Engines of the brain: The computational instruction set of human cognition

Richard Granger

Brain Engineering Laboratory
6207 Moore Hall, Dartmouth College, Hanover, NH 03755
(Formerly University of California, Irvine)
Richard.Granger@Dartmouth.edu; www.BrainEngineering.org

Abstract

Vast information from the neurosciences may enable bottom-up understanding of human intelligence; i.e., derivation of function from mechanism. This paper describes such a research program: simulation and analysis of the circuits of the brain has led to derivation of a specific set of elemental and composed operations emerging from individual and combined circuits. We forward the hypothesis that these operations constitute the "instruction set" of the brain, i.e., the basic mental operations from which all complex behavioral and cognitive abilities are constructed, establishing a unified formalism for description of human faculties ranging from perception and learning to reasoning and language, and representing a novel and potentially fruitful research path for the construction of human-level intelligence.

Introduction

There are no instances of human-level intelligence other than ourselves. Attempts to construct intelligent systems are strongly impeded by the lack of formal specifications of natural intelligence, which is defined solely in terms of observed and measured human (or animal) abilities, so candidate computational descriptions of human-level intelligence are necessarily under-constrained. This simple fact underlies Turing's proposed test for intelligence: lacking any specification to test against, the sole measures at that time were empirical observations of behavior, even though such behaviors may be fitted by multiple different hypotheses and simulated by many different proposed architectures.

Now, however, there is a large and growing body of knowledge about the actual machinery that solely computes the operations of human intelligence, i.e., human brain circuitry. By studying the structural (anatomical) and functional (physiological) mechanisms of particular brain structures, the operations that emerge from them may be identified via bottom-up analysis. The resulting algorithms often have unforeseen characteristics, including hierarchical structure, embedded sequences, hash coding, and others (see, e.g., Granger et al., 1994; Kilborn et al., 1996; Shimono et al., 2000; Rodriguez et al., 2004). Considered initially in isolation, the anatomical system-level layout of these circuits in turn establishes how the individual operators are composed into larger routines. It is hypothesized that these operators, comprising the "instruction set" of the brain, constitute the basic mental procedures from which allmajor behavioral and cognitive operations are assembled. The resulting constructs give rise to unexpected, and unexpectedly powerful, approaches to complex problems ranging from perception to higher cognition.

The following sections introduce minimal relevant background from neuroscience, to provide a "primer" for those neurobiological components from which computational abstractions will be constructed. The emphasis is on deriving constraints that limit hypotheses to those concordant with known biology. Conforming hypotheses are then presented, and sample computational realizations of these are introduced and characterized.

Organization of the human brain

Figure 1 depicts primary elements of the mammalian forebrain (telencephalon), shared across all mammal species and growing to become far and away the dominant set of structures in human brain. In the figure, sensory input is received by posterior cortex (PC), via diencephalic (non-forebrain) thalamic nuclei (T), whereas motor outputs are produced via interactions between anterior cortex (AC) and the elements of the striatal complex or basal ganglia (S, striatum; P, pallidum). Mammalian brains scale across several orders of magnitude (from milligrams to kilograms; mice to mammoths), yet overwhelmingly retain their structural design characteristics. As the ratio of brain size to body size grows, particular allometric changes occur, defining differences between bigger and smaller brain designs. As in parallel computers, connections among components are among the most expensive attributes, strongly constraining design. As the brain grows, those structures and connection pathways that grow disproportionately large are highly likely to be the most indispensable machinery, as well as developing into the key components of human brain that may set human intelligence apart from that of other mammals. Figure 1b illustrates the three largest changes that occur:

1) Connection pathways between anterior and posterior cortex ("fasciculi") grow large.
2) Output pathways from striatal complex change relative size: the recurrent pathway back to cortex via thalamus increases relative to the descending motor pathway.

3) Descending output from anterior cortex to brainstem motor systems (pyramidal tract) grows large.

These changes grow disproportionately with increased brain-body ratio, becoming notably outsized in humans. In relatively small-brained mammals such as mice, the primary motor area of neocortex is an adjunct to the striatally driven motor system. Whereas damage to motor cortex in mice causes subtle behavioral motor impairments, damage to motor cortex in humans causes complete paralysis. In this example of "encephalization of function" (Jackson, 1925; Ferrier, 1876; Karten, 1991; Aboitiz, 1993; Striedter 1997) motor operations are increasingly 'taken over' by cortex as the size of the pyramidal tract overtakes that of the descending striatal system. In mammals with large brain-body ratios, the role of the striatal complex is presumably altered to reflect that its primary inputs and outputs are now anterior neocortex; in other words, it is now primarily a tool or "subroutine" available for query by anterior cortex. For computational purposes, its operations then are most profitably analyzed in light of its dual utility as organizer of complex motor sequences (in small brained mammals) and as informant to anterior cortex (in large brained mammals).

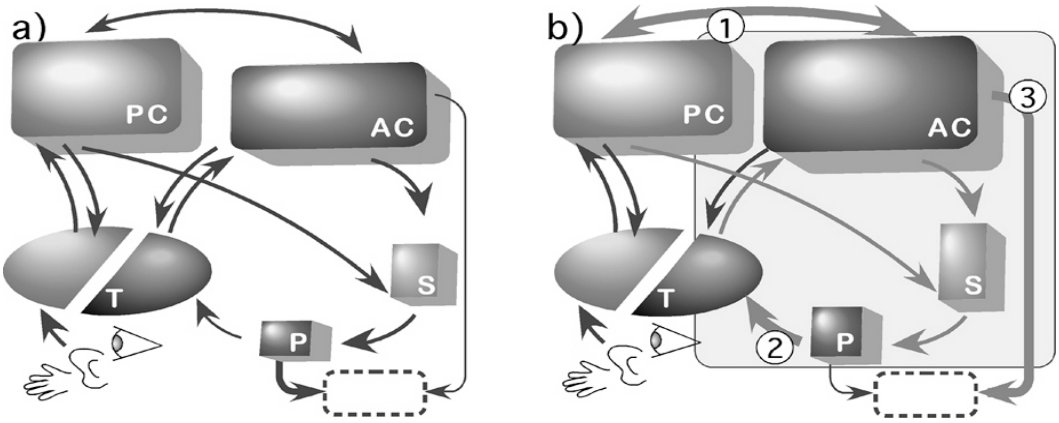

Figure 1. Telencephalic organization in small-brained (a) and large-brained (b) mammals. Posterior cortex (PC) receives primary sensory input (vision, audition, touch) via thalamus and interacts with anterior cortex (AC), which in turn forms loops with striatum (S) and pallidum (P). Pallidum and anterior cortex both produce movement output (dotted box). Brain growth results in differential (allometric) growth of components and interconnects. In particular, disproportionate growth occurs in posterior-anterior connections (1); recurrent return paths form a strong AC→S→P→T→AC loop (2), and motor functions increasingly depend on cortical input (3).

