhas. (C) Filogenia das 7 espécies de abelhas a partir de *ache2*. (*) suporte estatístico pelos métodos de MP e ML.

Fragmentos de 369 pb correspondentes à região 3' do gene *or83b* também foram cuidadosamente alinhados e as mesmas análises realizadas no *ache2* foram conduzidas. As seqüências deste gene são muito conservadas nestas espécies (84-95% de identidade em nt e 96-99% de identidade em aa; Figura 3.18, B). A composição de nucleotídeos mostra as mesmas tendências observadas no *ache2*, ou seja, maior conservação no primeiro e segundo sítios e maior variação na terceira posição, embora não significativa ao nível de 5% (Figura 3.18, A e B). A reconstrução

filogenética a partir do *or83b* também agrupa as abelhas da tribo Apini, mas com a ordem um pouco diferente, colocando *A. florea* mais próxima de *A. cerana* e *A. mellifera*. Os representantes das outras três tribos foram agrupados com relações semelhantes ao *ache2*, exceto por *Melipona* e *Bombus* terem origem a partir de um ancestral comum (Figura 3.18, C).

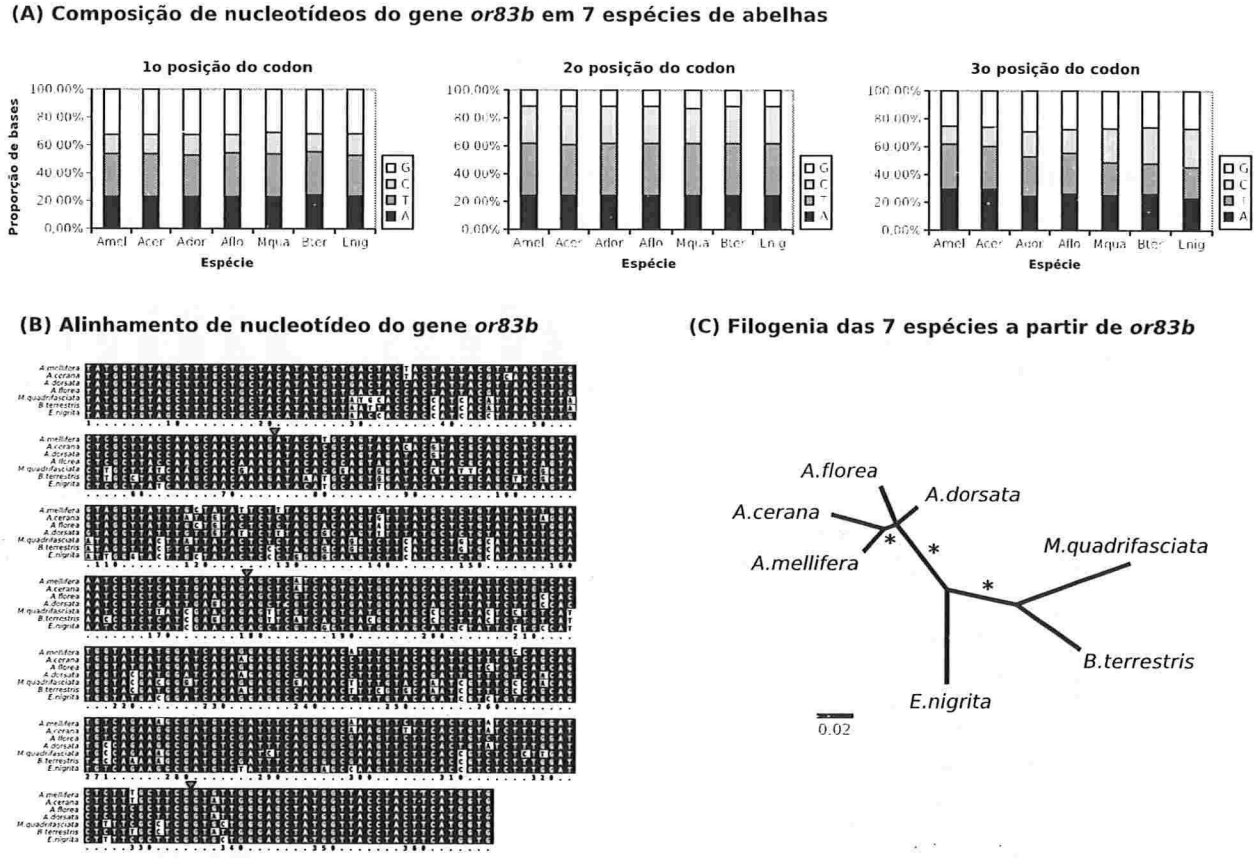


Figura 3.18: Análise das seqüências de DNA do gene *or83b* em 7 espécies de abelhas. (A) Composição de nucleotídeos em cada posição do códon. (B) Alinhamento múltiplo reverso das seqüências de DNA das 7 espécies de abelhas. (C) Filogenia das 7 espécies de abelhas a partir de *or83b*. (*) suporte estatístico pelos métodos de MP e ML.

Ainda que os dados de seqüência do gene *lw-rh* não sejam uma contribuição deste estudo, a forma como os dados foram analisados é inédita e, por esta razão, apresentamos os resultados sob as mesmas perspectivas dos outros dois genes (*ache2* e *or83b*). O grau de similaridade deste gene nas espécies de abelhas estudadas é muito alto, assim como nos outros dois genes já descritos (81-96% de identidade em nt e 85-99% de identidade em aa). A composição de nucleotídeos também apresentou maior variação na terceira posição do códon, embora o teste de homogeneidade não rejeite a hipótese nula (Figura 3.19, A e B). A filogenia de *lw-rh* mantém

uma topologia semelhante a *or83b*, onde espécies da tribo Apini foram agrupadas em um clado *A. mellifera*/*A. cerana* e um clado *A. florea*/*A. dorsata*, e as outras espécies representantes das três tribos (Meliponini, Bombini e Euglossini) se agrupam em clados mais distantes de Apini (Figura 3.19, C). Duas diferenças entre estas filogenias estão na resolução dos ramos. São elas: (1) *A. mellifera* e *A. cerana* seriam as espécies mais próximas do tipo ancestral; e (2) *Melipona* sp. e *E. nigrita* teriam se especiado de um ancestral comum já separado de Bombini (Figura 3.19, C).

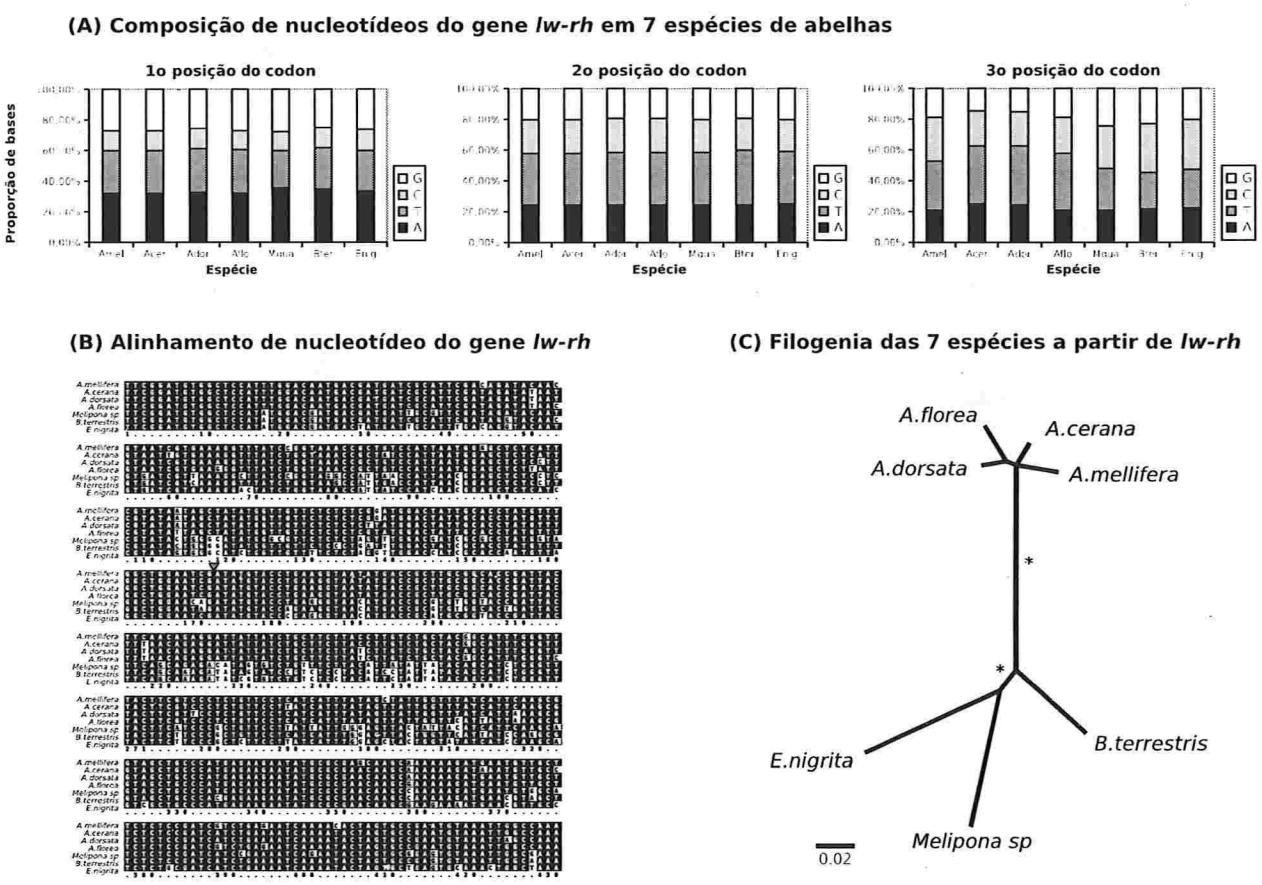


Figura 3.19: Análise das seqüências de DNA do gene *lw-rh* em 7 espécies de abelhas. (A) Composição de nucleotídeos em cada posição do códon. (B) Alinhamento múltiplo reverso das seqüências de DNA das 7 espécies de abelhas. (C) Filogenia das 7 espécies de abelhas a partir de *lw-rh*. (*) suporte estatístico pelos métodos de MP e ML.

Nos estudos de evolução molecular, as seqüências de DNA são mais informativas do que as de proteínas porque (1) existe DNA que não é traduzido (íntron, promotor) e (2) o código genético é degenerado. Por causa dessa degeneração, algumas substituições de nucleotídeos

são silenciosas e não resultam na substituição de aminoácido (KIMURA, 1977). As taxas de mutações sinônimas na terceira posição do códon devem ser maiores que as taxas de substituição de aminoácidos se a evolução neutra está ocorrendo (JUKES; CANTOR, 1969). As mutações no primeiro e segundo sítios do códon, na grande maioria da vezes, alteram o aminoácido codificado e o gene pode sofrer seleção purificadora já que mutações vantajosas são muito raras. Além disso, as taxas de substituições sinônimas ocorrem em taxas altas e aproximadamente constantes em diferentes tipos de genes (MIYATA; YASUNAGA; NISHIDA, 1980). Observamos variações nas seqüências de DNA de três genes diferentes entre as sete espécies de abelhas estudadas. A distribuição de nucleotídeos é mais conservada entre os genes ortólogos (polimorfismos) do que entre os diferentes genes (locos diferentes) de uma espécie.

As diferenças na filogenia reconstruída a partir da seqüência de cada gene podem ser resultado dos processos aleatórios da evolução, bem como de diferenças na história evolutiva das espécies (seleção, adaptação). As árvores de cada gene apresentam diferenças na organização de algumas ramificações (Figura 3.17, 3.18 e 3.19, C). As incongruências são observadas principalmente na resolução das ramificações mais derivadas. Um método eficiente para resolver estas discordâncias considera a informação genética de cada espécie em arranjos de seqüências concatenadas (ROKAS et al., 2003). As análises filogenéticas dos três genes concatenados (*ache2/or83b/lw-rh* com 1431 pb) resultaram em uma topologia única reconstruída por dois métodos, máxima parcimônia (MP) e máxima verossimilhança (ML) (Figura 3.20, D).

