Chapter 6

Biology Review

6.1 Introduction

The task which our robot attempts to solve — acquiring information which allows it to identify places in a novel environment, and reliably travel between them — is one which needs to be solved by animals as well. It is therefore interesting to compare and contrast what is known of animal behavior, and the machinery which supports it, with the corresponding behaviors and machinery in our robot.

Any such discussion regarding brain functions of mammals necessarily has a highly speculative component. Nevertheless, there is brain activity, particularly in the forebrains of rodents, which has at least an apparent connection with navigation, and also has interesting parallels to some of the structures and relationships which are built by our robot.

In this chapter, we survey research results concerning the animal brain activity in question. In chapter 7, we go on to consider similarities and differences between the data structures and algorithms actually implemented by our robot, and the computations which *appear* to be taking place in the brain, which we now outline below.

6.2 Place-cell activity in the hippocampus

The most common study subject for neural studies of navigation in mammals, in recent years, has been the rat (and more recently, mice, in which very similar phenomenology appear to be present). This is because of the well-known finding that the rat hippocampus (specifically, the CA1 and, to a lesser extent, CA3 gyri) contains "place cells" [36], which respond most strongly when the animal is in a particular place within a familiar environment. The "place fields" of these cells tend to be tightly confined, convex patches of space; cells tend to fire vigorously whenever the animal is in the place field, and much less so when it wanders outside. What makes this activity particularly compelling is that when the animal is roaming freely in a familiar environment, a very substantial fraction (sometimes, more than half) of excitatory, complex-spike units within appear to demonstrate some form of reliable place-field activity within that environment.

We begin this discussion of place field activity by very briefly reviewing experimental findings relevant to place fields themselves. We then discuss evidence regarding non-spatial tasks, and briefly speculate on interpretations of the data which might offer a single consistent explanation for both types of data.

We begin by considering the place-field activity itself.

6.2.1 Place-field basics

The basic finding of the place cell research is that when animals are trained to perform tasks which involve traversing an environment, or even allowed to range freely through one, then cells are found in the hippocampi of the rats (specifically, in the CA3 and CA1 gyri) with properties which appear to be specific to the layout of the environment — that is, they are active only when the animal is in a specific location.

For instance, in the early work described by O'Keefe and Nadel [36], rats were allowed to range freely over an environment with various cues located on the borders to provide landmarks which might allow the animal to distinguish one place from another. So, for instance, an environment might be a space surrounded by a red disk to the north, a buzzer to the east, and a green square to the southwest.

In these circumstances, when recordings are made from cells in the CA1 gyrus of the hippocampi of the free-ranging animals, the preponderance of cells which respond differentially at all appear to fire most vigorously when the animal is in a particular place. These place fields have several properties which have been consistently observed, and which are worth noting here explicitly. They tend to delimit small, compact regions of space — the regions the animal occupies when a cell's activity is above threshold tend to be compact circular blobs, not elongated regions nor star shapes. (In particular, they do not show some of the irregularities one might expect of an aspect-graph based representation). The same cell can have multiple place fields in different environments, if the animal has been trained on both, but in a given environment, it will not be active except when the animal is occupying the unit's place field.

As to the relationships between cells having place-fields, two facts emerge fairly directly. One is that it is generally quite easy to find cells whose place fields overlap — they are *not* forming a tiling of the space the animal has occupied. A more interesting point has to do with topography — while it would be nice to find a topographic representation of space within the (apparently) topographically specific responses can be situated, no one has ever been able to identify one; the relationship between the physical locus of a unit, and its place field (if, indeed, there is any causal relationship at all) is still a mystery.

6.2.2 Degrees of relation to sensory responses

The simplest interpretation of place cell responses is that they have nothing to do with location *per se*, but are simply recording the constellation of sensory stimuli which can be viewed at a given location — information which can be useful in building a map (particularly one of the aspect-graph type), but has no necessary relationship to mapping per se.

The easiest way to tell the difference is to put the animal in the same place but remove or alter the sensory cues. For instance, one approach, followed by Muller and Kubie [34] is to add a barrier to the environment which blocks the view from some particular place to a visible landmark, cells with place fields there will not be affected unless the barrier actually falls within the place field itself. Furthermore, a barrier which *does* cross a unit's place field will abolish the response of the unit even if the barrier itself is transparent plexiglass and all landmarks remain visible.

An particuarly elegant approach to this is found in the work of Best and Thompson [8], who had animals freely explore an environment with visual landmarks, and then turned off the lights. The result was that as the animal continues to wander in the dark (an activity consistent with the rat's natural habits), the activity of cells with place fields is essentially unmodified — the cells fire at an elevated rate whenever the animal is in the physical locations consistent with its place field as determined in the light. Therefore, the responses of the cells are *not* being determined solely by response to a learned combination of certain sensory cues that are immediately present.

What is more important, the continuation of correct place-field responses in the absence of sensory cues demonstrates that the cells exhibiting these responses either are themselves, or are driven by, some sort of a genuine mapping system — that is, one which stores enough information about not only the animal's current location, but its relationships to other locations which the animal has occupied in the past and might travel to again to support a useful dead reckoning capability.

6.2.3 Explicit representation of routing information

The experiments described previously seem to establish that the place-field activity is tied to some notion of "place" which is distinguishable from any property of the animal's direct afferent input itself — it really does have to do with memory for specific places. However, that is not, in and of itself, enough to make up a useful map — one also needs some sort of information about relative connections between the places, which would ideally (at least for the physiologists) be directly manifest as routing-dependant neural activity.

Such activity might take several forms. Indeed, several forms have been found.

Directional place-field responses

Interestingly, when the animal's travels in a particular area tend to be confined to a small set of directions — most typically two, as when the place is a narrow corridor, and the animal can only run along the corridor's axis in one direction or the other — the place field activity is most typically directional as well [32]. That is, a given unit will be active when the animal is at a given place *and* moving in a given direction. However, if the animal is allowed to freely explore an open field, then directional behavior is typically seen only briefly, as the animal learns the environment, and tends to give way to the more frequently discussed purely location-based pattern of behavior, in which a cell's response is more or less the same whenever the animal is within the cell's place field, with little variation in response with changes in the animal's direction of motion.

