Information Dynamics of the Gene Regulatory Network Underlying Mammalian Cortical Area Development

Harrison B. Smith

Arizona, United States

Abstract

Information measures have been proposed as a method to quantitatively distinguish biotic networks from abiotic or random networks [1]. Additionally, there has been increased interest in understanding the underlying control kernels that can map biologically unfavorable systems into biologically favorable ones [2]. Here we use the boolean network representation of mammalian cortical area development to show that information measures can be used to distinguish biologically relevant gene expression trajectories from all possible mammalian cortical area gene expression trajectories. Using dynamic information measures to distinguish non-biologically viable from biologically viable cortical gene expression networks, and the mechanisms to control each has implications for understanding the development of all mammalian brains.

1. Introduction

The information dynamics of yeast and budding yeast cell cycle networks reveal key differences between biological networks and their randomly generated counterparts [1]. Here I am interested in investigating if these differences are present between categories of trajectories of a network interpreted to be biologically relevant. In order to investigate this, I use a boolean model of the gene regulatory network underlying mammalian cortical area development [3].

This particular system is studied because the healthy development of a mammalian adult brain (specifically a mouse brain) is recognized by a particular expression gradient of five genes. Additionally, this healthy adult expression gradient will only be realized if, on embryonic day 8, there is a particular (but different) expression gradient of these same five genes. Thus,

while the initial and desired biological network states are specified by these empirically measured gradients, the rules governing the transition from the initial to final states must be determined through dynamical models.

Previous research has determined the optimal interaction rules necessary to reach the biologically desired final state from the specified initial biological state [3]. I use these interaction rules to generate the state transition map of the gene regulatory network underlying mammalian cortical area development, and analyze how the transfer entropy (TE) and active information (AI) of nodes in the network differs between the biologically relevant trajectory, trajectories within the biologically relevant primary attractor, and trajectories within any attractor of the state transition network as a whole. This study is motivated by previous research by Kim et al., 2015, which found that biological systems are distinctive in their informational architecture [1]. An open question of this research that I address here is: does the distribution of TE among causally connected node pairs in networks processing spatial information (such as the mammalian cortical network in this study) match networks processing temporal information (such as the cell-cycle networks) [1]?

In the same research, it was determined that the information processing unique to biology is a result of the regulation of function by a few key nodes. These nodes also happen to be the same subset of nodes that regulate the attractor dynamics associated with biological function, known as the control kernel [1][2]. In this study, I determine the presence of control kernels in the mammalian cortical network, compare to previous results on this networks' control kernels, and attempt to identify how they could play a role in shaping the information measures we observe in these networks [2].

2. Model Description

2.1. Network and dynamics

The boolean model of the gene regulatory network underlying mammalian cortical area development is described by ten nodes, *Fgf8*, *Emx2*, *Pax6*, *Coup-tfi*, *Sp8*, Fgf8, Emx2, Pax6, Coup-tfi Sp8. The italicized nodes represent genes, and the non-italicized nodes represent proteins (for clarity, I will herein denote the ten above nodes as gF, gE, gP, gC, gS, pF, pE, pP, pC, pS). Each node can either be in state 0 (inactive) or 1 (active). Nodes can be connected by an activation link or inhibition link. The interactions between nodes was determined in Giacomantonio and Goodhill, 2010 [3] by iterating through

all possible connections of nodes, where the best interaction network was deemed to be the one that ended up in the biologically desired state with the greatest probability. The rule for updating a nodes state is as follows: a node is activated if in the previous time step all nodes regulating it though activation links are activated, and all nodes regulating it through inhibition links are inactive. In other words, all conditions must be met for a node to become activated. In the original work used to determine the most likely interaction network [3], the network update rules are deterministic, but were implemented stochastically. This was to avoid synchronous node updating, seen by the authors as unrealistic. Additionally, research on mammalian cortical area development actually recognizes different initial and desired expression gradients in the anterior and posterior areas of the cortex.

