Identification of Animal Behavioral Strategy by Inverse Reinforcement Learning ~ its Application to Thermotaxis in *C. elegans* ~

Short title: Identification of Behavioral Strategy

Shoichiro Yamaguchi[a], Honda Naoki[b], Muneki Ikeda[c], Yuki Tsukada[c], Shunji Nakano[c], Ikue Mori[c] and Shin Ishii[a]

[a] Graduate School of Informatics, Kyoto University, Sakyo, Kyoto, Kyoto, Japan
[b] Graduate School of Biostudies, Kyoto University, Sakyo, Kyoto, Kyoto, Japan
[c] Graduate School of Science, Nagoya University, Chikusa, Nagoya, Aichi, Japan

Corresponding author: Honda Naoki
Address: Graduate School of Biostudies, Kyoto University, Yoshidakonoecho, Sakyo-ku, Kyoto, Kyoto 606-8315, Japan
Tel.: +81-75-753-9450
Fax: +81-75-753-4698
E-mail: n-honda@sys.i.kyoto-u.ac.jp

Abstract

Animals are able to flexibly adapt to new environments by controlling different behavioral patterns. Identification of the strategy used for this control (behavioral strategy) is important for understanding animals' decision making, but methods available for quantifying such behavioral strategies have not been fully established. In this study, we propose a computational approach to identify an animal's behavioral strategy from behavioral time-series data. To this end, we utilized inverse reinforcement learning (IRL) of a linearly-solvable Markov decision process (LMDP), with the assumption that animals behave optimally by minimizing costs, i.e., state cost and control cost. As a particular target, we focused on the thermotactic behaviors in *C. elegans* under a thermal gradient. After identifying the behavioral strategy dependent on thermosensory state, we found it comprised mixture of two strategies: directed migration (DM) and isothermal migration (IM). First, the DM is a strategy that the worms efficiently reach to specific temperature, which not only explained observation that the worms migrate toward the cultivated temperature, but also clarifies how the worms control thermosensory state through the migration. Second, the IM is a strategy that the worms track along a constant temperature, which reflects isothermal tracking well observed in previous studies. Furthermore, we applied our method to thermosensory neuron-deficient worms, which then identified neural basis of the DM strategy. Therefore, we believe this novel approach can quantitatively visualize hidden strategies extracted from the behavioral time-series data.

Keywords: Machine learning; Optimal control; Reward; LMDP, KL control

Significance Statement

Understanding animal decision-making has been a fundamental problem in neuroscience and behavioral ecology. Many studies analyze actions that represent decision-making in behavioral tasks, in which rewards are artificially designed with specific objectives. However, it is impossible to extend this artificially designed experiment to a natural environment, because in a natural environment, the rewards for freely-behaving animals cannot be clearly defined. To this end, we must reverse the current paradigm so that rewards are identified from behavioral data. Here, we propose a new reverse-engineering approach that can estimate a behavioral strategy from time-series data of freely-behaving animals. By applying this technique with thermotaxis in *C. elegans*, we successfully identified the reward-based behavioral strategy.

Introduction

Animals develop a behavioral strategy, a set of sequential decisions necessary for organizing the proper actions in response to environmental stimuli, to ensure their survival and reproduction. Such strategies lead the animal to its preferred states. For example, foraging animals are known to optimize their strategy to most efficiently exploit food sources (1). Although the behavioral sequence can reflect the overall behavioral strategy, a method to quantitatively identify the behavioral strategy from the behavioral time-series data has not been well established. To this end, we here propose a new computational framework based on the idea of reinforcement learning (RL).

RL is a mathematical paradigm to represent how animals adaptively learn behavioral strategy to maximize cumulative rewards via trial and error (2) (upper panel in **Fig. 1A**). Neuroscience studies over the last two decades have clarified that ventral tegmental area (VTA) dopaminergic neural activity encodes prediction error of reward (3), which is similar to temporal difference (TD) learning in RL (4). Thus, it is widely believed that RL is localized to the basal ganglia, a group of brain nuclei heavily innervated by VTA dopaminergic neurons (5–8). Therefore, an animal's behavioral strategy should be associated with reward-based representation in its neural system.

Here, our aim was to identify the reward-based representation for the animal's behavioral strategy. In particular, we utilized inverse reinforcement learning (IRL), which is a recently-developed machine learning-related framework that solves an inverse problem of RL (lower panel in **Fig. 1A**) (9, 10). IRL estimates state-dependent rewards from the history of an agent's actions and states, working from the assumption that the agent has already acquired the optimal strategy to maximize its cumulative rewards. One application is apprenticeship learning. For example, the seminal studies of IRL employed a radio-controlled helicopter, for which the state-dependent rewards of an expert were estimated by using the observed time-series of both the human expert's manipulation and the helicopter's state. Consequently, autonomous control of the helicopter was achieved by a (forward) RL that utilized the estimated rewards (11, 12). This engineering application prompted studies in which IRL was used to identify the behavioral strategies of animals, including humans. Recently, application studies of IRL have emerged, mostly regarding human behaviors, with a particular interest in constructing artificially intelligent systems that mimic human behaviors (13–15). In these studies, the experimenters designed the behavioral tasks with specific objectives, and hence the behavioral strategies therein are usually expected.

In contrast to these advanced applications, IRL applications to animal behavior in a natural environment are far from established. In the helicopter control example above, the measurable variables of the helicopter (e.g., location, velocity, and acceleration) were fully monitored for IRL. In human behavior experiments, the dimensions of the state and reward spaces, which represent behavioral strategies, are restricted by the artificially-designed task, and thus can be easily set for the RL and IRL. However, the natural environment in which animals live is quite different; animals always face a degree of uncertainty, and can only partially observe a current state due to sensory constraints. Moreover, animals exhibit stochastic behaviors even in the same condition. This makes it very difficult to set the state and reward space, as well as the dimension to be used. This difficulty has limited the applicability of IRL to the field of basic bioscience. In this study, we tried to identify the hidden strategies of freely-moving animals.

To this end, we chose freely migrating *Caenorhabditis elegans*, because this nematode is a well-studied model animal whose whole-body movements are tractable. In addition, a tracking system for freely migrating *C. elegans* has been established, and provides behavioral time-series data that are required for IRL (16). *C. elegans* have a behavioral strategy by which the worms sense external environments and migrate to preferred places; we especially focused on thermotaxis, in which worms cultivated at a certain temperature tend to migrate toward that temperature when placed within a thermal gradient (17, 18). Although the worms are not aware of either the spatial temperature profile or the location of the place with the target temperature, they often reach the preferred (target) place anyway.

In this study, we propose a new IRL-based framework for animal behavior analyses. Applying this framework to *C. elegans* behavioral data measured by the worm tracking system, we successfully identified the reward-based representation associated with the worms' thermotactic behavior strategy.

Results

Conceptualization of the IRL-based approach

To identify an animal's behavioral strategy based on IRL, we initially assume that the animal's behaviors are produced as a balance between two factors: passive dynamics (blue worm in **Fig. 1B**) and reward-maximizing dynamics (red worm in **Fig. 1B**). These factors correspond to inertia-based and purpose-driven body movements, respectively. That is, even if a worm moving in a straight line wants to make a purpose-driven turn towards a reward, it cannot turn suddenly due to the inertia of its already-moving body. Thus, it is reasonable to consider that animals' behaviors are optimized by taking both factors into account, namely by minimizing resistance to the passive dynamics and maximizing approach to the destination (reward). Such behavioral strategy has recently been modeled as a linearly-solvable Markov decision process (LMDP) (19), in which the agent requires not only a state-dependent cost (i.e., a negative reward), but also a control cost for quantifying resistance to the passive dynamics (**Fig. 1C**) (see **Materials and Methods**). Importantly, the optimal strategy in the LMDP is analytically obtained as a probability of controlled state transition:

$$\pi\left(s_{t+1} \mid s_t\right) = \frac{P(s_{t+1} \mid s_t)\exp\left\{-v(s_{t+1})\right\}}{\sum_s P(s \mid s_t)\exp\left\{-v(s)\right\}}, \qquad [1]$$

where s_t and $v(s)$ indicate the animal's state at time step t and a value function defined as the expected sum of state-dependent costs, $q(s)$, and control cost, $KL\left[\pi(\cdot \mid s) \middle\| p(\cdot \mid s)\right]$, from state s toward the future , respectively; $P(s_{t+1}|s_t)$ represents a probability of uncontrolled state transition, which represents the passive dynamics from s_t to s_{t+1}. In this equation, the entire set of $v(s)$ represents the behavioral strategy. Thus, the identification of the animal's behavioral strategy is equivalent to an estimation of the value function $v(s)$ based on observed behavioral data ($s_1, s_2, \ldots s_t, \ldots s_T$). This estimation was performed by maximum likelihood estimation (MLE) (20), and is an instance of IRL. In this study, introduction of smoothness constraints to the value function enabled us to stably estimate the behavioral strategy of *C. elegans*, in terms of the value function $v(s)$, when applied to behavioral data measured during the worms' thermotactic migration (lower panel in **Fig. 1A**).

