Distributed and dynamic intracellular organization of extracellular information

Distributed and dynamic intracellular organization of extracellular information

Alejandro A. Granados*, Julian M. J. Pietsch†, Sarah A. Cepeda-Humerez‡, Iseabail L. Farquhar†, Gašper Tkačik‡, Peter S. Swain†

Although cells respond specifically to environments, how environmental identity is encoded intracellularly is not understood. Here we study this organization of information in budding yeast by estimating the mutual information between environmental transitions and the dynamics of nuclear translocation for ten transcription factors. Our method of estimation is general, scalable, and based on decoding from single cells. The dynamics of the transcription factors are necessary to encode the highest amounts of extracellular information, and we show that information is transduced through two channels: generalists (Msn2/4, Tod6 and Dot6, Maf1, and Sfp1) can encode the nature of multiple stresses but only if stress is high; specialists (Hog1, Yap1, and Mig1/2) encode one particular stress, but do so more quickly and for a wider range of magnitudes. In particular, Dot6 encodes almost as much information as Msn2, the master regulator of the environmental stress response. Each transcription factor reports differently, and it is only their collective behavior that distinguishes between multiple environmental states. Changes in the dynamics of the localization of transcription factors thus constitute a precise, distributed internal representation of extracellular change. We predict that such multi-dimensional representations are common in cellular decision-making.

All organisms sense their environment and internally represent the information gained to elicit a change in behavior [1]. Much is understood about such representations in neural systems [2], but single cells must perform an analogous task [1, 3], encoding intracellularly the information about extracellular environments, and yet little is known about the nature of their encoding.

The activation of transcription factors is thought to provide an internal representation of a cell's environment [4, 5, 6, 7, 8, 9, 10], but how information is encoded dynamically, whether information is spread across multiple factors, and how information is read downstream all remain unclear (Fig. 1A). We do know that the biochemical implementation of such representations is likely to be stochastic [11] and that the same biochemistry may be used to sense disparate environments. Furthermore, cells typically have just 'one shot' at mounting the appropriate response from these internal representations, with competition being unforgiving for those that delay, at least among microbes [12, 13, 14]. Here we use information theory to investigate how eukaryotic cells answer these challenges.

To do so, we turn to budding yeast and to environmental changes for which we expect information encoding to be key: stresses that compromise growth and evoke adaptive gene expression [15]. In yeast, extracellular changes are sensed by signaling networks that regulate the activity of transcription factors, often by their translocation either into or out of the nucleus [16], analogous to p53 and NF-κB in mammalian cells [17, 18]. We therefore consider the movement of these transcription factors as a cell's internal representation of an environmental transition. The translocations are dynamic and stochastic, and the information available from the full time-series of the response could be substantially higher than that available from any temporal snapshot [9] (Fig. 1A).

Tens of transcription factors translocate in yeast [16], and we focus on a representative subset: Msn2 and its paralog Msn4 which drive the environmental stress response; Mig1 and its paralog Mig2 which respond to low glucose; Hog1, (a kinase) which responds to hyperosmotic stress; Yap1 which responds to oxidative stress; Sfp1 which promotes and Dot6 and its paralog Tod6 which repress the biogenesis of ribosomes; and Maf1 which represses the synthesis of tRNAs. We include Dot6 and Tod6, which are little studied, to determine if our approach can help determine their physiological importance. Some

*: Department of Bioengineering, Imperial College London, SW7 2AZ, UK; † SynthSys – Synthetic & Systems Biology, School of Biological Science, University of Edinburgh, EH9 3BF, UK; ‡ Institute of Science and Technology Austria, 3400 Klosterneuburg, Austria

