Behavior-specific changes in transcriptional modules lead to distinct and predictable neurogenic states

Sriram Chandrasekaran*a,†, Seth A. Amenta,b,1, James A. Eddyb, Sandra L. Rodriguez-Zasc, Bruce R. Schatzb,e,f, Nathan D. Pricea,b,c,g,1,3, and Gene E. Robinsonb,e,h,3

*Center for Biophysics and Computational Biology, †Neuroscience Program, Department of Bioengineering, ‡Department of Animal Sciences, †Institute for Genomic Biology, Department of Medical Information Science, §Department of Chemical and Biomolecular Engineering, and †Department of Entomology, University of Illinois at Urbana–Champaign, Urbana, IL 61801

Contributed by Gene E. Robinson, September 1, 2011 (sent for review June 29, 2011)

Using brain transcriptomic profiles from 853 individual honey bees exhibiting 48 distinct behavioral phenotypes in naturalistic contexts, we report that behavior-specific neurogenic states can be inferred from the coordinated action of transcription factors (TFs) and their predicted target genes. Unsupervised hierarchical clustering of these transcriptomic profiles showed three clusters that correspond to three ecologically important behavioral categories: aggression, maturation, and foraging. To explore the genetic influences potentially regulating these behavior-specific neurogenic states, we reconstructed a brain transcriptional regulatory network (TRN) model. This brain TRN quantitatively correlates with high accuracy gene expression changes of more than 2,000 genes involved in behavior, even for behavioral phenotypes on which it was not trained, suggesting that there is a core set of TFs that regulates behavior-specific gene expression in the bee brain, and other TFs more specific to particular categories. TFs playing key roles in the TRN include well-known regulators of neural and behavioral plasticity, e.g., Creb, as well as TFs better known in other biological contexts, e.g., NF-κB (immunity). Our results reveal three insights concerning the relationship between genes and behavior. First, distinct behaviors are subserved by distinct neurogenic states in the brain. Second, the neurogenic states underlying different behaviors rely upon both shared and distinct transcriptional modules. Third, despite the complexity of the brain, simple linear relationships between TFs and their putative target genes are a surprisingly prominent feature of the networks underlying behavior.

Apis mellifera | gene regulation | social behavior | systems biology

Behavior is influenced by both heritable and environmental factors, sometimes via massive changes in brain transcriptomes (1). An emerging insight is that these changes induce shifts in “neurogenic states” rather than activation of particular genes only in local neural circuits (2). This has led to the idea that distinct neurogenic states underlie distinct behaviors (1), but it is not known how these states are defined or maintained. Further, the regulatory architecture of behaviorally relevant neurogenic states has not been studied, and it is not known whether behavior is subserved by the kinds of transcriptional regulatory networks (TRNs) known for other phenotypes (3–6).

We applied tools and perspectives from molecular systems biology—used to study transcriptional regulation in the brain and elsewhere (3–6)—to transcript profiles from the BeeSpace Project, which used microarray analysis to study hereditary and environmental influences on brain gene expression and social behavior (Methods). This provided a unique aggregate dataset from a single laboratory (G.E.R.), using the same analytical platform, protocols, and analysis procedures (7). Because the natural behavioral repertoire of the honey bee (Apis mellifera) is perhaps the best studied of any nonhuman animal (8), we were able to analyze a rich set of naturalistic behavioral states.

We chose 27 pairwise comparisons among the 48 distinct behavior states that directly surveyed transcriptome responses to a heritable or environmental factor related to one or more of three ecologically important behavioral categories: aggression, hive defense; maturation, from working in the hive to foraging for nectar and pollen; and different foraging predispositions or types of experience, e.g., scouting for new floral resources or forming specific spatiotemporal memories for known resources, respectively (Methods and SI Appendix, Table S1). Nearly all genes were differentially expressed in at least one comparison; this broad survey thus captured natural variation across most of the brain transcriptome, without experimental genetic perturbation. SI Appendix, Tables S1 and S2 summarize the number of bees, experiments, and microarrays in this dataset and indicate which experiments have been published or will be described in forthcoming publications. Additional details of these experiments are in the SI Appendix, SI Methods.