Striatal Complex

The striatal complex or basal ganglia, the primary brain system in reptiles and second-largest structure in humans, is a collection of disparate but interacting structures. Figure 2 schematically illustrates the primary components included in the modeling efforts described herein. Distinct components of the basal ganglia exhibit different, apparently specialized designs. For comparative purposes, note that "S" in figure 1 corresponds to all that is labeled "matrisomes" and "striosomes" in figure 2, and "P" in figure 1 corresponds to all that is labeled "GPe" and "GPi" (pallidum, or globus pallidus, pars interna and externa) in figure 2. Three additional small but crucial components of basal ganglia shown in figure 2 are subthalamic nucleus (STN), tonically active cholinergic neurons (TANs), and substantia nigra pars compacta (SNc). These modules are connected via a set of varied neurotransmitter pathways including GABA, glutamate (Glu), dopamine (DA), acetylcholine (ACh), and Substance P (Sp) among others, each affecting multiple receptor subtypes. Neurotransmitter-receptor pathways can be roughly classified as excitatory (i.e., activating their targets), inhibitory (suppressing activity in their targets) and modulatory (altering the strength of the other two types.

The entire striatal system can be understood in terms of four subassemblies: i) cortex → matrisome projections (action); ii) cortex → striosome projections (evaluation); iii) SNc dopamine (DA) projections to both matrisomes and striosomes (learning); and iv) TAN projections to matrisomes (exploration).

i) Cortex → matrisomes (action):

Two separate pathways from cortex through matrisomes involve different subpopulations of cells: i) MSN1 neurons project to GPi → thalamus → cortex; ii) MSN2 neurons insert an extra step: GPe → GPi → thalamus → cortex. MSN and GP projections are inhibitory (GABAergic), such that cortical excitatory activation of MSN1s causes inhibition of GPi cells, which otherwise inhibit thalamic and brainstem regions. Hence MSN1

cells dis-inhibit, or enhance, cortical and brainstem activity. In contrast, the extra inhibitory link intercalated in the MSN2 pathway causes MSN2s to decrease the activity of cortex and brainstem neurons. These two pathways through MSN1 and MSN2 cells are thus termed "go" and "stop" paths, respectively, for their opposing effects on their ultimate cortical and motor targets. Coordinated operation over time of these pathways can yield a complex combination of activated (go) and withheld (stop) motor responses (e.g., to stand, walk, throw), or correspondingly complex "thought" (cortical) responses. These action responses will be subsequently modified by calculations based on action outcomes, as described below.

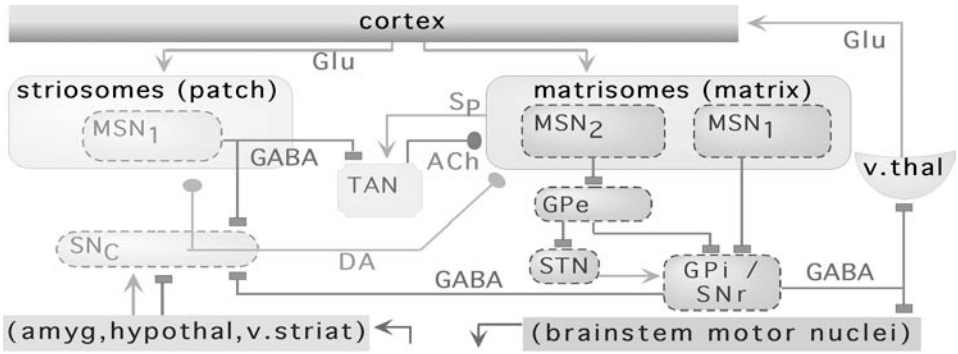

Figure 2. Schematic diagram of components and connection systems of striatal complex (basal ganglia). Medium spiny neurons (MSNs) in matrisome and striosome systems each receive cortical inputs. Striosomes form an inhibitory loop with SNc; matrisomes output to GP and thence back to cortex via thalamus (v.thal). Primary connections are denoted as excitatory (arrows), inhibitory (blocks) or modulatory (circles). (See text).

ii) Cortex → striosomes (evaluation):

The cortex → striosome path initially triggers what can be thought of as an "evaluation" signal corresponding to the "expected" reward from a given action. As with cortex → matrisomes, these expected reward responses can be initially "pre-set" to built-in default values but will be modified by experience (e.g., sensor measures). Each cortically-triggered action (via cortical-matrisome path) will activate a corresponding "reward expectation" via striosomes. Striosomes will then inhibit SNc as a function of that expected reward.

iii) SNc feedback → matrisomes & striosomes (learning):

In addition to input from striosomes just described, SNc receives input from the environment conveying "good" or "bad" state measurement information; i.e., if the action just performed resulted in a good outcome, SNc's activity is increased ("reward") whereas if the action resulted in an undesired state, SNc is decreased ("punishment"). SNc simply compares (e.g., subtracts) this input against its input from striosomes. The resultant difference between the actual reward and the striosomal "expectation," either a positive or negative resultant, becomes input back to both striosomes and matrisomes. In both cases, this calculated positive or negative resultant from SNc increases or decreases the strength of connections from cortex to MSN units. In matrisomes, if connection strength is increased, then the same input will tend to select the same output action, with increased probability. If decreased, then that cortical input's tendency to select that action will diminish, and other possible actions will compete to be the outcome from this cortical input. Similarly, in striosomes, strengthening or weakening connections between active cortical inputs and their striosomal targets will either increase or decrease the size of the future "expectation" produced by striosomes from this cortical input. Thus the system adjusts its selection of actions based on its experience of outcomes from those actions.

iv) TANs → matrisomes (exploration):

TANs receive inhibitory inputs from striosomes, and provide input to matrisomes. TANs can be thought of as a semi-random or "X" factor affecting matrisomes' choice of action from a given cortical input. For actions that have a negative expected reward (or a relatively small positive one), the inhibitory effect from striosomes onto TANs will be correspondingly small, and TANs modulatory effect on matrisomal action-selection will be unimpeded, leading to high variability in the matrisomal process of selecting actions from their cortical input. For actions that elicit a strongly positive expected reward from striosomes, the result will be strong striosomal inhibition of TANs, reducing their "X-factor" effect on matrisomes, lessening the variability of response (i.e., increasing the reliability with which an action will be selected by cortical inputs alone, without TANs' outside influence). The resulting behavior should appear "exploratory," involving a range of different responses to a given input. The mechanism provides a form of

"sensitivity analysis," testing the effects of slight variations in the process of selecting actions from input states.

The overall system can be thought of in terms of an adaptive controller, beginning with pre-set responses to inputs, tracking the outcomes of those responses, and altering behavior to continually improve those outcomes, as in reinforcement learning algorithms (Schultz et al., 1997; Schultz 2000; Dayan et al., 2000; see Table 1).

Table 1. Simplified basal ganglia algorithm

1) Choose action A. Set reward_estimate ← 0
   Set max_randomness ← R > 0
2) randomness ← max_randomness – reward_estimate
3) reward ← Eval(A + randomness)
4) If reward > reward_estimate then
   A ← A + randomness
   reward_estimate ← reward
5) goto step 2)

Thalamocortical system

Neurons throughout neocortex are organized into relatively stereotypical architectures (Figure 3a). Although cortical studies describe some (subtle but potentially crucial) differences among various cortical regions (e.g., Galuske et al., 2000; Gazzaniga, 2000), the overwhelmingly shared characteristics justify longstanding attempts to identify common basic functionality, which may be augmented by special purpose capabilities in some regions (Lorente de No, 1938; Szentagothai, 1975; Keller & White, 1989; Rockel et al., 1980; Castro-Alamancos & Connors, 1997; Braitenberg & Schuz, 1998; Valverde, 2002).

