Graph-Theoretic Structure Identification in Dynamic Link Neural Architectures

Richard B. Wells

Nov. 24 2003

Abstract. A new graph-theoretic approach to neural structure identification with implications for information-theoretic optimization is proposed for dynamic link architecture neural networks. The method addresses the organization and interconnection of feature-representing multicellular units (MCUs) at the network architecture level. It incorporates indices for the identification of both dynamic and static data pathway and binding code links between neurons within an MCU and between different MCUs. It also proposes a new objective function for optimization of a measure of conditional entropies and mutual information properties of the system. The method is based on an extension of a method of graph-theoretic information flow analysis recently proposed by Akuzawa and Ohnishi, here called Akuzawa-Ohnishi Analysis (AOA).

Background. Damasio [1]-[2] has proposed a general theoretical framework for understanding the neural basis of memory and consciousness at the systems level. His model proposes that time-locked multiregional retroactivation of feature-representing networks in sensory and motor cortices, coordinated by binding codes from convergence zone cell assemblies, constitutes a fundamental neural substrate for higher cognitive functions in the central nervous system (CNS). In this model objects (entities and events) are put together from representational fragments of activities, and no one feature-representing network gives the whole representation of an object. Object representations are constructed via binding code signals that feed back and coordinate a multiplicity of specific feature-representing networks. These representations consist of synchronized, time-locked sequences of firing patterns in multiple subnetworks, each of which presents a fragment of the object representation. However, Damasio's system-level model is non-specific with regard to synaptic-level connections and must assume the prior existence of the feature-representing networks acted upon by convergence zone assemblies. This raises the questions of how such feature-representing networks come to be in the first place and how they are organized.

It is well known that sensory-path neurons at almost all levels display various forms of stimulus selectivity in mature animals, including *Homo sapiens*. Furthermore, there is strong supporting evidence for the hypothesis that this selectivity depends on sensory activity during a transient critical period in early postnatal life [3]-[5]. Young neural networks tend to display very coarse connectivity structures which become increasingly fine-tuned during the critical period in

response to patterns of neural activity. Axons participating in the activity pattern tend to form more and stronger synaptic connections, while those not participating in the activity tend to retract. Synchronous firing of afferents within the activity pattern appears to be necessary for the tuning of the network, and the time-order, and frequency of occurrence and repetition of the afferent patterns determines which pattern or patterns will be learned by a specific network.

One hypothesis for explaining this effect is that postsynaptic cells during the critical period of development secrete neurotrophic factors when NMDA channels are opened by synchronous firing patterns converging on that neuron. It is not strictly necessary for the postsynaptic cell to respond to stimulus with an action potential. All that would be required is for the membrane potential in the vicinity of the synapses to depolarize sufficiently to open the NMDA gates. It is also known that the neurotransmitter phenotype developed in a presynaptic cell, and indeed the survival of that cell, may to some degree be dependent upon that cell's targets [6]. These findings suggest activity-dependent secretion of neurotrophic factors induced at other than glutaminergic synapses. There may be other activity-dependent mechanisms as well. The well-known Bienenstock-Cooper-Munro (BCM) model of synaptic plasticity [7] is a mathematical model of development inspired by these hypotheses.

The BCM model regards the neuron as performing spatial integration but synaptic modification as temporal competition between input patterns. It addresses the evolution of plastic synaptic weight changes at the level of the synapse. The BCM theory employs several assumptions regarding the probability distribution of synaptic inputs converging on the target cell as well as assumptions regarding the degree of statistical linear independence among these input signals. It constitutes a temporal competition model at the synaptic level for individual postsynaptic target cells, but it does not address the issue of how this competition affects the formation of whole cell assemblies, the specifics of neural network structure that may emerge from this temporal competition, nor how pre-conditions on initial coarse network connectivity might affect the final connectivity structure as feature networks are formed. These are questions that speak to the problem of identifying suitable structures for feature-representing networks.

The method used in establishing the initial configuration of a neural network prior to any adaptation, i.e. its base structure, has always been something of an art in artificial neural network theory. Options have ranged from specific topological arrangements (e.g. the generic connectionist feedforward network, Hopfield networks, etc.) to probabilistic initial interconnects, e.g. [8]. Studies of pulse-coded neural networks and spiking-neuron-model networks have usually employed specific topologies, as in the case of Malsburg and Schneider [9] or Eckhorn et al. [10].

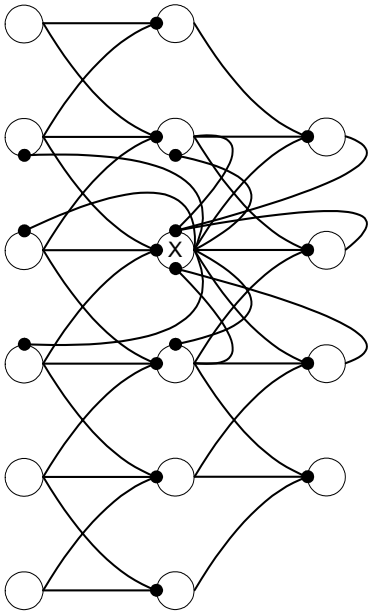

Figure 1: Example network. The specific starting configuration of the network and its subsequent evolution can be represented as a graph or as a connection matrix. Neuron X illustrates the generic scheme of connection. Other neurons may be assumed to employ the same or similar connectivity in the initial state. Filled circles denote synaptic connections, and each circle denotes more than one synapse. Lines issuing from the right-middle portion of the neuron (without a filled circle) represent extruded axons.