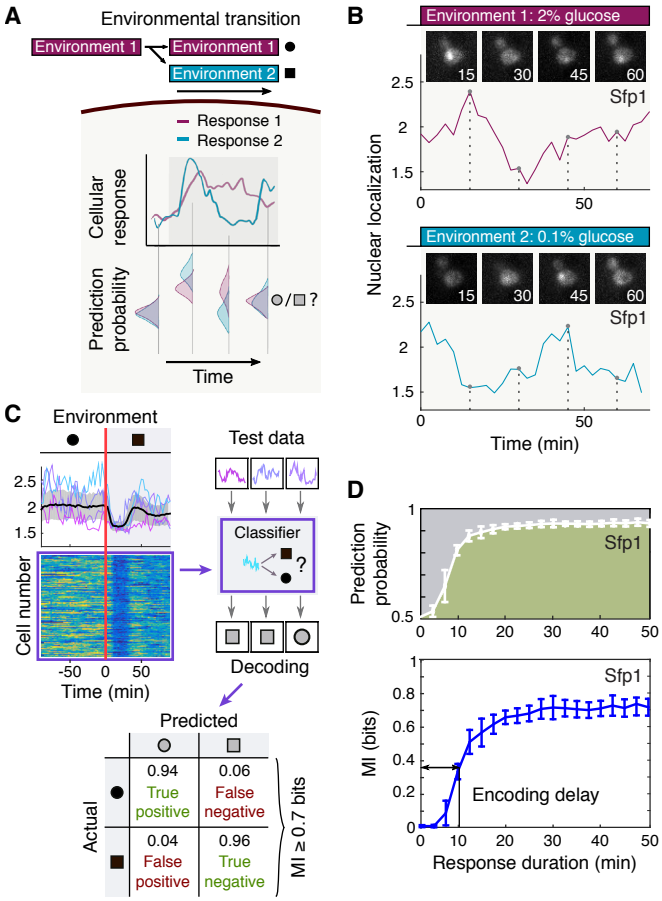

Figure 1: Intracellular responses carry information on extracellular change. (A) Environmental transitions trigger dynamic intracellular responses (purple and blue time-series), which may be the only indications to the cell that the environment has altered. Stochasticity can make intracellular responses a poor readout of the new environment, and a probabilistic representation of the possible environmental state based on the instantaneous intracellular response may vary over time. (B) Our approach compares single-cell time-series of nuclear translocation (Sfp1 is shown) in different environments. The transcription factors are tagged with Green Fluorescent Protein to reveal their dynamics (inset & Fig. S2). (C) We estimate the mutual information between the time-series and the environment by using 70% of the data to train a classifier to classify the time-series into the two environments (each coloured line in the heat map is a single cell with greater nuclear localization indicated by yellow and less by blue). With the remaining data, we determine the confusion matrix (the probability of correctly and incorrectly identifying the environment from a time-series) and, from this matrix, a lower bound on the mutual information. (D) By increasing the length of the time-series used by the classifier (each time-series has the same initial point), we estimate both the probability of correct and incorrect predictions and the mutual information as a function of the duration of the response. Here the actual environment is 0.1% glucose (green area indicates the fraction of correctly classified time-series). At $t = 0$, the two environments cannot be distinguished, and the best prediction is a random guess, which has a prediction probability of 0.5, corresponding to 0 bits of information. Error bars are S.D. across biological replicates ($n = 6$).

of these factors (Msn2/4, Mig1/2, and Dot6/Tod6) have pulsatile dynamics, with stochastic bursts of nuclear localization even without stress [19].

We consider environmental shifts from rich media (2% glucose) into either carbon stress (0.1% glucose), hyperosmotic, or oxidative stress. Using fluorescent tagging and microfluidics [20], we measure the degree of nuclear localization of the transcription factors in hundreds of single cells both before and after the stress is applied (Fig. 1B).

Results

To quantify the information available to the cell, we developed a general and scalable methodology to estimate the mutual information between the time-series of cellular responses and the state of the extracellular environment (SI Appendix; Fig. 1C & Figs. S4–S7). Mutual information, a measure of statistical dependency [21], allows us both to capture the effects of biochemical stochasticity, which can drive individual responses far from the average, and to avoid *a priori* assumptions about which features of the response are relevant, such as its magnitude or duration.

Our method involves training a classifier to predict the state of the environment from the time-series of as few as 100 cells. The classifier's output on the test data can then be used to estimate the mutual information (Fig. 1C). Formally, this approach provides a lower bound on the true information (SI Appendix), but, biologically, we are quantifying the information that a cell could plausibly recover and act upon after observing a single time-series of its response. By varying the duration of the time-series used by the classifier, we can determine how quickly cells accumulate information (Fig. 1D). In addition, errors made by the classifier indicate environments that are likely to be confused, giving insight into tasks potentially challenging for the cell. Although the final estimate of mutual information is determined by the signal-to-noise ratio of the data, this noise may not be all biological, again making our estimates a lower bound.

Mutual information is also determined by the choice of input distribution. Our focus is to use mutual information to quantify the signal-to-noise ratios in the single-cell time-series, and therefore we choose a uniform distribution, which does not favor any one input. When we repeat the estimations using the input distribution that maximizes the mutual information, we see few changes (Fig. S18). Indeed, in the true (natural) input distribution, stresses may not even occur one at a time as we have assumed.

Detecting environmental change

Considering transitions into one of three environments, all of which reduce growth (Fig. 2A), the 10 transcription factors have diverse dynamics (Fig. 2B). The mutual information we calculate addresses whether an environmental transition can be detected from a typical time-series of nuclear localization and is a number between 0 bits (indicating no detectable statistical differences between the dynamics of localization before and after the transition) and 1 bit (the dynamics of localization before and after the transition are distinct).