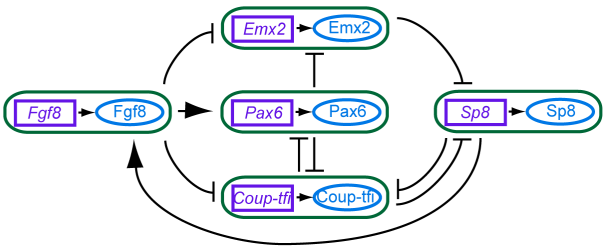

Figure 1: The most likely representation of the anterior mammalian cortical network, used for all analyses in this work. Arrows indicate activation edges, and bars indicate inhibition edges. Gene nodes (rectangular/purple) and protein nodes (elliptical/blue) are encapsulated for clarity (figure adopted from Giacomantonio and Goodhill, 2010 [3]).

When determining system dynamics for my information analyses, I followed strictly deterministic update rules using the best interaction network as determined by previous work [3] (See Fig. 1). As a consequence of following deterministic update rules, it is impossible to reach the desired posterior steady state from the posterior initial state 3, as they are in different attractor basins. Thus, my control kernel analyses and information dynamic analyses reference only the anterior network.

The biological initial state active nodes and biologically desired steady state active nodes are shown side-by-side with all nodes in Table 1.

2.2. Information measures

The information measures used this study are transfer entropy (TE) and active information (AI). Transfer entropy is defined as the deviation from

Anterior Mammalian Cortical Network: Initial and Desired States										
	gF	gE	gP	gC	gS	pF	pE	pP	pC	pS
Biological initial state	■					■				
Desired steady state	■		■		■	■		■		■

Table 1: All network nodes from Fig. 1 are shown in the top row, with abbreviations corresponding to the genes *Fgf8*, *Emx2*, *Pax6*, *Coup-tfi*, *Sp8*, and the proteins Fgf8, Emx2, Pax6, Coup-tfi Sp8 respectively. Nodes active in the biological anterior initial state, and biologically desired anterior steady state are shaded gray.

independence (in bits) of the state transition of an information destination X from the previous state of an information source Y [4].

$$T_{Y\to X}(k) = \sum_{x_{n+1},x_n^{(k)},y_n} p(x_{n+1}, x_n^{(k)}, y_n) \log_2 \frac{p(x_{n+1}|x_n^{(k)}, y_n)}{p(x_{n+1}|x_n^{(k)})}$$

Here n is a time index, $x_n^{(k)}$ refers to the k states of X up to and including x_n. Transfer entropy can be viewed as conditional mutual information. Because it is conditional, it is directional ($T_{Y\to X}(k) \neq T_{X\to Y}(k)$), unlike mutual information.

The active information of X is defined as the average mutual information between its semi-infinite past $x_n^{(k)}$ (as $k \to \infty$) and its next state x_{n+1} at time step $n+1$ [5].

$$A_X(k) = \sum_{x_{n+1},x_n^{(k)}} p(x_{n+1}, x_n^{(k)}) \log_2 \frac{p(x_{n+1}|x_n^{(k)})}{p(x_{n+1})p(x_n^{(k)})}$$

As with transfer entropy, n is a time index, $x_n^{(k)}$ refers to the k states of X up to and including x_n.

The active information can be interpreted as the average information in the next state of the destination x_{n+1} that is contained in the destination's past k, while the transfer entropy can be interpreted as the average information in the source y_n about the next state of the destination x_{n+1} that was not already contained in the active information $A_X(k)$. Both TE and AI are a measure of observed correlation, and not direct effect. A source can store information regardless of whether it is causally connected with itself (in the case of active information), or the destination (in the case of transfer en-

tropy). Thus, information storage can be facilitated in a distributed fashion via ones neighbors [5].

3. Results

3.1. Attractor

The state transition diagram of a network under deterministic update rules shows how each network state maps to another network state. Because I have 10 nodes which can each take on two states, there are $2^{10} = 1024$ possible network states in my state transition diagram. It consists of 3 basins of attractions (Fig. 2). The largest basin consists of 841 states and contains the initial state as well as the biologically desired steady state, which is a fixed attractor. The second largest basin consists of 174 states, with a two-state cyclic attractor. The smallest basin consists of 9 states, with a fixed attractor.

3.2. Control Nodes

Control kernels were identified using two different definitions (denoted here as type I and type II). In the first (type I), the control kernel is defined as the minimal set of network nodes needing to be regulated to drive the network state to converge to any desired attractor regardless of the systems initial state [2]. Using this definition, previous work [2] found there to be six control kernels of one node each (See Table 2).