Routing information in relative timing

So far, the results we have been discussing have all concerned the aggregate responses of place-field units over time. However, neural activity in the hippocampus has a significant timing-related component — specifically, wide-field recordings from the structure in waking animals consistently show a rhythm (called the θ-rhythm) at about 10 Hz. This suggests that the timing of the action potentials of units relative to both each other and the θ-rhythm as a whole may carry information. Recent work shows that this may well be the case.

Specifically, the work of Skaggs et al. [43] shows the following pattern: when an animal is leaving the region of space covered by a unit's place field, that unit will of course begin to fire less often. However, when it does fire, it turns out that those spikes almost always occur in the early phase of the θ-rhythm. Likewise, when the animal is entering the region of space corresponding to a unit's place-field, the unit's action potentials will tend to occur late in the θ cycle. The pattern holds as well when the animal is right in the middle of a unit's place field — that unit will be more active, but its action potentials do tend to occur mostly in the middle of the θ cycle.

In this manner, the activity during a single θ cycle in fact represents not only the *location* which the animal currently occupies, but the *direction* in which it is travelling — with the units representing the origin of the current segment firing first, and those representing the destination firing last.

(As a sidelight, similar pattern may be observed in sleeping animals. In sleep, the hippocampus' long-range potential is still dominated by rhythmic activity, though that rhythm is much faster than in the waking state. If multiple place cells are recorded simultaneously, their relative firing rates, when they are active, will tend to represent physically feasible routes, even though with the animal sleeping they are obviously not correlated directly to its current location or changes thereof).

It is interesting to note that the relatively tight timing of the activity seen here (involving time differences on the order of 30 msec. or so) do not allow for very much time for place-related activity to leave the hippocampus, travel to some other part of the brain (perhaps the parietal lobe), and return. It still remains possible that the activity is driven by the parietal lobe indirectly, but the machinery to support that hypothesis (perhaps some structure of phase-locked loops) would have to be rather complex.

6.2.4 The hippocampus as a spatial organ?

Let us now consider what all this might mean for the function of the hippocampus as a whole. The most straightforward explanation for this sort of activity is that the hippocampus is specialized, at least in the hippocampus, as a mapping organ; indeed, that is the explanation which was posited by O'Keefe and Nadel. However, the evidence available is too scant to make this conclusive. Suppose, for example (as one might, given the consequences of damage to homologous structures in higher animals) that the hippocampus is an organ with a broad role in the formation of new declarative memories. Then, when the animal is trained on navigation tasks (as in most place field experiments), one would certainly expect to see activity related to navigation tasks (just as one would expect to see task-activity whatever the animal was being trained to do).

Furthermore, there are aspects of "place cell" activity which are not directly consistent with the mapping hypothesis. In particular, there is no evident topographic relationship between cells and their place fields. Likewise, when cells have multiple place fields in different pieces of experimental apparatus, the place fields of pairs of cells do not show any consistent relationship.

So, to try to get a complete picture of what is going on in this area, it is useful to look at what is known of activity in this area during tasks with no overt connection to navigation.

6.3 Evidence regarding non-spatial tasks

While responses of the hippocampal gyri containing the place cells have been most extensively studied in spatial tasks, there have also been studies regarding other tasks, which suggest that the hippocampus may be performing some computation with no relation to spatial processing (which, if true, would imply that it is not strictly a spatial-processing organ). Unfortunately, while that evidence is highly suggestive regarding the consequences of hippocampal lesions in primates, it is less suggestive regarding the rats and mice which are of immediate concern to us. However, the evidence is still at least suggestive, and therefore worthy of some study.

6.3.1 Amnesia induced by hippocampal lesion

We begin by considering the well-known consequences of hippocampal lesion in higher mammals, including humans [54]. In monkeys, experiment has shown repeatedly that lesion of the hippocampus brings on severe amnesia, characterized in particular by the failure to learn associations between cues, or sets of cues, and given responses [2].

In humans (in particular, in the famous case of the patient HM, whose hippocampi were lesioned bilaterally in an attempt to control epilepsy, an operation which has never been performed again due to the consequences), these phenomena are seen, but they are seen as aspects of a more general deficit — namely, a dramatic failure of what has come to be called declarative or episodic memory. HM is capable of carrying on conversations normally, so long as his interlocutor remains in the room. However, if they leave the room even briefly, he loses all memory of having ever seen them — an amnesia which is instant, vast and permanent. He has no memory of people who have been caring for him for years; he is, for all intents and purposes, living in the year of his operation, more than thirty years ago.

Similar deficits may be occuring in the monkeys as well; however, declarative memory is difficult to test directly in the absence of language, so the question is hard to assess. There is anecdotal evidence that the animals are socially impaired, which would be a natural sequel of such a deficit.

These deficits, together with similar case reports of humans with naturally occuring lesions including the hippocampus (though not, of course, with surgical precision), have resulted in a widespread view of the hippocampus as a structure specifically associated with the formation of declarative memory, whether the memories in question are of a particularly spatial nature, or not.

One problem with such an interpretation, of course, is that a lesion of the hippocampus does not eliminate declarative memories which have already been formed. Interestingly, there are lesion syndromes in which previously known information is lost, but they are typically associated with lesions elsewhere in the cortex, often in an area which is commonly associated with processing the specific type of information — vision or language, to cite common examples — whose memories have been lost. For instance, in some reports of color agnosias, the patient will report not just a failure to perceive color, but an inability to remember what color was ever like. So, on the basis of lesion evidence, at least, it seems to be that fully formed declarative memory actually resides not in the hippocampus (which can be destroyed without disturbing it), but elsewhere.

A common, though still speculative, interpretation of these data is that the role of the hippocampus is to coordinate the "laying down" of memories elsewhere, by some as yet undetermined mechanism.

6.3.2 Conditioning

Classical conditioning with complex cues is a very well studied behavioral paradigm, and there is a considerable, though controversial, literature on responses in these areas of the hippocampus during classical conditioning with complex cues.

The basic finding is that in straightforward classical conditioning tasks, hippocampal cells acquire a predictive, apparently premotor activity regarding the response, in the presence of the conditioned stimulus [7]. There is also evidence that hippocampal cells can develop responses to particular *combinations* of cues in classical conditioning experiments [53]. Indeed, when Foster et al. [21] recorded from conditioning in an audio task, complex spiking neurons (the same class which give rise to place fields when the animal has navigational tasks) were instead found to be sensitive to the history of the tones presented, with statistically significant interactions extending over the course of at least a few separate trials.