The glucose specialists Mig1/2 perform almost optimally in carbon stress with almost the maximum possible information of 1 bit (Fig. 2C). We note that different transcription factors encode information in different ways. For example, Dot6 and Sfp1 have dissimilar dynamics (Fig. 2A), yet encode the same amount of bits (Fig. 2C). Paralogs, however, do not necessarily carry equal information (compare Tod6 and Dot6).

Information can be encoded within minutes of the environmental transition with no trade-off: the speed of encoding typically increases the more information is encoded. Defining the encoding delay as the time for the mutual information to reach 50% of its maximum, information plateaus earliest in the time-series of Mig1/2 (Fig. 2D). Fast responders are therefore typically more accurate, at least for such large stresses.

These general observations hold for transitions into osmotic and oxidative stress (Fig. 2E & 2F), establishing a hierarchy: in terms of the information and encoding delays, specialists (in blue) are followed by the environmental stress response (in pink), which in turn is followed by the others. The details of this hierarchy, however, are stress-specific, indicating that the dynamics of the transcription factors may encode not only the presence but also the nature of the environmental transition.

Detecting the nature of environmental change

We therefore extended our method to calculate the mutual information between a single-cell time-series and the four environmental states: rich media (before the environmental transition) and the three stresses (after the transition). Not only do we estimate the mutual information (Fig. 3A), but

