Genome scan for cis-regulatory DNA motifs associated with social behavior in honey bees

Saurabh Sinha^{*†‡}, Xu Ling^{*}, Charles W. Whitfield^{†§¶}, Chengxiang Zhai^{*†}, and Gene E. Robinson^{†§¶||}

Departments of ^{*}Computer Science and [§]Entomology, [†]Institute of Genomic Biology, and [¶]Neuroscience Program, University of Illinois at Urbana-Champaign, Urbana, IL 61801

Contributed by Gene E. Robinson, August 28, 2006

Honey bees (*Apis mellifera*) undergo an age-related, socially regulated transition from working in the hive to foraging, which is associated with changes in the expression of thousands of genes in the brain. To begin to study the cis-regulatory code underlying this massive social regulation of gene expression, we used the newly sequenced honey bee genome to scan the promoter regions of eight sets of behaviorally related genes differentially expressed in the brain in the context of division of labor among worker bees, for 41 cis-regulatory motifs previously characterized in *Drosophila melanogaster*. Binding sites for the transcription factors Hairy, GAGA, Adf1, Cf1, Snail, and Dri, known to function in nervous system development, olfactory learning, or hormone binding in *Drosophila*, were significantly associated with one or more gene sets. The presence of some binding sites also predicted expression patterns for as many as 71% of the genes in some gene sets. These results suggest that there is a robust relationship between cis and social regulation of brain gene expression, especially considering that we studied <15% of all known transcription factors. These results also suggest that transcriptional networks involved in the regulation of development in *Drosophila* are used to regulate behavioral development in adult honey bees. However, differences in gene regulation between these two processes are suggested by the finding that the promoter regions for the behaviorally related bee genes differed in both motif occurrence and G/C content relative to their *Drosophila* orthologs.

Apis mellifera | gene regulation | microarray | position weight matrix | transcription factors

Behavioral development in honey bees (*Apis mellifera*) gives rise to an intricate division of labor in the honey bee colony (1). Adult worker honey bees perform a series of tasks in the hive, such as brood care (“nursing”), when they are young and then shift to foraging for nectar and pollen outside the hive when they are 2–3 weeks of age for the remainder of their 5- to 7-week life. The transition from hive work to foraging involves changes in endocrine activity, brain chemistry, and brain structure. This transition is socially regulated and responsive to the needs of the colony, which are communicated via pheromones and other means. These social factors act directly or indirectly on physiological mechanisms that influence behavioral maturation, including mechanisms involving juvenile hormone (JH), the *foraging* and *malvolio* genes, and other genes still to be identified.

The transition from hive work to foraging in honey bee colonies also involves changes in the expression of thousands of genes in the brain (2, 3). This finding suggests that honey bees will be useful in elucidating the mechanisms by which social factors regulate gene expression in the brain, one frontier in the study of genes and social behavior.

To begin to study the cis-regulatory code underlying this massive social regulation of brain gene expression, we used the newly sequenced honey bee genome (4) to scan for cis-regulatory motifs in the promoter regions of genes previously shown to be differentially expressed in the brain in the context of socially regulated division of labor (2, 3). We searched for 41 motifs that were previously well characterized in *Drosophila melanogaster*, primarily

for embryonic development, which represents <15% of all known transcription factors (www.godatabase.org). Starting with these experimentally validated motifs allowed us to employ solely bioinformatics methods for this study; identification of motifs *de novo* would require additional molecular validation. With this approach, we also were able to begin to explore the hypothesis that transcriptional networks involved in the regulation of embryonic development in *Drosophila* are also used to regulate adult behavioral development in honey bees. The premise of this hypothesis is that many transcription factors show high levels of functional conservation across broad evolutionary distances (5).

Microarray analysis has revealed that nurse and forager honey bees show differences in brain mRNA abundance for about one-third of the genes analyzed, and subsequent experiments indicated that many of these genes are socially regulated and predictive of behavioral status (2). Whitfield *et al.* (3) generated additional sets of genes implicated in social regulation by performing microarray analyses to identify genes regulated in the brain by physiological, ontogenetic, and genetic factors that are also known to influence the age at onset of foraging. We used these sets of “behaviorally related” genes in our study.