The difficulty in interpreting these findings is that lesion of the hippocampus is not disruptive to classical conditioning per se — animals without hippocampi are perfectly capable of performing the task. Now, this does not, by itself, demonstrate that the hippocampal function is irrelevant to the task — it could be the case that the hippocampus does play a role in conditioning in normal animals, but that in lesioned animals, another system is capable of taking the role. However, it does place very strong conditions on what would be required to produce an experimental demonstration of such a role.

(In detail: An analogy might be drawn to the dual role of the superior colliculus and the frontal eye fields in rhesus monkeys. Either of these is capable of driving near-normal saccadic eye movements in the absence of the others, but a combined lesion of both areas abolishes saccades completely. However, in order to demonstrate a similar role for the hippocampus in conditioning, one would need to identify another area or areas with an equivalent role, and to perform the same set of demonstrations — that lesion of the hippocampus, or of the other area, alone did not block conditioning, but that a combined lesion does. No such experiment has yet been done, and any list of possible "partner areas" would, of course, be highly speculative).

Of course, what is of most interest to us here is not whether the hippocampus in rats and mice is involved in conditioning, but whether it is involved in *any* computation which is distinct from spatial modeling per se (whether that computation is relevant to a particular experimental task or not). Since the hippocampal responses do demonstrably change in conditioning tasks which have no spatial component, that can be taken as evidence that there are non-spatial computations in the hippocampus. Unfortunately, this too is weak evidence, because there has not, yet, been any clear demonstration that the changes in hippocampal responses are due to a computation occuring within the hippocampus itself (as opposed to one of the numerous areas — most of the sensory neocortex and much else besides — which supplies input to the hippocampus). Investigation of hippocampal responses within the mutant mice of Tsien et al. [50, 51], which have LTP blocked in the CA1 gyrus of the hippocampus but are otherwise normal, might help to elucidate this question.

6.4 Hippocampal activity — approaches to interpretation

Let's begin by discussing the actual neural activity which is known to occur in the hippocampus. So far, we have discussed two sorts of activity; one related to navigation and navigational tasks, and one related to classical conditioning, in which the cells appear to form responses which relate specifically to the temporal, rather than spatial sequences of stimuli in the experiment. As described above, there is not yet definitive evidence that the physiological responses seen in classical conditioning tasks are attributable to computations performed by the hippocampus itself. If, however, they are, then the hippocampus in these rodents would clearly be implicated in computation which is *not* purely spatial, contradicting the view of O'Keefe and Nadel [36] that it is purely a mapping organ.

In other words, if the responses in conditioning tasks are, in fact, an expression of the function of the hippocampus, then we would have to suppose that the function of the hippocampus is one that can manifest as "place-related" activity when the animal is exploring a static environment, but can also take the form of evolved responses to combinations of cues in non-spatial conditioning tasks. It is interesting to speculate on what such a function might be.

A natural place to begin such speculation is with the notion of a temporal sequence — such as the units reported on by Foster et al. [21]. Even in the trivial case of a single, simple cue, classical conditioning is concerned with having the animal produce a proper response (the conditioned response) when it sees the conditioned stimulus — a simple temporal sequence. Furthermore, the responses appear evolve additional structure when the cue itself becomes temporally sequenced (though, once again, there is a question as to whether those responses evolve within the hippocampus itself)..

Navigation also is concerned with the association of a set series of responses to a temporal series of cues. There are, of course, several key differences between the sequences of cues involved in navigation, and those involved in the classical conditioning paradigm. The simplest of these, of course, is that the cues being sequenced in navigation include the animal's own recent past, and perhaps possible future actions. More generally, the associations that have to be learned in the fully general place-field case have a higher intrinsic dimensionality than the associations among cues in the classical conditioning experiment — the one dimension of time in the conditioning case, versus the two dimensions of the surface which the animal explores (or three, if time is counted as a relevant dimension in the navigational case).

However, the differences between the two cues can perhaps be bridged by postulating an intermediate sort of experiment. Imagine, if you will, a rat on a roller coaster, with visual and other cues nearby allowing it to distinguish particular bends in the tracks. It is moving, and it knows that it's moving, but it has no choice in the direction of its motion. However, if it formed place fields representing locations along the track, then the sequence of those fields would have some resemblance to the sequence of items learned in classical conditioning — a single, set sequence of cues, which always fire off in the same order, with the samed learned cue or response appropriate at the end. The major difference between such a series, and the cues presented in the sequenced classical conditioning experiments is simply that in the latter case, the animal does not move.

What this "rat on a roller coaster" example is meant to demonstrate is that a structure capable of learning associations as general as those required for navigational tasks is also capable of handling the particular associations required in sequenced classical conditioning tasks as a simple special case, so long as there is nothing wired into the system which *limits* its use to sets of cues in which the animal is making autonomous choices. It is therefore interesting to speculate that the hippocampus is, or contains, such a mechanism, one capable of learning and representing series of cues or actions (including, necessarily, links between the elements of the sequences in question).

This hypothesis is at least consistent with the *absence* of certain features which one might expect to see in an area which is truly specialized for navigation exclusively, such as a predictable relationship between at least local features of the topography of the environment, and that of the area itself. It is also consistent with the observation that, across a considerable spectrum of tasks, experimenters find a very large fraction of units in the hippocampus which have apparently developed responses with a high degree of specificity towards the task in question.

(Finally, it is interesting to speculate that a structure which is capable of learning arbitrary graph structures, might be useful in learning graph structures which don't directly represent an aspect of the real world, including, in the limit, the graph structures of language — but such speculations are way beyond the scope of this thesis).

So much for review and analysis of the data concerning the hippocampus. We now turn to consideration of possible relationships between this activity, and the structures it represents (such as they are), and our robot-based mapping system.

Chapter 7

Biological parallels

7.1 Introduction

The previous chapter has briefly reviewed some research findings concerning activity in the mammalian hippocampus, which seems almost certain to be performing computations related to the modeling of space in rats and mice. (It is possible that the hippocampus is involved in other functions as well even in these animals, and that is quite likely the case in higher organisms — particularly considering the available data from human subjects such as patient HM — but there is no definitive data available).