Behavioral data acquired in C. elegans

For identifying the behavioral strategy of *C. elegans* thermotactic behavior specifically, behavioral data were collected through population thermotaxis assays, in which 80–150 worms that had been cultivated at 20 °C were placed on the surface of an agar plate with controlled thermal gradients (see **Fig. 2A**). Behavioral crosstalk is negligible, because the rate of physical contacts is low at this worm density. We prepared three different thermal gradients centered at 17 °C, 20 °C, and 23 °C to collect behavioral data; these gradients would encourage ascent up the gradient, movement around the center, and descent down the gradient, respectively. We confirmed that the fed worms aggregated around the cultivated temperature in all gradients (**Fig. 2B**). Using multi-worm tracking software (16), we tracked the trajectories of individual worms over 60 min within each gradient (**Fig. 2C**), and also obtained time-series data indicating the temperature that each individual worm experienced (upper panel in **Fig. 2D**).

Groundwork for the IRL

In order to apply our IRL-based approach to the behavioral data, we had to define two elements of LMDP: state representation and passive dynamics, which are signified by s and $P(s_{t+1}|s_t)$ in equation [1], respectively.

First, we characterized a state, which represents the sensory information that the worms process during thermotaxis. Because other studies as well as ours have previously shown that the nematode's AFD thermosensory neuron encodes the temporal derivative of temperature (21, 22), we assumed that the worm made decisions to select appropriate actions (i.e., migration direction and speed) based not only on temperature, but also on its temporal derivative. We then represented a state by two-dimensional (2D) sensory space, $s=(T, dT)$, where T and dT denote temperature and its temporal derivative, respectively (**Fig. 2D**). This means that the value function in equation [1] is given as a function of T and dT, $v(T, dT)$.

Second, we characterized passive dynamics, which is given by state transitions as a consequence of unpurposed behaviors. We considered that *C. elegans* likely migrated in a persistent direction, but in a sometimes fluctuating manner, in unpurposed situations. Thus, it is reasonable to define the passive transition from a state $s_t=(T_t, dT_t)$ to the next state $s_{t+1}=(T_{t+1}, dT_{t+1})$, where dT_{t+1} maintains dT_t with white noise and T_{t+1} is updated as T_t+dT_t with white noise. Thus, $P(s_{t+1}|s_t)$ to be simply modeled as a normal distribution (see **Materials and Methods**). Notice that the usages of T and t were discriminated throughout this study.

Thermotactic strategy identified by the IRL

Using the defined state, $s=(T, dT)$, and passive dynamics, $P(s_{t+1}|s_t)$, we performed the IRL (the estimation of the value function $v(s)$) based on behavioral data. In this estimation, we modified an existing estimation method called OptV (20) by introducing a smoothness constraint (see **Materials and Methods**). We confirmed that this smoothness constraint was indeed effective in accurately estimating the value function when applied to artificial data simulated by equation [1] (**Fig. S1**). Since the method was able to powerfully estimate the behavioral strategy based on artificial data, we next applied it to behavioral data of fed *C. elegans*.

Our method successfully estimated the value function of T and dT, $v(T, dT)$ (**Fig. 3A**), and visualized $\exp(-v(T, dT))$, which is called the desirability function (**Fig. 3B**). Because both the value and desirability functions essentially represent the same thermotactic strategy of *C. elegans*, we discussed only the desirability function below. We found that the identified desirability function maximized at T=20 °C and

dT=0, encouraging the worms to reach and stay close to the cultivated temperature; moreover, we identified both diagonal and horizontal components. The diagonal component represents directed migration (DM), which is a strategy that the worms efficiently reach to the cultivated temperature. For example, when at the lower temperatures than the cultivated temperature, the more positive dT is favored, whereas at the higher temperatures. This DM strategy is consistent with observation that the worms migrate toward the cultivated temperature, and also clarifies how the worms control thermosensory state through the migration. On the other hand, the horizontal component represents isothermal migration (IM), which is a strategy that the worms track along a particular temperature (i.e., isothermal line). This IM strategy generally explains a well-known characteristic called isothermal tracking; the worms typically exhibit circular migration under a concentric thermal gradient. Note that although we used the linear gradient, but not the concentric gradient, in our thermotaxis assay, our IRL successfully extracted the isothermal tracking-related IM strategy. We further revealed that the IM strategy worked not only at cultivated temperature, but also the other temperatures. It must be stressed that the identified desirability function (**Fig. 3B**) is not equivalent to the state distribution of T and dT (**Fig. S2**), but rather represents the desirability of the next state.

Moreover, the reward function, which is equivalent to the negative state cost function, could be calculated from the identified desirability function using equation [4] (**Fig. 3C**). The reward function primarily represents the worms' preference, and is thus a cause of the worms' developing a behavior strategy; the desirability function represents the behavioral strategy, and is thus a result of optimizing cumulative reward and control cost. Taken together, our method quantitatively visualized the behavior strategy of the cultivated *C. elegans.*

Reliability of the identified strategy

We examined the reliability of the identified DM and IM strategies by means of surrogate method-based statistical testing. Specifically, we checked whether the DM and IM strategies were not obtained by chance, under the null hypothesis that the worms randomly migrated under a thermal gradient with no behavioral strategy. We first constructed a set of artificial time-series of temperature that could be observed under the null hypothesis. By using the Iterated Amplitude Adjusted Fourier Transform (IAAFT) method (24), we prepared 1000 surrogate datasets by shuffling observed time-series of temperature (**Fig. 4A**), while preserving the autocorrelation of the original time-series (**Fig. 4B**). We then applied our IRL to this surrogate dataset to estimate the desirability functions (**Fig. 4C**). To assess the significance of the DM and IM strategies, we calculated, as test statistics, sums of the estimated desirability functions within the previously-described horizontal and diagonal regions, respectively (**Fig. 4D**). Empirical distributions of these test statistics for the surrogate datasets could then serve as null distributions (**Fig. 4E**). In both the DM and IM cases, the test statistic derived by the original desirability function was located above the empirical null distribution (p=0 for the DM strategy; p=0 for the IM strategy), indicating that the DM and IM strategies were not obtained by chance, but reflected an actual strategy of thermotaxis.

In addition, we estimated behavioral strategy based on a one-dimensional (1D) state representation, i.e., s=(T). Comparing between the 1D and 2D cases, we used cross-validation (**Materials and Methods**) to confirm that prediction ability for a future state transition in the 2D-behavioral strategy was significantly higher than that in the 1D-behavioral strategy (p=0.0002; Mann-Whitney U test) (**Fig. S3**).

Strategies of starved worms and thermosensory neuron-deficient worms

We also estimated the desirability and value functions in the starved condition, in which worms disperse on the surface (**Fig. 5A and Fig. S4**). We found that the DM strategy was lost and the IM strategy remained in this case (**Fig. 5Ab**), compared with the desirability function of the fed worm (**Fig. 3B**). This suggests that the starved worms were not using directed migration. Interestingly, the desirability function at the cultivated temperature was lower than surrounding temperatures, suggesting that the worms avoid the cultivated temperature. These data indicate that our method could distinguish the difference between strategies of normally fed and starved *C. elegans*.

Next, we performed the IRL on behavioral data from two *C. elegans* strains in which one of the two thermosensory neurons, AWC or AFD, were deficient (17, 18, 23). AWC has been genetically ablated via cell-specific expression of caspases **(Materials and Methods)**, and we employed *ttx-1* mutants as AFD-deficient animals (24). The AWC-deficient worms appeared to show normal thermotaxis (**Fig. 5Ba**). We also estimated the desirability function similar to that of WT (**Fig. 5Bb**), suggesting that the AWC did not play essential role in the thermotaxis.