We divided eight gene sets (2, 3), drawn from a total of 3,129 genes (estimated to represent ≈25% of the genes in the bee genome), into those sets that were up-regulated or down-regulated in the brain in response to a specific condition (Table 1). Each gene set contained 100–871 genes. Three additional pairs of gene sets were derived from these gene sets (see *Methods*) to capture specific combinations of conditions. We identified and quantified patterns of occurrence of the 41 cis-regulatory motifs, listed in Table 5, which is published as supporting information on the PNAS web site, in the promoter regions of these gene sets.

We scanned for these motifs by modeling the binding specificity of each transcription factor by using a position-specific weight matrix (PWM). We used the computer algorithm Stubb (6) to score a promoter (2,000-bp region upstream of the translation start site for each gene) for matches to PWMs. The Stubb algorithm was previously found to accurately predict cis-regulatory modules involved in the segmentation pathway in *Drosophila* (7).

The general scheme for this study is illustrated in Fig. 1. We performed “enrichment analysis” by using hypergeometric tests to analyze each gene set for statistical enrichment for each of the 41 motifs, compared with the entire complement of 3,129 genes. We also performed “correlation analysis,” analyzing each pair of up- and down-regulated gene sets for correlation between motif counts and gene expression. Because all of the gene sets are related to

Author contributions: S.S., C.W.W., C.Z., and G.E.R. designed research; S.S. and X.L. performed research; S.S., C.W.W., and C.Z. contributed new reagents/analytic tools; S.S. and X.L. analyzed data; and S.S. and G.E.R. wrote the paper.

The authors declare no conflict of interest.

Abbreviations: GO, Gene Ontology; JH, juvenile hormone; PWM, position-specific weight matrix; SVM, support vector machine.

[†]To whom correspondence may be addressed at: University of Illinois, 201 North Goodwin Avenue, Urbana, IL 61801. E-mail: sinhas@cs.uiuc.edu.

^{||}To whom correspondence may be addressed. E-mail: generobi@life.uiuc.edu.

© 2006 by The National Academy of Sciences of the USA

classifier, and found four such cases (Table 4; and Tables 15 and 16, which are published as supporting information on the PNAS web site). These cases involved three different motifs: Hairy and Snail (again) and GAGA. The presence or absence of the GAGA motif alone correctly classified the expression (up- or down-regulated) of 71% of 100 of the most differentially expressed genes in the Hive bee-to-forager transition $\uparrow$ vs. Hive bee-to-forager transition $\downarrow$ set, compared with 50% expected on the basis of chance alone ($P = 1.6\text{E-}5$, binomial test). Similar results were obtained for Hairy and the response of genes to methoprene (JH analog), and for Snail and the response of genes to cGMP (related to the *foraging* gene; see ref. 1).

Using all 41 motifs in our compendium and a support vector machine (SVM) classifier (<http://svmlight.joachims.org>), we found one more case: 2-fold cross-validation analysis correctly classified the expression pattern of 71% of the genes up-regulated in foragers and by methoprene compared with genes up-regulated in nurses and down-regulated by methoprene ($n = 78$ genes, 39 of each; $P = 0.00018$, binomial test). G/C content alone had insignificant predictive power ($P > 0.1$, binomial test). These findings are remarkable, considering that our motif compendium comprises $<15\%$ of the estimated transcription factors in the *Drosophila* (www.godatabase.org) or honey bee genomes.

cis-Regulatory Motifs Show Combinatorial Regulation of Honey Bee Behaviorally Related Genes. The transcriptional pathways involved in embryonic development are well known for their combined action of multiple transcription factors on common gene targets (21). The combinatorial nature of transcriptional pathways is a highly conserved feature of cis-regulation, so we expected to be able to detect this in the context of social regulation. Pairs of cis-regulatory binding sites do indeed co-occur in the promoter regions of genes in our gene sets.

Focusing on the Forager $\uparrow$ and Nurse $\uparrow$ gene sets, we counted the common gene targets of each pair of transcription factors and found this number to be significantly higher than by chance for 78 of the 820 pairs ($P < 0.001$, $Q = 0.01$, hypergeometric test; see ref. 22), even after controlling for the effect of local background composition. In comparison, a negative control experiment involving randomly permuted versions of the promoters produced only five pairs of transcription factors. After generating a pairwise interaction graph from these data and performing further analysis, we found a cohesive set of 7 transcription factors (Adf1, Hairy, Cfl and GAGA again, plus AbdB, Zeste, and Eve) of the 41 studied, with almost every pair having significant patterns of interaction (see Fig. 4, which is published as supporting information on the PNAS web site).