In this work we propose a high-level method for establishing a topological schema based on the idea of information flow within the network. The method is applicable to recurrent networks, such as the network illustrated by figure 1, and aims at developing finely-structured sub-networks (called multicellular units or MCUs) from an initial coarse structure. We assume a structure with *m* inputs projecting into a network of *n* neurons. Neurons are allowed to make feed-forward projections to "downstream" layers, lateral projections to neurons in the same layer, and feed-back projections to "upstream" neurons. The general problem to be addressed is identification of the synaptic connections that evolve under experience-dependent input afferents as well as identification of where those connections require "dynamic" processing, i.e. modulated responses from the target neuron on a short-term basis. Such dynamic connections are here called "elastic" synapses and imply rapid, reversible synaptic plasticity [19]. The method makes no *a priori* assumptions on the model neurons in the network, and so is applicable to a variety of neural network models, including pulse-coded neural networks (PCNNs).

General Considerations in Synapse Modeling. In the generic connectionist neural network (GCNN) model, as well as in the majority of other models, inputs to the neural network are usually regarded as connecting to every neuron in a first layer of the network. The synaptic connections they represent are those of ionotropic "data path" channels [11]. The initial synaptic

weights are often set to small random starting values. The adaptation algorithm then establishes which connections are excitatory and which are inhibitory. In addition the adaptation algorithm "prunes" some of the connections by evolving their corresponding weights to near-zero values such that inputs have little or no significant effect on the postsynaptic neuron. While this approach is probably as good as any in the absence of more specific *a priori* information about the network and its function, it is unsatisfying from a biological perspective. First, biological synapses are specifically excitatory or inhibitory, and in some cases may be both ionotropic and metabotropic, depending on the types of receptors expressed in the juxtaposed membrane of the postsynaptic cell [11]. Excitatory synapses are generally made on dendrites whereas most (about two-thirds) inhibitory synapses occur on the cell body with the remaining inhibitory synapses being made on dendrites. Dendrites are widely regarded as fairly sophisticated computing elements in their own right, and are treated as such in higher-order neuron models such as the sigma-pi, clusteron, and cluster models [12], as well as in the Eckhorn model [10].

This biological arrangement seems to be at odds with network learning approaches based on the traditional adaptation schemes, e.g. Perceptron rule, Widrow-Hoff rule, backpropagation, and others. These learning rules make no *a priori* assumptions as to whether or not a particular synapse is excitatory or inhibitory, although they do implicitly take them to be ionotropic channels. The algorithms merely aim to optimize some objective function without regard to the "sign" of the synaptic weights. In biological neural networks there is good reason to think that excitatory and inhibitory ionotropic channels may employ different adaptation schemes.

Excitatory ionotropic synapses, particularly glutaminergic synapses, fit well with the neurotrophic factors hypothesis discussed earlier. It is presently thought that all glutaminergic channels express NMDA receptors and are therefore capable of supporting the sort of metabotropic second messenger cascade reactions [11] believed necessary for secretion of neurotrophic factors involved in early sensory network formulation. The BCM algorithm in particular is aimed at excitatory synapses that support a mechanism of this sort, although BCM makes no explicit assumptions regarding the details of the physiological mechanisms other than the reasonable supposition that synaptic weight modification takes place over a time scale slower than that of the firing rate of the activity patterns. Although it is common practice in BCM modeling to use the firing rate of the postsynaptic neuron as a measure of its activity, the theory itself allows for other measures of postsynaptic activity, e.g. a measure of membrane depolarization in response to local excitatory inputs. BCM is therefore applicable to other Hebbian models such as the calcium-control model [13].

Inhibitory synapses, on the other hand, cannot call upon the same biophysical mechanism for synaptic modification because their effect on the postsynaptic cell is hyperpolarizing and does not involve synapses where NMDA receptors are expressed. Theories for treating inhibitory synapse adaptation are problematical at this time. Many researchers believe that inhibitory channels probably play a necessary role in critical-period fine tuning of feature networks [7], [9]. However, it is not presently clear what role this would be (other than preventing weight saturation) nor if the inhibitory synapses are themselves subject to long-term ("plastic") adaptive modification. Furthermore, since any given neuron is thought to have either only one neurotransmitter phenotype (which may involve one or two ligand neurotransmitters and possibly one or a few neuropeptide neuromodulators) or at most one initial phenotype with the capacity for developing a different mature phenotype, it is likely that most inhibitory synapses within a feature network are mediated by inhibitory interneurons. The approach taken in this work acknowledges the following hypothesis: The role of inhibitory pathways within a feature network is *network* segmentation rather than pattern segmentation. This does not mean that there are no inhibitory afferents coming into the network, but it does imply that synaptic modification involving inhibitory synapses is to be treated differently from that involving excitatory synapses.