We now discuss some of the obvious similarities to our robot, and ways in which these similarities reflect the underlying nature of the problems faced by both systems. We then proceed to consider manifest differences, the computational resources required to simulate more brain-like systems (and how that might improve performance), and how this all may reflect the differences between the resources available to the brain and the underlying hardware (or, in the case of the brain, "wetware").

Our robot-based system builds a graph of place representations, which have links to representations of neighboring places. The rodent brain, likewise, contains units which are active when the rodent is in a particular place. There is still no definitive evidence of the exact function of these responses. However, they appear not to result from any direct combination of sensory cues, as evidenced by the continuation of place-cell responses in the absence of cues (as in darkness). Indeed, performance in darkness indicates that the neural system contains some representation of the distances between the represented places as well, or it would be impossible for the representation of the animal's current location to stay current in a "lights out" environment. (We have not demonstrated, of course, that such a representation of relative displacements exists *in the hippocampus* — it might well exist in one of the many areas to which the hippocampal formation has reciprocal connections, or be a property of one of those connections. However, for the observed pattern of activity to be possible at all, it must exist someplace).

To illustrate the analogy between place fields and the place representations built by our system, we can imagine a hypothetical "unit" in our robot which would be active whenever the given place representation. Such a unit would then have a "place field" — that is, it would be active when the robot was near that place, and inactive otherwise; see figure 7.1.

In short, the both the rodent brain and our system contain units which appear to

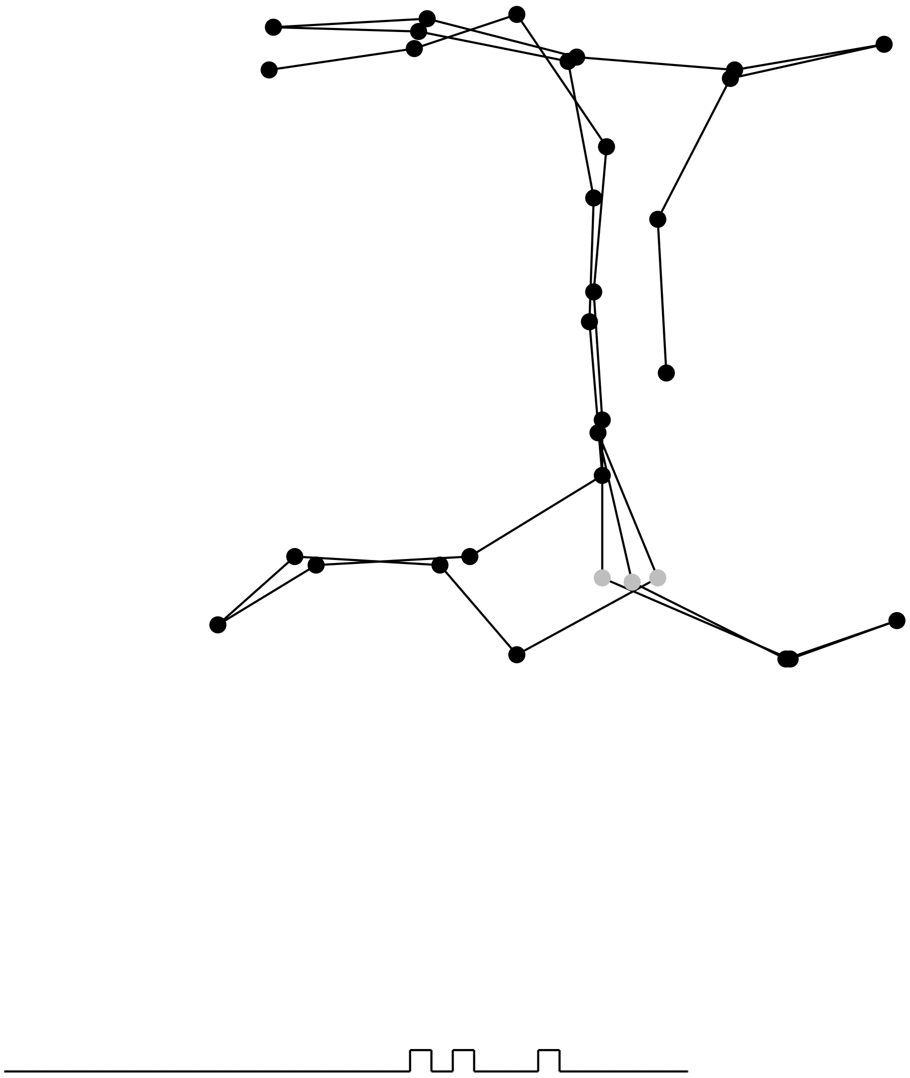

Figure 7-1: "Activity" of a single place representation in the cross environment. Above, we see the position fixes at which the chosen place representation is active, giving some notion of the extent of its "place field". Below, we see "activity" against time, the closest correspondant in this analogy to a physiological trace of unit activity. There are three instances of activity, corresponding to three separate visits to this physical place during the traverse shown.

represent the presence of the traveler at a particular place, and to represent explicitly their relative displacements in sufficient detail to allow it to tell which place representation, or units, to activate after at least a short stretch of dead-reckoned navigation, with fair accuracy. However, there are some differences, which we now discuss.

7.2 Directionality

One immediately striking difference between our place representations and the place cell activity in the brain is that the latter can be directional. That is, the place-cell units in the brain will often be active only if the animal is in a given place *and moving in a given direction.* In fact, as discussed above, directional activity is highly likely to develop when the directions in which the animal moves are constrained (as in a narrow corridor), and it is the non-directional activity which would most closely correspond to our place representations (associated with no particular direction of transit) which tends to emerge only later and more slowly after the animal has been allowed to explore environments without such constraints.

It is interesting to determine whether this reflects any fundamental distinction between the two models — in particular, can our graph structure sustain directional representations? The answer is that it can, but only with extensive modifications. The reason for this is that the links between place representations in our graph are not themselves directional. This can lead to inappropriate results in searches based on directional nodes — in particular, in retrace mode, the without some notion of the directionality of links, there will be nothing to prevent the robot from attempting to attempting to transit back to a place which will not be recognized because the representation has the wrong polarity.

Nevertheless, such a modification is possible, and with the modification in place, the robot does indeed build a pair of linked linear graphs of place representations, one for each direction, when run down a corridor (as shown in figure 7.2).