Conclusions

These results demonstrate a robust association for social behavior, brain gene expression, and distributions of transcription factor binding sites throughout the honey bee genome and are consistent with results from studies of social behavior in rodents, which have revealed important roles for cis regulation for single genes (23, 24). Our finding of motifs by using PWMs primarily derived from studies of *Drosophila* developments demonstrates that transcriptional networks involved in the regulation of embryonic development are reused by nature for adult behavioral functions. We also detected differences in motif–gene set associations and promoter G/C content between honey bee and *Drosophila* that might reflect unique aspects of gene regulation associated with social regulation. Social behavior is a highly derived trait, and we predict that the evolution of transcriptional combinatorial codes for socially regulated gene expression in the brain will involve both conserved and novel motifs that await discovery.

Methods

Scoring Genes for Motif Occurrence. The Stubb algorithm (6) computes a statistical (log-likelihood ratio) score for the clustering of binding sites of one or more transcription factors in a sequence, accounting for varying numbers and strengths of the sites. The software was run with a window size of 200, shift of 100, and a zeroth-order background Markov model. The training sequence for the Markov model was either the entire complement of 2-kbp promoters (“global background”) or the sequence of the promoter being analyzed (“local background”). (The local background of 2 kbp provides sufficient statistics to train a zeroth-order Markov model of four states.) The best-scoring 200-bp window in a gene’s promoter was used to score that gene. We masked short tandem repeats (putative micro- and minisatellites) in the sequence to prevent the Stubb algorithm from confounding these repeats with weak binding sites. Tandem repeats were masked using the program Tandem Repeat Finder (25), with recommended parameters 2 7 7 80 10 25 500. The heat shock element was represented by a set of eight PWMs from TRANSFAC (www.gene-regulation.com/pub/databases.html#transfac), corresponding to trimers of the AGAAN site in all possible orientations (26). Other PWM motifs, corresponding to binding sites in fruitfly, were obtained from TRANSFAC, JASPAR (27), and Schroeder *et al.* (28), and redundant matrices were removed manually to obtain a set of 41 motifs for 37 distinct transcription factors. Unless specified otherwise, a gene was considered as containing binding sites for a particular transcription factor if it was among the top 200 genes ranked by Stubb score for the corresponding motif; these genes are referred to as “targets” of a motif. Thus, motif targets are decided based on clustering of strong and weak motif occurrences in their promoters. The G/C content of a promoter was measured as the fraction of the promoter that is Gs or Cs.

Gene Sets. Eight sets of genes established as being significantly up-regulated or down-regulated in the brain in response to specific conditions (behavioral development, genetic variation, or neuroactive treatment) were taken from refs. 2 and 3. Each set was partitioned into up- and down-regulated subsets, as listed in Table 1. Treatments with cGMP, methoprene (JH analog), or manganese are known to induce precocious foraging in the honey bee. Hence, we intersected the Forager $\uparrow$ set with each of the cGMP $\uparrow$, Methoprene $\uparrow$, and Manganese $\uparrow$ sets, and likewise the Nurse $\uparrow$ set with each of cGMP $\downarrow$, Methoprene $\downarrow$, and Manganese $\downarrow$ sets to obtain six additional sets (set sizes as given in Table 17, which is published as supporting information on the PNAS web site), for a total of 22 sets. Enrichment analysis was done for each of these 22 sets separately.

Enrichment Analyses. Hypergeometric test. The set of genes targeted by a motif was intersected with a given gene set, and the P value of the intersection size was calculated with the hypergeometric test: $H(N, g, m, i)$, where N is the total number of genes (3,129), g is the size of the given gene set, m is the number of genes targeted by motif (200), and i is the intersection size. For motif enrichment vs. G/C enrichment, let $i_{G/C}$ be the size of the intersection of the gene set with the top m G + C-rich genes (promoters). We heuristically computed the P value of the hypergeometric distribution $H(N, i_{G/C}N/m, m, i)$ and required this P value to be ≤ 0.05 as an additional criterion for reporting the motif as significantly associated with the gene set. Note that the original hypergeometric test $H(N, g, m, i)$ compares the fraction i/m (the specificity of predicting gene set membership based on motif score) with the fraction g/N , which is the baseline specificity of a random predictor. In the second statistical test, we change the baseline performance to be that achieved by a G/C content-based predictor, $i_{G/C}/m$, and compare the fraction i/m with this fraction, using $H(N, i_{G/C}N/m, m, i)$.