To put this hypothesis in other words, a distinction is made between the ability to distinguish between classes of input activity patterns and the ability to form networks dedicated to representing feature fragments. The coarse connectivity structure in young neural networks appears to have the benefit of providing the network with a great many "options" for pattern discrimination, i.e. it provides the potential a large statistical capacity [49] in the target neural system. Network *selectivity* for feature representation, on the other hand, has for its neurological basis the partitioning of the neural system into specialized subnetworks, each of which has a reduced statistical capacity for feature discrimination. A feature-discriminating network must, under Damasio's model, produce synchronized and time-locked firing patterns in response to a class of input patterns it "recognizes" and represents, and this is what excitatory connections sponsor. Differentiation among different classes of inputs, on the other hand, requires accommodation of the network structure, i.e. *limitations* imposed by the formation of distinct cell assemblies. This is a task for which inhibitory connections are well-suited, as demonstrated by "winner-take-all" topologies such as the MAXNET and Mexican Hat networks. One important research question from this point of view is whether or not changes to inhibitory connections should be regarded primarily as *disinhibition* and, if so, whether disinhibition should be regarded as a mechanism for setting up *binding code pathways* rather than sensory data pathways.

As for plastic changes of synaptic weight at inhibitory synapses, it has not been established that long-term plastic change acting to *increase* inhibitory synaptic weight (i.e. long-term potentiation) actually exists, although long-term *depression* (LTD) of inhibitory synaptic weights has been demonstrated [14]-[15]. Even so, LTD involving inhibitory synapses seems more commonly to be the result of LTD at excitatory inputs to the inhibitory neuron as, e.g., in the case of the Purkinje cell [16]. There are, however, a number of non-permanent ("elastic") modulation mechanisms that can temporarily increase or decrease the weight at an ionotropic inhibitory synapse, and these involve either presynaptic mechanisms (axo-axonal synapses) or metabotropic modulations [17].

Putting this all together, excitatory and inhibitory synapses are presumed in this work to play different but complementary roles in the initial formation of feature-detecting networks. Excitatory synapses are presumed to principally function to discriminate activity patterns and produce time-locked firing patterns in response to the selected class of input patterns "recognized" by the feature network. Inhibitory synapses are presumed to principally function to segment networks, and disinhibition is regarded as an enabling mechanism for the formation of binding code patterns within the network. The model of the formation process thus divides naturally into: 1) an optimum filtering problem with respect to data signal pathways; and 2) a structure identification problem with respect to signals that lead to differentiation of subnetworks.

From the point of view of structure identification (as opposed to parameter estimation), what matters is not so much whether a particular synapse is excitatory or inhibitory. What matters is the number of different ways that a neuron is able to respond to its collective synaptic inputs. We regard the neuron as performing a mapping function (cf. eq. (10) below) from a set of input activities to an output activity. The greater is the cardinality of possible responses, the greater is the information-processing capacity of the neuron. The structure identification problem is concerned with determining what this capacity must be for the network as a whole to properly do its function. Therefore this work is not directly concerned with weight adaptation but rather with identifying the cardinality of each neuron's input-output relation. This property of the network's neurons implicates, at a lower level of analysis, the type of synaptic weight change behavior a specific implementation must be capable of providing.

Information Theoretic Considerations in the Structure of Feature Networks. The problem considered here is one instance of the general problem of "compositionality" in neural networks [18]. Furthermore, the development of feature-detecting neural structures is the earliest step in the evolution of a dynamic link architecture (DLA) [19], and as such its natural mode of

mathematical expression is graph-theoretic. Specifically, the problem is one of graph re-configuration from an initial coarse structure into multiple, dynamically-linked fine structures that: 1) exhibit sufficient selectivity to serve as MCUs representing feature-fragment; and 2) contain in their organization sufficient controllability to participate in the representation of objects (entities and events) under the application of binding codes fed back from downstream convergence zone assemblies. The task at hand is a constrained optimization problem, but one for which both the appropriate optimality criteria and the proper constraints are far from clear cut. The general principle of optimal structure construction according to modern system theory is Bellman's Principle, i.e., "An optimal policy has the property that no matter what the previous decisions have been, the remaining decisions must constitute an optimum policy with regard to the state resulting from those previous decisions." However, for the particular task here at hand it is far from obvious what sort of objective function and performance index are appropriate for a mathematical formulation of the adaptation problem in the form of a Hamilton-Jacobi-Bellman (HJB) equation. Whatever form it takes, the task is the same: to limit the number of potentially optimal decisions that must be investigated.

In the abstract, MCUs are information-processing structures and here it is useful to quote a comment made long ago by Weaver [20]: "This word 'information' in communication theory relates not so much to what you *do* say as to what you *could* say. . . . The concept of information applies not to individual messages . . . but rather to the situation as a whole." There are two complementary factors implicit in this description. The first is the idea that an information source produces, and an information channel transmits, a "message" taken from a *set* of possible messages. The second is the idea that the measure of information for the system is a measure that applies globally to the system as a whole. It is not "message-centric" but rather "capacity centric."