Two parallel circuit types occur, involving topographic projections of certain restricted thalamic populations and broad, diffuse projections from the remaining thalamic neurons. These two populations of thalamic cells, respectively termed "core" and "matrix" (no relation, confusingly enough, with "matrix" in striatum), are distinguishable by their targets, topography, and chemistries (Jones, 2001).

These two loops are activated as follows: peripheral inputs activate thalamic core cells, which in turn participate in topographic activation of middle cortical layers; e.g., ear → cochlea → auditory brainstem nuclei → ventral subdivision of medial geniculate (MGv) → primary auditory cortex (A1) (see Freund et al., 1985; 1989; Peters & Payne, 1993). Other cortical layers are then activated in a stereotypical vertically organized pattern: middle layers → superficial → deep layers. Finally, deep layer (layer VI) projections return topographically to the originating core thalamic nucleus, both directly and via an inhibitory intermediary (the nucleus reticularis). This overall "core" loop pathway is depicted in Figure 3b.

In contrast, matrix nuclei receive little or no peripheral sensory input, and are instead most strongly driven only by corticothalamic feedback (Diamond et al. 1992). Thus, once sensory inputs activate the core loop, then feedback from deep layers activates both core and matrix thalamic nuclei via these corticothalamic projections (Mountcastle 1957; Hubel & Wiesel 1977; Di et al. 1990); the matrix thalamus then provides further inputs to cortex (Figure 3c). Unlike core thalamic input, both feedforward and feedback pathways between cortex and matrix thalamus are broad and diffuse rather than strongly topographic (Killackey & Ebner, 1972, 1973; Herkenham 1986; Jones 1998).

Three primary modes of operating activity have typically been reported for thalamic neurons in these corticothalamic loops: tonic, rhythmic and arrhythmic bursting. The latter appears predominantly during non-REM sleep whereas the first two appear during waking behavior (McCarley et al., 1983; Steriade & Llinas, 1988; McCormick & Bal, 1994). There is strong evidence for ascending influences from ancient conserved brain components (e.g., basal forebrain) affecting the probability of neuronal response during the peaks and troughs of such "clocked" cycles. The most excitable cells will tend to fire in response even to slight afferent activity whereas less excitable neurons will only be added in response to stronger input; this excitability gradient selectively determines the order in which neurons will be recruited to respond to inputs of any given intensity (see, e.g., Anton et al., 1991) during any particular active cycle during this clocked or synchronous behavior.

Axons of inhibitory interneurons densely terminate preferentially on the bodies, initial axon segments, and proximal apical dendrites of excitatory pyramidal cells in cortex, and thus are well situated to exert powerful control over the activity of target excitatory neurons. When a field of excitatory neurons receives afferent stimulation, those that are most responsive will activate the local inhibitory cells in their neighborhood, which will in turn inhibit local excitatory cells. The typical time course of an excitatory (depolarizing) postsynaptic potential (PSP) at normal resting potential, in vivo, is brief (15-20 msec),

whereas corresponding GABAergic inhibitory PSPs can last roughly an order of magnitude longer (80-150 msec) (Castro-Alamancos and Connors, 1997). Thus excitation tends to be brief, sparse, and curtailed by longer and stronger feedback lateral inhibition (Coultrip et al., 1992).

Based on the biological regularities specified, a greatly simplified set of operations has been posited (Rodriguez et al., 2004). Distinct algorithms arise from simulation and analysis of core vs. matrix loops (see Tables 2 & 3).

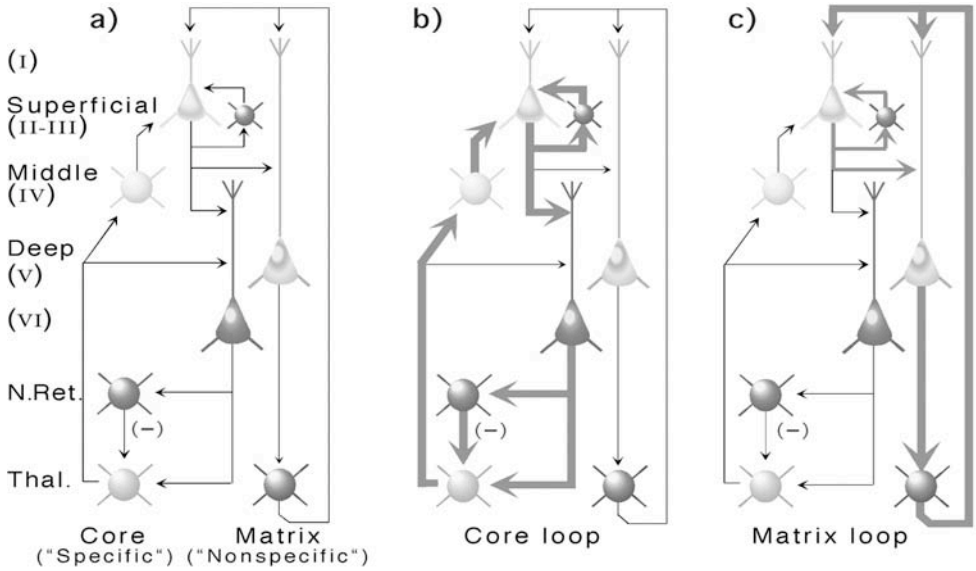

Figure 3. Thalamocortical loops. Complex circuitry (a) can be thought of in terms of two embedded loops: one (b) largely topographic, and incorporating negative feedback (–); the other (c) largely non-topographic, and driven by positive feedback (see text).

Thalamocortical "core" circuits. In the core loop, simulated superficial cells that initially respond to a particular input pattern become increasingly responsive not only to that input but also to a range of similar inputs (those that share many active lines; i.e., small Hamming distances from each other), such that similar but distinguishable inputs will come to elicit identical patterns of layer II-III cell output, even though these inputs would have given rise to slightly different output patterns before synaptic potentiation. These effects can be described in terms of the formal operation of clustering, in which sufficiently similar inputs are placed into a single category or cluster. This can yield useful generalization properties, but somewhat counterintuitively, it prevents the system from making fine distinctions among members of a cluster. For instance, four similar inputs may initially elicit four slightly different patterns of cell firing activity in layer II-III cells but after repeated learning / synaptic potentiation episodes, all four inputs may elicit identical cortical activation patterns. Results of this kind have been obtained in a number of different models with related characteristics (von der Malsburg, 1973; Grossberg, 1976; Rumelhart & Zipser, 1985; Coultrip et al., 1992; Kilborn et al., 1996).