To go into more details: the modification requires both place representations themselves and the links between them to be typed as "directional". For place representations, this means that the representation has to be marked with the particular direction in which the robot was facing when the representation was created (i.e., when the rosette was scanned), so that matches can be made to fail when the robot is in the same physical location but oriented differently. (This is a natural restriction for the robot, but for our robot, it is somewhat artificial). Links have to be marked with a designation of which end is the origin, and which the destination — otherwise, the graph-search procedures which direct exploration will try to direct the robot to previously viewed places where it has been, but oriented differently at the time, resulting in serious confusion. (The search procedures themselves have to be updated likewise).

In the simple corridor case, this has no major functional implications; increasing the number of representations by a factor of two will not have a devastating impact on performance. However, in situations where the robot is exploring an open field in which any given place can be approached from a variety of angles, the blowup

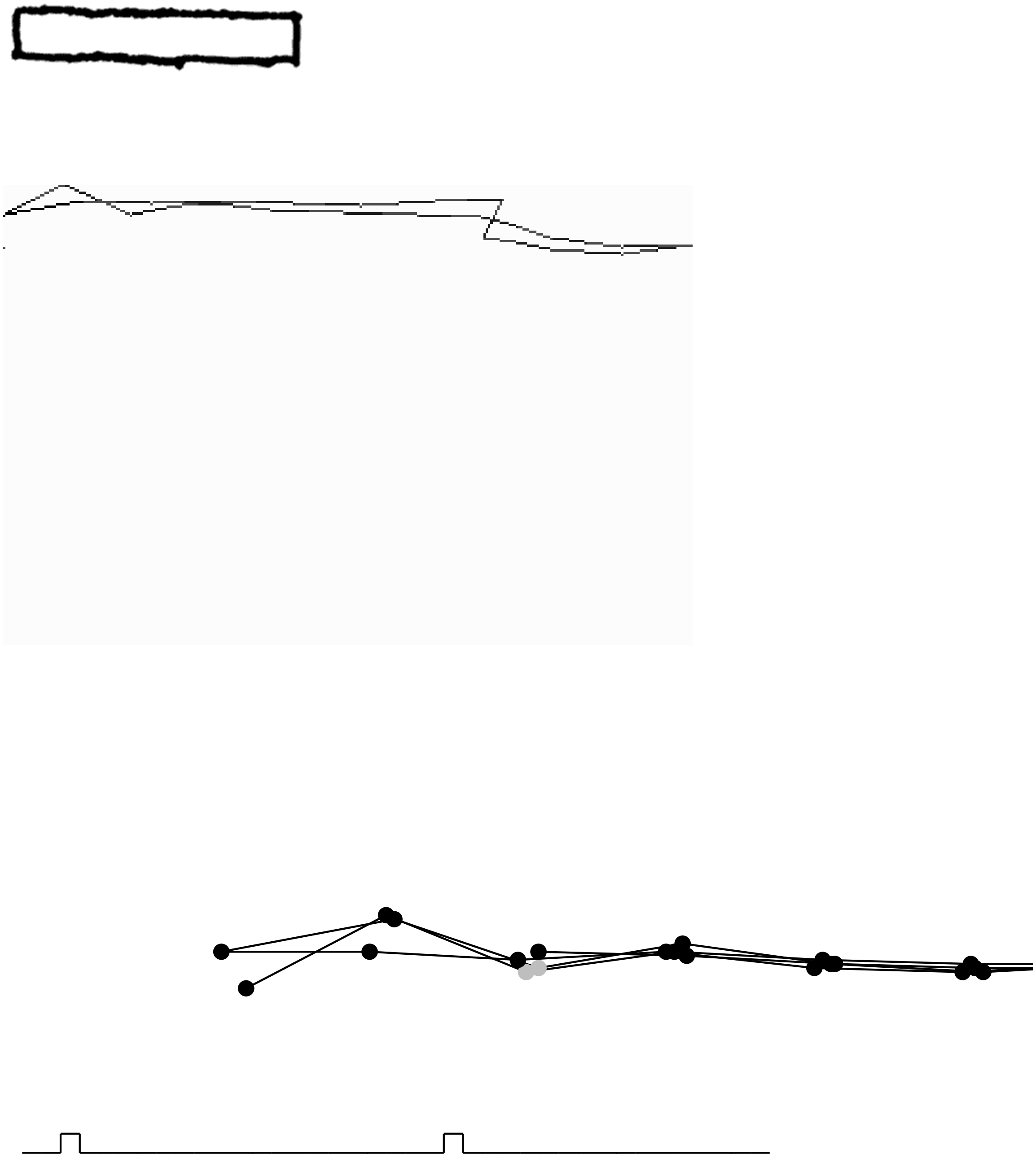

Figure 7-2: A simple corridor environment, and the directional graph built by a modified version of the algorithm, with directional place representations and links. Also shown are place fixes and activity trace of the "unit" corresponding to one directional place representation. Note that the simulated robot traverses the corridor back and forth twice, meaning that it is in each physical locale four times (twice each, out and back). The "unit" is only active twice, showing that it is active selectively depending on direction of transit.

gets more severe. The most direct cause of this is that the number of representations required increases more (with the number of headings from which each place must be approached). However, there is another effect which is even more severe — each of those multiple representations will, when first created, have unexplored exits in the directions of the other, prior approaches, which will yield, at least, an awful lot of redundant exploration to establish links. Both of these effects are on display in the excursion around such a wide area in figure 7.2.

Some of this additional complexity might be eased, of course, by explicitly representing rotation as a separate motor action on the part of the robot, which would correspond internally to transitioning between one of the redundant representations of a given place and another. However, there is no evidence of any such representation in the rodent, and it is interesting to note the reason why. The open-field case in which the directional representations pose the most severe problems for the robot is *also* the case in which rodents acquire hippocampal cells with *non*-directional place fields, in contrast to the apparently more primitive directional place fields seen in simple corridors. It is interesting to speculate that the adoption of nondirectional place fields is a strategy that the brain might adopt for dealing with the same issues of avoiding redundancy and minimizing search which arise when our robot tries to use directional representations in those contexts.

However, even in the simple case where the directional representations seem to work (for both the brain and the robot), they don't seem to confer any immediately obvious advantage. Indeed, in some respects, it seems to confer the opposite — most notably, it multiplies the total number of representations (with the associated increase in total number of nodes, and the probability of blowup due to redundant explorations starting from multiple representations of the same place). We may ask, then, why the brain does things this way.