On the other hand, the AFD-deficient worms demonstrate cryophilic thermotaxis (**Fig. 5Ca**). Consistently, the desirability function increased with a decrease in temperature (**Fig. 5Cb**). We further found that the desirability function lacked the dT-dependent structure of the desirability function, indicating that the DM strategy observed in the wild type (WT) worms disappeared. This indicated that the AFD-deficient worms inefficiently reached to lower temperature by a strategy primarily depending on the absolute temperature T, but not on the temporal derivative of temperature, dT (**Fig. 5Cb**). Moreover, the loss of the the dT-dependent structure was supported by the fact that the AFD encodes temporal derivative of temperature (21). Taken together, AFD, but not AWC, is essential for efficiently navigating to the cultivated temperature.

Discussion

We proposed an IRL-based framework to identify animals' behavioral strategies based on collected behavioral time-series data. We validated the framework using artificial data, and then applied it to *C. elegans* behavioral data collected during thermotaxis experiments in wild type worms. We quantitatively identified the worms' thermotactic strategy, which was represented by the desirability function of 2D sensory space (but not 1D space), i.e., absolute temperature and its temporal derivative. We then visualized the properties of the thermotactic strategy in terms of the desirability function, which successfully identified what states are pleasant and unpleasant for *C. elegans*. Finally, we demonstrated the ability of this technique to discriminate alterations in components within a strategy by comparing the desirability functions of two strains of the worm with impaired thermosensory neuron function; we found that AFD neuron (but not AWC) is fundamental to efficient navigation guided to the cultivated temperature.

Validity of LMDP

It is worth comparing the LMDP and the animals' behaviors to determine if the starting assumption of an LMDP is suitable. First, animals' movements are usually restricted by external constraints such as inertia and

gravity, and by internal (musculoskeletal) constraints, including the body's own passivity. These constraints were reflected by the passive dynamics in the LMDP. Second, animals resist entering unlikely states in which these restrictions are more powerful; that is, they prefer natural, unrestricted movements. This feature was reflected by the cost of resistance to the passive dynamics, and represented by the KL divergence (see equation [2]). Because it can deal with these issues satisfactorily, we believe the LMDP is suited for modeling the animals' behavioral strategy.

Validity of the identified strategies

We applied our IRL-based approach to several cases of worms (WT and two strains), and confirmed that identified behavioral strategies, i.e., the desirability functions, showed no discrepancy in thermotactic behaviors, as followed. Fed WT worms aggregated at the cultivated temperature (**Fig. 2B**), which can be explained by highest amplitudes of the desirability function at the cultivated temperature (**Fig. 3B**). Starved WT worms disperse around the cultivated temperature (**Fig. 5Aa**), accompanied by lowered amplitudes of the desirability function at the cultivated temperature (**Fig. 5Ab**). The AWC-deficient worms show normal thermotaxis (**Fig. 5Ba**), and consistently the desirability function was similar to that of WT (**Fig. 5Bb**). The AWC-deficient worms demonstrated cryophilic thermotaxis (**Fig. 5Ca**), which agreed with higher amplitudes of the desirability function at lower temperatures (**Fig. 5Cb**). In summary, these results demonstrated validity of our approach, and provided the potential of the method for determining changes in behavior strategy.

WT strategy

We found that the WT worms had a thermotactic strategy consisting of two components, DM and IM strategies (**Fig. 3B**). What is a functional meaning of these two strategies? We propose that the existence of these two strategies could be interpreted in terms of balancing exploration and exploitation. Exploitation is the use of pre-acquired knowledge in a greedy effort to obtain rewards, and exploration is the effort of searching for possibly greater rewards. For example, the worm knows that food is associated with the cultivated temperature, and can exploit that association. On the other hand, the worm could explore different temperatures to search for a larger cache of food than is available at the cultivated temperature. In an uncertain environment, animals usually face an 'exploration-exploitation dilemma' (25); exploitative behaviors reduce the chance to explore for greater rewards, whereas exploratory behaviors disrupt the collection of the already-available reward. Therefore, an appropriate balance between exploration and exploitation is important for controlling behavioral strategy. We offer a hypothesis that the DM strategy generates exploitative behaviors, whereas the IM strategy generates explorative ones; the worms basically, through the DM strategy, exploit the cultivated temperature, during which the worms explore the reward (food) through the IM strategy with each change in temperature.

How does the worms acquire these two strategies? We found that in the starved condition, temperature and feeding were dissociated, and as a result the DM strategy disappeared, whereas the IM strategy remained (**Fig. 5Ab**). According to these findings, we hypothesized that the DM strategy emerges as a consequence of associative learning between the cultivated temperature and food access; the IM strategy, however, could be innate. Further investigation could be expected for these hypotheses in the future.

Comparison between strains and WT

In addition to WT worms, we identified the desirability functions of the AWC- and AFD-deficient worms (**Fig. 5B and C**). The AWC and AFD neurons are both known to sense the temporal derivative of temperature, *dT* (17)(21). However, the AWC-deficient worms were endowed with similar profile of the desirability function to that of WT worms (**Fig. 5Bb**), whereas the AFD-deficient worms had different profile (**Fig. 5Cb**); the profile lost the DM's diagonal component and became rather symmetrical and unbiased along the *dT* axis. It then can be implied that impaired AFD neuron prevents the worm from deciding whether an increase or decrease in temperature is favorable, which could leads to inefficient thermotactic migration. Thus, AFD, but not AWC, produces oriented behaviors based on temporal changes in the temperature.

Advantages of our IRL-based method

Our IRL-based approach has several advantages. First, it is generally applicable to behavioral data not only of *C. elegans*, but also that of any animal, as long as suitable modeling of state and passive dynamics can be accomplished. Thus, our approach has potential use in other biological fields like ecology and ethology. Second, this approach can be applied independently of experimental conditions. Our approach is especially suitable for analyzing animals' behaviors in natural conditions where target animals are freely behaving. To the best of our knowledge, this is the first study to identify the behavioral strategy of a freely-behaving animal by IRL. Third, our approach is able to identify the behavioral strategy in terms of the desirability function, of which the neural substrates are expected to comprise many different functionally networked cortical (prefrontal cortex) and subcortical (basal ganglia) areas (5, 6). The approach herein thus allows analyses of neural correlates, such as comparing regional neural activities of freely-behaving animals with strategy-related variables calculated by our IRL. In an era where high-throughput experiments and "big data" analyses produce massive amounts of the behavioral data required for our IRL-based approach, it has the potential to become a fundamental tool with broad applicability in neuroscience, especially for the study of the neural mechanisms underlying behaviors and behavior strategies.

Materials and Methods

Reinforcement learning

Reinforcement learning (RL) is a machine learning model that describes how agents learn to obtain an optimal policy, i.e., behavioral strategy, in a given environment. An RL consists of several constituents: an agent, an environment and a reward function. The agent learns and makes decisions, and the environment is defined by everything else. The agent continuously interacts with the environment, in which the state of the agent transits based on its action (behavior), and the agent gets a reward at the new state according to the reward function. The aim of the agent is identify an optimal strategy (policy) that maximizes cumulative rewards in the long term.

In this study, the environment and the agent's behavioral strategy were modeled as LMDP, which is one of the settings of RL. An LMDP is characterized by passive dynamics of the environment in the absence of control, and controlled dynamics that reflect the behavioral strategy. Passive and controlled dynamics were

defined by transition probabilities from state s to s', $p(s'|s)$ and $\pi(s'|s)$, respectively. At each state, the agent not only acquires a cost (negative reward), but also receives resistance to the passive dynamics (**Fig. 1C**). Thus, net cost is described as

$$l\left(s,\pi(\cdot\mid s)\right)=q(s)+KL\left[\pi(\cdot\mid s)\middle\|p(\cdot\mid s)\right], \qquad [2]$$

where $q(s)$ denotes a state cost and $KL[\pi(.|s)||p(.|s)]$ indicates Kullback–Leibler (KL) divergence between $\pi(.|s)$ and $p(.|s)$; this represents the resistance to the passive dynamics.