Repeating the analysis, I also found there to be six control kernels of one node each (any of which must be active to reach the desired attractor state), but I found the Pax6 gene and protein to be control kernel nodes instead of the Emx2 gene and protein (See Table 2).

Anterior Mammalian Cortical Network: Type I Control Kernels										
	gF	gE	gP	gC	gS	pF	pE	pP	pC	pS
Kim et al., 2013 [2]	■	■			■	■	■			■
This study	■		■		■	■		■		■

Table 2: All network nodes from Fig. 1 are shown in the top row, with abbreviations corresponding to the genes *Fgf8*, *Emx2*, *Pax6*, *Coup-tfi*, *Sp8*, and the proteins Fgf8, Emx2, Pax6, Coup-tfi Sp8 respectively. Single node control kernels for the primary, biological attractor of this network as determined by Kim et al., 2013 [2] and this study are shaded gray.

My result is consistent with the desired steady state of my system, where the Emx2 and Coup-tfi genes and protein nodes are all inactive, and all other nodes are active. I believe the previously identified type I control kernel by Kim et al., 2013 [2] to be erroneous, because by their own definition of the control kernel, the desired attractor state is impossible to obtain by pinning the Emx2 gene or protein to the active state, and here was found not to converge on the primary attractor when pinned to the inactive state.

Besides driving the network to the primary attractor basin, I found there to be type I control kernels that drive the network state to the smallest basin of attraction, but not the remaining intermediate-size basin of attraction. This intermediate-size basin differs from the others in that it has a cyclic attractor comprised of two states instead of a single fixed attractor state. Perhaps this difference affects the causal power of the control kernel, preventing it from arriving in this attractor basin. Further analysis is necessary to either support or reject this hypothesis.

I also investigated a second definition of a control kernel (type II), defined by the minimal set of nodes needed to be active in order to guarantee that, regardless of the activity of the other nodes, the state will be in the primary attractor and thus converge on the desired steady state. Unlike the type I control kernels, regulating a network to a specific attractor basin using these type II kernels does not require intervening every time step to update a node. Instead, these type II control kernels define the minimum set of nodes that must be regulated only once in order to reach a specific attractor. Abiding by this definition, I found nine control kernels of two nodes each (See Table 3). Each type II control kernel is comprised exclusively of nodes that also happen to be a type I control kernel, and are active as part of the biologically desired steady state.

While there is no precedent that I am aware of for a control kernel with this definition, I believe it to be a useful concept when thinking about how biological systems ensure a trajectory to the functional attractor state. If an error results in a network state that pushes the system to a non-desired attractor basin, there is only a single intervention required for the system to move back to the primary attractor basin (i.e. making sure that one of the 9 type II control kernels are active). In contrast, to guarantee the system arrive at the functional attractor state using the type I control kernel, an intervention would be required every time the system nodes get updated.

Anterior Mammalian Cortical Network: Type II Control Kernels										
	gF	gE	gP	gC	gS	pF	pE	pP	pC	pS
Kernel 1	■					■				
Kernel 2	■							■		
Kernel 3	■									■
Kernel 4			■			■				
Kernel 5			■					■		
Kernel 6			■							■
Kernel 7					■	■				
Kernel 8					■			■		
Kernel 9					■					■

Table 3: All network nodes from Fig. 1 are shown in the top row, with abbreviations corresponding to the genes *Fgf8*, *Emx2*, *Pax6*, *Coup-tfi*, *Sp8*, and the proteins Fgf8, Emx2, Pax6, Coup-tfi Sp8 respectively. The type II control kernels for the primary, biological attractor of this network as determined in this study are shaded gray. Each control kernel is comprised of two nodes.

3.3. Information Dynamics

I used a time series length of 22 to calculate transfer entropy and active information because this is the maximum number of steps needed to drive any network state to its basins attractor. Because the biological initial network state converges to the desired steady state attractor in only 7 time steps, the history length was limited to 6 at the highest. History lengths between 2 and 5 yielded similar results for rank ordered transfer entropy and active information among each of the datasets. A history length of 2 was chosen for all analyses because it is the shortest history length, relative to the biological trajectory length, that still yields similar results to longer history lengths for the aforementioned distributions.

When comparing rank ordered TE to AI across the different trajectory regimes, there are a few things that stand out (Fig. 3). The biological trajectory has the lowest TE and highest AI, while TE is similar across trajectories in the primary attractor and cumulative across the network. Despite the similarity in TE, nodes in trajectories of the primary attractor show much lower AI than across all trajectories. Regardless, the absolute AI is lower than the absolute TE across all ranks and regimes.