Two possibilities come to mind. The first is simply a possible limitation on the matching capabilities of the animals. There is experimental evidence that animals do not immediately recognize a place which they have entered before if they are approaching it from an angle which is more than about 90 degrees away from the angle from which the animal had approached the place previously.

If we may momentarily return to the speculations of chapter 6 on the interpretation of the hippocampal activity generally, another explanation suggests itself. Suppose that the hippocampus is, in fact, a general structure for learning associations between situations and the consequences of resulting actions. Then, what may become significant is that the animal is in always in one of two different states in the closed-corridor case (running left *or* running right, but not having to decide which direction to move in). In the open-field case, however, the animal has to choose where it will move from a given place; furthermore, it faces the same choice at any given place no matter from which direction it originally approached that place. So, in the corridor case, the rodent is in two distinct situations with not much in common while at a given location; in the open-field case, it only has one (larger) set of options at any given place. If the representations built by the hippocampus are, in fact, associated with sets of options (and links to possible consequences), this may underlie the difference between the two distinct representations seen in the corridor case, and the single representation seen

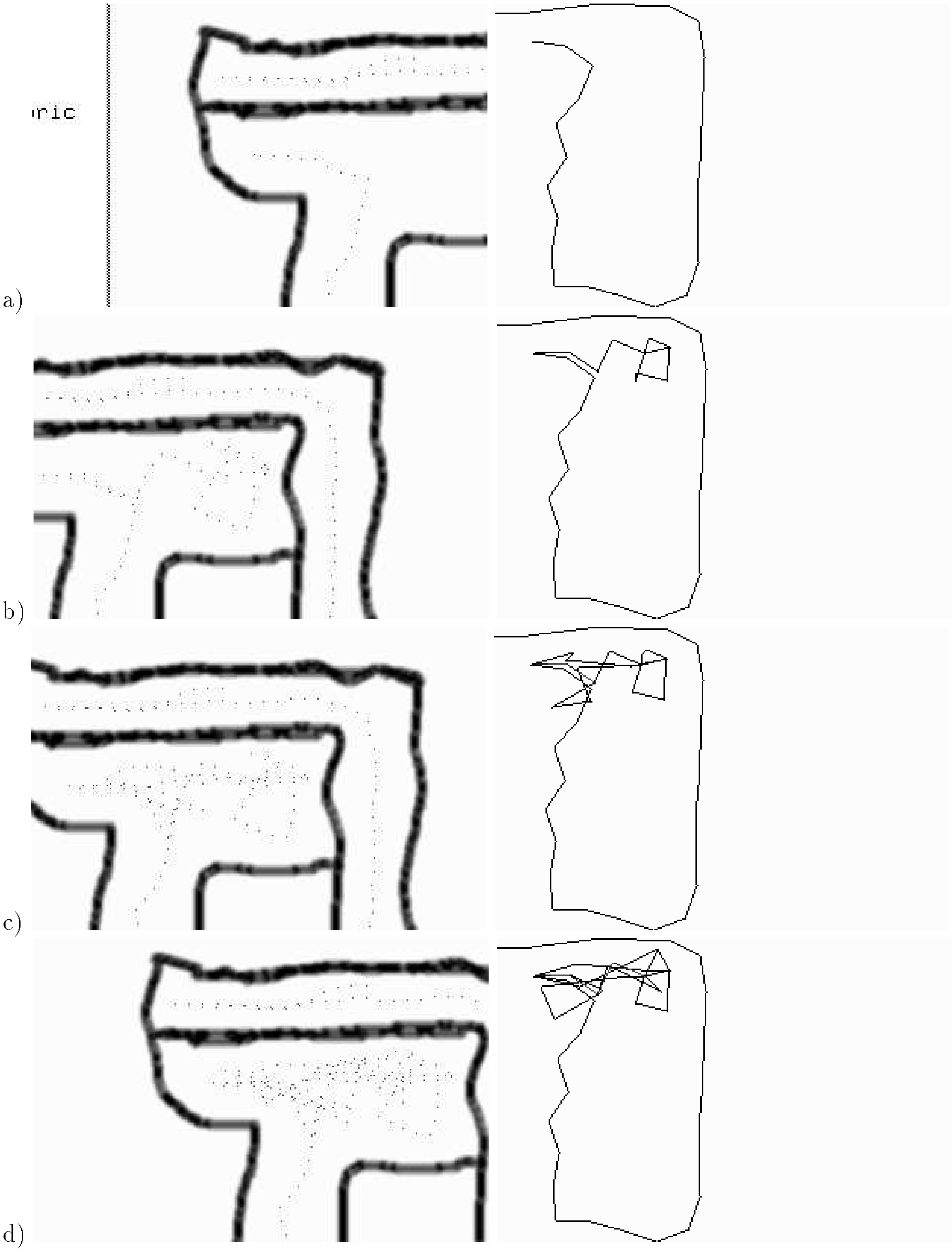

Figure 7-3: Performance of the directional graph algorithm in the case of exploring a large, open area. Note that there are cases, as in the point of map and graph b, where the robot does arrive near enough to a place it has previously encountered, *and facing the same direction*, to recognize it. However, these tend to be very few; the robot quickly acquires a fairly large stock of redundant representations for the same place; exploring all of their unexplored exits would compound the redundancy more before the representation finally stabilized at a massively redundant map.

in the open-field case.

7.3 More brain-like representations

A more systemic difference between the two cases is that our robot algorithms take extensive precautions to avoid building multiple representations of the same place, in order to avoid combinatorial blowup in the total number of representations built (as newly constructed redundant representations have, perforce, unexplored exits which lead to the creation of yet more of them). The brain, by contrast, thoroughly covers the space with an overabundance of representations, and makes no attempt to preserve uniqueness.

This means that the structures residing in the hippocampus of a trained animal (and very possibly elsewhere) are similar in some respects to our system (representing locations and links between them), and different in others (the massive overcommitment). In chapter 5, we have, in effect, extensively discussed the computatational properties of the representations which are built by our robot. In the following sections, we discuss the computational properties a more brain-like version of our representation might possess, and the additional costs and benefits which might be involved in its creation. We then briefly discuss how the differences between the two types of representations, and the associated costs and benefits, might reflect the differences between the underlying computational structures of computers and brains.