Superficial layer responses activate deep layers (V and VI). Output from layer VI initiates feedback activation of nucleus reticularis (N.Ret) (Liu and Jones 1999), which in turn inhibits the core thalamic nucleus (Fig 3b). Since, as described, topography is preserved through this sequence of projections, the portions of the core nucleus that become inhibited will correspond topographically to those portions of L.II-III that were active. On the next cycle of thalamocortical activity, the input will arrive at the core against the background of the inhibitory feedback from N.Ret, which has been shown to last for hundreds of milliseconds (Cox et al., 1997; Zhang et al., 1997). Thus it is hypothesized that the predominant component of the next input to cortex is only the un-inhibited remainder of the input, whereupon the same operations as before are performed. Thus the second cortical response will consist of a quite distinct set of neurons from the initial response, since many of the input components giving rise to that first response are now inhibited. Analysis of the second (and ensuing) responses in computational models has shown successive sub-clustering of an input: the first cycle of response identifies the input's membership in a general category of similar objects (e.g., flowers), the next response (a fraction of a second later) identifies its membership in a particular subcluster (e.g., thin flowers; flowers missing a petal), then sub-sub-cluster, etc. Thus the system repetitively samples across time, differentially activating specific target neurons at successive time points, to discriminate among inputs. An initial version of this derived algorithm arose from studies of feedforward excitation and feedback inhibition in the olfactory

paleocortex and bulb (Ambros-Ingerson et al., 1990; Gluck & Granger 1993), and was readily generalized to non-olfactory modalities (vision, audition) whose superficial layers are closely related to those of olfactory cortex, evolutionarily and structurally (Kilborn et al., 1996). The method can be characterized as an algorithm (Table 2).

Analysis reveals the algorithm's time and space costs. The three time costs for processing of a given input X are: i) summation of inputs on dendrites; ii) computation of "winning" (responding) cells C; iii) synaptic weight modification. For n learned inputs of dimensionality N, in a serial processor, summation is performed in O(nN) time, computation of winners takes O(n) time, and weight modification is O(N log n). With appropriate parallel hardware, these three times reduce to O(log N), O(log n), and constant time respectively, i.e., better than linear time. Space costs are similarly calculated: given a weight matrix W, to achieve complete separability of n cues, the bottom of the constructed hierarchy will contain at least n units, as the leaves of a tree with log Bn hierarchical layers, where B is the average branching factor at each level. Thus the complete hierarchy will contain ~ n[B/(B-1)] units, i.e., requiring linear space to learn n cues (Rodriguez et al., 2004).

These costs compare favorably with those in the (extensive) literature on such methods (Rodriguez et al., 2004). Elaboration of the algorithm has given rise to families of computational signal processing methods whose performance on complex signal classification tasks has consistently equaled or outperformed those of comparable methods (Coultrip and Granger, 1994; Kowtha et al., 1994; Granger et al., 1997; Benvenuto et al., 2002; Rodriguez et al., 2004).

Table 2. Simplified Thalamocortical Core Algorithm

```
for input X
    for C ∈ win(X,W)
        W_j ⇐ W_j + k(X − C)
    end_for
X ⇐ X − mean(win(X,W))
end_for
```

where

X = input activity pattern (vector); W = layer I synaptic weight matrix;
C = responding superficial layer cells (col vector); k = learning rate parameter;
win(X,W) = column vector in W most responsive to X before lateral inhibition [$\forall \mathbf{j}$, $\max(X \cdot W_j)$]

Thalamocortical "matrix" circuits. In contrast to the topography-preserving projections in the "core" loop between core thalamus and cortex, the diffuse projections from layer V to matrix nuclei, and from matrix nuclei back to cortex (Fig 3c) are modeled as sparsifying and orthogonalizing their inputs, such that any structural relationships that may obtain among inputs are not retained in the resulting projections. Thus input patterns in matrix or in layer V that are similar may result in very different output patterns, and vice versa. As has been shown in previously published studies, due to the nontopographic nature of layer V and matrix thalamus, synapses in layer V are very sparsely selected to potentiate, i.e., relatively few storage locations (synapses) are used per storage/learning event (Granger et al., 1994; Aleksandrovsky et al., 1996; Rodriguez et al., 2004). For purposes of analysis, synapses are assumed to be binary (i.e., assume the lowest possible precision: synapses that are either naïve or potentiated). A sequence of length L elicits a pattern of response according to the algorithm given previously for superficial layer cells. Each activated superficial cell C in turn activates deep layer cells. Feedforward activity from the matrix thalamic nucleus also activates layer V. Synapses on cells activated by both sources (the intersection of the two inputs) become potentiated, and the activity pattern in layer V is fed back to matrix. The loop repeats for each of the L items in the sequence, with the input activity from each item interacting with the activity in matrix from the previous step (see Rodriguez et al., 2004).

Activation of layer V in rapid sequence via superficial layers (in response to an element of a sequence) and via matrix thalamus (corresponding to feedback from a previous element in a sequence) selects responding cells sparsely from the most activated cells in the layer (Coultrip et al., 1992) and selects synapses on those cells sparsely as a function of the sequential pattern of inputs arriving at the cells. Thus the synapses potentiated at a given step in layer V correspond both to the input occurring at that time step together with orthogonalized feedback arising from the input just prior to that time step. The overall effect is "chaining" of elements in the input sequence via the "links" created due to coincident layer V activity corresponding to current and prior input elements. The sparse synaptic potentiation enables layer V cells to act as a novelty detector, selectively responding to those sequential strings that have previously been presented (Granger et al., 1994). The implicit data structures created are trees in which initial sequence elements branch to their multiple possible

continuations ("tries," Knuth, 1997). Sufficient information therefore exists in the stored memories to permit completion of arbitrarily long sequences from just prefixes that uniquely identify the sequence. Thus the sequence "Once upon a time" may elicit (or "prime") many possible continuations whereas "Four score and seven" elicits a specific continuation.

The resulting algorithm (see Table 3) can be characterized in terms of computational storage methods that are used when the number of actual items that occur are far fewer than those that in principle could occur. The number of possible eight-letter sequences in English is 26^8 ≈ 200,000,000,000, yet the eight-letter words that actually occur in English number less than 10,000, i.e., fewer than one ten-millionth of the possible words. The method belongs to the family of widely-used and well-studied data storage techniques of "scatter storage" or "hash" functions, known for the ability to store large amounts of data with great efficiency. Both analytical results and empirical studies have found that the derived matrix loop method requires an average of less than two bits (e.g., just two low-precision synapses) per complex item of information stored. The method exhibits storage and successful retrieval of very large amounts of information at this rate of storage requirement, leading to extremely high estimates of the storage capacity of even small regions of cortex. Moreover, the space complexity of the algorithm is linear, or O(nN) for n input strings of dimensionality N; i.e., the required storage grows linearly with the number of strings to be stored (Granger et al., 1994; Aleksandrovsky et al., 1996; Rodriguez et al., 2004).

Table 3. Simplified Thalamocortical Matrix Algorithm

```
for input sequence X(L)
        for C ∈ TopographicSuperficialResponse( X( L ))
                for V(s) ∈ C ∩ NNtResponse( X( L-1 ))
        Potentiate( V(s) )
   NNt( L ) ⇐ NontopographicDeepResponse(V)
        end_for
   end_for
end_for
```

where L = length of input sequence;
C = columnar modules activated at step X(L);
V(s) = synaptic vector of responding layer V cell,
NNt(L)= response of nonspecific thalamic nucleus to feedback from layer V.