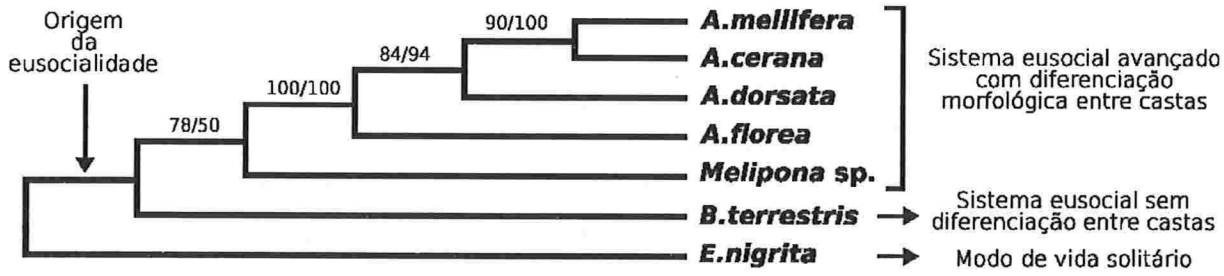


Figura 3.20: Filogenia das 7 espécies de abelhas corbiculadas pelo método de genes concatenados. A eusocialidade deve ter surgido no ancestral comum Bombini+Meliponini+Apini. Os valores de *bootstrap* dão suporte em todas as ramificações da árvore evolutiva por dois métodos (MP e ML).

O consenso filogenético obtido pelo método dos genes concatenados sugere que a eusocialidade nas abelhas corbiculadas teve origem no ancestral comum de Bombini, Meliponini e Apini e está em concordância com estudos anteriores (MICHENER, 1944; ENGEL, 2001, 2001;

ASCHER; DANFORTH; JI, 2001) (Figura 3.20). Embora não esteja entre os objetivos deste trabalho propor uma filogenia destas espécies de abelhas, é importante verificar que os genes usados neste estudo reforçam o modelo de evolução monofilética da eusocialidade nestas abelhas já proposto em trabalhos anteriores especializados na filogenia (DANFORTH, 2007). Outras premissas, tais como seleção neutra e homogeneidade nas taxas de mutação entre espécies são, a seguir, verificadas antes da aplicação das taxas de mutação na datação relativa dos eventos de duplicação gênica em *A. mellifera*.

3.3.4 Teste de neutralidade e de taxa relativa de mutação

Com a finalidade de usar as taxas de substituições sinônimas de nucleotídeos como relógio molecular em abelhas é importante verificar se os genes usados como calibradores atendem às premissas do modelo de evolução neutra. Um método para revelar a influência da seleção natural nas seqüências codificadoras testa a distribuição das variantes de uma população contra o modelo neutro (TAJIMA, 1989; FU; LI, 1993). Além da influência da seleção, os testes de taxa relativa são importantes para verificar se as taxas de mutação e o tempo de divergência entre espécies se acumulam de forma similares (WU; LI, 1985; TAJIMA, 1993).

Os 7 fragmentos dos 3 genes estudados (*ache2*, *or83b* e *lw-rh*) foram submetidos ao teste de neutralidade e nenhum destes genes apresentou evidências estatísticas para a rejeição da hipótese nula de evolução neutra (Tabela 3.10)

Tabela 3.10: Diversidade de nucleotídeos nos genes *ache2*, *or83b* e *lw-rh* de sete espécies de abelhas. A estatística D não rejeita a hipótese nula ao nível de 5% (TAJIMA, 1989; FU; LI, 1993). m, número de seqüências. n, número de sítios segregando. Eta, número total de mutações. π, diversidade de nucleotídeo por sítio. $\theta = \frac{S}{\sum_{i=1}^{n-1} \frac{1}{i}}$.

Gene	m	n	S	Eta	π	θ	D	
							Tajima	Fu-Li
ache2	7	630	183	261	0,1484	0,1691	-0,7203	-0,1744
or83b	7	369	86	94	0,1053	0,1040	0,0745	0,1801
lw-rh	7	432	122	141	0,1286	0,1332	-0,2015	-0,0723

Os testes de taxa relativa (TAJIMA, 1989) foram conduzidos a partir das seqüências (concatenadas) com diferentes níveis de relacionamento evolutivo. Os resultados dos testes de

taxa relativa para 5 pares de espécies está na Tabela 3.11. Nenhuma combinação de espécies demonstrou rejeição da hipótese nula de que a mutação se acumula aproximadamente constante entre as diversas espécies de abelhas corbiculadas.

Tabela 3.11: Teste de taxa relativa do relógio molecular em abelhas com corbícula.

Espécie1	Espécie2	Grupo externo	χ^2	df	P-valor
A.mellifera	*A.cerana*	*Melipona* sp.	0,02	1	0,9
A.mellifera	*A.dorsata*	*Melipona* sp.	0,69	1	0,4
A.mellifera	*A.florea*	*Melipona* sp.	2,14	1	0,14
Melipona sp.	*B.terrestris*	*E.nigrita*	2,38	1	0,12
A.mellifera	*A.florea*	*E.nigrita*	0,13	1	0,718

Assim, estes 3 locos gênicos não parecem sofrer influência significativa da seleção natural e evoluem sob seleção neutra. Além disso, a taxa de mutação destas abelhas também parece se manter similar ao longo da evolução, permitindo inferências sobre o tempo de divergência entre espécies e duplicações gênicas a partir da quantidade de mutações acumuladas entre as espécies de abelhas comparadas.

3.3.5 Cálculo de substituições de nucleotídeos e tempo de divergência entre as espécies de abelhas

Nos estudos de divergência evolutiva de seqüências de DNA, é geralmente necessário estimar o número de substituições sinônimas (silenciosas) e não sinônimas (mudança de aminoácido) separadamente. As taxas de substituições silenciosas são muito mais altas do que as de substituições não sinônimas. Além disso, a taxa de mutação sinônima é similar para muitos genes e espécies diferentes e pode ser usada para datar o tempo evolutivo entre espécies próximas (KIMURA, 1977; MIYATA; YASUNAGA; NISHIDA, 1980).

As taxas de substituição sinônima e não sinônima entre os genes *ache2*, *or83b* e *lw-rh* em 7 espécies de abelhas foram estimadas pelo método de Nei-Gojobori (NEI; GOJOBORI, 1986) e estão apresentadas na Figura 3.21 (A-C). Como esperado pelo modelo de evolução neutra, observamos que as taxas de substituição sinônima (K_s) são maiores dos que as taxas de substituição (K_n) não sinônima nos 3 genes considerados neste estudo. Além disso, os valores de K_s são menores entre as espécies de Apini e aumentam gradualmente em direção a Euglossini.

Tal observação parece estar em concordância com as distâncias evolutivas entre essas espécies de abelhas.

A calibragem de um relógio molecular foi feita pela aproximação de uma reta a partir de dois pontos com o tempo aferido por registro fóssil. O tempo de divergência entre as duas tribos de abelhas eusociais, Apini e Meliponini, foi estimado a partir de um fóssil de abelha da tribo Meliponini que data do final do Cretáceo (*Cretotrigona Prisca*, Época Mastrichiana ≈65 Ma) (ENGEL, 2000; DANFORTH, 2007). Desta maneira, a divergência entre as tribos de abelhas com corbícula deve ter se iniciado há mais de 75 Ma ainda no período Cretáceo, quando as angiospermas se diversificavam (ENGEL, 2001, 2001).

Assumindo que as taxas de substituição sinônima se comportam de forma aproximadamente constante ao longo da evolução destas abelhas, uma função linear pode ser usada para calcular o tempo de divergência a partir de K_s. A equação da reta que descreve esta função foi obtida a partir de dois pontos ($t_{Am/Am} = 0, K_s^{Am/Am} = 0.0$; $t_{Am/Me} = 70, K_s^{Am/Me} = 1,062$), onde $t_{Am/Am}$ e $K_s^{Am/Am}$ são os valores teóricos esperados para uma comparação Am/Am e $t_{Am/Me}$ e $K_s^{Am/Me}$ são valores estimados pelo tempo de divergência entre Apini e Meliponini e $K_s^{Am/Me}$ pela média das taxas de substituição sinônima entre Am/Me (Tabela 3.12).

Tabela 3.12: Taxas de substituição sinônima de cada gene e média em relação a *A. mellifera*.

Espécies	K_s^{ache2}	K_s^{or83b}	K_s^{lwrh}	Média±SE
Ac	0,275	0,187	0,166	0,209±0,033
Ad	0,406	0,201	0,155	0,254±0,077
Af	0,489	0,248	0,166	0,301±0,096
Me	1,285	0,999	0,903	1,062±0,114
Bt	1,779	0,721	0,752	1,084±0,347
En	3,651	0,637	1,185	1,824±0,927

A função linear que se aproxima do relógio molecular de abelhas com corbícula pode então ser expressa pela equação $T = 65,88K_s^{ij}$, onde T é o tempo de divergência (em Ma) e K_s^{ij} a taxa de substituição sinônima para um par de seqüências i, j (Figura 3.22, A). A partir do tempo de divergência e K_s estimados para cada uma das seis comparações com *A. mellifera* é possível verificar que *ache2* tem uma taxa de substituição mais acelerada que os outros dois genes (*or83b* e *lw-rh*) (Figura 3.22, B). Uma árvore filogenética *neighbor-joining* linearizada pela taxa evolutiva absoluta calculada como $V_s = \frac{K_s}{2T}$ (KIMURA, 1980), é construída assumindo taxas constantes

(A) Taxa de substituição de nucleotídeos em ache2 de abelhas

	Amel	Acer	Ador	Aflo	Mqua	Bter	Enig
Amel		0,0084	0,0052	0,0095	0,0255	0,0377	0,0255
Acer	0,2756		0,0052	0,0095	0,0244	0,0333	0,0234
Ador	0,4063	0,3296		0,0063	0,0244	0,0322	0,0223
Aflo	0,4888	0,4382	0,3992		0,0201	0,0235	0,0159
Mqua	1,2854	1,3324	1,2306	0,9225		0,0127	0,0084
Bter	1,7791	1,9092	1,5649	1,9518	1,2353		0,0170
Enig	3,6515	2,1865	3,2796	2,8119	2,0091	0,9263	

Ka — **Ks**

(B) Taxa de substituição de nucleotídeos em or83b de abelhas

	Amel	Acer	Ador	Aflo	Mqua	Bter	Enig
Amel		0,0072	0,0000	0,0000	0,0219	0,0145	0,0036
Acer	0,1869		0,0072	0,0072	0,0293	0,0219	0,0109
Ador	0,2012	0,1722		0,0000	0,0219	0,0145	0,0036
Aflo	0,2479	0,2167	0,1736		0,0218	0,0145	0,0036
Mqua	0,9988	1,1263	0,9129	1,0065		0,0256	0,0182
Bter	0,7211	0,8406	0,8751	0,7255	0,4348		0,0109
Enig	0,6373	0,7736	0,4733	0,6949	0,6876	0,8050	

Ka — **Ks**

(C) Taxa de substituição de nucleotídeos em lw-rh de abelhas

	Amel	Acer	Ador	Aflo	Mel	Bter	Enig
Amel		0,0030	0,0152	0,0091	0,0863	0,0663	0,0611
Acer	0,1661		0,0183	0,0122	0,0898	0,0697	0,0645
Ador	0,1550	0,0732		0,0121	0,0846	0,0608	0,0726
Aflo	0,1667	0,1538	0,1190		0,0813	0,0609	0,0661
Mqua	0,9036	0,7236	0,8302	1,0293		0,0385	0,0435
Bter	0,7526	0,7214	0,6946	0,7435	0,5680		0,0213
Enig	1,1849	0,9678	1,1703	1,5153	0,6573	0,8214	

Ka — **Ks**

Figura 3.21: Taxa de substituições de nucleotídeos dos genes *ache2*, *or83b* e *lw-rh* em 7 espécies de abelhas com corbícula. As substituições sinônimas (K_s) estão representadas abaixo da linha diagonal da matriz e as não sinônimas (K_n) estão acima desta linha . Taxa de substituição sinônima e não sinônima do gene *ache2* (A), *or83b* (B) e *lw-rh* (C). O retângulo cinza evidencia as taxas de substituição silenciosa em relação a *A. mellifera*. Amel, *A. mellifera*. Acer, *A. cerana*. Ador, *A. dorsata*. Aflo, *A. florea*. Mqua, *M. quadrifasciata*. Mel, *Melipona* sp.. Bter, *B. terrestris*. Enig, *E. nigrita*.

entre as linhagens e *E. nigrita* como grupo externo (Figura 3.22, C). Os valores máximos e mínimos de tempo de divergência são calculados pelas equações, $T_{min} = 79,12(K_s^{ij} - e)$ e $T_{max} = 55,21(K_s^{ij} + e)$, respectivamente, onde e é o erro padrão da média de K_s^{ij} dos três genes estudados (Figura 3.22, C).