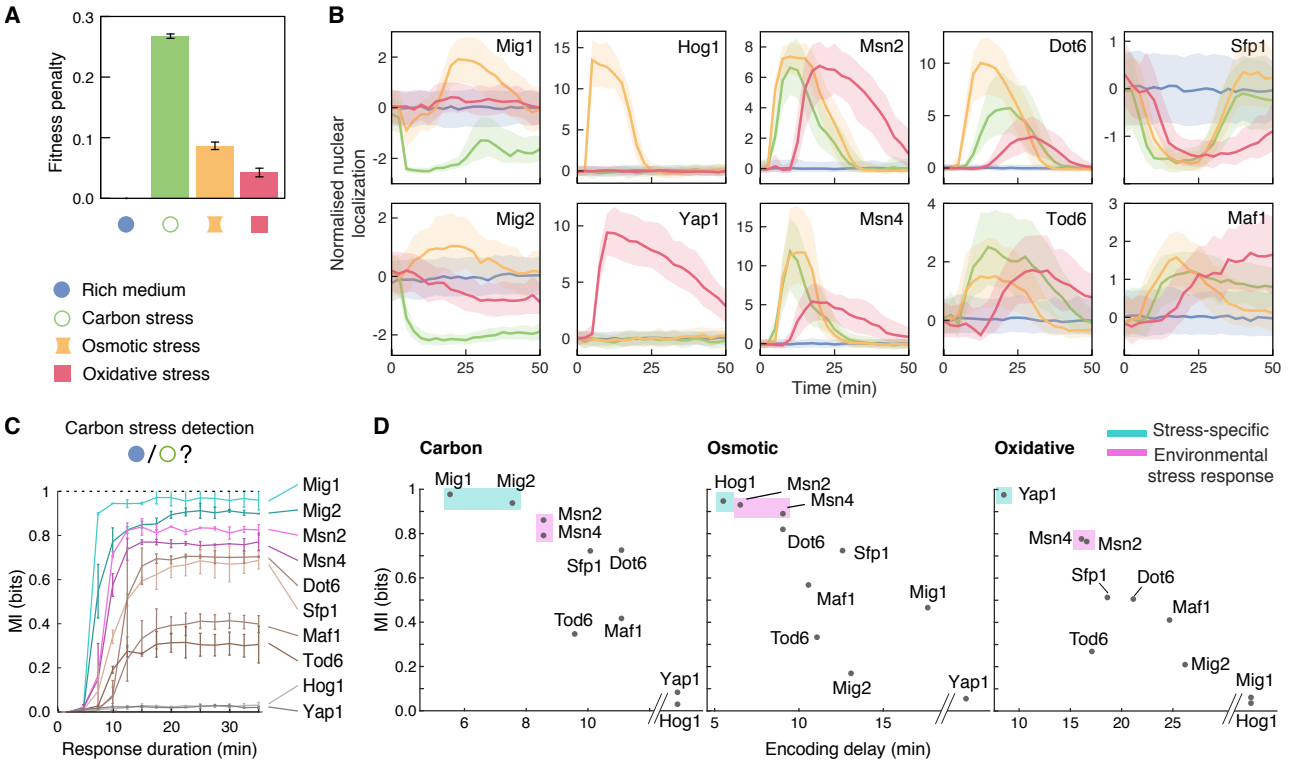

Figure 2: A hierarchy of information encoding, both in bits and encoding delays, holds across environmental transitions. (A) From population measurements (Fig. S3), each new environment, either carbon stress (0.1% glucose), osmotic stress (0.4M NaCl), or oxidative stress (0.5mM H_2O_2), reduces growth compared to growth in rich media (2% glucose). (B) For 10 transcription factors, we quantify nuclear localization across 4 environments using a step-change from rich media to stress at $t = 0$. The median and interquartile range are shown. (C) In response to carbon stress, the mutual information shows a hierarchy. The maximum possible information is 1 bit (dotted line). Mean and SD of 2 experiments per transcription factor. (D) The hierarchy's order is maintained for transitions into other stresses: specialists (blue) encode the most information and are fastest, followed by the environmental stress response (pink).

we can also predict how a typical time-series is likely to be classified as a function of the duration of the environment (Figs. 3B, S11 & S12).

Although no single transcription factor reaches the maximum of 2 bits, the time-series of Msn2, Msn4 and, unexpectedly, Dot6, carry sufficient information to identify three environmental categories (for example, two environmental states and the remaining two states lumped together) (Fig. 3A). We observe, however, that the information is instead 'spread' so that all environmental states are eventually classified with a greater than 80% accuracy (Fig. 3B: Dot6). In contrast, an ideal specialist should perfectly discriminate one environmental state and lump together the remaining states to encode 0.8 bits (SI Appendix). Indeed, Hog1 and Yap1 do encode this much information (Fig. 3A), and their signalling networks operate nearly optimally in these high stresses. After only a few minutes, both specialists unequivocally identify their associated stress and never report false positives (Fig. 3B; Yap1).

Conditioning the mutual information on the identity of the environmental states delineates specialists from the other transcription factors (Figs. 3A & S15), which we term generalists because they encode information on multiple types of stress. Nevertheless, these groups are not mutually exclusive: Mig1 is not only a specialist for carbon stress, but also carries information on the other environmental states at late times, particularly osmotic stress, for which the probability of correctly identifying the environment is more than twice the 25% probability of a random choice (Fig. 3B).