We used a systems approach to create, test, and interpret a network model that predicts the expression of a large number of genes in the bee brain given the expression values of the TFs. Our strategy consisted of the following steps: (i) Integrate previous measurements of brain gene expression over a wide range of behavior states, which consist of behavioral responses to various genetic and environmental perturbations. (ii) Integrate expression from various conditions and develop a predictive behavioral TRN model from training data that links expression of putative target genes to the expression values of the TFs. (iii) Evaluate the TRN model in new test conditions and also on the basis of comparison with genomic data. (iv) Analyze the properties of the TRN with respect to other networks inferred for other phenotypes. (v) Identify key regulators/hubs in the TRN that predict behavior and may drive behavioral changes. (vi) Compare various subnetworks associated with aggression, maturation, and foraging.

Results and Discussion

Neurogenic States Related to Behavior. We first used hierarchical clustering to determine whether brain transcriptomic profiles tracked behavior. Each comparison was designed a priori to measure transcriptome responses to a heritable or environmental factor related to aggression, maturation, or foraging. Hierarchical clustering recapitulated this pattern; we identified three distinct clusters of transcriptomic profiles, corresponding to aggression, maturation, and foraging. This result demonstrates a strong relationship be-


The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

1Present address: Institute for Systems Biology, 401 Terry Avenue North, Seattle, WA 98109.

‡Present address: Department of Molecular and Cellular Biology, University of California, 16 Barker Hall, Berkeley, CA 94720.

‡To whom correspondence may be addressed. E-mail: generobi@illinois.edu or nprice@systemsbiology.org.

This article contains supporting information online at www.pnas.org/cgi/doi/10.1073/pnas.1114093108

www.pnas.org/cgi/doi/10.1073/pnas.1114093108
between behavior and brain gene expression, as it was obtained via unsupervised clustering (Fig. L4 and SI Appendix, Fig. S1).

Clustering analysis also revealed differences between transcriptome responses to environmental or heritable factors (Fig. L4). We found, for example, distinct subclusters of aggression-related comparisons corresponding to environmental (colony environment), or hereditary (individual genotype) factors. SI Appendix, Table S2 describes each of the comparisons and the degree of gene expression differences between the states.

The two results obtained with hierarchical clustering also were obtained independently with multidimensional scaling (MDS) (Fig. 1B). These findings indicate that distinct neurogenomic states underlie distinct behaviors and hint at differential effects of “nature” and “nurture.”

**Brain Transcriptional Regulatory Network.** To explore the mechanisms potentially regulating behavior-specific neurogenomic states, we reconstructed a brain TRN model using an approach we call Analyzing Subsets of Transcriptional Regulators Influen
cing eXpression (ASTRIX). This approach is based on a combination of two well-known algorithms used for network inference (Fig. 2A and Methods). We generated a network of high-confidence putative TF (“target”)–gene interactions using Accurate Reconstruction of Cellular Networks (ARACNE) (9) and leveraged these interactions to predict expression in new conditions with Least Angle Regression (LARS) (10). In the training set, a subset of TF–gene relationships are identified that can accurately predict the target gene’s expression quantitatively, and then these interactions are moved forward to the test set for validation. Interactions were inferred from expression profiles and a list of 236 predicted bee TFs based on orthology with *Drosophila melanogaster*.

The bee brain TRN quantitatively predicts gene expression changes with remarkably high accuracy, even for behavioral states on which it was not trained (Fig. 2B and Table 1 and SI Appendix, Tables S2 and S11). It was able to model ~25% (2,176) of the genes profiled (chosen on the basis of strong fit in the training set, Pearson correlation *r* > 0.8) and quantitatively predicted their expression with an average correlation of 0.87 in test sets (root mean square deviation, RMSD = 0.35, both in training and test sets). By contrast, a control experiment with TF expression permuted across states resulted in a 0.22 correlation, and zero genes predicted with >0.8 correlation, even in the training set (SI Appendix, Fig. S4). Stringent leave-one-state-out cross validation, with model prediction tested on behavior states for which it had not been trained, yielded similarly accurate results (RMSD_{train} = 0.33, RMSD_{test} = 0.32); likewise for the 10-fold cross-validation test (Fig. 2C and SI Appendix, Figs. S2 and S3). The model’s success in predicting expression in new behavioral states suggests that there is a core set of TFs that regulate behavior-specific gene expression in the bee brain.