Combined telencephalic algorithm operation and the emergence of complex specializations. In combination with time dilation and compression algorithms arising from amygdala and hippocampal models (Granger & Lynch, 1991; Granger et al., 1994; Kilborn et al., 1996), a rich range of operations is available for composition into complex behaviors. From the operation of thalamocortical loops arises the learning of similarity-based clusters (Table 2) and brief sequences (Table 3), yielding the primary data structure of thalamocortical circuitry: sequences of clusters. These are embedded into thalamo-cortico-striatal loops which enable reinforcement-based learning of these sequences of clusters. The output of any given cortical area becomes input (divergent and convergent) to other cortical areas, as well as receiving feedback from those cortical areas. Each such region in the thalamo-cortico-striatal architecture performs the same processing on its inputs, generating learned nested sequences of clusters of sequences of clusters.

Auditory cue processing. Figure 4a illustrates a spectrogram (simplified cochleogram) of a voice stream (the spoken word "blue"), as might be processed by presumed auditory "front end" input structures. Proceeding left to right (i.e., in temporal order) and identifying "edges" that are readily detected (by simple thresholding) leads to creation of brief sequences / segments corresponding to these edges as in Figure 4b.

The learned cortical sequences (characterized as line segments) correspond to constituents of the signal. As multiple instances of the signal are learned, some features will be strengthened more than others, corresponding to a statistical average of the signal rather than of any specific instance. Outputs from cortical areas are input to other cortical areas, combining individual pairwise sequences into sequences of sequences (actually sequences of clusters of sequences of clusters, etc.), and statistics are accreted for these by the same mechanisms. The result is a widely distributed set of synaptic weights that arise as a result of training on instances of this kind. (There is contention in the literature as to whether such learned internal patterns of synaptic weights are "representations," a term that has baggage from other fields. Without engaging this controversy, we use the expression as a term of convenience for these patterns of weights.) These differ from many other types of

representations, in that they are not strict images of their inputs but rather are statistical "filters" that note their sequence of features (or sequence of sequences) in a novel input, and compete against other feature filters to identify a "best partial match" to the input. It is notable that since each sequence pair simply defines relative positions between the pair, they are independent of particular frequencies or exact time durations.

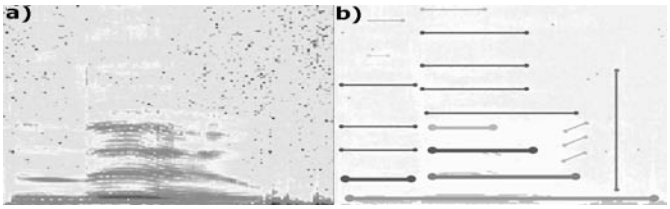

Figure 4. Spectrogram and sample illustration of learned cortical sequences

Figure 5 illustrates two different instances of the utterance "blue" that, after learning, can be recognized by the algorithm as members of the same category, since they contain many of the same organization of relational elements (sequences of clusters, and sequences of clusters of sequences of clusters), whereas other utterances contain distinguishing differences. These representations, arising simply from distributed patterns of synaptic strengthening in the described brain circuit networks, have desirable properties for recognition tasks.

The "best partial match" process can pick out candidate matches from a stream of inputs. Thus the detector for "blue" and that for "bird" identify their respective targets in a continuous utterance (e.g., "the blue bird"). Recognition systems traditionally have difficulty with segmentation, i.e., division of a stream into parts. In the proposed recognition scheme, recognition and segmentation are iteratively interleaved: identification of the sequence components of a candidate word in the stream gives rise to a candidate segmentation of the stream. Competing segmentations (e.g., from sequence components of other words overlapping) may overrule one segmentation in favor of an alternative.

The figure illustrates the nested nature of the operation of the thalamo-cortico-striatal loops. Initial processing of input a) involves special-purpose "front ends" that in the model are carried out by (well-studied) Gabor filters and edge detection methods, producing a first internal representation of sequences as seen in Figure 4. Each successive stage of processing takes as input some combination of the outputs of prior stages. Thus the brief sequences in Figure 4b become input to a copy of the same mechanism, which identifies sequences of the sequences (5b). Downstream regions then identify sequences of those sequences, and so on (5c,d). With learning, the resulting set of relative feature positions comes to share substantial commonalities that are partial-matched, as in the two different utterances of the word "blue" in the top and bottom frames of Figure 5.

Visual image processing. Once past the initial, specialized "primary" cortical sensory regions, thalamocortical circuits are remarkably similar (though, as mentioned, differences have been found, with unknown implications). Moreover, the vast majority of cortical areas appear to receive inputs not originating just from a single sensory modality but from conjunctions of two or more, begging the question of whether different internal "representations" can possibly be used for different modalities.

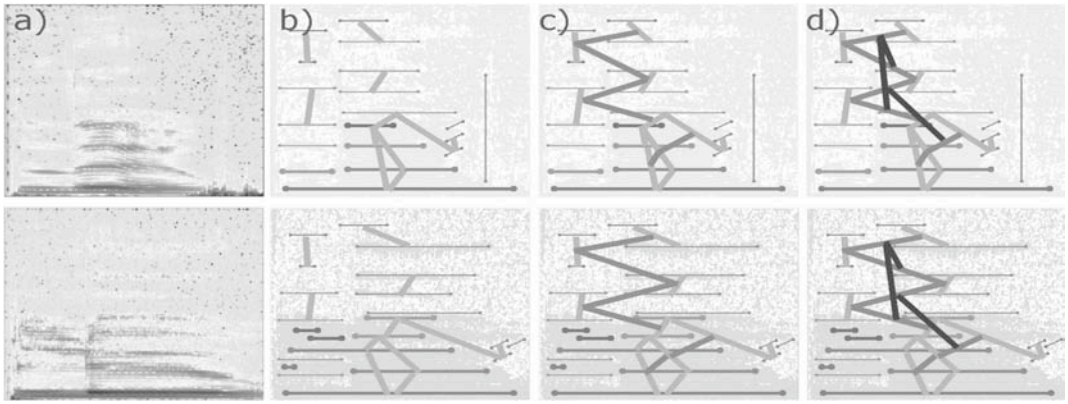

Figure 5. Two utterances and illustration of learned nested sequences (see text). Spectrogram input (a) is processed by an auditory front end (primarily Gabor filters) for edge detection (Figure 4b); the resulting edges are treated as short sequences (b); subsequently, sequences of those sequences, and sequences of sequences of sequences, etc. (b, c, d) are successively identified. The resulting learned downstream data structures are used for partial matching in recognition.