The optimal policy that minimizes the cumulative net cost has been analytically obtained as

$$\pi^*\left(s'\mid s\right)=\frac{P(s'\mid s)\exp\left\{-v(s')\right\}}{\sum_{s'}P(s'\mid s)\exp\left\{-v(s')\right\}}, \qquad [3]$$

where $v(s)$ is a value function, i.e., the expected cumulative net costs from state s toward the future, which satisfies Bellman's self-consistency:

$$\exp(-v(s))=\exp(-q(s))\sum_{s'}P(s'\mid s)\exp\left\{-v(s')\right\}. \qquad [4]$$

Inverse reinforcement learning (*estimation of the value function*)

To estimate the value function $v(s)$, we assumed that the observed sequential state transitions $\{s_t, s_{t+1}\}_{t=1:T}$ were generated by the optimal policy π^*. We then maximized the likelihood of the sequential state transition:

$$L=\prod_t \pi^*\left(s_{t+1}\mid s_t\right), \qquad [5]$$

where $\pi^*(s_{t+1}|s_t)$ corresponds to equation [3]. This maximum likelihood estimation (MLE) was called OptV [17]. Based on the estimated value function, the primary cost function, $q(s)$, can be calculated by using equation [4].

It is reasonable to assume that animals have value functions that are smooth in their state space in order to compensate noisy sensory systems. To obtain smooth value functions, we regularized MLE as

$$\hat{v}(s)=\underset{v(s)}{\arg\min}\left[-\log L(v(s))+\lambda\sum_s\sum_{s'\in\chi(s)}\left|v(s)-v(s')\right|^2\right], \qquad [6]$$

where the first term represents negative log-likelihood, and the second term represents a smoothness constraint introduced to the value function; a positive constant λ indicates the strength of the constraint, and $\chi(s)$ indicates a set of neighboring states of s in the state space. Notice that the cost function, the regularized negative log-likelihood, is convex with respect to $v(s)$, which means there are no local minima in its optimization procedure.

Passive dynamics of thermotaxis in C. elegans

To apply the IRL to thermotactic behaviors of *C. elegans*, state s and passive dynamics $p(s'|s)$ must be defined. We previously found that the thermosensory AFD neuron encodes the temporal derivative of the environmental temperature (21), so we assumed that the worm can sense not only absolute temperature T, but also the temporal derivative of temperature dT/dt. Thus, we set a 2D state representation as (T, dT). Note that dT/dt is simply denoted as dT.

The passive dynamics were described by the transition probability of a state (T, dT) as

$$P\left((T',dT')\middle|(T,dT)\right)=\mathcal{N}\left(T'\middle|T+dT\Delta t,\sigma_T\right)\mathcal{N}\left(dT'\middle|dT,\sigma_{dT}\right), \quad [7]$$

where $\mathcal{N}(x|\mu, \sigma)$ indicates a Gaussian distribution of variable x with mean μ and variance σ, and Δt indicates the time interval of monitoring in behavioral experiments. This passive dynamics aspect can be loosely interpreted that the worms inertially migrate in a short time interval under a thermal gradient, but is also perturbed by white noise.

Artificial data

We confirmed that our regularized version of OptV (equation [6]) provided a good estimation of the value function using simulation data. First, we designed the value function of T and dT as the ground truth (**Fig. S2A**), and passive dynamics through equation [7]. Thus, the optimal policy was defined by equation [3]. Second, we generated a time-series of state transitions according to the optimal policy, and separated these time series into training and test datasets. After that, we estimated the value function from the training dataset, varying the regularization parameter λ in equation [6] (**Fig. S2B**). We then evaluated the squared error between the behavioral strategy based on the ground truth and the estimated value function, using the test dataset. Since the squared error on the test data was substantially reduced (by 88.1%) due to regularization, we deemed it effective for avoiding overfitting (**Fig. S2C**).

Cross-validation

In estimation of the value function, we performed cross-validation to determine λ in equation [6], and σ_T and σ_{dT} in equation [7], with which the prediction ability is maximized. We divided the behavioral time-series data equally into nine parts. We then independently performed estimation of the value function nine times; for each estimation, eight of the nine parts of the data were used for estimation, while the remaining part was used to evaluate the prediction ability of the estimated value function by the likelihood (equation [5]). We then optimized those parameters at which we obtained the lowest negative log-likelihood as averaged from the nine estimations.

C. elegans *preparation*

All worms were hermaphrodites and cultivated on OP50 as bacterial food using standard techniques (26). The following strains were used: N2 wild-type Bristol strain, IK0615 *ttx-1(p767)*, IK2808 *njIs79*[*ceh-36p::cz::caspase-3(p17)*, *ceh-36p::caspase-3(p12)::nz*, *ges-1p::NLS::GFP*]. The AWC-ablated strain (IK2808) was generated by the expression of reconstituted caspases (27). Plasmids carrying the reconstituted caspases were injected at 25 ng/μl with the injection marker pKDK66 (*ges-1p::NLS::GFP*) (50 ng/μl). Extrachromosomal arrays were integrated into the genome by gamma irradiation, and the resulting strains were outcrossed four times before analyses. To assess the efficiency of cell killing by the caspase transgenes, the integrated transgenes were crossed into integrated reporters that expressed GFPs in several neurons, including the neuron of interest, as follows: IK2811 *njIs82*[*ceh-36p::GFP*, *glr-3p::GFP*] for AWC. Neuronal loss was confirmed by the disappearance of fluorescence; 100% of *njIs80* animals displayed the loss of AFD, and 98.4% of *njIs79* animals displayed the loss of AWC.

Thermotaxis assay

Thermotaxis (TTX) assays were performed as previously described (28). Animals cultivated at 20 °C were placed on the center of an assay plate (14 cm × 10 cm, 1.45 cm height) containing 18 ml of TTX medium with 2% agar, and were allowed to freely move for 60 min. The center of the plate was adjusted to 17 °C, 20 °C, or 23 °C, to create three different gradient conditions, and the plates then maintained at a linear thermal gradient of approximately 0.45 °C/cm.

Behavioral recording

Worm behaviors were recorded using a Multi-Worm Tracker (16) with a CMOS sensor camera-link camera (8 bits, 4,096 × 3,072 pixels; CSC12M25BMP19-01B; Toshiba-Teli), a Line-Scan Lens (35 mm, f/2.8; YF3528; PENTAX), and a camera-link frame grabber (PCIe-1433; National Instruments). The camera was mounted at a distance above the assay plate that consistently produced an image with 33.2 μm per pixel. The frame rate of recordings was approximately 13.5 Hz. Images were captured and processed by custom software written in LabView (National Instruments), and a custom image analysis library written in C++, to detect worm bodies and measure behavioral parameters such as the position of the centroid.

Acknowledgements

We thank Drs. Eiji Uchibe, Masataka Yamao, and Shin-ichi Maeda for their valuable comments. We are also grateful to Dr. Shigeyuki Oba for giving an advice on statistical testing. This study was supported by the Platform Project for Supporting in Drug Discovery and Life Science Research (Platform for Dynamic Approaches to Living System) (H.N. and S.I.) from Japan Agency for Medical Research and development (AMED) and the Strategic Research Program for Brain Sciences (H.N., S.N, Y.T, I.M, and S.I.) from the Ministry of Education, Culture, Sports, Science, and Technology, Japan.

Author contributions

H.N. and S.I. conceived the project. S.Y. performed the computational analysis. M.I., Y.T, and S.N performed the experiments. H.N. and S.Y. wrote the draft, and H.N., S.Y., M.I., S.N., I.M., and S.I. prepared the final version of the manuscript.

Additional information

Competing financial interests: The authors declare no competing financial interests.