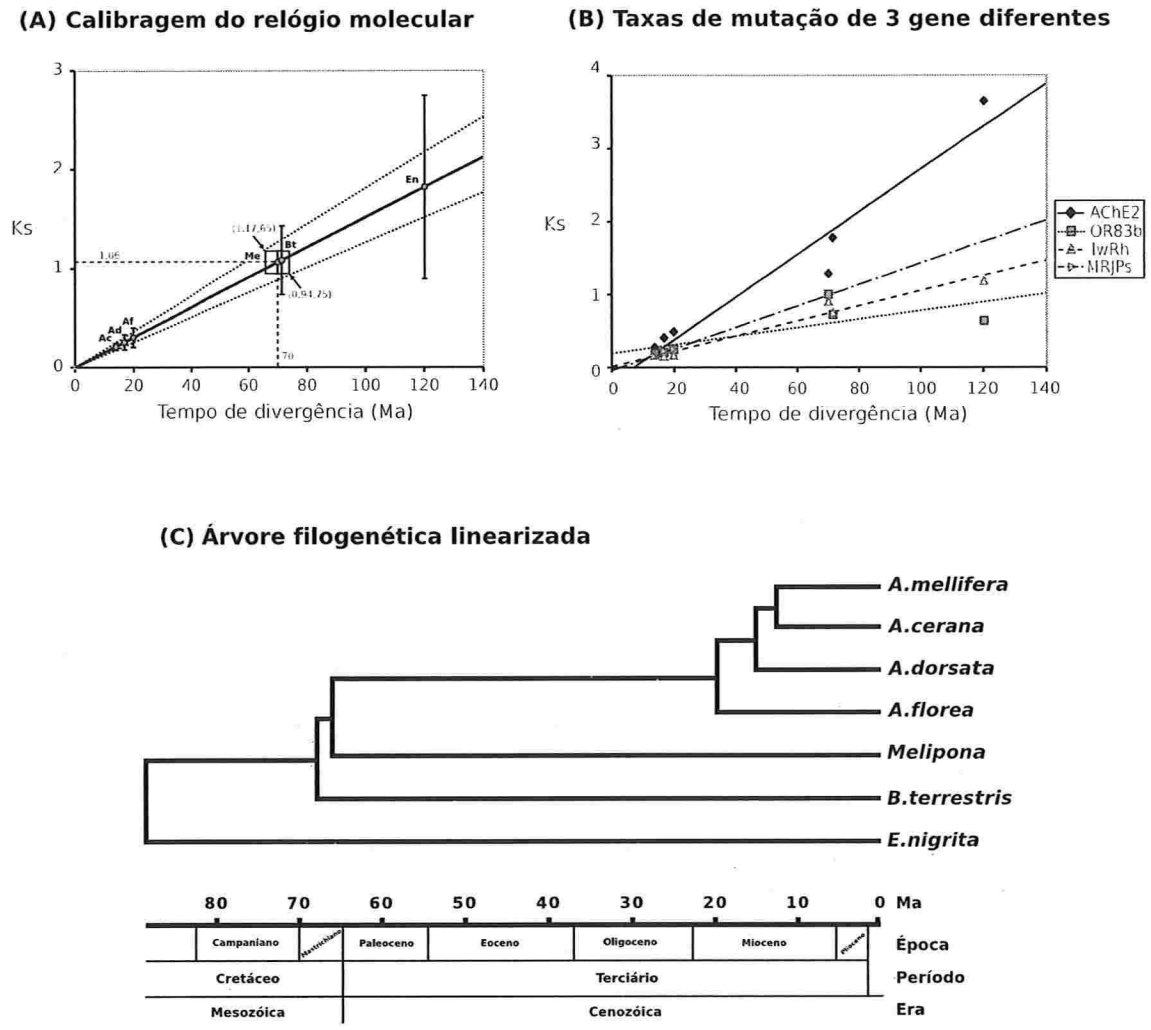


Figura 3.22: Tempo de divergência entre as sete espécies de abelhas com corbículas. (A) A equação da reta, $T(K_s) = 65,88K_s$, é obtida a partir de dois pontos calibradores Am/Am (0,0) e Am/Me (1.06,70). As outras duas retas formando um cone representam os dois extremos considerando os erros de tempo e K_s. (B) Taxas de mutação dos genes estudados apresentam maiores diferenças entre as espécies mais distantes evolutivamente. (C) Árvore filogenética *neighbor-joining* linearizada pela taxa evolutiva absoluta estimado pelos valores de $\bar{K}_s$ e tempo de divergência estimado entre as 6 espécies e *A. mellifera*

Os dados moleculares revelam que o tempo de divergência entre *A. mellifera* e *A. cerana*

ocorreu há ≈11,5-16,5 Ma, *A. mellifera* e *A. dorsata* se separaram há ≈14-20 Ma e finalmente *A. florea* possivelmente se especiou há ≈16,5-23,8 Ma. Estes resultados sugerem que a diversificação das espécies de Apini consideradas neste estudo ocorreu durante a primeira metade do Mioceno (Figura 3.22, C). A divergência entre *A. mellifera* e *Melipona* sp deve ter ocorrido há ≈58,6-84 Ma, entre *A. mellifera* e *B. terrestris* há ≈59,8-85,7 Ma e, finalmente, *A. mellifera* e *E. nigrita* há ≈100,7-144,3 Ma. Tais resultados sugerem que as tribos Bombini, Meliponini e Apini devem ter se radiado em um período muito próximo durante o fim do Cretáceo, quando as angiospermas ainda experimentavam uma grande diversificação. A divergência da tribo Euglossini, de modo de vida solitário, é datada há pelo menos 100 Ma, quando as abelhas com corbícula tiveram sua origem (Figura 3.22, C). Desta maneira, sugerimos que a evolução do sistema social em abelhas corbiculadas surge entre 70-80 Ma.

Estudos anteriores a partir de registros fósseis (ENGEL, 2001, 2001, 2006) sugerem que Apini deve ter se originado no início do Oligoceno e se radiou pela Eurásia durante este período. A distribuição deste grupo deve ter sido limitada pelas condições climáticas já que os ninhos eram construídos em espaços abertos expostos às condições do tempo. O comportamento de construção do ninho em cavidades é observado em *A. mellifera* e *A. cerana*. Esta apomorfia (caráter derivado) deve ter evoluído no ancestral destas abelhas durante o período de resfriamento climático da terra, que se acentuou no meio do Mioceno (≈14 Ma) permitindo que estas linhagens ocupassem regiões de clima temperado. A origem das abelhas é sugerida como tendo ocorrido a ≈125 Ma e as abelhas com corbícula teriam surgido a ≈ 90 Ma (ENGEL, 2001, 2006). Os tempos de divergências destas abelhas estimados pela datação molecular estão muito próximos das datações por registros fósseis e podemos então usar este modelo para datar duplicações gênicas ocorridas no genoma de *A. mellifera*.

3.3.6 Aplicando o relógio molecular para datação de duplicações gênicas em *A. mellifera*

Algumas famílias multigênicas apresentam diferenças no número de duplicações (parálogos) quando comparadas principalmente a outros insetos, tais como *D. melanogaster* e *A. gambiae* (The Honeybee Genome Sequencing Consortium, 2006). Os genes de maior interesse neste estudo são aqueles com potencial influência na evolução dos diferentes níveis de eusocialidade

observados nas abelhas das tribos Bombini, Meliponini e Apini. Genes da família das P450, principalmente do clado CYP3 (subdividido em CYP6 e CYP9), sofreram eventos de duplicações recentes e são expressos em vários tecidos, tais como cérebro e antenas, além de alguns homólogos em outros insetos serem importantes na resistência a substâncias tóxicas (ex: piretróides, DDT, neonicotinóides) (CLAUDIANOS et al., 2006). Outra família com um grande número de duplicações específicas de abelhas é a de receptores de olfato. Esta expansão tem sido relacionada com a notável habilidade olfatória destes insetos que inclui a percepção de várias combinações de feromônios, de sinais de reconhecimento de parentesco e da diversidade de odores florais (ROBERTSON; WANNER, 2006). Uma das famílias mais interessantes e importantes na evolução da eusocialidade avançada são as MRJPs. Estes genes da família MRJP/Yellow se originaram de *yellow-e3* e apresentam padrões de expressão complexos em diferentes estágios do desenvolvimento dos diferentes sexos e castas de *A. mellifera* (DRAPEAU et al., 2006a).

As taxas de substituição sinônimas (K_s) foram estimadas para todos os pares de alinhamento dos grupos MRJP, CYP6, CYP9 e OR. O tempo de divergência dos genes homólogos conservados foi então calculado por $T = 65,88K_s$. A taxa absoluta de substituição sinônima ou taxa evolutiva é calculada por $V_s = \frac{K_s}{2T}$ (KIMURA, 1980) (nas abelhas estudas $V_s = 7,6 \cdot 10^{-9}$ substituições por nucleotídeo/ano). As árvores filogenéticas destes grupos de genes foram reconstruídas por *neighbor-joining* considerando somente a terceira posição do códon e o modelo de substituição de nucleotídeos com correção para múltiplas substituições (JUKES; CANTOR, 1969) e, em seguida, foram linearizadas pela taxa evolutiva, mostrando o tempo médio de divergência entre os parálogos (Figura 3.23, 3.24).

O grupo CYP6 parece ter 2 duplicações muito recentes que podem ter se originado após a divergência de *A. cerana* e *A. mellifera* (6AS1/6AS2 e 6AS17/6AS18; Figura 3.23, A). Outras 5 duplicações parecem ter ocorrido especificamente nas abelhas da tribo Apini (Figura 3.23, A). As CYP9 são um pouco mais antigas e datam de um período logo após a divergência de Meliponini+Apini (Figura 3.23, B).

Os genes da família MRJP provavelmente se duplicaram recentemente, após a divergência Meliponini+Apini e suporta o fato não termos identificado estes genes em nenhuma espécie fora da tribo Apini (Figura 3.23, C). Estas proteínas compõem 90% das proteínas da geléia real e esta substância é produzida pela glândula mandibular, secretada por operárias nutrizes

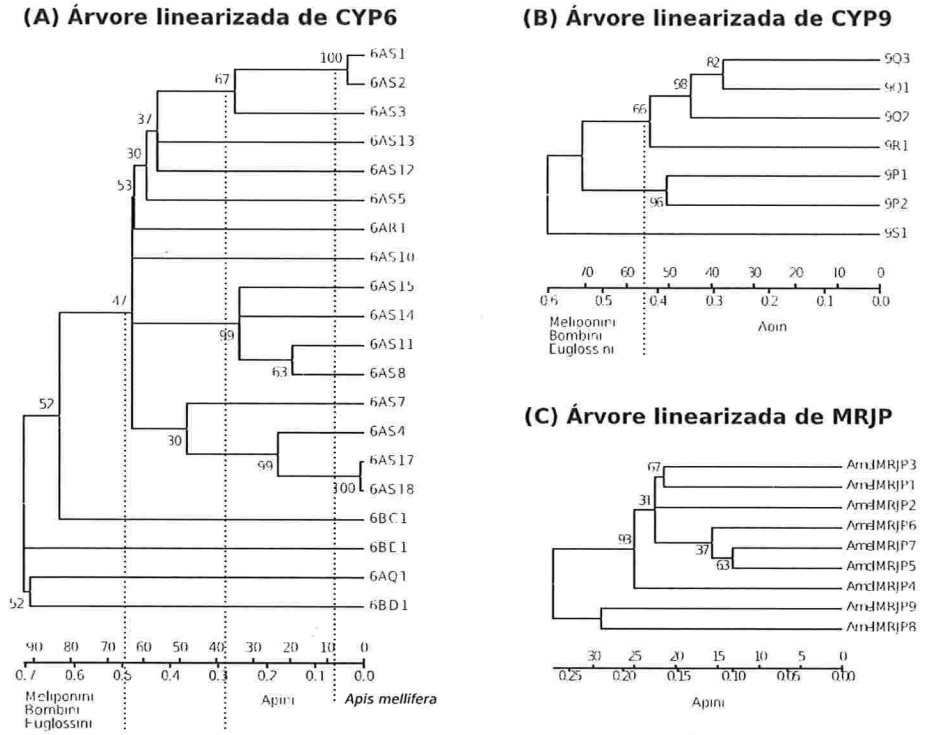


Figura 3.23: Tempo de divergência das duplicações de CYP6, CYP9 e MRJP

que alimentam continuamente as larvas em desenvolvimento (SCHMITZOVA et al., 1998). Em Meliponini, o alimento larval também é produzido pela secreção de uma glândula de operárias que é misturada com pólen e então regurgitado em uma célula. A rainha oviposita sobre o alimento e a célula é em seguida fechada até o momento da emergência do adulto. Esta diferença na estratégia reprodutiva entre Meliponini e Apini ilustra uma das várias alternativas de manutenção de um sistema eusocial avançado. Embora alguns genes homólogos de MRJP tenham sido encontrados na vespa *N. vitripennis*, eles não são ortólogos de nenhuma MRJP e devem ter evoluído independentemente a partir de um ancestral comum destas MRJP. Além disso, nenhum gene das MRJP foi identificado em *M. quadrifasciata* e as proteínas sintetizadas pelas glândulas deste Meliponini não seriam ortólogas, mas talvez alguns genes homólogos podem ter evoluído como observado em *Nasonia*.