Here we have two complementary aspects for describing the development of feature-fragment representing networks. A highly-tuned feature sub-network (MCU) represents a restricted set of possible "messages" from among the afferent patterns entering the system as a whole. It can "say" only a relatively few things. This is necessary if object features are to be made *distinct* and therefore "recognizable." This means that the appropriate performance measure for tuning an MCU is vested in the character of its *information loss* properties. In mathematical terms, this is called the MCU's equivocation, denoted $H\left(X\middle|Y\right)$ where X is the set of inputs and Y is the set of outputs. It is related to the input and output entropies, $H(X)$ and $H(Y)$ respectively, by a well-known theorem (entropy of a function of a random variable) as

$$H\left(X|Y\right)=H\left(X\right)-H\left(Y\right). \tag{1}$$

A highly selective feature network will have a relatively high-valued equivocation, implying maximization of (1), but the character of its equivocation must at the same time be carefully sculpted. To use an analogy, an adder circuit is an example of a carefully-crafted lossy network. For the adder to work properly, some set of inputs must map to one output, e.g. 1 + 4 = 5, 2 + 3 = 5, etc., while others map to a different output, e.g. 1 + 1 ≠ 5. It performs a number of different many-to-one mappings and therefore has high equivocation since knowledge of its output is not sufficient to determine its inputs. Likewise, an MCU must map a large set of input patterns, *X*, into a much smaller set of output patterns, *Y*, including "null patterns" that represent cases where $x \in X$ is not a feature represented by the network. A feature network discriminates.

On the other hand, the "network of networks" that develops from the original coarse structure must have a large capacity for representing many features in *X*. Each feature network can "say" only a relatively few things, but the network of networks can "say" a great many things. Thus the totality of the developed structure constituted by *n* feature networks is characterized by a relatively small equivocation

$$H\left(X|Y_1, Y_2, \cdots, Y_n\right)=H\left(X\right)-H\left(Y_1, Y_2, \cdots, Y_n\right). \tag{2}$$

Equation (2) is therefore an objective function for the system-as-a-whole that is to be minimized. However, how does this square with the idea that equation (1) is a performance function to be maximized? Here some common sense is required because (1) can be maximized in many ways. An absolute maximum is achieved by simply turning off the network entirely (*Y* = null), which is an obvious absurdity. (1) can be maximized subject to the constraint $H(Y) \neq 0$ by making the feature network so selective that it can distinguish only *one* input pattern $x \in X$ while producing a null response to all other input patterns. This is a logical contradiction because for the system as a whole the quantity of information it can represent is given by its output entropy,

$$H\left(Y_1, Y_2, \cdots, Y_n\right) \leq \sum_{j=1}^{n} H\left(Y_j\right), \tag{3}$$

and with a finite number of feature networks (limited by the number of neurons contained in the structure), maximizing (1) using *static* feature-representing networks also means minimizing (3), which is contrary to minimizing (2).

The resolution of this contradiction is where the DLA concept comes to its application to feature networks. DLA calls upon two network characteristics not considered in more traditional

neural network approaches. The first is the capability of individual neurons to exhibit fast-acting short-term changes in synaptic weight (called "elastic modulation" here and "reversible synaptic plasticity by von der Malsburg [19]). Physiological considerations for modulation channels in artificial neuron models has been reviewed in [11] as well as within the references cited in [19]. The second characteristic is the capacity for an MCU to cooperate with other MCUs such that the connectivity within these MCUs is dynamically altered in response both to X and to binding codes feeding back to the MCUs from "downstream" neural structures ("convergence zone assemblies") [19]. Binding codes represent another class of signals, the control class, and they perform feature linking. Such feature linking does, however, depend upon the prior formation of specialized connectivity patterns among MCUs. Connections running between MCUs at the same neurological level can be regarded as carrying representations of what here will be called "context states." The idea of a context state is this: An MCU in a given context state exhibits a constrained maximal equivocation, denoted $H(X|Y,S)$ where S is a set of context states. This maximizes (1) in any given context state, but by having a multiplicity of "potential" context states the network *as a whole* is able to minimize (2).

Proof of this assertion is as follows. With the addition of context states the quantity to be maximized in place of (1) becomes

$$H(X|Y,S) = H(X|S) - H(Y|S) \tag{4}$$

where in the derivation of (4) we have used the equality

$$H(Y|X,S) = 0 .$$

The first term on the right-hand side of (4) is the equivocation of the input with respect to context states, and the second is the equivocation of the output with respect to context states. Two conditions are required to maximize (4). First, the set of possible outputs for any given context state must be small compared to $|X|$. Second, it must be computationally intractable to determine $x \in X$ given only the knowledge of a context state. The structure of an MCU must promote the property of a large equivocation $H(X|S)$. This is possible if the network maps a large number of inputs into the same context state, such that there are relatively few context states in an MCU, and for each context state the neuron produces distinct outputs Y as a function of only those inputs X that constitute data path inputs and not control path (S-establishing) inputs.