First, let us briefly sketch out what a more "brain-like" version of our place graphs might look like. The major difference between the structure itself and the graphs which our actual system builds is, of course, their redundancy. A second immediate difference concerns how the structure itself is used. In our system, the robot represents its own location, in effect, as a single distinguished place representation, and the estimated displacement from its center. The brain is clearly doing something very different — at any given point in time, while the animal is exploring the space it has learned, a very large number of units will be active (some earlier and some later in the theta rhythm in the direction-related pattern discussed earlier). In the graph structure, the most directly analogous sort of representation would be to assign each node in the graph a numerical activity value, and possibly a phase.

The obvious way to get a single estimated location out of such a structure is by averaging the centers of the active representations, weighted by their activations, as discussed earlier. However, that is not necessarily the most useful way to consider such an assignment. It might be more useful to consider the set of activations as a probability distribution over the place representations, in which the activation of each representation indicates a current estimate of the probability that the robot is at (or near) the place represented by that node.

To be processed effectively, such a graph would likewise require a fairly dense set of links. It is interesting to consider the notion of assigning these a probabilistic interpretation as well, in terms of the probability of arriving at a particular place when leaving a particular other place while moving in a particular direction.

Such a "graph with probabilties" would not, of course, directly model the phenomenology of the theta-rhythm. We are neglecting it momentarily because the author feels that the distributed nature of the representation, and its massive redundancy, are the differences with the most impact on the computational tractability of the structures themselves, as discussed below.

As discussed below, the actual construction of such a probabilistic model, with full theoretical rigor, would require the development of a fairly elaborate theory, due to the several levels of dependancies that would be involved, and no instances of such a theory are known in the literature. However, even if researchers don't have such a theory immediately, the brain itself might, and it is interesting to consider the advantages that such a theory might confer.

7.3.1 Attractive features of the "brain-like" approach

One immediately interesting feature of the hypothetical probability-web representation of space is that it allows graceful handling of a set of nasty cases which our existing representation cannot handle easily, by spreading the activation among multiple nodes. For example, consider a probability distribution in which there are two local peaks — this would directly represent a case in which the robot could, on the basis of prior data, be in one of two distinct places, which cannot be distinguished on the basis of immediate sensory data.

The usefulness of this possibility becomes apparent when we consider allowing the robot to move around while having such a bimodal notion of its own location. So long as it moved only among spaces which are, once again, indistinguishable (imagine that the robot knows, on some basis, that it is in one of two long corridors, but has no reliable information as to which), the two peaks will be maintained, and will move correspondingly in the graph (as the robot's estimated positions transit up and down the two corridors). However, when the robot reaches a situation where the two peaks correspond to different sensory situations (say, at the ends of the corridors, where the geometry happens finally, to be different), it is easy to determine by direct sensory inspection which of the two representations is incorrect, and the corresponding peak will be suppressed by the failure to fit the sensory data.

In short, the use of probability distributions, rather than a single distinguished position estimate, would allow the robot to maintain and reason about multiple possible hypotheses about its current position. Indeed, there is nothing which limits us to two peaks in the distribution. A limiting case of this is the (actually quite plausible) scenario of the robot starting up knowing that it is in an environment which it has previously occupied, but not knowing where in that environment it has been placed. With the probability-distribution representation, the robot can construct an initial probability distribution based solely on the sensory resemblances of its current location to *each* of the stored representations, and eliminate the ambiguities by exporing, while suppressing peaks which are inconsistent with sensory data acquired at subsequent locations.

(It is interesting to note, in passing, that given a multimodal probability distribution of the sort we are considering, it is absolutely inappropriate to sum the estimates to give a "global position estimate" — that would yield a position midway between the two corridors, which is one place where, given the full probability distribution, the robot most probably is not).

Incidentally, we have just specified a procedure, in the initial stage of recovery from the "unknown location" condition, which is far more expensive than anything which our actual robot does — in particular, we have specified that in the first step of recovery, building the initial distribution, the robot should compute the prior likelihood of each place representation based solely on available sensory data; in effect, an alignment. What makes that procedure expensive for the robot, of course, is that its single microcontroller in the robot can only consider a single place representation at a time.

In our own procedures (in chapter 5), we have tried to avoid matching the available sensory data to anything more than a subset of the available place representations. Furthermore, we are not dealing now with one of our own graphs, with their relatively small number of place representations, but with a multiply redundant version with a higher number of place representations per unit area. This is a very sizable computational problem — beyond what we would expect from the robot itself (though we might consider reducing the cost by using a cruder match, hoping that the redundancy of the representation, leading to mutual support for correct matches, will quickly suppress any mistakes). However, to do several thousand matches is entirely reasonable for the brain, as one or a few neurons can function as a powerful matching engine in their own right. As we will discuss below, this is an instance of a general pattern.

7.3.2 Iterative processes vs. "brain-like" LTP

One immediately striking difference between the two systems is that it is difficult to imagine any direct biological parallel to the link creation procedure of section 5.4.4. Recall that the goal of this procedure is to establish links between two separate chains of representations of the same general physical region. The procedure itself calls for immediate establishment of full-strength connections between designated place nodes which meet a particular, fairly elaborate criterion, which is not something that biological neurons typically do.

However, it is possible that imposing a Hebb-rule like activity upon the network would have the same effect, over the course of several repeated traverses. That is, suppose two separate sets of representations of the same locale arose in a network of the "brain-like" type. By "separate sets", we mean precisely that the two sets of representations are well-connected internally each within each, but that units in one set have no functionally strong connections to units in the other set.

Now, suppose the rodent repeatedly moves through the doubly represented space while Hebb-style reinforcement applies to the (initially weak) connections between the two sets of nodes. Since they both, in fact, represent the same region of the environment, they will tend to fire in coincident fashion. Thus, strong connections between them will be produced by the Hebb-like process in precisely those cases where analogous to the cases where graph links are introduced in our robot by the procedure of section 5.4.4.

Of course, the consequence of somehow disabling the Hebb-rule learning of such a system would be catastrophic — the system might be able to form useful representations of individual places, but it would be unable to form any useful notion of their relationships, since the mechanism for doing so would be blocked. Therefore, even if one could find active place representations inside the system, and they seemed to be precise enough to be useful, it still could not be used for actual navigation.