A função do gene *yellow* em *Drosophila* tem sido atribuída ao processo de pigmentação e comportamento de corte do macho, sendo controlado, pelo menos em parte, pela via de determinação do sexo (DRAPEAU et al., 2003). As 9 MRJPs e um pseudogene (ψ*mrjp*) tiveram sua origem de *yellow-e3* e alguns destes têm papel importante na nutrição das larvas *A.*

mellifera. Entrentanto, é possível que estes genes tenham evoluído funções adicionais. Algumas MRJPs são expressas em rainhas, operárias e zangões (*mrjp1* e *mrjp3*), outras são expressas exclusivamente na glândula hipofaringeana de operárias (*mrjp2*, *mrjp4*, *mrjp5*, *mrjp6* e *mrjp7*) e a MRJP9 tem sido encontrada também na glândula de veneno de operárias. Os genes das MRJPs foram inicialmente descritos como relacionados com a nutrição, mas atualmente sabe-se que estes genes apresentam padrões de expressão específicos de fases do desenvolvimento, tecido, casta e sexo. A diversidade de funções dos genes da família MRJPs parece influenciar alguns pontos críticos da organização social das abelhas, modulando a fertilidade sexo-específica dos indivíduos de uma colônia, bem com a transição do comportamento das operárias nutrizes e forrageiras (DRAPEAU et al., 2006a, 2006b).

Os receptores de olfato (OR) apresentam uma notável quantidade de duplicações em *A. mellifera*, que devem ter ocorrido recentemente na evolução destas abelhas. Pelo menos 8 eventos de duplicação podem ter ocorrido especificamente em *A. mellifera* (Figura 3.24). Outras várias duplicações também indicam que estes eventos devem ter acontecido ao longo da evolução de Apini e muitos destes novos genes podem estar relacionados à capacidade destas abelhas de manutenção da estrutura social de suas colônias por meio de comunicação modulada pelo olfato bem como com a discriminação de um largo espectro de odores florais que podem variar com os diferentes ambientes e sazonalidade. Este grande número de genes OR pode estar intimamente relacionado com a evolução da vida social avançada nestes insetos.

3.3.7 Conclusão

Os estudos das taxas evolutivas em abelhas tiveram como principal objetivo a datação dos eventos de duplicação gênica através das taxas de substituição sinônima em genes de abelhas corbiculadas. Os genes, *mrjps*, *ache2*, *or83b* e *lw-rh*, muito distintos em estrutura, função e localização foram escolhidos como marcadores para a estimativa de taxas evolutivas. Nenhum dos nove genes da família MRJP foram encontrados nas abelhas fora da tribo Apini. Além disso, apenas alguns ortólogos foram encontrados nas quatro espécies de *Apis*. As taxas médias de mutação silenciosa dos genes *ache2*, *or83b* e *lw-rh* nas sete espécies de abelhas estudadas foram usadas para calcular o tempo de divergência entre estas espécies, assim como os eventos de duplicação gênica mais recentes em *A. mellifera*.

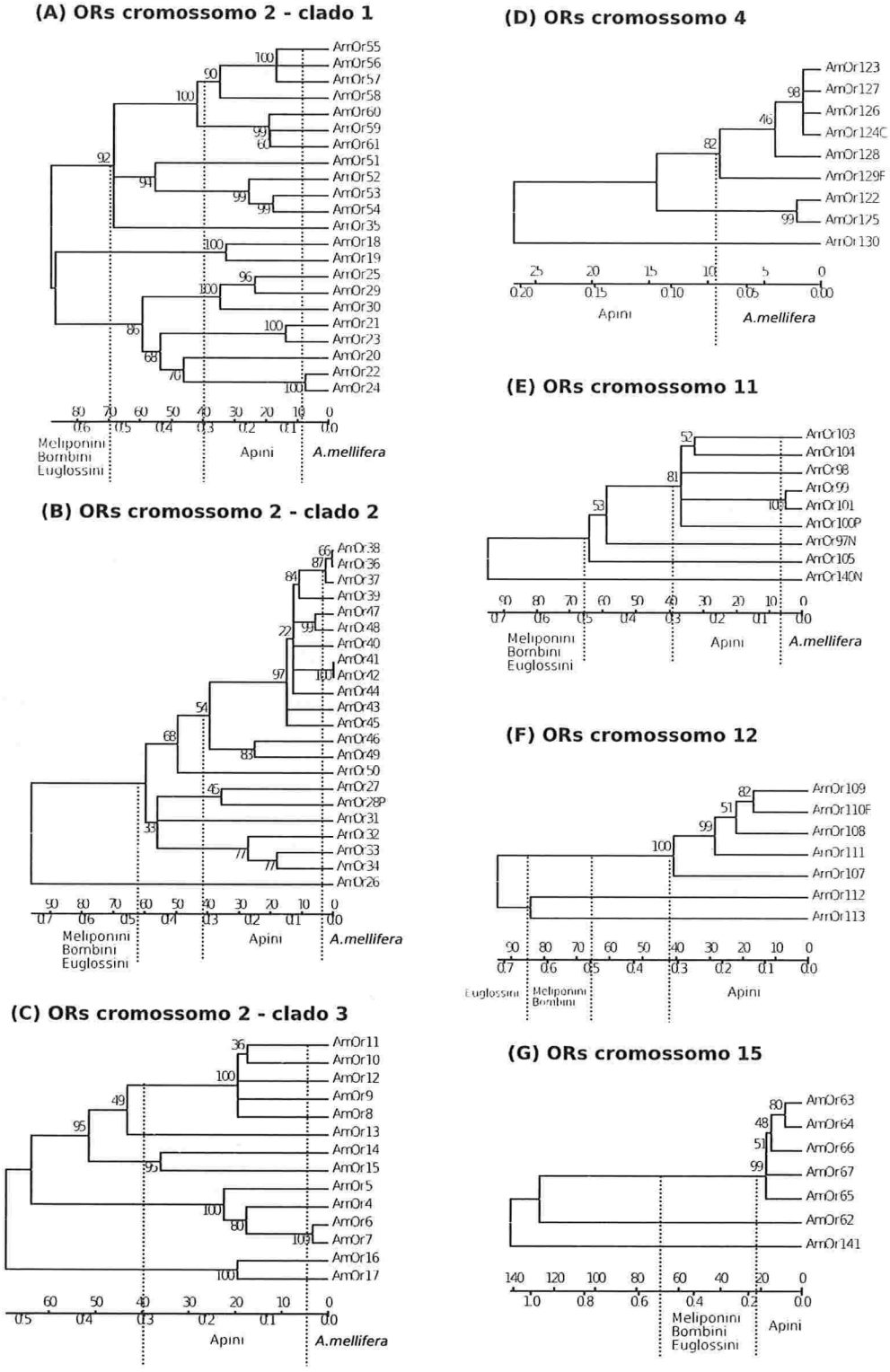


Figura 3.24: Tempo de divergência das duplicações de ORs

O tempo de divergência estimado pelo relógio molecular destas abelhas está em total concordância com as estimativas feitas a partir dos registros fósseis. Desta maneira, os resultados nos permitem reforçar a hipótese de origem comum do sistema social nestas abelhas a partir de um ancestral de Bombini, Meliponini e Apini. Além disso, podemos dizer que as abelhas corbiculadas devem ter surgido no final do Cretáceo juntamente com a grande expansão das angiospermas, enquanto as representantes atuais da tribo Apini evoluíram na primeira metade do Mioceno, quando o clima na terra ainda estava aquecido. A origem do comportamento de construção de ninho em cavidade (*A. mellifera* e *A. cerana*) deve ter surgido pouco antes do meio do Mioceno, quando a temperatura do planeta cai drasticamente e possibilita que estas abelhas ocupem regiões de clima temperado como a Europa.

As duplicações de alguns genes da família P450 (clados CYP6 e CYP9), MRJP e OR devem ter ocorrido principalmente após a divergência de Bombini, Meliponini e Apini. Além disso, alguns genes das CYP6 e OR teriam se originado especificamente em *A. mellifera*. Embora o perfil funcional destes novos genes ainda seja pouco conhecido, algumas evidências apontam para funções relacionadas à degradação de hormônios, feromônios e substâncias tóxicas (P450s), a nutrição e comportamento (MRJPs) e a comunicação e interação com o ambiente (ORs). Muitos dos genes destas famílias são expressos em várias partes do corpo ou tecidos específicos (ex: antenas, glândula hipofaringeana, corpos cogumelares, glândula de veneno, entre outros). Além da expressão diferencial em tecidos dos diferentes sexos e castas, estes genes também mostram uma diversidade de expressão em diferentes estágios do desenvolvimento (CLAUDIANOS et al., 2006; DRAPEAU et al., 2006a; ROBERTSON; WANNER, 2006).

Contudo, evidências empíricas resultantes deste estudo sugerem que um grande número das duplicações mais recentes no genoma de *A. mellifera* está estritamente relacionado à estrutura eusocial avançada destes insetos, principalmente na tribo Apini. Entretanto, isto não significa que Meliponini e Bombini ou mesmo Euglossini não tenham duplicações gênicas destas famílias, e possivelmente isto deve ter acontecido sob os mecanismos estocásticos da evolução que são filtrados pela seleção natural, moldando o modo de vida destas abelhas coletoras de pólen. Entender os elementos genéticos hereditários que evoluem gradualmente por mutação, duplicação e recombinação para a realização da estrutura social nestes insetos é uma das possibilidades mais concretas para se entender como genes e seus produtos se integram para a formação de padrões fenotípicos complexos.

4 *Considerações finais*

O presente estudo foi conduzido principalmente sob as perspectivas da bioinformática com ênfase nos determinantes genéticos envolvidos na manutenção de um sistema reprodutivo que possibilita a manifestação do padrão eusocial em *A. mellifera*. Os três temas abordados neste estudo envolvem a determinação do sexo e de castas, e a evolução de famílias gênicas em abelhas solitárias e eusociais. Os mecanismo genéticos e processos moleculares que determinam geneticamente o sexo em abelhas por haplodiploidia (ausência de cromossomos sexuais), a plasticidade fenotípica na formação das castas como resultado das diferenças na composição do alimento larval e a evolução de genes com implicações na interação com o ambiente influenciando o modo de vida dos indivíduos são algumas das questões abordadas. As análises experimentais foram sempre delineadas com a finalidade de validar os resultados hipotéticos obtidos por bioinformática. Uma parte dos resultados deste estudo já estão publicados em periódicos científicos (CRISTINO et al., 2006a, 2006b; The Honeybee Genome Sequencing Consortium, 2006).