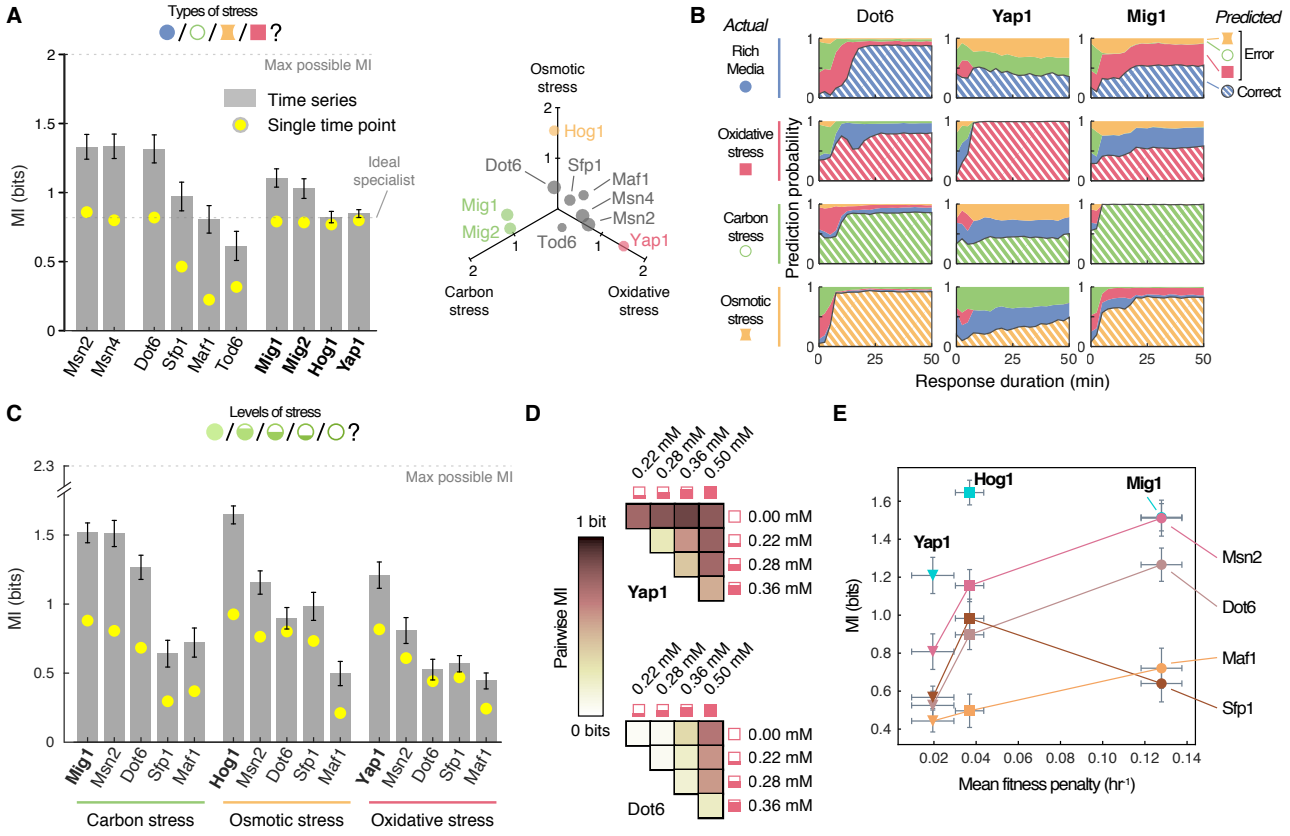

Figure 3: Generalists can identify the nature of large environmental transitions but only specialists distinguish smaller changes. (*A*) The mutual information between a time-series of translocation and the 4 environmental states shows that multiple transcription factors can encode the nature of the environment (errors are SD over bootstrap replicates). The highest mutual information that can be found using a single time point is always lower (yellow dots). Inset: Conditioning the mutual information on the environmental states differentiates the specialists (in color). (*B*) Our methodology also estimates the typical probability of inferring each environmental state from a time-series. The true state is marked with hash lines. Inference generally improves with the duration of the new environment, and rich media and oxidative stress are confused most often. (*C*) The mutual information between a time-series of translocation and 4 levels of one type of stress and rich media. Specialists perform best and encode with their entire time-series (compare yellow dots with those in *A*). (*D*) Comparing the mutual information between the time-series and all pairs of environments with the same type of stress at different levels, generalists can identify high stress, but poorly distinguish adjacent levels compared to specialists. (*E*) The more a stress reduces growth, the more information is typically encoded. Comparing the mutual information from *C* with the mean fitness penalty (averaged over the different levels of stress) for each type of stress, there is a strong correlation (¿ 0.9) for all generalists except Sfp1. Triangles, squares, and circles denote oxidative, osmotic, and carbon stress. Errors are SD.

Detecting the magnitude of environmental change

In such high stresses, specialists appear unnecessary because the generalists identify stress so well, but this situation changes if we consider transitions into stresses of lower magnitude (Figs. 3C & S10). From rich media, we applied four different levels of the same type of stress and estimated the mutual information between the time-series of translocation and the five environmental states. Specialists now outperform generalists. Considering the mutual information between the time-series and all pairs of the different levels of stress (Figs. 3D & S13), we see that distinguishing between adjacent levels is most challenging, and generalists, but not specialists, can often only identify high stress.

Generalists and specialists also encode information differently: generalists often use their entire time-series whereas specialists only do so to distinguish stresses of lower magnitude. By calculating