Given this, the next requirement is to minimize the total network equivocation

$$H\left(X\middle|\langle Y_1, S_1\rangle, \langle Y_2, S_2\rangle, \cdots, \langle Y_n, S_n\rangle\right).$$

To simplify our notation, let ζ_i denote the ordered pair $\langle Y_i, S_i\rangle$. (2) then becomes

$$H\left(X\middle|\zeta_1, \zeta_2, \cdots, \zeta_n\right) = H(X) - H(\zeta_1, \zeta_2, \cdots, \zeta_n) \tag{5}$$

and this equation is to be minimized. By the chain rule for entropies,

$$H(\zeta_1, \cdots, \zeta_n) = H(\zeta_1) + \sum_{i=2}^{n} H\left(\zeta_i\middle|\zeta_1, \cdots, \zeta_{i-1}\right) \leq \sum_{i=1}^{n} H(\zeta_i) \tag{6}$$

with equality in the right-most term if and only if each ζ_i is statistically independent of the other terms ζ_j. This is equivalent to saying that the rules by which context states and their associated output set are determined for any one MCU are independent of the rules by which these are determined in other MCUs. As independence for the setting of individual context states in different MCUs is approached, (6) approaches a maximum, upper bounded by $H(X)$, and (5) is minimized. This is where cooperation between MCUs in a DLA comes into play, but in a peculiar manner. MCU-to-MCU interactions must decorrelate the determination of context states during the formation of MCUs so as to maximize (6) in the totality of the mature DLA structure.

Equations (4)-(6) constitute three objective functions to be optimized in the network structure. Of these, (4) and (6) are to be maximized while (5) is minimized. Maximizing (4) for each feature network maximizes the selectivity of the individual MCUs. Because each MCU has a corresponding objective function (4), the optimization problem requires all equations (4) in the system plus (6) plus (5) to be optimized in regard to an overall objective function

$$J\left[H\left(X|\zeta_1\right), \cdots, H\left(X|\zeta_n\right), H\left(X|\zeta_1, \cdots, \zeta_n\right), H(\zeta_1, \cdots, \zeta_n)\right].$$

Note that initially $n = 1$ if no MCU subnetworks initially exist. Furthermore, the context states S_i develop during the course of structuring the network. Therefore the objective function J will itself undergo significant changes as the structuring process progresses. A general algorithm for optimizing J has not been presented.

Algorithmic Constraints and Assumptions. Algorithms derived from information-theoretic arguments have been slowly gaining in popularity for several years now. Entropy arguments have been used to produce a variety of algorithms for unsupervised learning based on global objective functions that these arguments provide [21]. The most popular of these approaches appear at this

time to be Independent Component Analysis (ICA) [22]-[23], the Infomax Principle (IP) [24]-[26], and algorithms based on points of intersection between ICA and IP [27], [22]. These approaches have in common the objective of maximizing the mutual information, $I(X;Y)$, between the network input and its output. A related approach, minimum redundancy encoding [28], has also been proposed which minimizes the sum of feature entropies. Another approach is the Imax learning procedure [29], which works to maximize the mutual information between *outputs* of different neural modules that receive inputs from different sensory modalities. ICA and IP share some points of similarity, especially when combined with maximum likelihood (ML) methods [22], [30]. Algorithms of this class can be called Infomax-ICA algorithms. A number of other variations also exist [31]. IP-based algorithms also, in particular cases, exhibit significant similarities to the BCM algorithm [26]. Relatively simple algorithms for implementing IP-ICA in networks using connectionist neuron models are known [22] although the application to pulse-coded neural networks is more complex [26].

The approaches and algorithms just cited all commonly make some sort of assumption regarding the probability distribution of X and on noise properties of neurons. It has been a popular practice to assume that biological neurons are intrinsically "noisy" elements in their firing characteristics. In terms of neurotransmitter release and postsynaptic membrane response there is experimental evidence that supports a stochastic neuron model in regard to synaptic behavior. On the other hand, when one considers that synchronized firing patterns seem to be the rule rather than the exception in sensory pathways, there is reason to doubt that noisy behavior at the synaptic level translates into noisy behavior at the neuronal output level. What might appear to be random behavior in the laboratory setting could just as well be due to our lack of understanding of the "neural code" as it could to "real" noisy behavior [32]. In any event, models of neural networks are deterministic functions and so it is common practice (and not biologically implausible) to assume $H(Y|X) = 0$; this assumption is employed in this paper.

At the same time, there is ample reason to regard the sensory input signals probabilistically in the analysis of neural systems. The most commonly made assumptions here include Gaussian distribution of input patterns, multi-modal Gaussian distribution of input patterns, Poisson distributions, or uniform distributions. There is little experimental basis for any of these assumptions, however. What data there is tends to suggest distributions with higher kurtosis and longer tails than are given by the Gaussian distribution [22]. These distributions have been termed "super-Gaussian" distributions. It is well known that non-Gaussian distributions are not completely characterized by first- and second-order statistics alone. A number of the algorithms cited above rely on or make arguments based on Gaussian assumptions, and that they succeed as

well as they do is attributed to the robustness conjecture [22]. In the case of pulse-coded neuron models and PCNNs, the assumption of a Poisson process for firing activity is normally invoked. This typically turns out to be too difficult to handle analytically, but it is possible to argue the conjecture that the Poisson distribution is adequately accounted for by assuming a Brownian diffusion process [26]. Even so, treatment of this case becomes substantially more complicated than is the case for non-spiking neuron models.