Os genes que formam os sinais iniciais da via de determinação do sexo estão pouco conservados mesmo em espécies próximas. Em *A. mellifera*, este sinal não depende de cromossomos sexuais, mas de apenas um loco gênico que codifica proteínas complementares, que regulam o *splicing* alternativo de um fator de transcrição conservado na cascata de determinação do sexo dos metazoários. Os fatores de transcrição estão entre as funções moleculares mais representadas entre os genes conservados da via de determinação do sexo. Estes genes reguladores são os principais responsáveis pelo controle da transcrição de genes que constituem a efetiva qualidade e quantidade de componentes genéticos responsáveis pela determinação e manutenção do fenótipo celular.

É importante observar que a diferenciação morfológica e comportamental específica de

cada sexo é, em primeira instância, determinada pelo controle de transcrição gênica alternativo. Em geral, estes fatores pleiotrópicos estão sob forte pressão de seleção já que alterações nas regiões codificadoras podem significar mudanças no controle de transcrição. As alterações nas regiões promotoras dos genes alvo (sob controle) destes fatores podem evoluir mais rapidamente resultando em disrupção de controle e conseqüentemente mudança de função. Por exemplo, a Vg não é uma proteína específica de fêmea em *A. mellifera* mas desempenha funções importantes no desenvolvimento de castas (diversidade dentro de um mesmo sexo), enquanto em *B. mori* esta proteína é expressa especificamente em fêmeas para fins reprodutivos. No entanto, um considerável número de genes alvo deve conservar elementos regulatórios de maneira que os padrões de desenvolvimento na formação dos órgão reprodutivos e outras estruturas possam garantir as diferenças complementares necessárias para completar o ciclo reprodutivo.

Na determinação de casta, o sinal inicial é dado pela composição de nutrientes do alimento larval que é sintetizado na glândula hipofaringeana de operárias. As operárias nutridoras decidem que larvas serão alimentadas para se tornarem rainhas férteis ou operárias estéreis. Diferenças nutricionais ativam programas genéticos distintos e ilustram quão importantes são as variáveis ambientais para a diferenciação de fenótipos alternativos a partir de um mesmo genótipo. Os genes da determinação de casta estão principalmente envolvidos com o metabolismo e os genes diferencialmente expressos em operária compartilham mais sítios conservados quando comparados aos genes de rainha. Tais diferenças nas regiões controladoras dos genes de casta demonstram que as redes genéticas de regulação variam entre as castas e que tal variação foi principalmente determinada por uma entrada diferencial de nutrientes seguida de diferenças nos títulos de hormônios (JH e Ecd).

A abelha *A. mellifera* vive exclusivamente em um sistema social altamente organizado. Alguns milhares de insetos, em sua grande maioria operárias estéreis, trabalham para a manutenção da colônia, desempenhando diversas tarefas ao longo do desenvolvimento adulto (polietismo etário). Apenas uma rainha produz ovos, sendo que os fecundados, em geral, se desenvolvem em fêmeas diplóides, enquanto os não fecundados se desenvolvem em machos haplóides. A produção de machos em uma colônia é controlada pelas operárias, que constroem células maiores para que a rainha oviposite um ovo não fecundado. Entretanto, as condições ambientais devem ser favoráveis. A produção de rainhas também é determinada pelas operárias. A alimentação de larvas de fêmeas exclusivamente com uma substância sintetizada na glândula

hipofaringeana de operárias resulta em uma rainha. Todos estes mecanismos de controle reprodutivo estão de alguma maneira codificados no genoma.

O sistema social das abelhas com corbícula evoluiu a partir de em um ancestral comum. Entre os representantes destes insetos polinizadores estão abelhas de modo de vida solitário, semi-social e eusocial. Algumas famílias multigênicas apresentam um considerável número de duplicações gênicas recentes no genoma de *A. mellifera*. Entre as expansões mais interessantes estão genes das famílias das ORs e MRJPs que estão relacionados com a interação, via olfato, com o ambiente e outros membros da colônia e com a nutrição das larvas para produção de rainha ou operária, respectivamente. As radiações destes genes parecem ter ocorrido principalmente na tribo Apini logo após a separação das abelhas da tribo Meliponini. Apesar disso, os Meliponini também se organizam em um sistema eusocial, mas com algumas diferenças marcantes, que vão da arquitetura dos ninhos ao comportamento social dos indivíduos, com implicações na reprodução das castas. Embora as funções destas duplicações recentes ainda sejam pouco conhecidas, algumas delas têm sido caracterizadas com perfis de expressão específicos de sexo e casta.

Contudo, entendemos que os resultados aqui apresentados caracterizam uma considerável contribuição para a compreensão da biologia destes magníficos insetos sociais. Além disso, as análises de bioinformática e os validações experimentais conduzidas neste estudo integraram vários conceitos teóricos em biologia, utilizando como organismo modelo a abelha *A. mellifera*. Esta integração de conceitos e ferramentas computacionais possibilita um gerenciamento efetivo e eficiente de grandes quantidades de dados biológicos transformando-os em informações úteis que aceleram o delineamento experimental adequado e a reunificação das ciências biológicas para a geração de um conhecimento científico aplicável a diversas situações nos sistemas vivos.

4.1 Perspectivas - Redes genéticas e os sistemas complexos

As perspectivas que se abrem a partir deste amplo estudo são várias. Experimentos que objetivem caracterizar a função dos genes identificados como potenciais participantes dos processos e mecanismos aqui descritos devem revelar acertos e erros de nossas predições orientando os próximos passos. Os estudos individualizados dos genes são muito importantes

para a compreensão da função de um gene em si, mas, possivelmente, não revelam como produtos de cada gene se integram para formar os padrões complexos dos fenótipos que observamos na natureza.

A partir da identificação de grupos de genes que compartilham funções moleculares e processos biológicos em comum, bem como grupos de genes de uma mesma família, pretendemos estudar como estes genes estão interconectados. A aquisição de uma nova função em uma duplicação gênica pode ser importante para a adaptação dos organismos e o surgimento de novidades funcionais que influenciam no modo de vida de um organismo. Entretanto, tais propriedades não surgem apenas de modificações da região codificadora, mas, principalmente, de variações nos elementos regulatórios responsáveis pelo controle de transcrição dos genes em diferentes estágios do desenvolvimento e tecidos que, em última instância, se manifestam diferencialmente nos sexos e castas.

Dois exemplos ilustram como os grupos funcionais ou famílias gênicas podem ser representados por redes complexas a partir da análise dos motivos regulatórios descobertos com o *pipeline* FindMotif desenvolvido neste estudo (Figura 4.1). Os motivos encontrados nos genes diferencialmente expressos nas castas são representados por uma rede bipartida onde genes e motivos estão interconectados de acordo com a ocorrência destes motivos nas regiões promotoras dos genes (Figura 4.1, A e B). As topologias destas duas redes são claramente distintas e quantificações das propriedades destas redes por métodos e conceitos de redes complexas serão utilizados para uma caracterização mais precisa destas diferenças. Entre os genes da famílias das MRJP também descobrimos motivos específicos destes genes (Figura 4.1, C) e podemos observar que os genes originados recentemente por eventos de duplicação compartilham um maior número de elementos regulatórios (Figura 4.1, D), o que é coerente com a filogenia reconstruída pela região codificadora (Figura 4.1, E).

A ilustração da Figura 4.1 é simplificada e seria necessária uma explicação mais detalhada para apresentar qualquer conclusão a partir dela. Os métodos aqui apresentados estão descritos em maiores detalhes em um trabalho recente desenvolvido por Barchuk e col. (submetido para BMC Developmental Biology). No referido trabalho, técnicas de *microarray* são empregadas para identificar genes diferencialmente expressos durante o período de mudança alimentar nas larvas e sob a presença de maiores concentrações de hormônio juvenil em larvas de operárias. Contudo, não faz parte dos objetivos deste estudo detalhar as análises das redes, mas apenas

ilustrar quais serão os próximos passos que orientarão estudos futuros em nossas pesquisas sobre expressão gênica e formação de padrões fenotípicos utilizando como organismo modelo a abelha *A. mellifera*.

4.2 Trabalhos publicados ou submetidos

Cristino *et al.*, **2006** A comparative analysis of highly conserved sex-determining genes between *Apis mellifera* and *Drosophila melanogaster*. Genet. Mol. Res., 5(1):154-68, 2006

Cristino *et al.*, **2006** Caste development and reproduction: a genome-wide analysis of hallmarks of insect eusociality. Insect Mol. Biol. 15(5): 703-14, 2006

Honeybee Genome Sequencing Consortium, 2006 Insights into social insects from the genome of the honeybee *Apis mellifera*. Nature, 443(7114): 931-949, 2006.

Barchuk *et al.*, **Submetido 2006** Molecular determinants of caste differentiation in the highly eusocial honeybee *Apis mellifera*. Submetido em 2006 para BMC Developmental Biology.

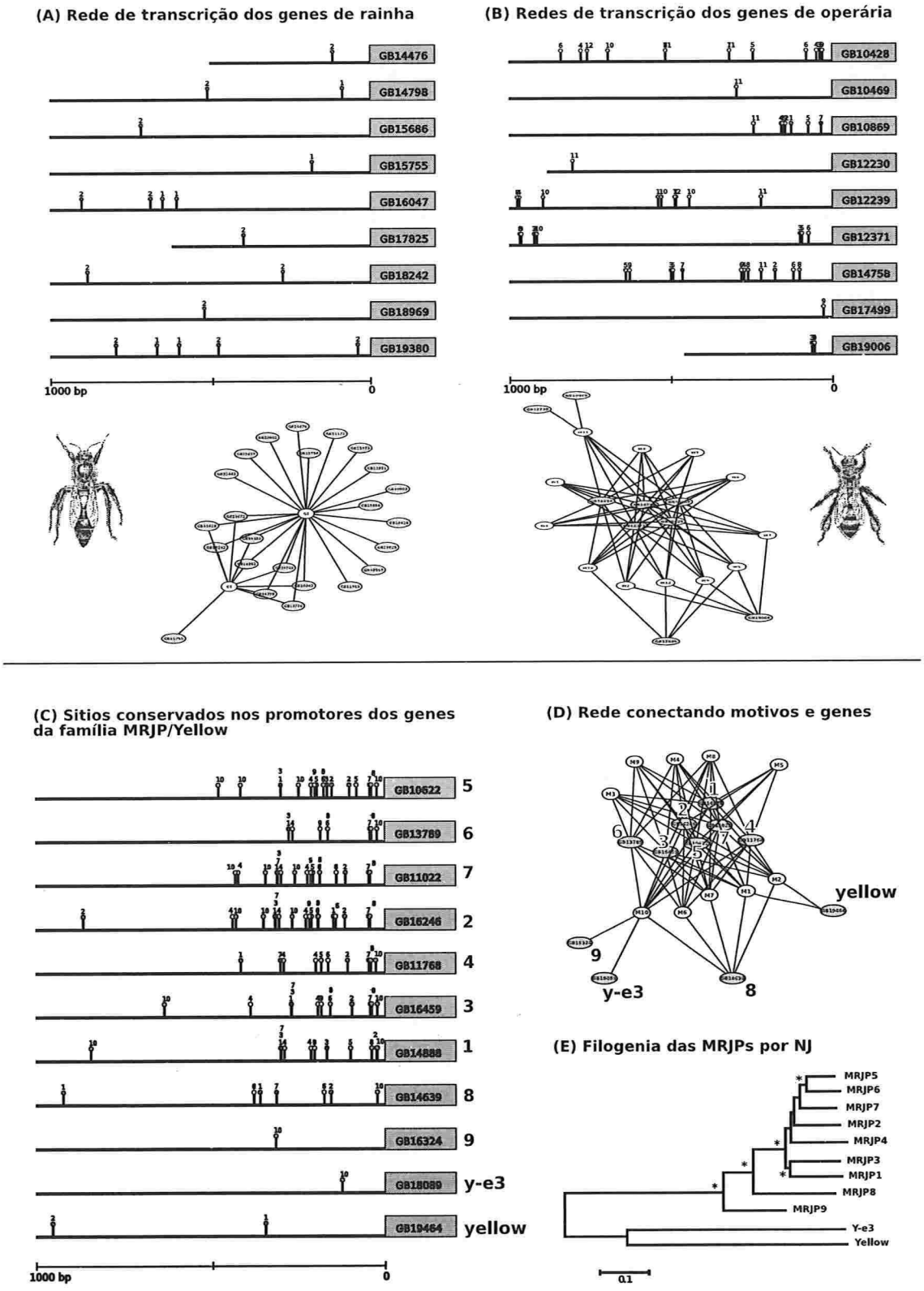


Figura 4.1: Diagrama ilustrando as perspectivas de estudos em redes genéticas por uma abordagem de redes complexas.