This TRN includes genes with known neurobiological and behavioral significance. Many neural-related Gene Ontology (GO) categories like synaptic transmission and neurotransmitter uptake were overrepresented among the TRN genes (SI Appendix, Table S4), indicating that features reflecting the general organization of the bee brain transcriptome are consistent with those of known brain processes. Genes in modules often shared annotated biological features: 87% (166 out of 190) of the modules were enriched for specific biological processes [based
on GO or Kyoto Encyclopedia of Genes and Genomes annotations; SI Appendix, Table S5; false discovery rate (FDR) < 0.1]. Genes in modules also often shared evolutionarily conserved cis-regulatory motifs in upstream sequences (FDR < 0.1; SI Appendix, Table S5). Some modules also showed high similarity with previously described Drosophila networks (SI Appendix, Table S5), consistent with the existence of conserved modules of coregulated genes related to specific biological functions (11). As expected, genes in the TRN were enriched (P < 1e-10) for those that were strongly differentially expressed between behavioral states.

Having demonstrated predictive ability and biological relevance, we analyzed the TRN to explore the hypothesis that behaviorally related neurogenomic states arise, in part, from the coordinated action of TFs and their predicted target genes; the predicted targets were defined as those genes that share very high mutual information (P < 10^-6) with a TF and have high predictive ability (r > 0.8). The top 20 most connected TFs (“hubs”) together were predicted to regulate 75% of the genes, and the top 3, 33% (SI Appendix, Figs. S5 and S6). Most of the top 20 hubs and many of their predicted targets were differentially expressed (FDR < 0.05) across many of the 27 comparisons (Fig. 3B and SI Appendix, Table S6). Only four TFs and their modules (sets of putative target genes predicted to be regulated by a specific TF) were active across all three behavioral categories (“global regulators”), i.e., their predicted target genes were significantly overrepresented among differentially expressed genes across all three behavioral categories. Other TFs and their modules were active only for a particular behavioral state within a category (Fig. 3C and SI Appendix, Table S7). State-specific TFs were significantly more common than predicted by chance (P < 1e-10), indicating that they also are a robust feature of this TRN. These results provide support for our hypothesis.

TFs playing key roles in the bee brain TRN include well-known regulators of neural and behavioral plasticity. Creb, an iconic regulator of neural plasticity in both invertebrates and vertebrates (12), is a top hub. Two other hubs are ftz-f1 and broad (br), which mediate responses to juvenile hormone (JH) (13), a hormone that regulates bee maturation, brain chemistry, and brain structure (14). The TRN also prominently features TFs better known in other contexts. For example, dorsal (dl), one of only four global regulators, is the insect ortholog of NF-κB (15), a well-known mediator of immune responses.

We were able to robustly and accurately model a large percentage (~25%) of the brain transcriptome from whole-brain data. This might seem surprising, because neuron-, circuit-, and brain region-specific patterns of gene expression are well known (16). However, some hormones and neuromodulators can have widespread transcriptional effects on the brain (17). We used the results of two