Most information-theoretic approaches have considered only a single network from an input-output basis. None have specifically been cast in terms of a DLA model, which means that the effect of context states and rules linking emerging MCU feature networks is as yet unexplored. The objective functions used in these earlier works are therefore significantly simpler than the J stated earlier, and the cost of this simplicity is inability to gain insight into the early development of the feature-representing MCUs and their interactions.

Another factor that must be kept in mind for Damasio's system structure is temporal sequencing. The term "object" refers to both entities and events. Entities at the first level of representation can be treated in an effectively "static" fashion merely by choice of the time interval T over which the averaging of afferent signal information is regarded as taking place. This is, for instance, an assumption invoked in the BCM model and implicit in the majority of previous information-theoretic treatments as well as in (1)-(6) above. In Damasio's model entities are put together from "type I" binding codes produced by convergence zones in the sensory association cortices [1]. Events, on the other hand, cannot be treated in this fashion. In Damasio's model events are formed by "type II" binding codes from convergence zones in motor-related cortices, and these binding codes aim to reproduce temporal sequences. The impact of this on early upstream feature networks is perhaps small because temporal sequences are themselves made up of successive entity fragments, but the issue cannot be ignored farther downstream (where convergence zones must form that generate type II binding codes) nor for the development of motor learning. Work has barely begun on compiling a comprehensive theory for temporal sequencing, although a few early steps have been made [33].

The issue of temporal sequencing has profound consequences for information-theoretic treatment of event representation. First, all the entropies employed in the objective function must now become entropy *rates* [34]. This is a significant complication. Moreover, there is utterly no reason to think that the weight-changing mechanisms of neurons have in any way the sort of statistical record-keeping mechanisms that would be required to estimate "on line" any direct measure of entropy rates. Thus, despite the mathematical attractions of "global" objective functions [33], any optimization scheme based on an information-theoretic J must be such that

the factors controlling adaptation must be *locally measurable* within some finite time frame. This is, of course, obvious.

Furthermore, the problem addressed in this work has the additional dynamic element that new MCUs and their context states are being constructed as the structure identification progresses. There are two primary considerations here. First, the objective function itself is therefore time-varying, which means that stationary-statistics arguments may not be invoked. All previous information-theoretic treatments implicitly invoke such arguments when they estimate entropies through time averages. Second, the fine-tuning process of sensory networks is contingent upon sensory experience, and this has implications for the strategy to be employed in accordance with Bellman's Principle.

It is generally accepted that Bellman's Principle implicates a strategy of optimization working "backwards" from goal to present situation. What, however, can constitute such a "goal"? In unsupervised learning the system has no "training exemplar" (object as entity) upon with to base an error-signaling process. The system must therefore build its own model (i.e. form objective feature-fragment representations) upon the basis of some other measure of the "quality" of its representations [21]. In the *ideal* this measure is given by the entropy functions (4)-(6), but in the presence of contingency in the evolution of the network structure these entropies are not *practically* observable. It follows from this that J should be a function based not upon the entropies themselves but rather on *properties* of entropy that are necessary for the possibility of a minimum in (5) and a maximum in (4) and (6). In addition, J must be based upon observables that can be plausibly connected with both the aforementioned properties and with the available neuronal mechanisms of synaptic plasticity. The proposal presented here is that the appropriate observables are related to possible information *flow* within the system.

Graphical Analysis of Neural Network Information Flow. A new type of metric for the evaluation of information flow within decentralized systems that must adapt their structure to accommodate environmental conditions was recently represented by Akuzawa and Ohnishi [35]. The central idea is that information concentration and distribution within a network can be evaluated by eigenvector analysis of the network graph. This "design indices" method was developed for application to decentralized robotic systems rather than for the neural network problem. One shortcoming of the method as it was presented is that it is applicable only to systems that can be represented by strongly-connected graphs [36]. The issue here is that nodes representing system inputs to a neural network cannot themselves be reached from neural nodes within the graph and therefore the system graph is not strongly connected. The second

shortcoming is that the analysis does not account for differences in the entropies of different input pathways or for possible redundancies contained in the set of inputs. The first shortcoming is easily overcome; overcoming the second requires an extension of the idea of the connection matrix of a graph.

A. Forcing Connectedness by Virtual Feedback. First it is shown that by a simple extension the Akuzawa-Ohnishi analysis (AOA) becomes applicable to the problem of structure identification of MCUs. Let V be a set of $m+n+1$ vertices where m is the number of system inputs and n is the number of neurons in the network. Let A be a set of arcs connecting the vertices and let A be described by a connection function ϕ. Let $V_M \subset V$ be a set of vertices v_i, $i = 1, 2, ..., m$ designating the input vertices, and let $V_N \subset V$ be a set of vertices v_j, $j = m + 1, ..., m+n$, designating the neurons in the network. Let $v_{m+n+1} \in V$ be a special vertex called the *virtual feedback vertex*. Let $G = [V, A, \phi]$ denote a directed multigraph that describes the system. Let T denote the connection matrix of the system with $t_{i,j} \in T$ denoting the connection (or its absence) from donator vertex j to receptor vertex i. Let B be an $n \times m$ submatrix of T denoting the connections from V_M to V_N. Let C be an $n \times n$ submatrix of T denoting connections from V_N to V_N. B and C are called the input distribution and network distribution matrices, respectively. Let F be an $m \times 1$ vector denoting connections from v_{m+n+1} to V_M. F is called the virtual feedback vector. We further assume there are no arcs connecting from V_N to V_M, i.e. that the inputs do not receive direct feedback from the neural network, and that no input directly connects to any other input. Finally, let every $v_j \in V_N$ connect unidirectionally to vertex v_{m+n+1} and assume the virtual feedback vertex has no loop. The connection matrix is then given by