References

1. Iwasa Y, Higashi M, Yamamura N (1981) Prey Distribution as a Factor Determining the Choice of Optimal Foraging Strategy. *Am Nat* 117(5):710.
2. Sutton RS, Barto AG (1998) Introduction to Reinforcement Learning. *Learning* 4(1996):1–5.
3. Schultz W, Dayan P, Montague PR (1997) A neural substrate of prediction and reward. *Science (80-)* 275(June 1994):1593–1599.
4. Montague PR, Dayan P, Sejnowski TJ (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J Neurosci* 16(5):1936–1947.
5. Tanaka SC, et al. (2016) Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. *Behavioral Economics of Preferences, Choices, and Happiness*, pp 593–616.
6. Samejima K, Ueda Y, Doya K, Kimura M (2005) Representation of action-specific reward values in the striatum. *Science* 310(5752):1337–40.
7. Doya K (2008) Modulators of decision making. *Nat Neurosci* 11(4):410–416.
8. Yagishita S, et al. (2014) A critical time window for dopamine actions on the structural plasticity of dendritic spines. *Science (80-)* 345(6204):1616–1620.
9. Russell S (1998) Learning agents for uncertain environments (extended abstract). *Proc 11th Annu Conf Comput Learn Theory*:101–103.
10. Ng A, Russell S (2000) Algorithms for inverse reinforcement learning. *Proc Seventeenth Int Conf Mach Learn* 0:663–670.
11. Abbeel P, Coates A, Ng AY (2010) Autonomous Helicopter Aerobatics through Apprenticeship Learning. *Int J Rob Res* 29(13):1608–1639.
12. Abbeel P, Coates A, Quigley M, Ng AY (2007) An application of reinforcement learning to aerobatic helicopter flight. *Education* 19:1.
13. Vu VH, et al. (2016) Adaptive use of interaction torque during arm reaching movement from the optimal control viewpoint. *Sci Rep* 6(1):38845.
14. Muelling K, Boularias A, Mohler B, Schölkopf B, Peters J (2014) Learning strategies in table tennis using inverse reinforcement learning. *Biol Cybern* 108(5):603–619.
15. Mohammed RAA, Staadt O (2015) Learning eye movements strategies on tiled Large High-Resolution Displays using inverse reinforcement learning. *2015 International Joint Conference on Neural Networks (IJCNN)* (IEEE), pp 1–7.
16. Swierczek NA, Giles AC, Rankin CH, Kerr RA (2011) High-throughput behavioral analysis in C. elegans. *Nat Methods* 8(7):592-U112.
17. Kuhara A, et al. (2008) Temperature sensing by an olfactory neuron in a circuit controlling behavior of C-elegans. *Science (80-)* 320(5877):803–807.
18. Mori I, Ohshima Y (1995) Neural regulation of thermotaxis in Caenorhabditis elegans. *Nature* 376(6538):344–348.
19. Todorov E (2006) Linearly-solvable Markov decision problems. *Adv Neural Inf Process Syst* (1):8.
20. Dvijotham K, Todorov E (2010) Inverse Optimal Control with Linearly-Solvable MDPs. *Proc 27th Int Conf Mach Learn*:335–342.

21. Tsukada Y, et al. (2016) Reconstruction of Spatial Thermal Gradient Encoded in Thermosensory Neuron AFD in Caenorhabditis elegans. *J Neurosci* 36(9):2571–81.
22. Ramot D, MacInnis BL, Goodman MB (2008) Bidirectional temperature-sensing by a single thermosensory neuron in C. elegans. *Nat Neurosci* 11(8):908–15.
23. Biron D, Wasserman S, Thomas JH, Samuel ADT, Sengupta P (2008) An olfactory neuron responds stochastically to temperature and modulates Caenorhabditis elegans thermotactic behavior. *Proc Natl Acad Sci U S A* 105(31):11002–11007.
24. Satterlee JS, et al. (2001) Specification of thermosensory neuron fate in C. elegans requires ttx-1, a homolog of otd/Otx. *Neuron* 31(6):943–956.
25. Ishii S, Yoshida W, Yoshimoto J (2002) Control of exploitation-exploration meta-parameter in reinforcement learning. *Neural Networks* 15(4–6):665–687.
26. Brenner S (1974) The genetics of Caenorhabditis elegans. *Genetics* 77(1):71–94.
27. Chelur DS, Chalfie M (2007) Targeted cell killing by reconstituted caspases. *Proc Natl Acad Sci U S A* 104(7):2283–8.
28. Ito H, Inada H, Mori I (2006) Quantitative analysis of thermotaxis in the nematode Caenorhabditis elegans. *J Neurosci Methods* 154(1–2):45–52.

Figure Legends

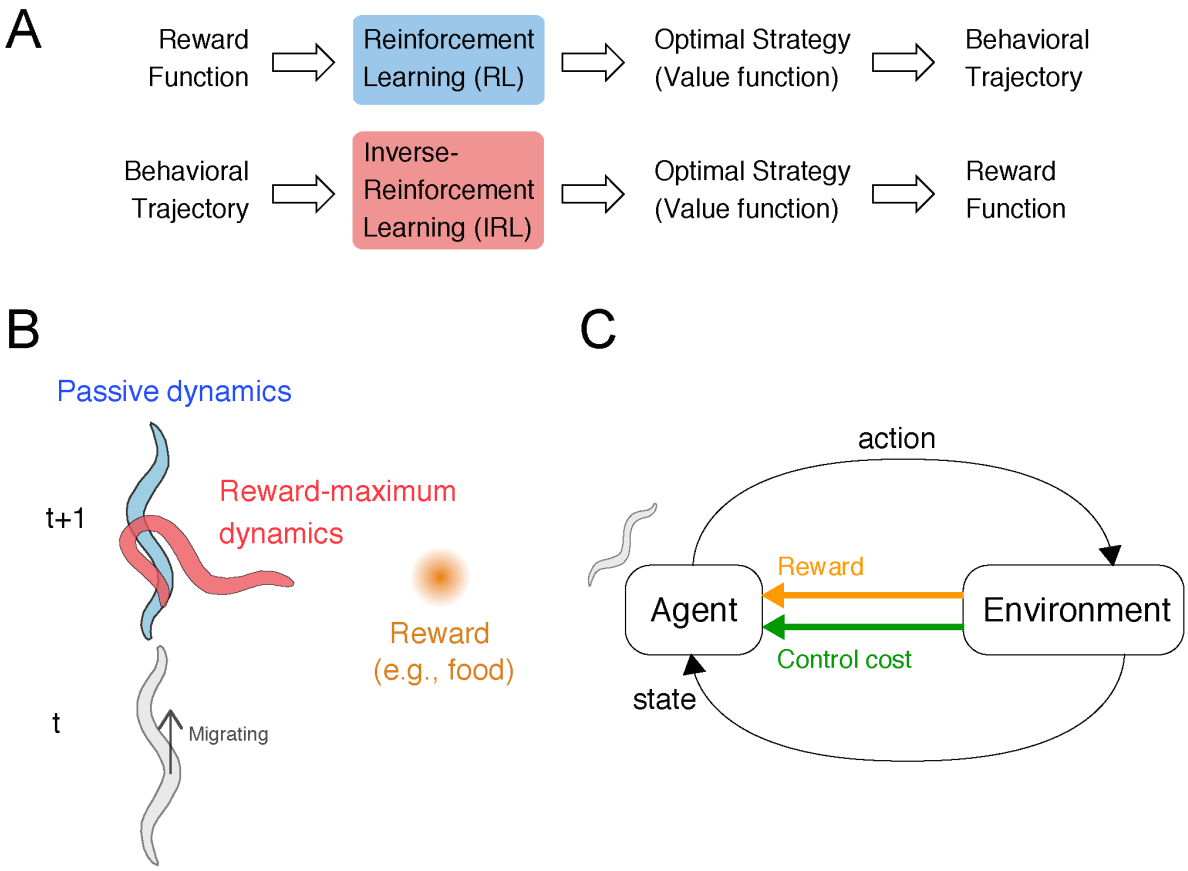

Figure 1: Concept and procedure of the inverse reinforcement learning (IRL)-based approach

A. The RL is a forward problem, in which a behavioral strategy is determined to maximize the cumulative reward given as a series of state-dependent reward. The IRL is an inverse problem, in which a behavioral strategy, or its underlying reward function, is estimated in order to reproduce an observed series of behaviors. In the IRL procedure used here, we first extracted a series of behaviors by tracking animals during behavioral experiments, then identified the value function and reward function from the behavioral series. The behavioral strategy is evaluated by the profiles of the identified functions.

B. Examples of passive dynamics and controlled dynamics. Here, an animal migrates upwards whereas the food (reward) is placed to its right. In this situation, if the animal continues migrate upwards, the food becomes distant. If the animal could take a harder body control, i.e., a change in its migrating direction toward the food, on the other hand, the food becomes nearer. Thus, the animal should make decisions via a tradeoff between the two different dynamics.