---

**Figure 2.** Reconstruction of a genome-scale model of a behaviorally related transcriptional regulatory network (TRN) from honey bee brain transcriptome profiles. (A) Flowchart outlining the ASTRIX approach: (1) input transcriptomic dataset; (2) “Skeletal” network inferred with ARACNE; high mutual information threshold (P < 10^-6) retains only strong (putatively direct) interactions; (3) Data processing inequality (DPI) eliminates indirect interactions; (4) LARS fits a regression model for the regulation of each gene by transcription factors (TFs); (5) model accuracy for each gene assessed in training and test sets; and (6) accurately modeled genes retained in a parsimonious, quantitative model that predicts gene expression from TF expression. (B) A total of 190 TFs (triangles outside the sphere) are predicted to regulate 2,176 genes via 6,757 interactions (mean indegree = 3; mean outdegree = 35, Table 1, SI Appendix, Tables S2 and S11). Black triangles at the top represent the predicted global regulators: br, lilli, dl, and GB13780. The genes unique to each behavioral category (subnetwork; same color coding as in Fig. 1A) are highlighted along with their corresponding subnetwork specific TFs (within each box). TFs predicted to be involved in the regulation of two subnetworks are labeled between the boxes. (C) Plots show the accurate fit of the model (red) to actual expression data (blue) in training and test samples (Left and Right of the blue dotted line, respectively), for nine randomly selected genes with training correlation >0.8.
BeeSpace experiments that profiled gene expression in specific brain regions to explore whether performance of the TRN might be enhanced for a more homogenous population of cells. By contrast, the TRN was a better predictor of gene expression for whole-brain transcriptome profiles than for brain-region-specific profiles (Methods, SI Appendix, Fig. S10 and Table S10). These results suggest that additional factors that functionally link multiple cell types and brain regions may drive gene expression across the brain, including hormones, immune/inflammatory responses, and synaptic connections between neurons. Such factors are represented in the TRN by prominent hub genes (ftz-f1 and br, all and Deaf1); and Creb, respectively. Other modules may reflect more general cellular functions, e.g., those involving general “housekeeping” (16): the largest module in the bee brain TRN is regulated by Log1, a TF linked to metabolism (18). These and likely other reasons made it possible for the TRN to capture coherent patterns of transcriptional regulation, even for averaged values of gene expression across the brain.

Two intriguing dynamic properties of this TRN were detected, but their biological significance is unclear. First, whereas modules generally showed consistent gene expression across individuals within behavioral states (SI Appendix, Fig. S7), some modules exhibited significant changes in relative gene expression ordering between states (Fig. 3D). This “shuffling” was greatest for modules associated with hub genes (SI Appendix, Fig. S8) such as ftz-f1, Creb, and Log1. Gene shuffling is associated with transcriptional “dysregulation” in diseased states (19), but such differences have not yet been detected in network analysis for normal states. Second, the shuffling varied depending on behavioral timescale. Increased shuffling was more often associated with longer-lasting behavioral states, e.g., those related to behavioral maturation, rather than more dynamic states, e.g., spatiotemporal floral memories (\( P = 0.008; \) Fig. 3D and SI Appendix, Table S8 and Fig. S9). Longer-lasting states also showed a significant increase in the number of TFs predicted to regulate each gene (\( P = 0.002; \) SI Appendix, Fig. S8). These results are reminiscent of the above-mentioned cluster analysis timescale differences and again hint at differences in transcriptional regulation over different timescales.

Validation of Transcriptional Regulatory Network. Comprehensive validation of the predictions of this TRN would require manipulation of brain expression of multiple TFs with RNAi across multiple behavioral states under natural conditions. This is not yet feasible, although RNAi treatment has been used to test the effects of single genes in individual behavioral contexts in the field (20). Instead, partial validation for some of our results was obtained by four comparative bioinformatics analyses. (i) Genes in the modules of 14 TFs in the bee brain TRN had significant overlap with the direct, physical targets of the orthologous TF in Drosophila (11), (SI App-
The bee TRN does not encompass all of the layers of complexity inherent in brain function, which includes indirect and combinatorial regulation of gene expression by TFs and regulation via noncoding RNAs, epigenetic mechanisms, and posttranscriptional processes. However, its ability to accurately model a surprisingly high percentage of the transcriptome absent information on nonlinear interactions between genes, physical interactions, time course, or brain subregion localization, and using only a partial (−30%) list of TFs (21)—suggests that the relationship between brain gene expression and behavior is both stronger and more predictable than previously imagined.