Bibliografia

ABOUHEIF, E.; WRAY, G. Evolution of the gene network underlying wing polyphenism in ants. *Science*, v. 297, n. 5579, p. 249–52, 2002.

AL-SHAHROUR, F.; DIAZ-URIARTE, R.; DOPAZO, J. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, v. 20, n. 4, p. 578–80, 2004.

ALBERT, R.; BARABÁSI, A. Statistical Mechanics of Complex Networks. *Rev. Modern Phys*, v. 74, n. 1, p. 47–97, 2002.

ALBERTOVA, V. et al. Organization and potential function of the *mrjp3* locus in four honeybee species. *J Agric Food Chem*, v. 53, n. 20, p. 8075–81, 2005.

ALBERTS, B. et al. *The molecular biology of the cell.* New York: Garland Publishing, 2002.

ALTSCHUL, S. et al. Basic local alignment search tool. *J Mol Biol*, v. 215, n. 3, p. 403–10, 1990.

AMDAM, G.; OMHOLT, S. The regulatory anatomy of honeybee lifespan. *J Theor Biol*, v. 216, n. 2, p. 209–28, 2002.

AMDAM, G. et al. Hormonal control of the yolk precursor vitellogenin regulates immune function and longevity in honeybees. *Exp Gerontol*, v. 39, n. 5, p. 767–73, 2004.

AMDAM, G. V. et al. Disruption of *vitellogenin* gene function in adult honeybees by intra-abdominal injection of double-stranded RNA. *BMC Biotechnol*, v. 3, n. 1, p. 1, 2003.

AN, W.; WENSINK, P. C. Integrating sex- and tissue-specific regulation within a single *Drosophila melanogaster*. *Genes Dev.*, v. 9, p. 256–266, 1995.

AN, W.; WENSINK, P. C. Three protein binding sites form an enhancer that regulates sex- and fat body-specific transcription of *Drosophila* yolk protein genes. *EMBO J.*, v. 14, p. 1221–1230, 1995.

ANTONIEWSKI, C. et al. Characterization of an EcR/USP heterodimer target site that mediates ecdysone responsiveness of the *Drosophila Lsp-2* gene. *Mol Gen Genet*, v. 249, n. 5, p. 545–56, 1995.

ASCHER, J.; DANFORTH, B.; JI, S. Phylogenetic utility of the major opsin in bees (Hymenoptera: Apoidea): a reassessment. *Mol Phylogenet Evol*, v. 19, n. 1, p. 76–93, 2001.

ASHBURNER, M. et al. Gene ontology: tool for the unification of biology. the Gene Ontology Consortium. *Nat. Genet.*, v. 25, p. 25–29, 2000.

BAILEY, T.; ELKAN, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol*, v. 2, p. 28–36, 1994.

BAKER, B. S.; RIDGE, K. A. Sex and the single cell. i. on the action of major loci affecting sex determinating in *Drosophila melanogaster*. *Genetics*, v. 94, p. 383–423, 1980.

BARABASI, A.; OLTVAI, Z. Network biology: understanding the cell's functional organization. *Nat Rev Genet*, v. 5, n. 2, p. 101–13, 2004.

BARASH, Y.; BEJERANO, G.; FRIEDMAN, N. A simple hyper-geometric approach for discovering putative transcription factor binding sites. *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, v. 2149, p. 278–293, 2001.

BARCHUK, A. R.; BITONDI, M. M. G.; SIMõES, Z. L. P. Effects of juvenile hormone and ecdysone on the timing of vitellogenin appearance in hemolymph of queen and worker pupae of *Apis mellifera*. *J. Insect Science*, v. 2, p. 1, 2002.

BARTON, N. H.; CHARLESWORTH, B. Why sex and recombination? *Science*, v. 281, p. 1986–1990, 1998.

BATEMAN, A. et al. The Pfam protein families database. *Nucleic Acids Res*, v. 32, n. Database issue, p. D138–41, 2004.

BEJARANO, F.; BUSTURIA, A. Function of the Trithorax-like gene during *Drosophila* development. *Dev Biol*, v. 268, n. 2, p. 327–41, 2004.

BEYE, M. The dice of fate: the *csd* gene and how its allelic composition regulates sexual development in the honey bee, *Apis mellifera*. *Bioessays*, v. 26, p. 1131–1139, 2004.

BEYE, M. et al. The gene csd is the primary signal for sexual development in the honeybee and encode an sr-type protein. *Cell*, v. 114, p. 419–429, 2003.

BHAT, K. et al. The GAGA factor is required in the early Drosophila embryo not only for transcriptional regulation but also for nuclear division. *Development*, v. 122, n. 4, p. 1113–24, 1996.

BITONDI, M. et al. Characterization and expression of the *Hex 110* gene encoding a glutamine-rich hexamerin in the honey bee, *Apis mellifera*. *Arch Insect Biochem Physiol*, v. 63, n. 2, p. 57–72, 2006.

BLANCHETTE, M.; TOMPA, M. Discovery of regulatory elements by a computational method for phylogenetic footpriting. *Genome Research*, v. 12, p. 739–748, 2002.

BOWNES, M. The regulation of the yolk protein genes, a family of sex differentiation genes in *Drosophila melanogaster*. *BioEssays*, v. 16, p. 745–752, 1994.

BRIDGES, C. B. Non-disjunction as proof of the chromosome theory of heredity. *Genetics*, v. 1, p. 1–52, 1916.

BULL, J. J. *Evolution of sex determining mechanisms*. Menlo Park, California: Benjamin/Cummings, 1983.

BULYK, M. L. Computational prediction of transcription-factor binding site locations. *Genome Biology*, v. 5, p. 201, 2003.

BURGE, C.; KARLIN, S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol*, v. 268, n. 1, p. 78–94, 1997.

BUSTIN, S. Quantification of mRNA using real-time reverse transcription PCR (RT-PCR): trends and problems. *J Mol Endocrinol*, v. 29, n. 1, p. 23–39, 2002.

CHARTRAND, G.; LESNIAK, L. *Graphs and Digraphs*. [S.l.]: Chapman & Hall/CRC,, 1996.

CHERVITZ, S. A. et al. Comparison of the complete protein sets of worm and yeast: orthology and divergence. *Science*, v. 282, p. 2022–2028, 1998.

CHRISTIANSEN, A. E. et al. Sex comes in from the cold: the integration of sex and pattern. *Trends Genet.*, v. 18, p. 510–516, 2002.

CLARKE, N.; GRANEK, J. Rank order metrics for quantifying the association of sequence features with gene regulation. *Bioinformatics*, v. 19, n. 2, p. 212–8, 2003.

CLAUDIANOS, C. et al. A deficit of detoxification enzymes: pesticide sensitivity and environmental response in the honeybee. *Insect Mol Biol*, v. 15, n. 5, p. 615–36, 2006.

CLINE, T. W.; MEYER, B. J. Vive la différence: males vs females in flies vs worms. *Annu. Rev. Genet.*, v. 30, p. 637–702, 1996.

CORONA, M.; ESTRADA, E.; ZURITA, M. Differential expression of mitochondrial genes between queens and workers during caste determination in the honeybee *Apis mellifera*. *J Exp Biol*, v. 202, n. Pt 8, p. 929–38, 1999.

COSCHIGANO, K. T.; WENSINK, P. C. Sex-specific transcriptional regulation by the male and female doublesex proteins of *Drosophila*. *Genes Dev.*, v. 7, p. 42–54, 1993.

COSTA, L. The hierarchical backbone of complex networks. *Phys Rev Lett*, v. 93, n. 9, p. 098702, 2004.

COSTA, L. F. Learning about knowledge: a complex network approach. *Phys Rev E Stat Nonlin Soft Matter Phys*, v. 74, n. 2 Pt 2, p. 026103, 2006.

COSTA, L. F.; ROCHA, L. E. C. A generalized approach to complex networks. *The European Physical Journal B - Condensed Matter and Complex Systems*, v. 50, p. 237–242, 2006.

COSTA, L. F.; SPORNS, O. Structured thalamocortical connectivity revealed by random walks on complex networks. *Applied Physics Letters*, v. 89, p. 013903, 2006.

CREWS, D. Sex determination: where environment and genetics meet. *Evolution and Development*, v. 5, p. 50–55, 2003.

CRISTINO, A. et al. A comparative analysis of highly conserved sex-determining genes between *Apis mellifera* and *Drosophila melanogaster*. *Genet Mol Res*, v. 5, n. 1, p. 154–68, 2006.

CRISTINO, A. et al. Caste development and reproduction: a genome-wide analysis of hallmarks of insect eusociality. *Insect Mol Biol*, v. 15, n. 5, p. 703–14, 2006.

CROOKS, G. et al. WebLogo: a sequence logo generator. *Genome Res*, v. 14, n. 6, p. 1188–90, 2004.

CROZIER, R. H. Heterozygosity and sex determination in haplodiploidy. *Amer. Nat.*, v. 105, p. 399–412, 1971.

CUNHA, A. et al. Molecular cloning and expression of a hexamerin cDNA from the honey bee, *Apis mellifera*. *J Insect Physiol*, v. 51, n. 10, p. 1135–47, 2005.

DANFORTH, B. Bees. *Curr Biol*, v. 17, n. 5, p. R156–61, 2007.

DANFORTH, B. et al. The history of early bee diversification based on five genes plus morphology. *Proc Natl Acad Sci U S A*, v. 103, n. 41, p. 15118–23, 2006.

DAUWALDER, B. et al. The *Drosophila takeout* gene is regulated by the somatic sex-determination pathway and affects male courtship behavior. *Genes Dev.*, v. 16, p. 2879–2892, 2002.

DAVIDSON, E. H. *Genomic regulatory systems: development and evolution.* San Diego, California: Academic Press, 2001.

DAYHOFF, M. O. *Atlas of protein sequence and structure, Volume 5, Supplement 3, 1978.* Washington, D.C.: National Biomedical Research Foundation., 1979. 348 p.

DRAPEAU, M. et al. Evolution of the Yellow/Major Royal Jelly Protein family and the emergence of social behavior in honey bees. *Genome Res*, v. 16, n. 11, p. 1385–94, 2006.

DRAPEAU, M. et al. A cis-regulatory sequence within the yellow locus of *Drosophila melanogaster* required for normal male mating success. *Genetics*, v. 172, n. 2, p. 1009–30, 2006.

DRAPEAU, M. et al. A gene necessary for normal male courtship, yellow, acts downstream of fruitless in the *Drosophila melanogaster* larval brain. *J Neurobiol*, v. 55, n. 1, p. 53–72, 2003.

DURBIN, R. et al. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. [S.l.]: Cambridge University Press, 1998.

DURHAM, A. et al. EGene: a configurable pipeline generation system for automated sequence analysis. *Bioinformatics*, v. 21, n. 12, p. 2812–3, 2005.

EDDY, S. Profile hidden Markov models. *Bioinformatics*, v. 14, n. 9, p. 755–63, 1998.

ENGEL, M. A new interpretation of the oldest fossil bee (Hymenoptera: Apidae). *Amer. Mus. Novitates*, v. 3296, p. 11, 2000.

ENGEL, M. Monophyly and extensive extinction of advanced eusocial bees: insights from an unexpected Eocene diversity. *Proc Natl Acad Sci U S A*, v. 98, n. 4, p. 1661–4, 2001.

ENGEL, M. A Giant Honey Bee from the Middle Miocene of Japan (Hymenoptera: Apidae). *Amer. Mus. Novitates*, v. 3504, p. 12, 2006.

ENGEL, M. S. A monograph on the Baltic amber bees and evolution of the Apoidea (Hymenoptera). *Bull. Amer. Mus. Natural History*, v. 259, p. 1–192, 2001.

ERDMAN, S. E.; BURTIS, K. C. The *Drosophila* doublesex proteins share a novel zinc finger related DNA binding domain. *EMBO J.*, v. 2, n. 12, p. 527–535, 1993.