Auditory cortical regions arise relatively early in mammalian evolution (consistent with the utility of non-visual distance senses for nocturnal animals) and may serve as prototypes for further cortical elaboration, including downstream (non-primary) visual areas. It is here hypothesized that, although primary cortical regions perform specialized processing, subsequent cortical regions treat all inputs the same, regardless of modality of origin. The physiological literature suggests particular visual front end processing (arising from retina, LGN, early cortical areas) resulting in oriented line and curve segments comprising an image. From there on, images may be processed as sounds, though due to recruitment of front end visual processing, arbitrary covert "movements" through an image are assumed to occur, rather than processing being limited to an arbitrary "left to right" corresponding to the flow of time in an auditory image. I.e., it is as though auditory processing were a callable subroutine of visual processing. Thus, after initial processing of an image (such as part of Figure 6a) (performed in this case via oriented Gabor filters (6b) at different frequency parameter settings, to roughly approximate what has been reported for visual front end processing from many sources over many years), the resulting segments (pairwise sequences) are composed into sequences of sequences (6c), etc until, over training trials, they become hierarchical statistical representations of the objects (e.g., letters) on which they have been trained (6d).

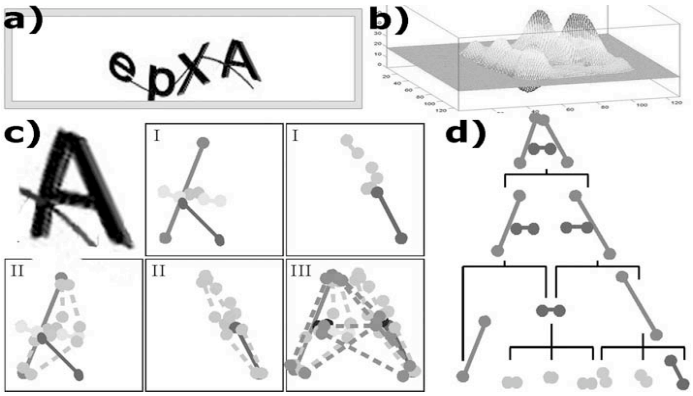

Figure 6. Nested sequences of clusters identified in images (see text). As in audition, inputs (a) scanned by filters (b) give rise to edges that are stored as sequences (c) and constructed into nested sequence structures (d).

As with auditory data, this method leads to representations that iteratively alternate recognition and segmentation; i.e., there exists no separate segmentation step but rather candidate segments emerge, as recognizers compete to identify best partial matches in an image. Further characteristics shared with auditory processing include a number of invariances: translation, scaling and distortion, as well as resistance to partial occlusion. Again, these invariances are not add-on processing routines but rather emerge as a result of the processing. Since the sequences, and sequences of sequences, record relative relationships as opposed to absolute locations, and since the front end filtering occurs across multiple size and frequency scales, recognition of a small A in a corner proceeds just like that of a large centered A. And since the result is merely a best partial match (Figure 7a), a partially distorted (Figure 7b) or occluded (7c) image may match to within threshold.

Figure 7. Emergent invariances from the derived methods. The nested (hierarchical) structure of the internal representations enables partial best-matching of distorted (a), scaled (b) or partially occluded (c) versions of the input.

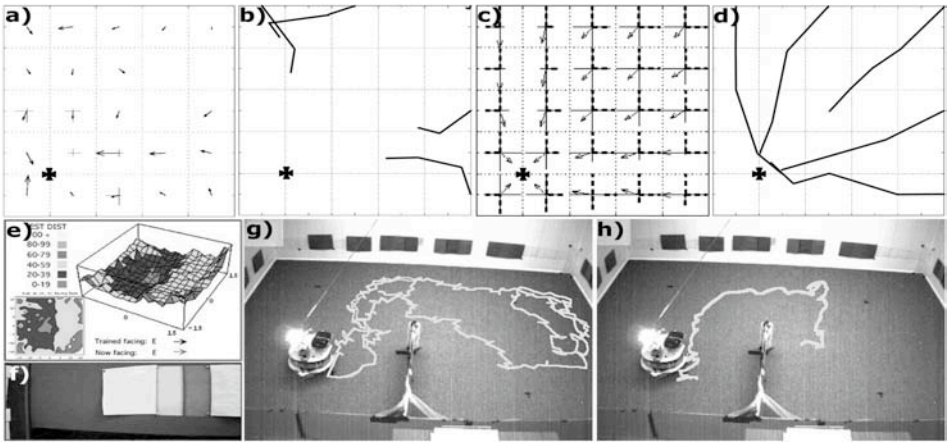

Figure 8. Learned internal representations, and trajectories (see text). Simulated cortico-striatal loops learn via trial and error to traverse a maze; shown are internal representations of learned reward and punishment information before (a, b) and after (c, d) ten thousand trials traversing the space from various starting points. A robot learns a similar internal map (e) of an environment with colored visual cues (f), shortening from initial random (g) to efficient (h) traversals.

Navigation. Presentation of locations containing a hard-coded artificial desirable "goal" state, and sequential reinforcement training from various starting locations, causes the system to improve its approaches to the goal from arbitrary starting points. Figure 8 shows the internal representations (a,c) constructed in the striatal complex as a result of training trials, and illustrates sample trajectories (b,d) to the goal from five starting points, both before (a,b) and after (c,d) this training. The representations correspond to the learned positive and negative "strengths" of four candidate movement directions (N,S,E,W), along with a resultant vector, at each location in the grid. Figures 8e-f show the corresponding internal representation (e) from photographs (f), enabling a robot navigating a simple visual environment to learn from initial reinforced trials (g) to improve its traversals from different starting locations (h).

Hierarchical grammatical structure. It is notable that the emergent data structure of the thalamo-cortico-striatal model, nested sequences of clusters, is a superset of the structures that constitute formal grammars, i.e., ordered sequences of "proto-grammatical" elements, such that each element represents either a category (in this case a cluster), or expands to another such element (nesting), just as rewrite rules establish new relations among grammatical elements.

The incremental nature of the data structure (nested sequences of clusters) enables it to grow simply by adding new copies of thalamo-cortico-striatal (TCS) loops, corresponding to the incremental addition of "rules" acquired by the grammar, adding to the power of the resulting behavior that the data structure can give rise to. As more telencephalic "real estate" is added, the data structures that are constructed correspond to both longer and more abstract sequences, due to iterative nesting. In the model, though all "regions" are identical in structure, they receive somewhat different (though overlapping) inputs (e.g., certain visual features; certain combinations of visual and auditory features). After exposure to multiple inputs, regional specializations of function (e.g., human voices vs. other sounds; round objects vs. angular objects) arise due to lateral competition among areas, giving rise to "downstream" regions that, although performing the same computational function, are selectively performing that function on different aspects of their "upstream" inputs, thus becoming increasingly dedicated to the processing of particular types of inputs. Within each such area, data structures become increasingly abstract, each one matching any of a number of different inputs depending not on their raw perceptual features but on the relations among them.

As these nested structures are built up incrementally, successively more complicated behaviors arise from their use. This is specifically seen in examples above. E.g., in Figure 5, successive processing of the input, carried out by increasingly downstream components of the model, identifies first a simple set of features and relations among those features; then successively more complex nested relations among relations. Thus small-brained mammals may acquire relatively small internal grammars, enabling learning of comparatively simple mental constructs, whereas larger-brained mammals may learn increasingly complex internal representations. That is, changing nothing of the structure of thalamocortical loops, only the number of them, can in this way give rise to new function.