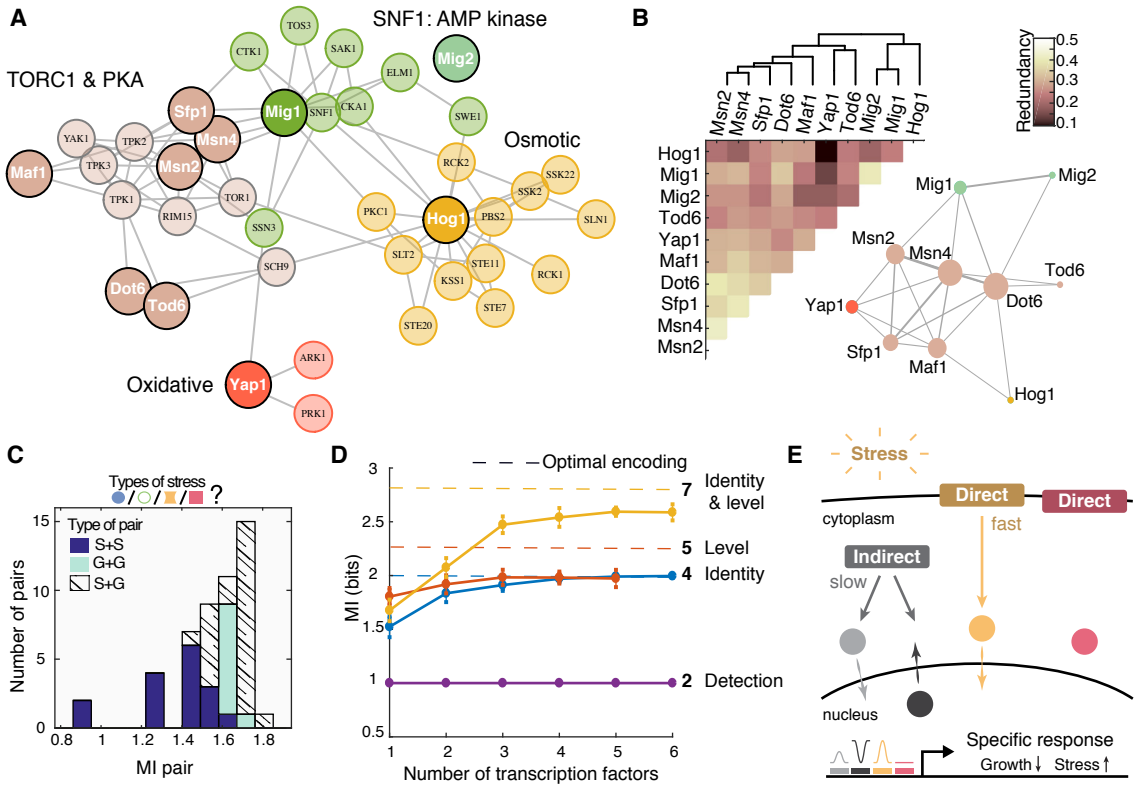

Figure 4: Collectively, transcription factors provide an internal representation of complex environments. (*A*) In the intracellular signaling network, generalists are co-located and specialists are distinct. Edges between kinase and substrate are weighted by the evidence for that interaction [22]. Clusters of highly interconnected components have different colors, and we include only the transcription factors in our study. (*B*) The redundancy in information between pairs of transcription factors reflects the structure of the signaling network in *A*. Edges are proportional to a pair's redundancy, $1 - \mathrm{MI}_{12}/(\mathrm{MI}_1 + \mathrm{MI}_2)$ [23], and the size of each node is determined by the number of its edges. Redundancy was calculated from an average of 2 datasets (6 experiments per transcription factor). (*C*) A generalist and a specialist typically encode the most information out of all possible pairs of transcription factors. For each pair, time-series from one transcription factor were concatenated randomly with the time-series of another to calculate the mutual information. Legend: S – specialist; G – generalist. (*D*) Only through combinations of transcription factors can information on complex environments be encoded. Different colors represent different numbers of environmental states (SI Appendix): orange – 7 states; red – 5 states; blue – 4 states; purple – 2 states. Plotted lines are maxima calculated from all combinations of concatenated time-series for a given number of transcription factors. (*E*) Cells transduce information through two types of channel—specialists and generalists—and four transcription factors are sufficient to encode all the available information. The direct channels (specialists) respond to extracellular change; the indirect channels (generalists) respond to intracellular change and so to broader categories of extracellular change.

the mutual information between summary features of the single-cell time-series and the environmental state (Figs. S16 & S17), we found that the amplitude of a specialist's initial translocation can identify its associated stress if that stress is sufficiently severe (yellow dots in Fig. 3A), explaining specialists' short encoding delays. For transitions into stresses of lower magnitude, however, information is encoded in the dynamics of the specialists' response (yellow dots in Fig. 3C). Generalists can encode twice the amount of information in their time-series compared to the highest information encoded by any single time point, and both the timing of their initial translocation, particularly for Msn2 and Dot6, and the shape of the times series can be important (SI Appendix).

For 3 out of the 4 generalists considered, there is a substantial correlation between the amount of mutual information encoded and the severity of the stress (estimated by its reduction of growth

compared to growth in rich media) (Fig. 3E). This correlation may reflect that generalists are typically involved in regulating growth, through, for example, affecting translation [16]. Specialists, in contrast, do not encode more information if their cognate stress is more severe (Fig. 3E).

Generalists versus specialists

To better understand the differences between generalists and specialists, we asked how the transcription factors are organized within the cell's network of signal transduction. Using data on the substrates of kinases [22], we confirmed [16] that the generalists are either directly or indirectly targets of protein kinase A (which has isoforms Tpk1-3) and TORC1 or its downstream kinase Sch9 (S6 kinase) (Fig. 4A). The generalists therefore do respond to the cell's potential for growth: protein kinase A orchestrates the cell's response to the availability of glucose [16] and TORC1 controls the response to the availability of nitrogen [16]. Similarly, Mig1 is sensitive to the levels of cellular ATP through its regulation by AMP kinase [16]. In contrast, the specialists Hog1 and Yap1 are mostly embedded in their own signalling networks.