$$T = \left[\begin{array}{ccc|c} & 0 & & 0 & F \\ \hline & B & & C & \begin{matrix}\vdots\\0\\\vdots\end{matrix} \\ \hline \cdots & 0 & \cdots & \cdots\ 1\ \cdots & 0 \end{array}\right] . \qquad (7)$$

Without loss of generality every column of B contains at least one non-zero element (otherwise the corresponding input connects to nothing). Likewise, every row of C contains at least one non-zero element (otherwise the corresponding neuron is not part of the network). T is a

non-negative matrix, and if every element of F is non-zero the graph is easily shown to be strongly connected. In this case T has a unique, real, and positive largest eigenvalue, λ , with corresponding positive normal eigenvector R by the Perron-Frobenius Theorem [35]. Akuzawa and Ohnishi demonstrated that under these conditions λ is a measure of the degree of information *concentration* in the system and the elements of R describe how this information is *distributed* within the system. Generally, the more input paths a vertex has, the larger will be its corresponding element of R with respect to the other vertices in the graph. Thus, its element, r_k, is a relative index of how much information is concentrated at that particular vertex.

As presented, AOA is non-rigorous with regard to the concept of "information" in the system. The assumptions implicit in the interpretation of λ and R just cited are discussed below. Before discussing them the concept of virtual feedback in the network must be discussed. Let F be a non-zero vector with one zero element f_i in row i. For any eigenvalue of T and its corresponding eigenvector we have

$$T R = \lambda R . \tag{8}$$

For the i^{th} row direct multiplication gives us $0 \cdot r_i + f_i \cdot r_{m+n+1} = \lambda \cdot r_i$, and for $f_i = 0$, $r_i = 0$. Therefore, any input vertex not receiving feedback from virtual feedback vertex v_{m+n+1} will have zero for its corresponding eigenvector element and that input's corresponding column in B makes no contribution to R in (8). G is no longer strongly connected in this case, but a strongly connected reduced graph G' is easily obtained by striking out the row and column for each input vertex for which $f_i = 0$ in F. For G' the conditions of the Perron-Frobenius Theorem again apply provided that F is not an all-zero vector, and a unique maximum positive eigenvalue and corresponding eigenvector are obtained. If we now *augment* this eigenvector by adding new rows corresponding to the previously-eliminated inputs and inserting zeroes into those rows, (8) is still satisfied and we retain unique values for λ and for R. But since the solution of (8) is unique (up to an arbitrary multiplicative constant for R), the eigenvalue and eigenvector obtained through this augmentation are the same as would have been obtained by direct solution to (8). All this is summarized in the following theorem.

Theorem 1: For connection matrix T as given in (7), the results of Akuzawa-Ohnishi analysis are unchanged if and only if F contains at least one non-zero element. Furthermore, G has a unique homomorphic image in some lower-order, irreducible, and strongly-connected graph G'.

B. Information Analysis Extension of AOA. The AOA terms "information concentration" and "information distribution" are not terms sanctioned by formal definitions in information theory. Consequently a direct connection between AOA and entropy measures is not established by [35]. For AOA to be justified for information-theoretic optimization neural network structure, this situation must be clarified. For normalized eigenvector $R = [r_1\ r_2\ \ldots\ r_{m+n+1}]^T$, AOA defines the "index of the degree of information concentration in the system" as

$$Q \overset{\Delta}{=} \lambda \sum_{k=1}^{m+n+1} (r_k)^2 = \lambda \tag{9a}$$

and the "index of degree of information concentration in the k^{th} element of the graph" as $q_k = r_k$. Neither of these terms specify quantity of information; quantity of information is measured by entropy and these indices are not entropies. AOA index q_k provides a measure of what we might term the "information potential" of a vertex, that is, an indication of the fraction of the total information capacity in the system that is potentially available at vertex v_k merely by virtue of the nature of connectivity of the graph. It takes into account no measure of source entropy, no measure of information loss in passing through a vertex, and no measure of redundancy among the data scattered throughout the vertices of the system. Formally,

$$q_k = \frac{1}{\lambda} \sum_{j=1}^{m+n+1} t_{k,j} \cdot r_j \tag{9b}$$

which can be regarded as the fraction of Q distributable from v_k in accordance with the connection function ϕ of the graph.