The extensible (generative) nature of human language has typically been explained in terms of

grammars of this kind: from a given grammar, a potentially infinite number of outputs (strings in the language) can be produced. Humans uniquely exhibit rapidly-acquired complex grammatical linguistic behavior, prompting the search for uniquely human brain regions that could explain the presence of this faculty in humans and its absence in other primates (see, e.g., Hauser et al., 2002; Fitch and Hauser 2004; O'Donnell et al., 2005; Preuss 1995; 2000; Galuske et al., 2000). The modeling described herein leads to a specific hypothesis: that human language arises in the brain as a function of the number of thalamo-cortico-striatal loops. With the addition of TCS modules, some become increasingly dedicated to communication due to their inputs, just as some other areas become increasingly dedicated to particular subsets of visual inputs. Rather than wholly new brain modules that differentially process language, the evolutionary addition of TCS modules leads to the incremental acquisition of linguistic abilities. This growth need not be linear; grammars have the property of exhibiting apparently new behaviors due to the addition of just a few rules. There is a fourfold difference in overall brain size between humans and our closest primate relations (chimps, bonobos), and a far greater size difference if just the anterior cortical areas underlying language abilities are considered. There are no living apes or hominids with brain sizes between those of humans and other primates. If human language arises directly from increased TCS loops, then the present "computational allometry" argument suggests that intermediate prelinguistic or protolinguistic abilities may have been present in early hominids, even though not in extant primates. The conjecture is consistent with a broad range of constraints that are argued to rule out alternative hypotheses (see, e.g., Pinker 1999; Pinker & Jackendoff 2005).

The processing of linguistic input, then, need not be a different function from that of other brain processing, but rather the same computational faculties present in smaller brains, now applied in far larger numbers. With an understanding of the specific nature of these computations, it is possible to see how they operate on simpler (e.g., perceptual) inputs as well as complex (linguistic) inputs, differing enormously in the depth of processing and thus the size of the constructed grammars.

Conclusions

Procedures that seem easy and natural to humans (e.g., language) and even to other animals (image recognition, sound recognition, tracking), have been notoriously difficult for artificial systems to perform. Many of these tasks are ill-specified, and the only reason that we know that our current engineering systems for vision and language can be outperformed is that natural systems outperform them.

Human brains arose via a series of intermediaries and under a range of different conditions, without any set of computational plans or top-down principles. Thus brains and their constituent circuits are not "optimized" for any particular task but represent earlier circuits co-opted to perform new jobs, as well as compromises across multiple tasks that a given circuit may have to participate in under different circumstances. Bottom up analysis of circuits, without targeting any "intended" or "optimized" functions, leads to a set of computational units that may comprise the complete "instruction set" of the brain, from which all other operations are composed. The overwhelming regularity of cortical structures, and of large loops through cortical and striatal telencephalon, suggests the universality of the resulting composite operations.

The basic algorithms that have been derived include many that are not typically included in proposed "primitive" or low-level sets: sequence completion, hierarchical clustering, retrieval trees, hash coding, compression, time dilation, reinforcement learning. Analysis indicates the algorithms' computational efficiency, showing that they scale well as brain size increases (Rodriguez et al., 2004). Application of these derived primitives gives rise to a set of unusual approaches to well-studied tasks ranging from perception to navigation, and illustrates how the same processes, successively re-applied, enable learning of data structures that account for generative human language capabilities.

Persistent questions of brain organization are addressed. For instance: How can replication of roughly the same (neocortical) circuit structure give rise to differences in kind rather than just in number? Thalamocortical and corticostriatal algorithms must be constituted such that making more of them enables interactions that confer more power to larger assemblies. This property is certainly not universal (e.g., backpropagation costs scale as the square of network size, and do not solve new kinds of problems simply by growing larger). As discussed, it is the nature of the particular data structures formed by the telencephalic algorithms, nested sequences of clusters, and their relation to grammars, that enables simple growth to generate new capabilities.

What relationships, if any, exist between early sensory operations and complex cognitive operations? The specific hypothesis is forwarded here that, beyond initial modality-specific "front end" processing, all telencephalic processing shares the same operations arising from successive thalamo-cortico-striatal loops. Complex "representations" (objects, spaces, grammars, relational dictionaries) are composed from simpler ones; "cognitive" operations on these complex objects are the same as the perceptual operations on simpler representations; and grammatical linguistic ability is

constructed directly from iterative application of these same operators.

Many systems that learn statistically and incrementally have been shown to be inadequate to the task of learning rule-like cognitive abilities (Pinker, 1999). We have here illustrated that unusual data structures of grammatical form arise directly from models that contain the anatomical architectures and physiological operations of actual brain circuits, demonstrating how this class of circuit architecture can avoid the problems of extant models and give rise to computational constructs of a power appropriate to the tasks of human cognition. Ongoing bottom-up analyses of brain circuit operation may continue to provide novel engineering approaches applicable to the seemingly intractable problems of cognition.

Acknowledgements: This work was supported in part by funding from the Office of Naval Research and the Defense Advanced Research Projects Agency.

References

Aboitiz F (1993) Further comments on the evolutionary origin of mammalian brain. Medical Hypotheses 41: 409-418.

Aleksandrovsky B, Whitson J, Garzotto A, Lynch G, Granger R (1996) An algorithm derived from thalamocortical circuitry stores & retrieves temporal sequences. IEEE Int'l Conf on Pattern Recognition, 550-554.

Ambros-Ingerson J, Granger R, Lynch G (1990) Simulation of paleocortex performs hierarchical clustering. Science 247:1344-1348.

Anton P, Lynch G, Granger R. (1991) Computation of frequency-to-spatial transform by olfactory bulb glomeruli. Biological Cybernetics., 65: 407-414.

Benvenuto J, Jin Y, Casale M, Lynch G, Granger R (2002) Identification of diagnostic evoked response potential segments in Alzheimer's Disease. Experimental Neurology 176:269-276.

Bourassa J, Deschenes M (1995) Corticothalamic projections from the primary visual cortex in rats: a single fiber study using biocytin as an anterograde tracer. Neuroscience 66:253-263.

Braitenberg V, Schüz A (1998) Cortex: statistics and geometry of neuronal connectivity, NY: Springer.

Castro-Alamancos M, Connors B (1997) Thalamocortical synapses. Prog Neurobiology 51:581-606.

Coultrip R, Granger R (1994) LTP learning rules in sparse networks approximate Bayes classifiers via Parzen's method. Neural Networks 7:463-476.

Coultrip R, Granger R, Lynch G (1992) A cortical model of winner-take-all competition via lateral inhibition. Neural Networks 5:47-54.

Cox C, Huguenard J, Prince D (1997) Nucleus reticularis neurons mediate diverse inhibitory effects in thalamus. Proceedings of the National Academy of Sciences 94:8854-59.

Dayan P, Kakade S, Montague P (2000) Learning and selective attention. Nature Neuroscience 3:1218-1223.

Deschenes M, Veinante P, Zhang ZW (1998) The organization of corticothalamic projections: reciprocity versus parity. Brain Research, Brain Research Reviews 28:286-308.

Di S, Baumgartner C, Barth D (1990) Laminar analysis of extracellular field potentials in rat barrel cortex. Journal of Neurophysiology 63:832-840.