We quantified the redundancy between pairs of transcription factors to determine if the cell's organization of information reflects the signalling network. If regulated by the same upstream signalling, two transcription factors may be completely redundant so that when paired together the amount of information does not increase above that from any one factor alone. By concatenating two time-series (Fig. S19), we can estimate the mutual information from simultaneously observing two transcription factors and so their redundancy (Fig. 4B).

Plotting the redundancy (Fig. 4B), we see a network similar to the network of signal transduction: the generalists are together in a core from which the specialists are distinct. Msn2/4 and Dot6 appear to coordinate the behavior of specialists with the general environmental stress response, having a substantial degree of redundancy with the highest number of transcription factors, including each other. Msn2, Msn4 and Dot6 have the highest number of redundant factors (size of the nodes in Fig. 4B). Specialists are not redundant with other specialists, but each is redundant with a distinct subset of the core generalists: Yap1 is partly redundant with Msn2/4 but not Dot6; Hog1 and Mig2 with Dot6 but not Msn2/4; and Mig1 is partly redundant with all three.

The redundancies imply that pairing a generalist with a specialist is best (Fig. 4C), and indeed such pairs typically encode the highest information (Fig. S20). With its distinct signal transduction (Fig. 4A), a specialist can identify the environmental state that is most poorly distinguished by a generalist. For example, Msn2 is best paired with Mig2 (Fig. S22).

As environments become more complex, multiple transcription factors are needed to generate an internal representation. Pooling the data to consider environments with different states (Fig. 4D), the maximum mutual information plateaus as the numbers of transcription factors increase, with four sufficing to generate $\sim$95% of the information. This increase comes both from the distinct dynamics of the transcription factors [24], such as differences in timing (Fig. S21), and from decreasing the effects of stochasticity by averaging the multiple readouts.

Discussion

In summary, we have shown that transcription factors can encode enough information in the dynamics of their nuclear translocations to unambiguously report an environmental change if that change is sufficiently large, that the nature of the change can also be encoded although with some degree of error, that how the information is encoded alters for changes of different magnitudes, and that no single transcription factor can accurately encode both the nature and magnitude of environmental change.

Information is transduced through two channels of specialists and generalists. Specialists are faster and can better identify a transition into their associated stress than generalists, but the variety of environments experienced by cells makes having a specialist for every environment implausible. We postulate that generalists avoid this constraint by providing an indirect channel that responds not to the extracellular signals sensed by specialists [25, 26] but to intracellular signals [27, 28], such as changes in cAMP, uncharged tRNAs, and the availability of amino acids [16] (Fig. 4E). By detecting physiological

perturbations, generalists respond to broader ranges of stress (Fig. S23) and are agnostic to the environment's precise nature. Generalists are therefore necessarily slower than specialists because they must wait for the environment to modify intracellular biochemistry. Indeed, we conjecture that the stochastic pulsing of the generalists in constant environments [19] is a response to spontaneous fluctuations in intracellular physiology.

Consistent with more recent interpretations [29], our data do not support a distinct environmental stress response controlled by Msn2/4, but show that aspects of the behavior of Msn2/4 are present in the dynamics of multiple transcription factors, such as Dot6, Sfp1, and Maf1. These latter factors act to determine rates of translation, consistent with the push-pull relationship between stress and growth [12]. In particular, we have demonstrated that Dot6, although not Tod6, encodes the nature of environmental change in its dynamics to an accuracy almost comparable with Msn2/4, implying that Dot6 may play a similar role in cellular physiology. We find too that Mig1/2, although considered glucose specialists, encode information on osmotic stress and might better be classed as generalists. Indeed Mig1/2 are activated by AMP kinase (Snf1, which responds to levels of ADP [16]) consistent with our proposal that generalists respond to an environmental change's intracellular effects.

Finally, our results show that it is only through the collective dynamics of multiple transcription factors that cells can encode sufficient information to generate specific downstream responses [15, 30]. Paralleling discoveries in neuroscience, we expect that such multi-dimensional internal representations are widespread within cellular biology and that their failures in encoding information, by causing dysfunctional decision-making, instigate deleterious behaviors and disease [31].

Methods

Time-lapse microscopy

BY strains with fluorescently tagged transcription factors [32] were grown in ALCATRAS microfluidic devices [20] in synthetic complete medium with 2% glucose for at least 3 hours and then exposed to stress for 5 hours (SI Appendix). Switching between media was via an external mixer and syringe pumps. Bright-field and fluorescence images were acquired every 2.5 min and cells were segmented using the DISCO algorithm [33].