C. The agent-environment interaction. The agent autonomously acts in the environment, observes the resultant state-transition through its sensory system, and receives not only the state reward but also the body control cost. The behavioral strategy is optimized to maximize the accumulation of net reward, which is given as state reward minus body control cost.

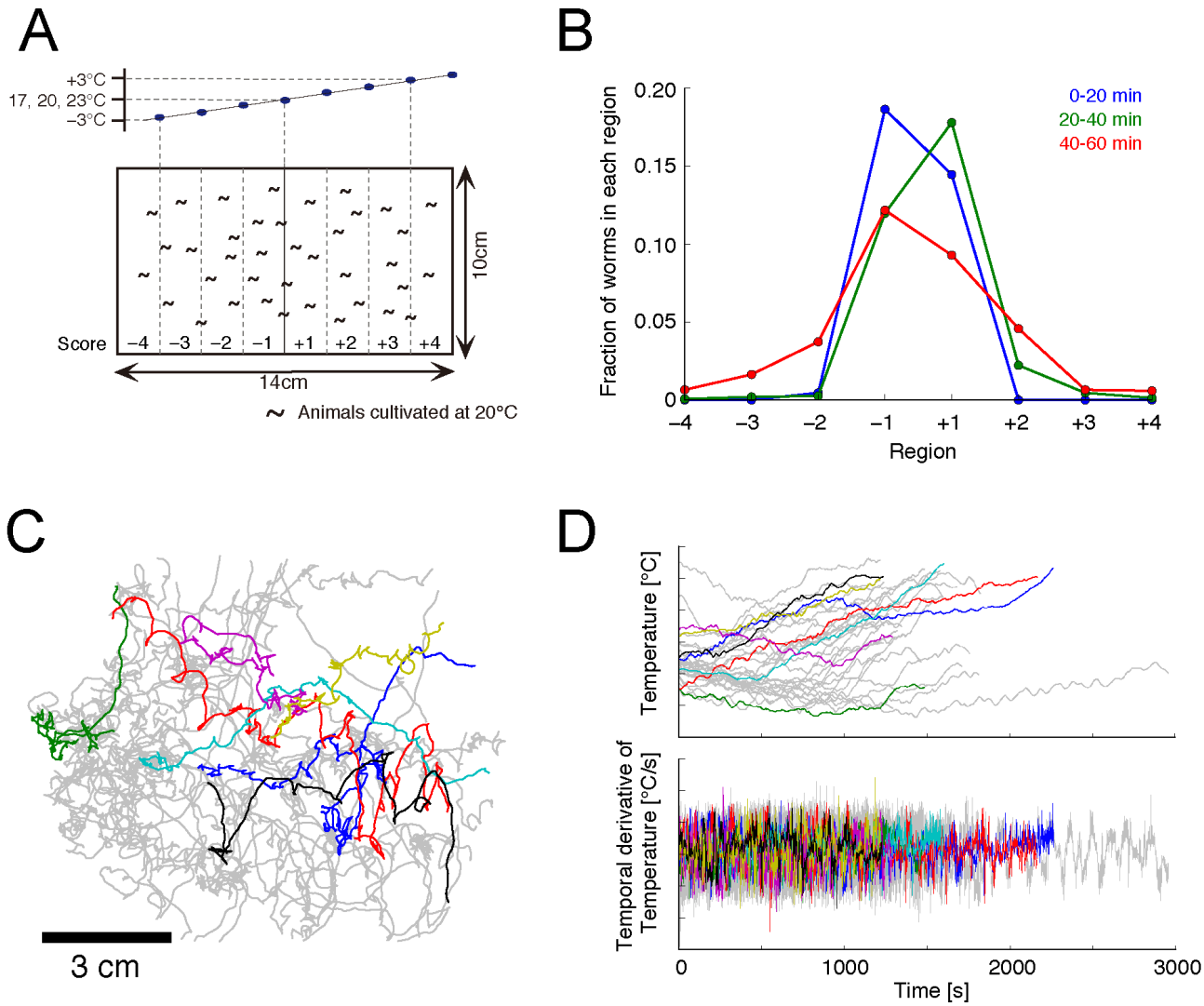

Figure 2: Data acquisition of *C. elegans* behaviors

A. Thermotaxis assay with a thermal gradient. In each assay, a linear temperature gradient was set along the agar surface, whose center was set at either of 17, 20, or 23 °C. At the onset of the assays, fed or starved worms were uniformly placed on the agar surface.

B. Temporal changes in the worms' spatial distribution under the 20 °C-centered thermal gradient in the fed condition.

C. Trajectories of a number of worms extracted by the multi-worm tracking system. Different colors indicate different individual worms.

D. Time series of the temperature experienced by the migrating worms shown in C (colors correspond to those in C) and its derivative.

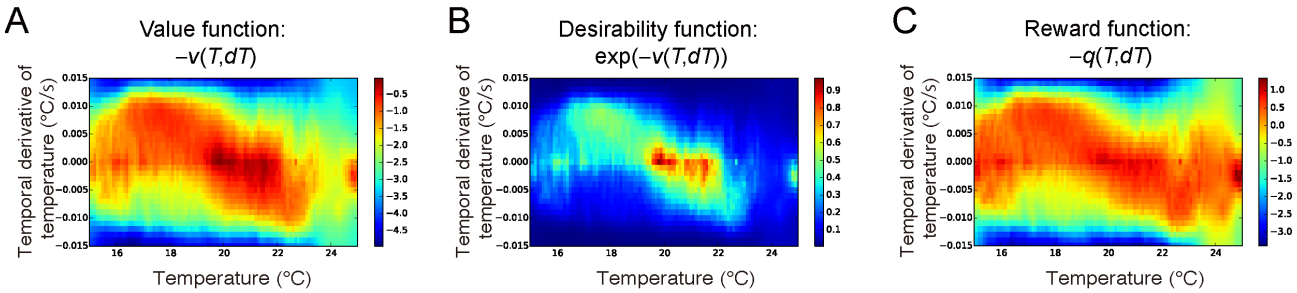

Figure 3: Behavioral strategy identified for fed wild type (WT) worms

The behavioral strategies of fed WT worms represented by the value (A), desirability (B) and reward (C) functions. The worms prefer and avoid the red- and blue-colored states, respectively.