Diamond M, Armstrong-James M, Ebner F (1992) Somatic sensory responses in the rostral sector of the posterior group (POm) and in the ventral posterior medial nucleus (VPM) of the rat thalamus. Journal of Comparative Neurology 318:462-476.

Ferrier D (1876) Functions of the brain. London: Smith, Elder.

Fitch T, Hauser M (2004) Computational constraints on syntactic processing in a nonhuman primate. Science 303: 377-380.

Freund T, Martin K, Whitteridge D (1985) Innervation of cat visual areas 17+18 by physiologically identified X & Y-type thalamic afferents. I. Arborization patterns & quantitative distribution of postsynaptic elements. Journal of Comparative Neurology 242:263-274.

Freund T, Martin K, Soltesz I, Somogyi P, Whitteridge D (1989) Arborisation pattern and postsynaptic targets of physiologically identified thalamocortical afferents in striate cortex of the macaque monkey. J Comparative Neurology 289:315-336.

Galuske RA, Schlote W, Bratzke H, Singer W (2000) Interhemispheric asymmetries of the modular structure in human temporal cortex. Science 289:1946-1949.

Gazzaniga MS (2000) Regional differences in cortical organization. Science 289:1887-1888.

Gluck M, Granger R (1993) Computational models of the neural bases of learning and memory. Annual Review Neuroscience 16:667-706.

Granger R, Lynch G (1991) Higher olfactory processes: perceptual learning and memory. Current Opinion Neurobiology 1:209-214.

Granger R, Whitson J, Larson J, Lynch G (1994) Non-Hebbian properties of long-term potentiation enable high-capacity encoding of temporal sequences. Proceedings of the National Academy of Sciences 91:10104-10108.

Granger R, Wiebe S, Taketani M, Ambros-Ingerson J, Lynch G (1997) Distinct memory circuits comprising the hippocampal region. Hippocampus 6:567-578.

Granger R (2002) Neural Computation: Olfactory cortex as a model for telencephalic processing. In: Learning & Memory (J. Byrne, Ed), MacMillan Reference Books, pp. 445-450

Grossberg S (1976) Adaptive pattern classification and universal recoding. Biological Cybernetics 23:121-134.

Hauser M, Chomsky N, Fitch T (2002). The language faculty: What is it, who has it, and how did it evolve? Science 298: 1569-1579.

Herkenham M (1986) New perspecties on the organization and evolution of nonspecific thalamocortical projections. In: Cerebral Cortex (Jones E, Peters, A., ed). NY: Plenum.

Hubel D, Wiesel T (1977) Functional architecture of macaque monkey visual cortex. Procedings of the Royal Society, Lond B Biol Sci 198:1-59.

Jackson JH (1925) Neurological fragments. London: Oxford Univ

Jones E (1998) A new view of specific and nonspecific thalamocortical connections. Advances in Neurology 77:49-71.

Karten H (1991) Homology and evolutionary origins of 'neocortex'. Brain Behavior Evolution, 38: 264-272.

Keller A, White E (1989) Triads: a synaptic network component in cerebral cortex. Brain Research 496:105-112.

Kilborn K, Granger R, Lynch G (1996) Effects of LTP on response selectivity of simulated cortical neurons. J Cognitive Neuroscience 8:338-353.

Killackey H, Ebner F (1972) Two different types of thalamocortical projections to a single cortical area in mammals. Brain Behav Evol 6:141-169.

Killackey H, Ebner F (1973) Convergent projection of three separate thalamic nuclei on to a single cortical area. Science 179:283-285.

Knuth D (1997) The art of computer programming. MA: Addison-Wesley.

Kowtha V, Satyanara P, Granger R, Stenger D (1994) Learning & classification in a noisy environment by a simulated cortical network. 3rd Ann Comp Neur Sys 245-50. Boston: Kluwer

Liu X, Jones E (1999) Predominance of corticothalamic synaptic inputs to thalamic reticular nucleus neurons in the rat. J Comparative Neurology 414:67-79.

Lorente de No R (1938) Cerebral cortex: Architecture, intracortical connections, motor projections. In: Physiology of the nervous system (Fulton J, ed), pp 291-340. London: Oxford.

McCarley R, Winkelman J, Duffy F (1983) Human cerebral potentials associated with REM sleep rapid eye movements: links to PGO waves and waking potentials. Brain Research 274:359-364.

McCormick D, Bal T (1994) Sensory gating mechanisms of the thalamus. Current Opinion Neurobiology 4:550-556.

Mountcastle V (1957) Modality and topographic properties of single neurons of cat's somatic sensory cortex. J Neurophysiology 20:408-434.

O'Donnell T, Hauser M, Fitch T (2004). Using mathematical models of language experimentally. Trends in Cognitive Sci, 9: 284-289.

Peters A, Payne B (1993) Numerical relationships between geniculocortical afferents and pyramidal cell modules in cat primary visual cortex. Cerebral Cortex 3:69-78.

Pinker S (1999). Words and rules: the ingredients of language. New York: HarperCollins.

Pinker S, Jackendoff R (2005). The faculty of language: what's special about it? Cognition, 95: 201-236.

Preuss T (1995). Do rats have prefrontal cortex? The Rose-Woolsey-Akert program reconsidered. J Cognitive Neuroscience, 7: 1-24.

Preuss T (2000). What's human about the human brain? In: The New Cognitive Neurosciences. M.Gazzaniga (Ed.), Cambridge, MA: MIT Press, pp.1219-1234.

Rockel AJ, Hiorns RW, Powell TPS (1980) Basic uniformity in structure of the neocortex. Brain 103:221-244.

Rodriguez A, Whitson J, Granger R (2004) Derivation & analysis of basic computational operations of thalamocortical circuits. J. Cognitive Neuroscience, 16: 856-877.

Rumelhart D, Zipser, D (1985) Feature discovery by competitive learning. Cognitive Science 9:75-112.

Schultz W (2002) Getting formal with dopamine and reward. Neuron 36:241-263.

Schultz W, Dayan P, Montague P (1997) A neural substrate of prediction and reward. Science 275:1593-9.

Shimono K, Brucher F, Granger R, Lynch G, Taketani M (2000) Origins and distribution of cholinergically induced beta rhythms in hippocampal slices. Journal of Neuroscience 20:8462-8473.

Steriade M, Llinas R (1988) The functional states of thalamus and associated neuronal interplay. Phys Rev 68:649-742.

Striedter G (1997) The telencephalon of tetrapods in evolution. Brain Behavior Evolution. 49:179-213.

Szentagothai J (1975) The 'module-concept' in cerebral cortex architecture. Brain Research 95:475-496.

Valverde F (2002) Structure of cerebral cortex. Intrinsic organization & comparative analysis. Revista de Neurología. 34:758-780.

von der Malsburg C (1973) Self-organization of orientation sensitive cells in striate cortex. Kybernetik 14:85-100.

White E, Peters A (1993) Cortical modules in the posteromedial barrel subfield (Sml) of the mouse. J Comparative Neurology 334:86-96.

Zhang S, Huguenard J, Prince D (1997) GABAa receptor mediated Cl- currents in rat thalamic reticular & relay neurons. J Neurophysiology 78:2280-2286.