Estimating the mutual information

Our algorithm involves (SI Appendix): (i) using principal component analysis to delineate a basis for the time-series; (ii) training a linear support vector machine to classify time-series in this basis; (iii) calculating a confusion matrix using the test data; (iv) estimating (a lower bound on) the mutual information by interpreting the entries of the confusion matrix as conditional probabilities.

Acknowledgements

We thank S. Jaramillo-Riveri, P. Thomas, M. Voliotis, E. Wallace, and the Swain laboratory for critical comments; Reiko Tanaka for her support and advice (AGG); and our funders: the BBSRC (JMJP, IF, & PSS), the EPSRC (AAG), and the Austrian Science Fund – FWF P28844 (GT).

References

[1] G. Tkačik, W. Bialek, *Annu Rev Condens Matter Phys* **7**, 89 (2016).

[2] J. I. Gold, M. N. Shadlen, *Annu Rev Neurosci* **30**, 535 (2007).

[3] T. J. Perkins, P. S. Swain, *Mol Syst Biol* **5**, 326 (2009).

[4] L. Cai, C. K. Dalal, M. B. Elowitz, *Nature* **455**, 485 (2008).

[5] R. Cheong, A. Rhee, C. J. Wang, I. Nemenman, A. Levchenko, *Science* **334**, 354 (2011).

[6] N. Hao, E. K. O'Shea, *Nat Struct Mol Biol* **19**, 31 (2011).

[7] N. Hao, B. A. Budnik, J. Gunawardena, E. K. O'Shea, *Science* **339**, 460 (2013).

[8] J. E. Purvis, G. Lahav, *Cell* **152**, 945 (2013).

[9] J. Selimkhanov, *et al.*, *Science* **346**, 1370 (2014).

[10] Y. Goulev, *et al.*, *eLife* **6**, e23971 (2017).

[11] M. B. Elowitz, A. J. Levine, E. D. Siggia, P. S. Swain, *Science* **297**, 1183 (2002).

[12] L. López-Maury, S. Marguerat, J. Bähler, *Nat Rev Genet* **9**, 583 (2008).

[13] A. Mitchell, P. Wei, W. A. Lim, *Science* **350**, 1379 (2015).

[14] A. A. Granados, *et al.*, *Elife* **6**, e21415 (2017).

[15] A. P. Gasch, *et al.*, *Mol Biol Cell* **11**, 4241 (2000).

[16] J. R. Broach, *Genetics* **192**, 73 (2012).

[17] J. E. Purvis, *et al.*, *Science* **336**, 1440 (2012).

[18] L. Ashall, *et al.*, *Science* **324**, 242 (2009).

[19] C. K. Dalal, L. Cai, Y. Lin, K. Rahbar, M. B. Elowitz, *Curr Biol* **24**, 2189 (2014).

[20] M. M. Crane, I. B. N. Clark, E. Bakker, S. Smith, P. S. Swain, *PLoS ONE* **9**, e100042 (2014).

[21] C. E. Shannon, W. Weaver, *The mathematical theory of communication* (University of Illinois Press, Urbana, Illinois, 1999).

[22] S. Sharifpoor, *et al.*, *Genome Biol* **12**, R39 (2011).

[23] D. J. C. MacKay, *Information theory, inference, and learning algorithms* (Oxford University Press, Oxford, U.K., 2003).

[24] J. O. Dubuis, G. Tkacik, E. F. Wieschaus, T. Gregor, W. Bialek, *Proc Natl Acad Sci U S A* **110**, 16301 (2013).

[25] A. Delaunay, A.-D. Isnard, M. B. Toledano, *EMBO J* **19**, 5157 (2000).

[26] V. Reiser, D. C. Raitt, H. Saito, *J Cell Biol* **161**, 1035 (2003).

[27] R. Dechant, S. Saad, A. J. Ibáñez, M. Peter, *Mol Cell* **55**, 409 (2014).

[28] M. Filteau, *et al.*, *Proc Natl Acad Sci U S A* **112**, 4501 (2015).

[29] A. P. Gasch, *et al.*, *PLoS Biol* **15**, e2004050 (2017).

[30] A. S. Hansen, E. K. O'Shea, *Mol Syst Biol* **9**, 704 (2013).

[31] Q. Luo, J. M. Beaver, Y. Liu, Z. Zhang, *Genes* **8**, 66 (2017).

[32] W.-K. Huh, *et al.*, *Nature* **425**, 686 (2003).

[33] E. Bakker, P. S. Swain, M. M. Crane, *Bioinformatics* **in press** (2017).