Looking at TE vs edge connections, it can be seen that in both the edge case and no edge case, the primary attractor and cumulative trajectories

show more node pairs with transfer entropy than without (Fig. 4). The majority of nodes in these two regimes are correlated via information transfer but without causal connection (the percentage of nodes with TE > 0 is higher in the no edge case than the edge case). The majority of node pairs in the biological state trajectory have no transfer entropy. Of those that do, more are correlated via information transfer without causal connection than with causal connection. This data provides an answer to the open question posed by Kim et al., 2015 and repeated in the introduction of this paper. The distribution of TE among causally connected node pairs in networks processing spatial information (such as this network) appear to match networks processing temporal information (such as the cell-cycle networks) [1].

The only nodes in the biological trajectory (Fig. 5**C**) which have non-zero AI are gF, gP, pF, and pP. Besides being control nodes, these are all the only gene/protein pairs that are induced by other nodes (Fig. 1). All node pairs with non-zero TE are also control nodes. While pS→pF may look out of place, but pS induces gF, which induces pF. The chain of pS→gF→pF are all inductive connections whose source nodes are not at all inhibited, which could explain why they all have identical TE when targeted nodes they directly connect to. When comparing the TE to the AI of the control nodes in the biological trajectory, the only target node whose present state is informed more by its own past state than by another node is gF. Interestingly, gF is also the morphogen, or the signaling molecule used during embryonic development to determine the location, differentiation and fate of surrounding cells [6]. Across the primary attractor trajectories (Fig. 5**B**), there are no target nodes whose past states are more predictive of its present state than another node (i.e. the TE is higher than the AI for at least one pair of nodes across all target nodes). While the control kernel nodes tend to have the highest TE both when acting as source nodes and target nodes, the distinction is not as clear cut as that of the biological attractor. There is a trend in that the highest TEs are from genes→proteins or proteins→genes, which are the only causally/directly connected nodes in the network. Now looking across all trajectories (Fig. 5**A**), we again see the AI of target nodes approach values of TE to those nodes, and in the case of gE its AI is even higher than its TE. Otherwise, this heatmap has a similar distribution to that of the primary attractor trajectories, where the TE still tends to be higher among control kernel nodes, and the highest TEs are from genes→proteins or proteins→genes.

The information processing unique to the biological trajectory thus ap-

pears to be very closely aligned with the biological control kernels. However, the information processing across all trajectories in the primary attractor basin, and across all trajectories in all attractor basins also appear to be somewhat aligned with the control kernels we determined. Perhaps confounding to these results is the fact that the primary attractor basin, containing the biologically desired steady state whose active nodes are identical to the type I control kernels, is also by far the largest attractor basin. Thus when looking at information dynamics averaged over all trajectories they will be strongly weighted to the dynamics of the primary attractor trajectories. A stronger case could be made for the control kernels uniquely aligning with the information processing in the biological trajectories if the information processing of the non-biological trajectories were analyzed independently of the primary attractor trajectories.

4. Summary/Discussion

In studying the dynamics of the boolean model of the gene regulatory network of mammalian cortical area development, I found six control kernel nodes, two of which vary from existing literature. I also laid out the definition for a second type of control kernel, characterized by the idea that there is a way to control the attractor basin for a network without causually intervening in every step of the network dynamics.

I also investigated how, in this network, the biologically favored trajectory differs from other trajectories within the biologically favored (primary) attractor, and ones that comprise the entire state transition diagram. The information processing unique to the biological trajectory appears to be very closely aligned with the biological control kernels. However, the information processing across all trajectories in the primary attractor basin, and across all trajectories in all attractor basins also appear to be somewhat aligned with the control kernels we determined. A stronger case could be made for the control kernels uniquely aligning with the information processing in the biological trajectories if the information processing of the non-biological trajectories were analyzed independently of the primary attractor trajectories.

It appears that when looking at node pairs which have non-zero transfer entropy across any trajectory regime, the information transfer is facilitated without casual connection (between nodes not directly connected by an edge). Thus, the distribution of TE among causally connected node pairs in networks processing spatial information (such as this network) appear to match

networks processing temporal information (such as the cell-cycle networks) [1].