ESTRADA, B.; CASARES, F.; SáNCHEZ-HERRERO, E. Development of the genitalia in *Drosophila melanogaster*. *Differentiation*, v. 71, p. 299–310, 2003.

EVANS, J.; WHEELER, D. Differential gene expression between developing queens and workers in the honey bee, *Apis mellifera*. *Proc Natl Acad Sci U S A*, v. 96, n. 10, p. 5575–80, 1999.

EVANS, J.; WHEELER, D. Expression profiles during honeybee caste determination. *Genome Biol*, v. 2, n. 1, p. RESEARCH0001, 2001.

EWING, B. et al. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res*, v. 8, n. 3, p. 175–85, 1998.

FALB, D.; MANIATIS, T. Drosophila transcriptional repressor protein that binds specifically to negative control elements in fat body enhancers. *Mol Cell Biol*, v. 12, n. 9, p. 4093–103, 1992.

FARKAS, G. et al. The *Trithorax*-like gene encodes the *Drosophila* GAGA factor. *Nature*, v. 371, n. 6500, p. 806–8, 1994.

FELSENSTEIN, J. PHYLIP: phylogeny inference package. version 3.2. *Cladistics*, v. 5, p. 164–166, 1989.

FERGUSON, M.; JOANEN, T. Temperature of egg incubation determines sex in *Alligator mississippiensis*. *Nature*, v. 296, n. 5860, p. 850–3, 1982.

FLAVELL, A.; DYSON, J.; ISH-HOROWICZ, D. A novel GC-rich dispersed repeat sequence in *Drosophila melanogaster*. *Nucleic Acids Res*, v. 15, n. 10, p. 4035–48, 1987.

FlyBase Consortium. The FlyBase database of the *Drosophila* genome projects and community literature. *Nucleic Acids Res*, v. 31, n. 1, p. 172–5, 2003.

FOUNDATION, P. S. *Python programming languange*. 2007. Http://www.python.org/.

FRAZER, K. A. et al. Cross-species sequence comparisons: a review of methods and available resources. *Genome Res.*, v. 13, p. 1–12, 2003.

FU, Y.; LI, W. Statistical tests of neutrality of mutations. *Genetics*, v. 133, n. 3, p. 693–709, 1993.

FUTUYMA, D. J. *Evolutionary Biology*. [S.l.]: Sinauer Associates, Inc., 1998.

GARRETT-ENGELE, C. M. et al. *intersex*, a gene required for female sexual development in *Drosophila*, is expressed in both sexes and functions together with *doublesex* to regulate terminal differentiation. *Development*, v. 129, p. 4661–4675, 2002.

GIBBS, A.; MCINTYRE, G. The diagram, a method for comparing sequences. Its use with amino acid and nucleotide sequences. *Eur J Biochem*, v. 16, n. 1, p. 1–11, 1970.

GIBSON, G.; MUSE, S. V. *A primer of Genome Science*. Second edition. [S.l.]: Sinauer Associates, Inc., 2004.

GILBERT, S. *Developmental Biology*. Sunderland, USA: Sinauer Associates, Inc, 2003. 838 p.

GOODISMAN, M. et al. Evolution of insect metamorphosis: a microarray-based study of larval and adult gene expression in the ant *Camponotus festinatus*. *Evolution Int J Org Evolution*, v. 59, n. 4, p. 858–70, 2005.

GORDON, D. et al. TAMO: a flexible, object-oriented framework for analyzing transcriptional regulation using DNA-sequence motifs. *Bioinformatics*, v. 21, n. 14, p. 3164–5, 2005.

GRAHAM, P.; PENN, J. K. M.; SCHEDL, P. Masters change, slaves remain. *Bioessays*, v. 25, p. 1–4, 2002.

GREENBERG, A.; YANOWITZ, J.; SCHEDL, P. The *Drosophila* GAGA factor is required for dosage compensation in males and for the formation of the male-specific-lethal complex chromatin entry site at 12DE. *Genetics*, v. 166, n. 1, p. 279–89, 2004.

GUIDUGLI, K.; HEPPERLE, C.; HARTFELDER, K. A member of the short-chain dehydrogenase/reductase (SDR) superfamily is a target of the ecdysone response in honey bee (*Apis mellifera*) caste development. *Apidologie*, v. 35, p. 37–47, 2004.

GUIDUGLI, K. et al. Vitellogenin regulates hormonal dynamics in the worker caste of a eusocial insect. *FEBS Lett*, v. 579, n. 22, p. 4961–5, 2005.

GUIDUGLI, K. et al. Vitellogenin expression in queen ovaries and in larvae of both sexes of *Apis mellifera*. *Arch Insect Biochem Physiol*, v. 59, n. 4, p. 211–8, 2005.

HARBISON, C. et al. Transcriptional regulatory code of a eukaryotic genome. *Nature*, v. 431, n. 7004, p. 99–104, 2004.

HARBISON, S. et al. Quantitative genomics of starvation stress resistance in *Drosophila*. *Genome Biol*, v. 6, n. 4, p. R36, 2005.

HARBISON, S. et al. Quantitative trait loci affecting starvation resistance in *Drosophila melanogaster*. *Genetics*, v. 166, n. 4, p. 1807–23, 2004.

HARSHMAN, L. G.; JAMES, A. A. Differential gene expression in insects: transcription control. *Annu. Rev. Entomol.*, v. 43, p. 671–700, 1998.

HARTFELDER, K.; EMLEN, D. Endocrine control of insect polyphenism. In: GILBERT, L. I.; LATROU, K.; GILL, S. (Ed.). *Comprehensive Insect Molecular Science*. [S.l.]: Elsevier, Oxford, 2005. v. 3, p. 651–703.

HARTFELDER, K.; ENGELS, W. Social insect polymorphism: hormonal regulation of plasticity in development and reproduction in the honeybee. *Curr Top Dev Biol.*, v. 40, p. 45–77, 1998.

HARVEY, P.; SLATKIN, M. Some like it hot: temperature-determined sex. *Nature*, v. 296, n. 5860, p. 807–8, 1982.

HAYDAK, M. Honey bee nutrition. *Annu. Rev. Entomol.*, v. 15, p. 143–156, 1970.

HAYDAK, M. H. Larval food and development of castes in the honeybee. *J Econ Entomol*, 1943.

HEDIGER, M. et al. Sex determination in *Drosophila melanogaster* and *Musca Domestica* converges at the level of the terminal regulator *doublesex*. *Dev. Genes Evol.*, v. 214, p. 29–42, 2004.

HENIKOFF, S.; HENIKOFF, J. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, v. 89, n. 22, p. 10915–9, 1992.

HEPPERLE, C.; HARTFELDER, K. Differentially expressed regulatory genes in honey bee caste development. *Naturwissenschaften*, v. 88, n. 3, p. 113–6, 2001.

HUANG, X.; MADAN, A. CAP3: A DNA sequence assembly program. *Genome Res*, v. 9, n. 9, p. 868–77, 1999.

HUELSENBECK, J.; RONQUIST, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, v. 17, n. 8, p. 754–5, 2001.

HUGHES, J. et al. Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J Mol Biol*, v. 296, n. 5, p. 1205–14, 2000.

HUGHEY, R.; KROGH, A. Hidden Markov models for sequence analysis: extension and analysis of the basic method. *Comput Appl Biosci*, v. 12, n. 2, p. 95–107, 1996.

HUNG, M.; WENSINK, P. The sequence of the Drosophila melanogaster gene for yolk protein 1. *Nucleic Acids Res*, v. 9, n. 23, p. 6407–19, 1981.

HUNT, J.; BUCK, N.; WHEELER, D. Storage proteins in vespid wasps: characterization, developmental pattern, and occurrence in adults. *J Insect Physiol*, v. 49, n. 8, p. 785–94, 2003.

JINWAL, U. et al. Sex-, stage- and tissue-specific regulation by a mosquito hexamerin promoter. *Insect Mol Biol*, v. 15, n. 3, p. 301–11, 2006.

JUKES, T. H.; CANTOR, C. R. Evolution of protein molecules. In: H. N. Munro (Ed.). *Mammalian Protein Metabolism*. [S.l.]: Academic Press, New York, 1969. p. 21–132.

KEISMAN, E. L.; BAKER, B. S. The *Drosophila* sex determination hierarchy modulates *wingless* and it decapentaplegic signaling to deploy *dachshund* sex-specifically in the genital imaginal disc. *Development*, v. 128, p. 1643–1656, 2001.

KEISMAN, E. L.; CHRISTIANSEN, A. E.; BAKER, B. S. The sex determination gene doublesex regulates the a/p organizer to direct sex-specific patterns of growth in the *Drosophila* genital imaginal disc. *Dev. Cell*, v. 1, p. 215–225, 2001.

KIMURA, M. Evolutionary rate at the molecular level. *Nature*, v. 217, n. 5129, p. 624–6, 1968.

KIMURA, M. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature*, v. 267, n. 5608, p. 275–6, 1977.

KIMURA, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol*, v. 16, n. 2, p. 111–20, 1980.

KING, J.; JUKES, T. Non-Darwinian evolution. *Science*, v. 164, n. 881, p. 788–98, 1969.

KROGH, A. et al. Hidden Markov models in computational biology. Applications to protein modeling. *J Mol Biol*, v. 235, n. 5, p. 1501–31, 1994.

LAI, E. Micro RNAs are complementary to 3' UTR sequence motifs that mediate negative post-transcriptional regulation. *Nat Genet*, v. 30, n. 4, p. 363–4, 2002.

LATCHMAN, D. S. *Eukaryotic transcription factors*. San Diego, California: Academic Press, 1998.

LAWRENCE, C. et al. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, v. 262, n. 5131, p. 208–14, 1993.

LAWRENCE, C.; REILLY, A. An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins*, v. 7, n. 1, p. 41–51, 1990.

LEE, T. et al. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, v. 298, n. 5594, p. 799–804, 2002.

LEVINE, M.; TJIAN, R. Transcription regulation and animal diversity. *Nature*, v. 424, p. 147–151, 2003.

LEWIN, B. *Genes VII.* Oxford: Oxford University Press, 2000.

LI, T.; WHITE, K. Tissue-specific gene expression and ecdysone-regulated genomic networks in *Drosophila*. *Dev Cell*, v. 5, n. 1, p. 59–72, 2003.

LI, W.; WU, C.; LUO, C. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol Biol Evol*, v. 2, n. 2, p. 150–74, 1985.

LI, W. H.; GRAUR, D. *Fundamentals of molecular evolution*. [S.l.]: Sinauer Associates, Inc., 1991.

LIU, X.; BRUTLAG, D.; LIU, J. An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat Biotechnol*, v. 20, n. 8, p. 835–9, 2002.

LONG, M. et al. The origin of new genes: glimpses from the young and old. *Nat Rev Genet*, v. 4, n. 11, p. 865–75, 2003.

LYON, M. Some milestones in the history of X-chromosome inactivation. *Annu Rev Genet*, v. 26, p. 16–28, 1992.

MACISAAC, K.; FRAENKEL, E. Practical strategies for discovering regulatory DNA sequence motifs. *PLoS Comput Biol*, v. 2, n. 4, p. e36, 2006.

MACKENSEN, O. Viability and sex determination in the honey bee (*Apis mellifera* L.). *Genetics*, v. 5, n. 36, p. 500–509, 1951.

MAJOROS, W. et al. GlimmerM, Exonomy and Unveil: three ab initio eukaryotic genefinders. *Nucleic Acids Res*, v. 31, n. 13, p. 3601–4, 2003.

MARDULYN, P.; CAMERON, S. The major opsin in bees (Insecta: Hymenoptera): A promising nuclear gene for higher level phylogenetics. *Mol Phylogenet Evol*, v. 12, n. 2, p. 168–76, 1999.

MARTINEZ, T. et al. Sequence and evolution of a hexamerin from the ant *Camponotus festinatus*. *Insect Mol Biol*, v. 9, n. 4, p. 427–31, 2000.

MARíN, I.; BAKER, B. S. The evolutionary dynamics of sex determination. *Science*, v. 281, p. 1990–1994, 1998.

MCCLUNG, C. E. The accessory chromosome - sex determinant? *Biol. Bull. Mar. Biol. Lab.*, v. 3, p. 43–84, 1902.

MEISE, M. et al. *Sex-lethal*, the master sex-determining gene in *Drosophila*, is not sex-specifically regulated in *Musca domestica*. *Development*, v. 125, n. 8, p. 1487–1494, 1998.