Interestingly, the actual hippocampi of rodents — and most specifically, the CA1 and CA3 gyri wherein place cells are most often found, have long been known to host a Hebb-like mechanism known as "long-term potentiation", which causes synapses between two neurons to increase in effective strength if the postsynapctic neuron fires shortly after it receives stimulation from the synapse [55].

Further, recent work has been on genetically engineered mice [50, 51, 31], which (thanks to some very clever molecular genetics) lack one of the key proteins for operation of the LTP mechanism — a specific type of membrane-bound glutamate receptor known as the NMDA receptor, on the postsynaptic side – in the CA1 gyrus of the hippocampus. (The brains of the mutant rats being apparently normal elsewhere). Interestingly, examination of these mice shows precisely the pattern above — normal place fields are found for cells in the CA3 field, and unusually large, but not otherwise abnormal, place fields are found in CA1 (perhaps parasitic on the place-field activity of the CA3 inputs), but the mice are simply unable to perform on simple navigation-related tasks, such as the standard "water maze".

7.3.3 Problems with the "brain-like" approach

The discussion above has presumed that the brain is, in effect, implementing some sort of probabilistic model, in which the activations of units can be seen as (representations of) a probability distribution over possible locations (or, e.g., behavioral states) of the rat, and the physiological connections between them made manifest by high cross-correlations in sleep are seen as (representations of) a probability distribution over the possible links between one place in the environment and another.

However, there are deep difficulties in constructing such a theory, particularly if the probabilistic view of links is adopted. The problem comes in trying to manage and update such a structure — simply stating the conditions which the various sets of probabilities would have to meet in order to be consistent with each other becomes quite difficult. Likewise, interpreting the state of such a system, should it fall into a state where all the probabilities are far from their extreme values (close to one or close to zero) might well be an extremely difficult problem. To put it another way, a model in which *everything* is subject to probabilistic interpretation, and *nothing* is known for certain, must be very carefully arranged to avoid a kind of ontological collapse.

However, even with the a sound theory in place, we would still encounter problems with the "brain-like" data structure due to its scale. We have already shown that for one possible application of this data structure, namely, "cold-starting" the robot at an unknown location in a previously known environment. However, the same problem arises in a slightly milder form in the course of ordinary navigation, since, with this structure, many different place representations have supra-threshold activations at any given time, and those at least must be matched against sensory data for their activations to be properly updated. For computers, again, this would pose a severe problem, due to the expense of doing an alignment (although, once again, it might be pointed out that this could be alleviated somewhat by the use of a cheaper, and less reliable, alignment and evaluation procedure). The brain, by contrast, would have no problem doing these operations — or indeed, matching the entire set of active and non-active representations — on a time-scale of tens of milliseconds.

So far, we have just been considering how to use a preexisting "brain-like" map, which has been learned by unspecified means. In particular, forming the graph requires formation of representations of links from one place representation to several (possibly very many) neighbors. Interestingly, this is one case where a purely navigation-related structure might wish to use the structure of the underlying space being explored (in fact, merely the constraint that it *is* a space, with some underlying metric, in which the places close to a given place are close to each other) in order to make a pre-built dense web of local connections. But, if the brain were doing this, we would expect to see that nearby place-field cells have nearby place-fields, and this is not (yet?) seen; no systematic relationship between the physical locations of cells and the locations of their place fields has (yet?) been discovered.

To put it another way — so far as can yet be determined representations of neighboring places, which we might need to link if only to sustain dead-recknoning performance, are not systematically related. Therefore, in order to build the map, the brain (including, possibly, extra-hippocampal structures) must somehow strengthen a link between two hippocampal units which, so far as we can tell, are arbitrarily chosen. It is tempting to merely attribute this to a process implementing something like the Hebb rule for modifying strengths of connections, such as the long-term potentiation which is known to be operating in the hippocampal formation. However, such a process can only operate to strengthen a connection which already physically exists, so this hypothesis would require the prior existence of a very dense web of connections, most of which would be inactive at any given time. The alternatives are that the brain has some way of physically growing new connections as required (by growing new dendrites), or that there *is* in fact some systematic relationship between cells representing neighboring places, which has simply eluded discovery to date.

Another interesting question regarding the maintenance of such a data structure is determining when to tear down old structure or old relationships. The observation that most cells in CA1 tend to acquire some place field in any given environment would seem to imply that there is a need to tear down old structure, lest memories of old environments interfere with learning of, and performance in, the environment currently being learned.

Experimental data seems to suggest that the brain itself may actually have this problem; work in the Wilson lab and elsewhere shows that in cases where the environment itself changes (or, in some cases, where the environment doesn't change but the rat has been disoriented) it appears that the map, as made manifest in the place fields shown by particular cells, can shift radically in fairly short periods of time.

Two questions arise in connection with this phenomenon. The first is the question of determining when to eliminate connections. Here again, the very redundancy of the network may be a saving grace — it may not be bad to be continually paring away at the network and rebuilding it, since the bits which have not been pared away will continue to function, and help to retrain the newly untrained parts; conversely, if the animal does enter a new or changed environment, the old representations will find little sensory support, the newly untrained regions would retrain themselves to suit the new environment, and the new environment would be acquired.

The second, more interesting question, concerns whether it is, in fact, appropriate to tear down known-stable representations. Common sense suggests this is risky; if the animal's environment has been stable for a long period of time, and suddenly changes, it may very well change back, and having to completely relearn a formerly familiar environment would expose the animal to unnecessary danger. Of course, the brain might well deal with this by "laying down" sufficiently stable hippocampal memories elsewhere in cortex, as implied by lesion data, through some as yet unknown mechanism.

7.4 Computers vs. brains?

A theme running through the past few pages has been that the potential advantages of the "brain-like" data structure which we have been envisioning are purchased at the cost of enormous computation — but computation which can proceed in parallel, with communication among units which have (or are acquiring) direct connections, but without substantial global communication. This renders it suitable for the brain, which is made up of units capable of performing fairly complex, but non-iterative, computations on several thousand inputs on a timescale of milliseconds to tens of milliseconds.

This organization contrasts strongly with the organization of modern computers, which is increasingly tending towards a very fast central processor, with a large memory. Interestingly, performance of such computers is increasingly constrained by the inability to get information into and out of the central processor; a key concern in the design of modern supercomputers — those specialized for high-speed evaluation of numerical models — is, in fact, bandwidth to memory. The brain's organization of lots of relatively slow, but highly flexible processors, seems to have advantages in this problem domain.