While it is hard to say anything confidently about what the implications of the other intricacies of the TE and AI mean for these systems, there are many more interesting questions to ask about this network, especially when taking its stochasticity into account. How would TE and AI across this deterministic implementation of the gene regulatory network underlying mammalian cortical area development differ from trajectories sampled from a stochastic implementation? How do the AI and TE of trajectories sampled from a stochastic implementation which successfully reaches the desired biological attractor differ from those trajectories which do not reach the desired biological attractor? How do the information dynamics differ between anterior and posterior stochastic trajectories?

5. References

[1] H. Kim, P. Davies, S. I. Walker, New scaling relation for information transfer in biological networks, Journal of The Royal Society Interface 12 (2015) 20150944.

[2] J. Kim, S.-M. Park, K.-H. Cho, Discovery of a kernel for controlling biomolecular regulatory networks, Scientific reports 3 (2013).

[3] C. E. Giacomantonio, G. J. Goodhill, A boolean model of the gene regulatory network underlying mammalian cortical area development, PLoS Comput Biol 6 (2010) e1000936.

[4] J. T. Lizier, M. Prokopenko, Differentiating information transfer and causal effect, The European Physical Journal B 73 (2010) 605–615.

[5] J. T. Lizier, S. Pritam, M. Prokopenko, Information dynamics in small-world boolean networks, Artificial Life 17 (2011) 293–314.

[6] J. Gurdon, P.-Y. Bourillot, Morphogen gradient interpretation, Nature 413 (2001) 797–803.

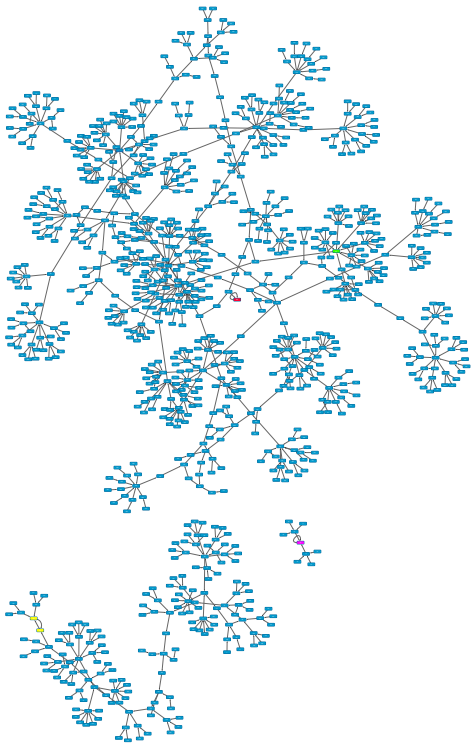

Figure 2: The deterministic state transition diagram of the anterior mammalian cortical network as modeled in this work (based off of Giamantonio and Goodhill, 2010 [3]). There are three attractor basins. The primary attractor basin is shown at the top, with the biological initial state colored green, and the desired biological attractor state colored red. The next largest attractor basin is shown in the lower right, and has a two-state cyclic attractor (colored yellow). The smallest attractor basin is shown in the lower right, and has a fixed attractor (colored pink).

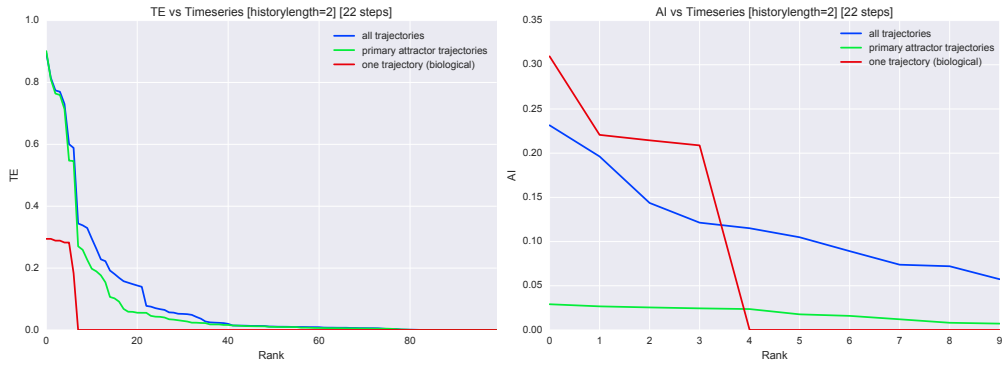

Figure 3: Rank ordered TE (left) and AI (right) across trajectory regimes. The biological trajectory has the lowest TE and highest AI, while TE is similar across trajectories in the primary attractor and cumulative across the network. Despite the similarity in TE, nodes in trajectories of the primary attractor show much lower AI than across all trajectories. Regardless, the absolute AI is lower than the absolute TE across all ranks and regimes.