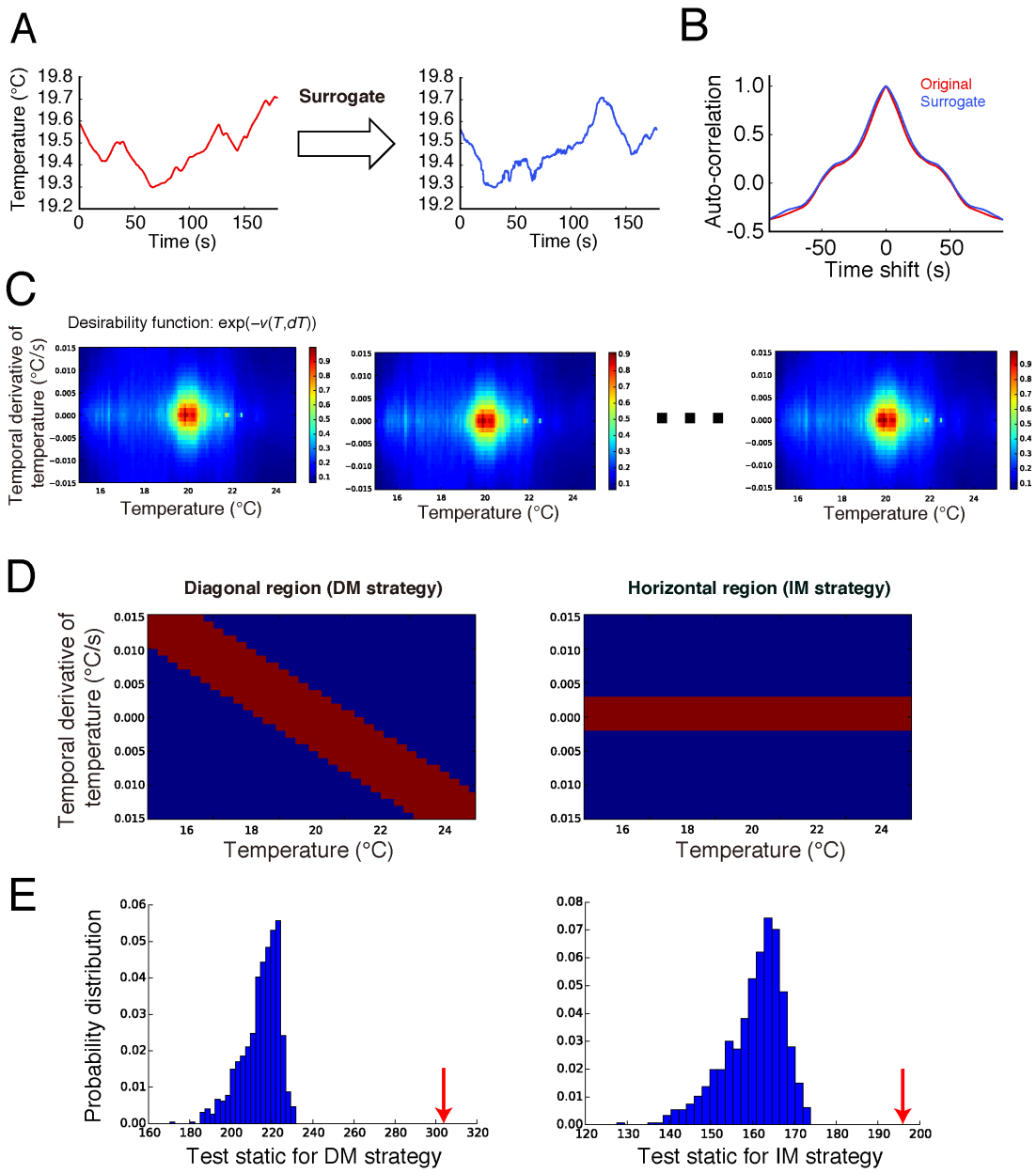

Figure 4: Statistical test for reliability of behavioral strategies with the surrogate method

The reliability of the directed migration (DM) and isothermal migration (IM) strategies (Fig. 3) was assessed by means of statistical testing with the null hypothesis that the worms randomly migrate with no behavioral strategy. (A) To generate time-series data under the null hypothesis, original time-series data of temperature (left panel) was surrogated by the IAAFT method (right panel). (B) Before and after the surrogation, the autocorrelations were almost preserved. (C) The desirability functions estimated from the surrogate datasets. (D) The DM and IM strategies correspond to the red-highlighted diagonal and horizontal regions of the desirability function, respectively. Within these regions, sums of the estimated desirability functions were calculated as test statistics. (E) Histograms of the empirical null distributions of the test statistics for the DM and IM strategies. The test statistics derived by the original desirability function (red arrows) are located above the empirical null distributions (p=0 for the PT strategy; p=0 for the IT strategy).

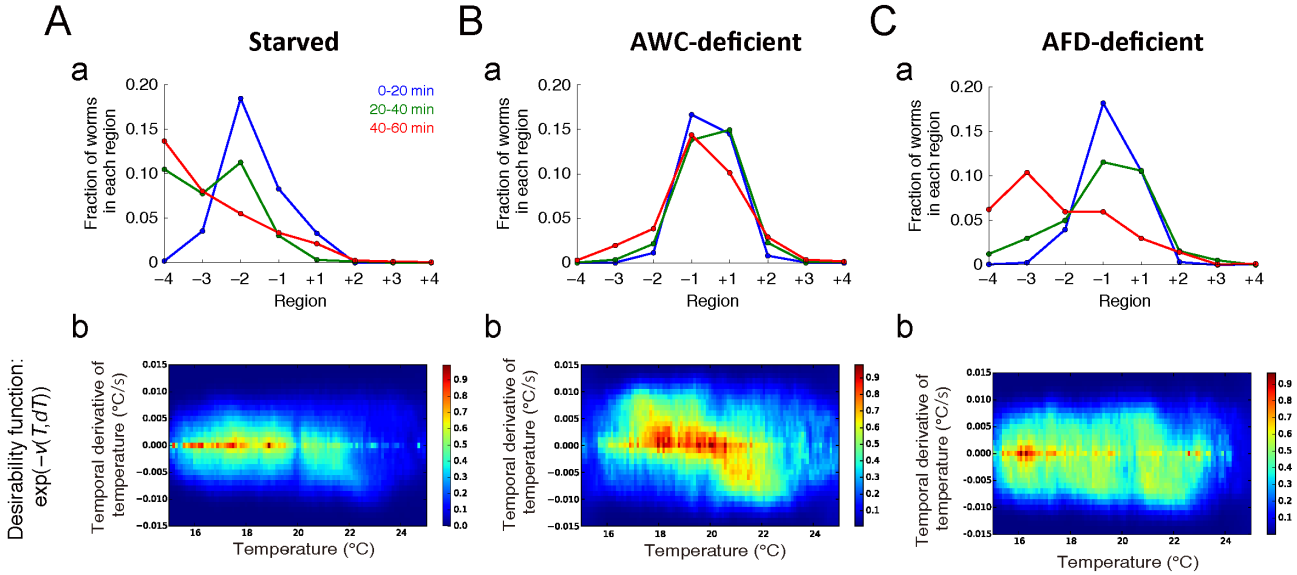

Figure 5: Inverse reinforcement learning (IRL) analyses of starved worms, AWC- and AFD-deficient worms

Temporal changes in distributions of starved worms, AWC-deficient worms and AFD-deficient worms in the 20 °C-centered thermal gradient after the behavior onset are graphed in column a of A, B and C, respectively. Starved worms disperse under a thermal gradient; the AWC-deficient worms migrate to the cultivated temperature similarly to fed wild type worms and the AFD-deficient worms show cryophilic thermotaxis. Corresponding desirability functions are shown in column b of A, B and C.

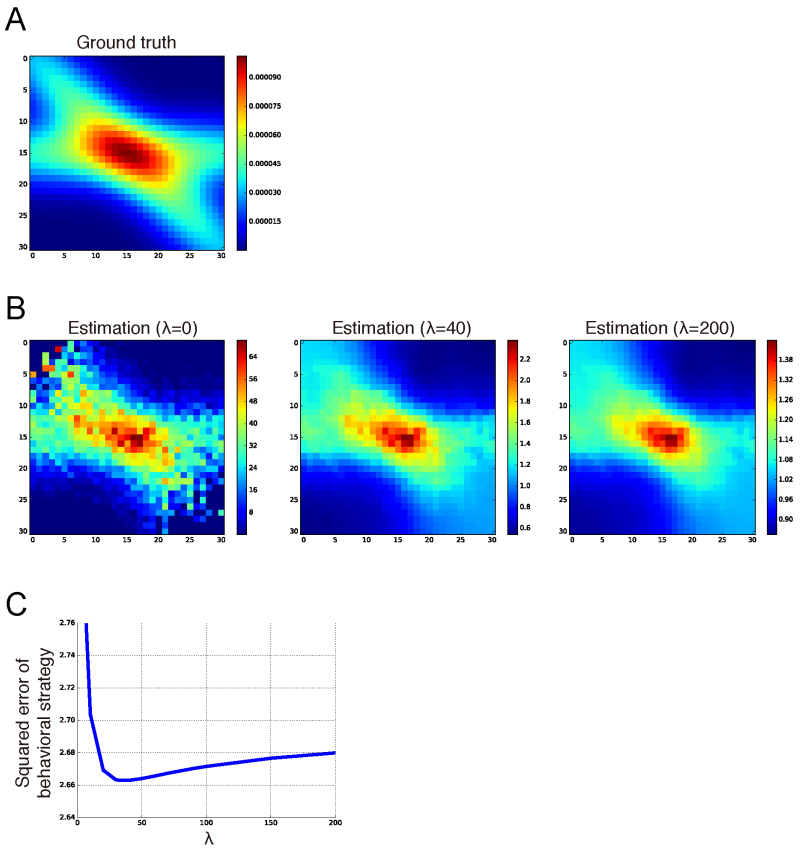

Figure S1: Validation of the regularized (OptV) estimation method in artificial data

(A) The desirability function corresponding to the ground truth value function used for generation of artificial data. Time-series data were artificially generated as training and test data sets by sampling equation [1] given the ground truth of the value function.

(B) The desirability functions described by equation [6] under 3 different regularization parameters (λ) were visualized from the estimated value functions.