MENG, A. et al. A *Drosophila* doublesex-related gene, *terra*, is involved in somitogenesis in vertebrates. *Development*, v. 126, n. 6, p. 1259–1268, 1999.

MICHEL-SALZAT, A.; CAMERON, S.; OLIVEIRA, M. Phylogeny of the orchid bees (Hymenoptera: Apinae: Euglossini): DNA and morphology yield equivalent patterns. *Mol Phylogenet Evol*, v. 32, n. 1, p. 309–23, 2004.

MICHENER, C. Comparative external morphology, phylogeny, and a classification of the bees. *Bull. American Mus. Nat. Hist.*, v. 82, p. 151–326, 1944.

MICHENER, C. D. *The social behavior of the bees. A comparative study.* [S.l.]: Harvard University Press, Massachusetts, 1974.

MICHENER, C. D. *The bees of the world.* [S.l.]: The Johns Hopkins University Press, 2000.

MIYATA, T.; YASUNAGA, T.; NISHIDA, T. Nucleotide sequence divergence and functional constraint in mRNA evolution. *Proc Natl Acad Sci U S A*, v. 77, n. 12, p. 7328–32, 1980.

MORGAN, T. H. Sex limited inheritance in *Drosophila*. *Science*, v. 32, p. 120–122, 1910.

MOUNT, D. *Bioinformatics: sequence and genome analysis.* USA: Cold Spring Harbor Laboratory Press, 2004. 692pp p.

MUKHERJEE, S. et al. Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat Genet*, v. 36, n. 12, p. 1331–9, 2004.

MULLIS, K.; FALOONA, F. Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Methods Enzymol*, v. 155, p. 335–50, 1987.

NEEDLEMAN, S. B.; WUNSCH, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, v. 48, p. 443–453, 1970.

NEI, M.; GOJOBORI, T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol*, v. 3, n. 5, p. 418–26, 1986.

NEI, M.; ROONEY, A. Concerted and birth-and-death evolution of multigene families. *Annu Rev Genet*, v. 39, p. 121–52, 2005.

NEWMAN, M. E. J. The Structure and Function of Complex Networks. *SIAM Review*, v. 45, n. 2, p. 167–256, 2003.

NOTREDAME, C.; HIGGINS, D.; HERINGA, J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*, v. 302, n. 1, p. 205–17, 2000.

OAKESHOTT, J. et al. Carboxyl/cholinesterases: a case study of the evolution of a successful multigene family. *Bioessays*, v. 21, n. 12, p. 1031–42, 1999.

OLTVAI, Z.; BARABASI, A. Systems biology. Life's complexity pyramid. *Science*, v. 298, n. 5594, p. 763–4, 2002.

OTTOLENGHI, C. et al. Novel paralogy relations among human chromosomes support a link between the phylogeny of *doublesex*-related genes and the evolution of sex determination. *Genomics*, v. 79, n. 3, p. 333–343, 2002.

PEREBOOM, J. et al. Differential gene expression in queen-worker caste determination in bumble-bees. *Proc Biol Sci*, v. 272, n. 1568, p. 1145–52, 2005.

PERERA, S. et al. Heterodimerization of ecdysone receptor and ultraspiracle on symmetric and asymmetric response elements. *Arch Insect Biochem Physiol*, v. 60, n. 2, p. 55–70, 2005.

PFAFFL, M. A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res*, v. 29, n. 9, p. e45, 2001.

PIULACHS, M. et al. The vitellogenin of the honey bee, *Apis mellifera*: structural analysis of the cDNA and expression studies. *Insect Biochem Mol Biol*, v. 33, n. 4, p. 459–65, 2003.

PROULX, S.; PROMISLOW, D.; PHILLIPS, P. Network thinking in ecology and evolution. *Trends Ecol Evol*, v. 20, n. 6, p. 345–53, 2005.

RANSON, H. et al. Evolution of supergene families associated with insecticide resistance. *Science*, v. 298, n. 5591, p. 179–81, 2002.

RAYMOND, C. S. et al. Evidence for evolutionary conservation of sex-determining genes. *Nature*, v. 391, p. 691–695, 1998.

RIDDIHOUGH, G.; PELHAM, H. An ecdysone response element in the *Drosophila hsp27* promoter. *EMBO J*, v. 6, n. 12, p. 3729–3734, 1987.

ROBERTSON, H.; WANNER, K. The chemoreceptor superfamily in the honey bee, *Apis mellifera*: expansion of the odorant, but not gustatory, receptor family. *Genome Res*, v. 16, n. 11, p. 1395–403, 2006.

ROKAS, A. et al. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, v. 425, n. 6960, p. 798–804, 2003.

ROTH, F. et al. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat Biotechnol*, v. 16, n. 10, p. 939–45, 1998.

ROZAS, J.; ROZAS, R. DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics*, v. 15, n. 2, p. 174–5, 1999.

RUTHERFORD, K. et al. Artemis: sequence visualization and annotation. *Bioinformatics*, v. 16, n. 10, p. 944–5, 2000.

SAITOU, N.; NEI, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, v. 4, p. 406–425, 1987.

SANGER, F.; NICKLEN, S.; COULSON, A. R. DNA sequencing with chain-terminating inhibitors. *Biotechnology*, v. 24, p. 104–108, 1977.

SCHMIDT, H. et al. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, v. 18, n. 3, p. 502–4, 2002.

SCHMITZOVA, J. et al. A family of major royal jelly proteins of the honeybee *Apis mellifera* L. *Cell Mol Life Sci*, v. 54, n. 9, p. 1020–30, 1998.

SCHNEIDER, T.; STEPHENS, R. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res*, v. 18, n. 20, p. 6097–100, 1990.

SCHüTT, C.; NöTHIGER, R. Structure, function and evolution of sex-determining systems in Dipteran insects. *Development*, v. 127, p. 667–677, 2000.

SIEGAL, M. L.; BAKER, B. S. Functional conservation and divergence of *intersex*, a gene required for female differentiation in *Drosophila melanogaster*. *Dev. Genes Evol.*, v. 215, p. 1–12, 2005.

SINCLAIR, A. H. et al. A gene from the human sex-determining region encodes a protein with homology to a conserved dna-binding motif. *Nature*, v. 346, n. 6281, p. 240–244, 1990.

SMITH, C.; SINCLAIR, A. Sex determination: insights from the chicken. *Bioessays*, v. 26, n. 2, p. 120–32, 2004.

SMITH, T. F.; WATERMAN, M. S. Identification of common molecular subsequences. *J. Mol. Biol.*, v. 147, p. 195–197, 1981.

SNELL, G. D. The determination of sex in *Habrobracon*. *PNAS*, v. 21, p. 446–453, 1935.

SONNHAMMER, E. et al. Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res*, v. 26, n. 1, p. 320–2, 1998.

STORMO, G. D. DNA binding sites: representation and discovery. *Bioinformatics*, v. 16, p. 16–23, 2000.

SU, S. et al. Molecular cloning and analysis of four cDNAs from the heads of *Apis cerana cerana* nurse honeybees coding for major royal jelly proteins. *Apidologie*, v. 36, n. 3, p. 389–401, 2005.

SULLIVAN, A.; THUMMEL, C. Temporal profiles of nuclear receptor gene expression reveal coordinate transcriptional responses during *Drosophila* development. *Mol Endocrinol*, v. 17, n. 11, p. 2125–37, 2003.

SUZUKI, M. G. et al. Analysis of the biological function of a *doublesex* homologue in *Bombyx mori*. *Dev. Genes Evol.*, v. 213, p. 345–354, 2003.

TAJIMA, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, v. 123, n. 3, p. 585–95, 1989.

TAJIMA, F. Simple methods for testing the molecular evolutionary clock hypothesis. *Genetics*, v. 135, n. 2, p. 599–607, 1993.

TAMURA, K.; NEI, M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol*, v. 10, n. 3, p. 512–26, 1993.

The Honeybee Genome Sequencing Consortium. Insights into social insects from the genome of the honeybee *apis mellifera*. *Nature*, v. 443, n. 7114, p. 931–949, 2006.

THOMPSON, J. D.; HIGGINS, D. G.; GIBSON, T. J. CLUSTAL W: improving the sensitiviry of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, v. 22, p. 4673–4680, 1994.

TOMPA, M. et al. Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol*, v. 23, n. 1, p. 137–44, 2005.

VLIEGHE, D. et al. A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Res*, v. 34, n. Database issue, p. D95–7, 2006.

VOLFF, J.-N.; ZARKOWER, V. J. B. D.; SCHARTL, M. Evolutionary dynamics of the dm domain gene family in metazoans. *J. Mol. Evol.*, v. 57, p. S241–S249, 2003.

WANG, T.; STORMO, G. Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics*, v. 19, n. 18, p. 2369–80, 2003.

WASSERMAN, W. W.; SANDELIN, A. Applied bioinformatics for the identification of regulatory elements. *Nature Rev. Genet.*, v. 5, p. 276–287, 2004.

WERNERSSON, R.; PEDERSEN, A. RevTrans: Multiple alignment of coding DNA from aligned amino acid sequences. *Nucleic Acids Res*, v. 31, n. 13, p. 3537–9, 2003.

WEST-EBERHARD, M. J. *Developmental plasticity and evolution.* [S.l.]: Oxford University Press, USA, 2003.

WHEELER, D.; NIJHOUT, H. A perspective for understanding the modes of juvenile hormone action as a lipid signaling system. *Bioessays*, v. 25, n. 10, p. 994–1001, 2003.

WHITE, M. J. D. *Animal cytology and evolution*. Cambridge, UK: Cambridge University Press, 1973.

WHITING, P. W. Selective fertilization and sex-determination in Hymenoptera. *Science*, v. 78, p. 537–538, 1933.

WHITING, P. W. Multiple alleles in complementary sex determination of *Habrobracon*. *Genetics*, v. 28, p. 365–382, 1943.

WILKINS, A. S. Moving up the hierarchy: A hypothesis on the evolution of a genetic sex determination pathway. *Bioessays*, v. 1, n. 17, p. 71–77, 1995.

WILSON, E. B. The chromosomes in relation to determination of sex in insects. *Science*, v. 22, p. 500–502, 1905.

WILSON, E. O. *Sociobiology: The new synthesis*. [S.l.]: Harvard University Press, Cambridge, 1975.

WINGENDER, E. et al. TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res.*, v. 1, n. 28, p. 316–319, 2000.

WINTERS-HILT, S. Hidden Markov Model Variants and their Application. *BMC Bioinformatics*, v. 7 Suppl 2, p. S14, 2006.

WRAY, G. A. et al. The evolution of transcriptional regulation in eukaryotes. *Mol. Biol. Evol.*, v. 20, p. 1377–1419, 2003.

WU, C.; LI, W. Evidence for higher rates of nucleotide substitution in rodents than in man. *Proc Natl Acad Sci U S A*, v. 82, n. 6, p. 1741–5, 1985.

WYATT, G. *Yolk proteins, invertebrates. Encyclopedia of reproduction*. Amsterdam: Elsivier, 1999.

YANO, K. et al. Structure and expression of mRNA for vitellogenin in *Bombyx mori*. *Biochim. Biophys Acta*, v. 1218, p. 1–10, 1994b.

ZAKHARKIN, S. O. et al. Female-specific expression of a hexamerin gene in larvae of an autogenous mosquito. *Eur. J. Biochem.*, v. 268, p. 5713–5722, 2001.

ZHOU, X.; OI, F.; SCHARF, M. Social exploitation of hexamerin: RNAi reveals a major caste-regulatory factor in termites. *Proc Natl Acad Sci U S A*, v. 103, n. 12, p. 4499–504, 2006.

ZHOU, X. et al. Two hexamerin genes from the termite *Reticulitermes flavipes*: Sequence, expression, and proposed functions in caste regulation. *Gene*, v. 376, n. 1, p. 47–58, 2006.

ZHU, L. et al. Sexual dimorphism in diverse metazoans is regulated by a novel class of intertwined zinc fingers. *Genes Dev.*, v. 14, p. 1750–1764, 2000.

ZUCKERKANDL, E.; PAULING, L. Molecules as documents of evolutionary history. *J Theor Biol*, v. 8, n. 2, p. 357–66, 1965.