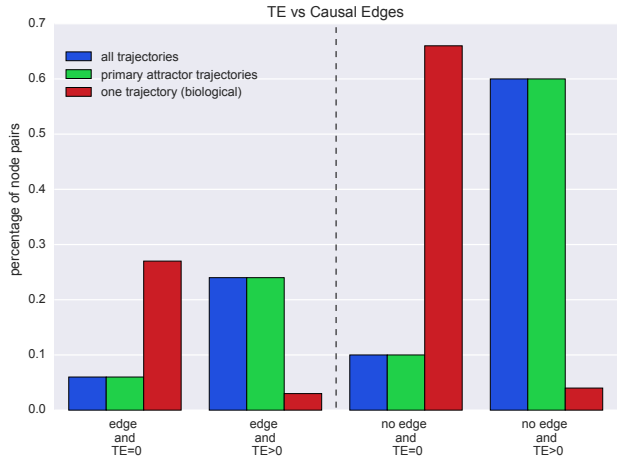

Figure 4: In both the edge case and no edge case, the primary attractor and cumulative trajectories show more node pairs with transfer entropy than without. The majority of nodes in these two regimes are correlated via information transfer, but without causal connection (the percentage of nodes with TE > 0 is higher in the no edge case than the edge case). The majority of node pairs in the biological state trajectory have no transfer entropy. Of those that do, more are correlated via information transfer without causal connection than with causal connection.

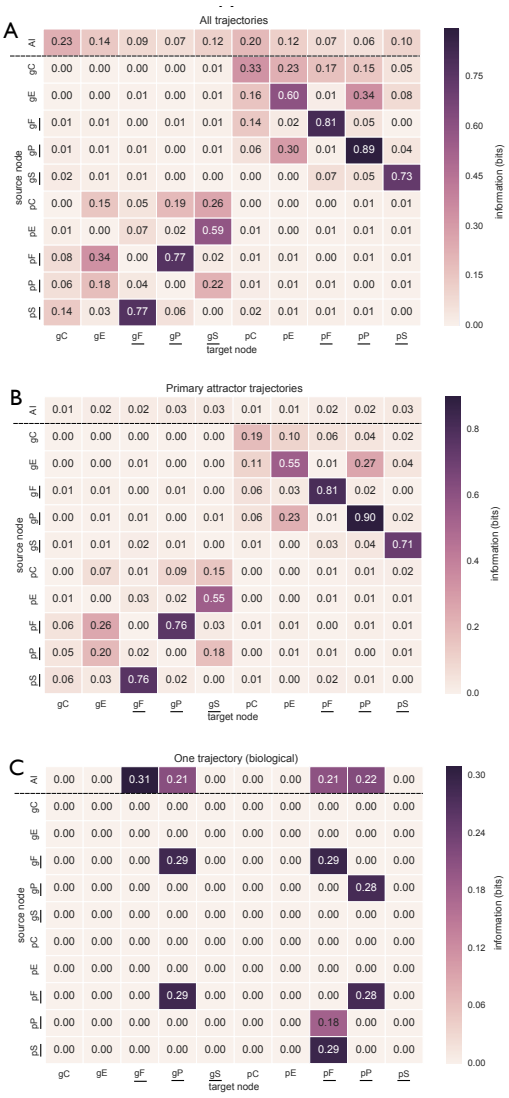

Figure 5: Heatmaps displaying TE from source to target nodes, and AI of target nodes (top row of each heatmap). All data is for a history length of two, and a time series length of 22 steps. Underlined nodes denote control kernels as determined in this study. TE and AI averaged across all attractor trajectories is shown in **A** (top). TE and AI averaged across all the primary attractor trajectories is shown in **B** (middle). TE and AI in the biological trajectory is shown in **C** (bottom). Each colorbar is scaled for its local heatmap.