Conclusions

Our ASTRIX method is built on two well-known methods to infer interactions and predict expression in new conditions. ARACNE is an information-theoretic method for identifying transcriptional interactions from microarrays or other gene expression data. It is specifically designed to scale up to the complexity of regulatory networks in mammalian cells (4). ARACNE has been shown to correctly predict the regulatory network of c-Myc in human B cells (4) and to accurately reconstruct the TRN responsible for epithelial-to-mesenchymal transformation in human brain tumors (3). In ARACNE, a TF and putative target gene are predicted to interact if the mutual information between their expression levels is above some set threshold. An additional feature of the ARACNE algorithm is that it uses the data processing inequality (DPI) technique to eliminate the majority of indirect interactions inferred by coexpression methods. In this way, ARACNE both predicts interactions between TFs and putative target genes and performs variable reduction, retaining only the strongest, direct interactions.

LARS is a regression algorithm used for inferring relationships between a dependent variable (in this case putative target gene expression) and one or more independent variables (predictors; in this case TFs inferred by ARACNE). LARS outputs a quantitative model, which can be used to predict the response variable on the basis of the states of the dependent variables. LARS is a model selection algorithm, which is a less greedy version of the traditional forward selection method. It selects a parsimonious set of predictors from a large collection of possible covariates for the prediction of a response variable (10). Given that there is almost no physical TF–target interaction data available for honey bees, it seemed prudent to emphasize parsimony at every step.

There are several advantages of our approach for network inference over using ARACNE or LARS alone. First, whereas ARACNE gives only the topology of the network, this combination outputs both the topology and also whether each interaction is activating or inhibitory. Second, LARS allows one to estimate the percentage of the variance in a target gene's expression, which is explained by each TF, and predict the expression of a target gene in new conditions. Third, by performing variable reduction and removing indirect interactions in ARACNE, we avoid problems faced by LARS with large numbers of interactions and highly correlated variables (10). Most importantly, ASTRIX only selects the subset of TF–gene relationships that can accurately predict the target gene's expression quantitatively in the training set, and only these interactions are moved forward to the test set for validation.

Cross-Validation and Accuracy of Fit. The quantitative output of the TRN model enabled us to test the accuracy of the predictions in new conditions. All data used in this procedure were normalized before network inference to have row variances of 1; this normalization means that for the influence of a TF on a given target gene, we can uniformly interpret the magnitude of the coefficients α_i and β_i and use their magnitudes to rank the individual interactions (29). We quantified the accuracy of the inferred model by measuring the RMSD. RMSD has the same units as variance, thus providing an estimate for the amount of variance of the gene explained by the model. We also quantified the Pearson correlation between the model's predictions and the actual expression data.

We performed two tests to determine the TRN's ability to predict the expression of each gene in new states. In 10-fold cross-validation using random subsets of samples as test sets, using the expression levels of the TFs alone, our model made highly accurate gene expression predictions with a mean RMSD of 0.35. We selected only the genes that we could predict well
in the training set with a RMSD threshold = 0.5, and the resulting mean RMSD$_{\text{train}}$ was 0.35. We repeated the analysis using correlation as a metric. Correlation also yielded the same result: the test correlation was as good as the correlation in the training set (0.87), suggesting a good fit to the data without overfitting. Second, we performed “leave-one-state-out” cross-validation, in which all of the bees from 1 of the 48 states were removed from the training set; this is a more stringent test in that the model is tested on a “new” behavioral state, i.e., for which it has not been trained. We again found an equally good accuracy in training and test set using a RMSD threshold of 0.5 in training for selecting genes. (RMSD$_{\text{train}}$ = 0.33; RMSD$_{\text{test}}$ = 0.32; SI Appendix, Fig. S4).

After estimating the accuracy of fit in cross-validation, we then used the entire dataset and inferred the final network using ARACNE and LARS. We then selected genes that could be predicted with correlation >0.8 to be in the final model—exactly as we had done on each training set in the error estimation stage. The mean correlation in the total set was 0.87. We built the final network on the basis of the LARS model for each gene. We chose the TF-target gene interactions that had a regression p-value greater than 0.1. The  β-values (regression coefficients) give an estimate of the amount of variance explained by each TF. The putative regulators with regression coefficients of less than 0.1 were pruned out and the final network was determined, which consists of 2,176 genes and 190 TFs.