(C) Squared error between the behavioral strategies based on the ground truth and estimated value functions using the test data set. The presence of an optimal λ, at which minimal squire error is obtained, indicates that the regularization was effective for accurately estimating the value function.

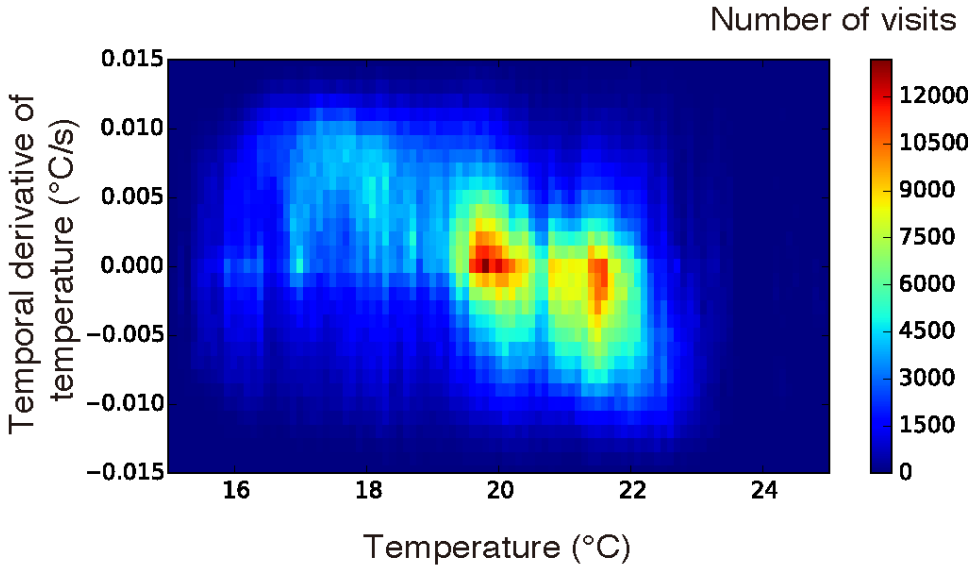

Figure S2: State distributions in the fed wild type (WT) worms

Observed distributions of T and dT in fed WT worms are shown by heat map. Notice that the distribution is substantially different from the desirability function (Fig. 3B).

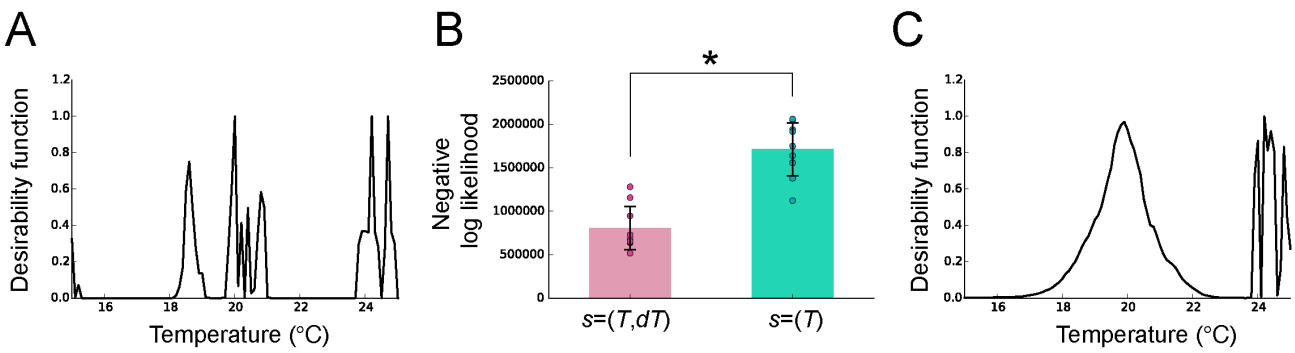

Figure S3: Inverse reinforcement learning (IRL) analysis with the one-dimensional state representation

IRL was analyzed with one-dimensional state representation (s=(T)).

(A) The desirability function was calculated by the estimated value function. In the estimation, the regularization parameter, λ, in equation [6] was optimized by cross-validation.

(B) Prediction ability was compared between IRLs with s=(T, dT) and s=(T) using a cross-validation dataset. The negative log-likelihood of the behavioral strategies (equation [1]) with the estimated value function of both of T and dT (Fig. 3B) was significantly smaller than that with the estimated value function of T alone (Fig. S3A) (p=0.0002; Mann-Whitney U test). Thus, the behavioral strategy with s=(T, dT), was more appropriate than that with s=(T).

(C) The desirability function became smoother as λ was increased. This desirability function peaks around the temperature to which the worm has been most exposed (i.e., the temperature in which they were cultivated, 20°C).

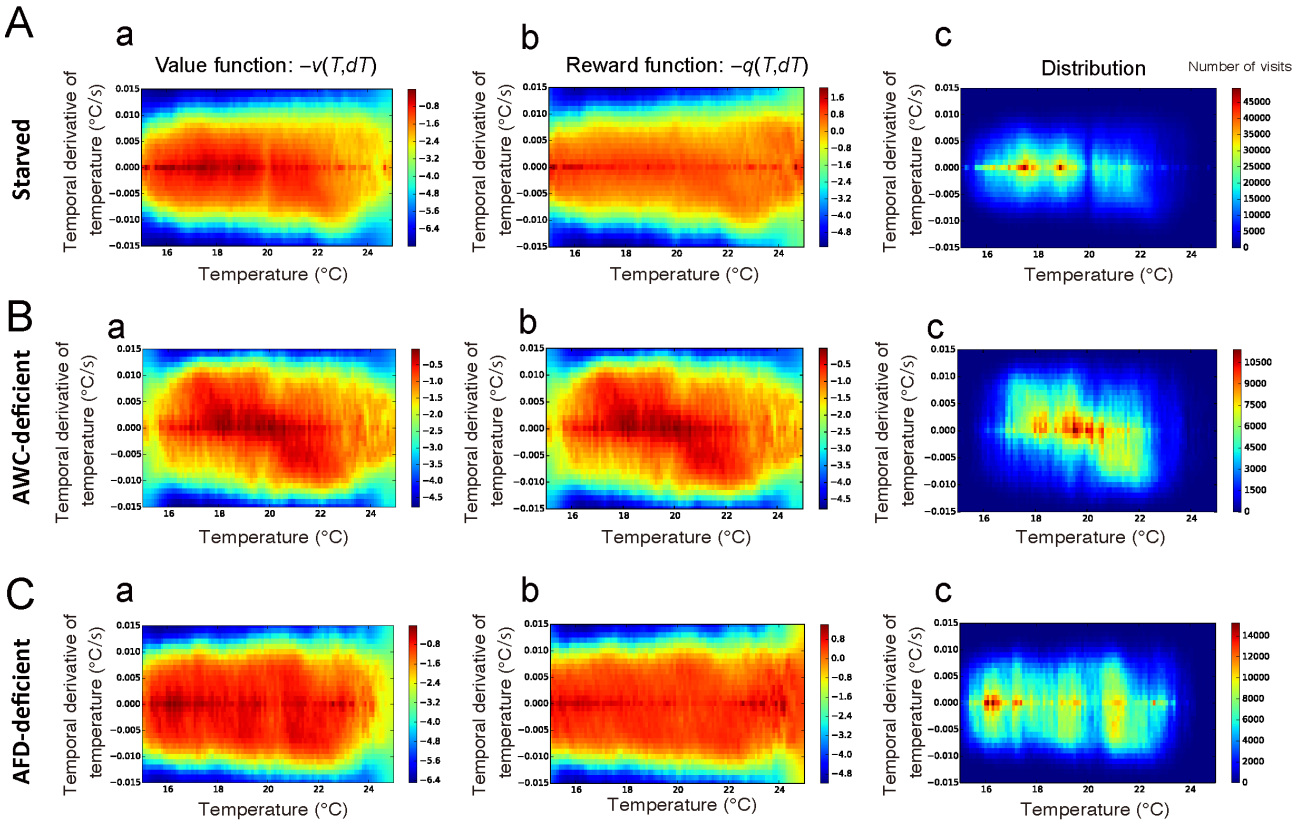

Figure S4: Estimated value/reward functions and state distributions

The estimated value functions (a), reward functions (b) and state distributions (c) were depicted for the starved wild type worm (A), the AWC-deficient worms (B) and the AFD-deficient worms (C).