**Estimation of Background Accuracy and Controls for Expression Prediction.** We performed three experiments to estimate the model’s accuracy on control conditions. (i) We inputted a random dataset, the same size as our dataset, normalized and preprocessed in the same way, and then tried to infer from it a network model. This control model did not predict any interactions significantly below the required mutual information threshold (P < 1E-6), suggesting that very few, if any, interactions obtained in the experimental bee brain TRN could have occurred by random chance. (ii) We used the honey bee expression data and then randomly shuffled the expression of TFs (but not putative targets) across experiments. Once again, no significant interactions were predicted. Although some interactions had high mutual information, none of the predicted interactions had a training correlation greater than 0.8. The average test set correlation of this control model across all 9,546 genes was only 0.22 in 10-fold cross-validation.

These results suggest that the accurate predictions of the bee brain TRN (r = 0.8; mean test set correlation, r = 0.87) were specifically due to interactions between TFs and their predicted target genes rather than to a cryptic structure in the dataset or the modeling approach used. (iii) We tested the ability of the bee brain TRN to model gene expression data in a prominent subregion of the brain, the mushroom bodies (SI Appendix, Fig. S9). We again trained the model on all whole-brain data, then used the mushroom body-specific foraging experience dataset (SI Appendix, Table S1) as a test set (r = 0.36). To determine the significance of the model’s performance, we measured the test prediction using the mean expression of the TFs alone instead of using the linear model predicted by our approach; the average test set prediction correlation was only 0.01.

**Measuring Relative Gene Shuffling.** We used differential rank conservation (DIRAC) (19) to identify changes in relative gene expression ordering in networks. DIRAC quantifies and assesses expression consistency by ranking the target genes within a selected module on the basis of expression.

**ACKNOWLEDGMENTS.** We extend special gratitude to Alissa Eisenberg for her outstanding microarray “spot finding” work throughout the BeeSpace Project. We also thank S. Sinha for the list of bee cis-regulatory elements; J. Drnevich and C. Fields for database construction; R. E. Page and G. V. Amdam for the high- and low-pollen-hoarding bees; E. Carefoot for contributions to Fig. 28; and D. F. Clayton, J. Earls, C. Fields, C. Funk, J. M. George, M. E. Hudson, T. Gernat, C. C. Lutz, S. Ma, Y. Ono, K. Scott, S. Sinha, M. B. Sokolovski, L. J. Stubbs, and S. Zhong for reviewing the manuscript. BeeSpace experiments funded by National Science Foundation Frontiers in Biological Research Award EF25852 (to B.R.S., Principal Investigator). Additional support for transcriptional regulatory network (TRN) analysis comes from: National Institutes of Health Director’s Pioneer Award 1DP1OD006416 (to G.E.R.) and National Cancer Institute Howard Temin Pathway to Independence Award, Department of Defense W81XWH-08-1-0420, Grand Duchy of Luxembourg, and the Roy J. Carver Charitable Trust (N.D.P.). Microarray data meet minimum information about microarray experiment (MIAME) standards, see SI Appendix, Table S10 for accession nos. for each experiment (http://www.ebi.ac.uk/arrayexpress). The TRN is publicly available at http://price.systemsbiology.net/downloads/.

20. Adryan B, Teichmann SA, FJ, Ay, Y, Ono, K, Scott, S, Sinha, M, B, Sokolovski, L, J, Stubbs, and S, Zhong for reviewing the manuscript. BeeSpace experiments funded by National Science Foundation Frontiers in Biological Research Award EF25852 (to B.R.S., Principal Investigator). Additional support for transcriptional regulatory network (TRN) analysis comes from: National Institutes of Health Director’s Pioneer Award 1DP1OD006416 (to G.E.R.) and National Cancer Institute Howard Temin Pathway to Independence Award, Department of Defense W81XWH-08-1-0420, Grand Duchy of Luxembourg, and the Roy J. Carver Charitable Trust (N.D.P.). Microarray data meet minimum information about microarray experiment (MIAME) standards, see SI Appendix, Table S10 for accession nos. for each experiment (http://www.ebi.ac.uk/arrayexpress). The TRN is publicly available at http://price.systemsbiology.net/downloads/.