As for Q, an understanding of what this index implies is based on appreciating that AOA indices are defined on strongly-connected graphs regarded as *closed* systems. Such a system has no "input." The vertices of the graph represent subsystems within a distributed system, and these subsystems are regarded not only as information processors but also as information sources with memory. Because the output of a vertex depends as much on the internal state of the processor it represents as it does on information it receives from other vertices, the network is not merely an information relay network. The information in an output message from a donator vertex also conveys state information to the other processors (the "receptors") to which that vertex is connected. This state information may (or may not) alter the internal state of the receptor

processors, depending on what their internal states may be, what "rules" they follow for processing signals from their connected donators, and what other information they are receiving from elsewhere in the system.

This AOA idea of "information flow within the system" is a generalization of the same idea that underlies the graph-theoretic treatment of block data translation codes, which is a well-known application of information theory in communications [34], [37]. Block codes are regarded by information theory as noiseless channels, and by a well-known theorem [34] the capacity in bits of such a "channel" is known to be $\log_2(\lambda)$, where λ is the largest positive eigenvalue of the connection matrix that describes permitted code sequences. Now, AOA networks are not data translation codes, but the AOA index $Q = \lambda$ is a measure that is in this same sense monotonically related to the information capacity of the network. From this it follows that the indices q_k are in the same sense relative measures of how this capacity is distributed within the network.

This interpretation of what the AOA indices convey leads to the following formal extension of the AOA approach for its application to neural networks. Define a *rule* as any assertion made under a set of conditions. Define a *decision* as the asserting of a rule. The synaptic input signals to a neuron constitute a condition, and the response of the neuron to this condition constitutes an assertion. Thus, a neuron can be formally regarded as a processor for rule evaluation and decision making. A neuron P with t synaptic inputs, each carrying signal σ_i , $i = 1, ..., t$, is presented with the condition $(\sigma_1, \sigma_2, \ldots, \sigma_t)$ and asserts a rule as an output signal σ_o. Using formal notation, $P(\sigma_1, \sigma_2, \cdots, \sigma_t) \mapsto \sigma_o$. If we let Σ denote the set of possible neuronal signals then the *rule structure* of a neuron can be regarded as a mapping

$$P : \underbrace{\Sigma \times \Sigma \times \cdots \Sigma}_{t-tuple} \to \Sigma . \tag{10}$$

In general, every neuron is capable of producing more than one output response, and every neuron providing it with inputs is capable of providing more than one σ_i . (10) is merely an input-output mapping, and P will depend upon the internal state of the neuron. It is possible for the output entropy of P to be greater than the entropy of its input signals because the output can be said to contain information about the internal state of P. (This is not an unusual situation in information theory; the property of dependence on internal state in the entropy of an output is what makes cipher systems work [34]). Conversely, it is also possible for the output entropy to be less than the entropy of the input set, and this can happen whenever P maps many input sets to the same output.

In [35] the elements in T were constrained to be either 0 or 1. Doing this leads to an under-estimation of Q in the system because it fails to account for different possible signals σ_o available from each neuron as well as for the significance that a particular output presents to another neuron as part of that neuron's Σ. This limitation can be overcome by introducing *parallel arcs* into the graph structure. In the conceptually simplest case if there are χ set-distinguishable signals that a donator j can transmit to a receptor i, this can be represented by setting $t_{i,j} = \chi$ in T. This formal mathematical trick is analogous to introducing parallel trellis paths in describing trellis-coded modulation systems [38]. Two outputs, σ_1 and σ_2, are set-distinguishable if the response of the receptor neuron can be different depending on which signal it receives. In this sense, a receptor neuron can be regarded as a "set-distinguishing receiver" or SDR [39]. Because χ is always non-negative, the conditions of the Perron-Frobenius Theorem are still met by T, only now $\log_2(\lambda)$ is a truer measure of the information capacity of the network at any particular time. Note, however, that as the network adapts, the mappings P change and therefore the elements of T will in general also change. Thus, Q and the q_k of the network evolve in time along with the structure of the neural network itself.

Virtual Feedback Connections. The virtual feedback vector F and the virtual node v_{m+n+1} in (7) are mathematical artifacts by which the inputs to the neural network are brought under the required conditions for analysis by AOA. How these components of the graph are to be interpreted is next examined.

The first interpretation to be made is what is represented by the eigenvector component r_{m+n+1}. Vertex v_{m+n+1} is not a neuron, yet r_{m+n+1} affects the other r_i terms in the normalized eigenvector R. If the terms in R are to be interpreted as a measure of the distribution of information capacity in the system yet v_{m+n+1} represents neither a neuron nor an input, it is clearly incorrect to regard the term r_{m+n+1} in the same way as the other eigenvector components are regarded. Now,

$$r_{m+n+1} = \frac{1}{\lambda} \sum_{k=m+1}^{m+n} r_k \ . \tag{11a}$$

Furthermore, for $k = 1, ..., m$ we have

$$r_k = f_k \cdot \frac{r_{m+n+1}}{\lambda} = \frac{f_k}{\lambda^2} \cdot \sum_{j=m+1}^{m+n} r_j \ . \tag{11b}$$

Substituting these expressions in (9a) gives

$$Q = \lambda \sum_{k=m+1}^{m+n} (r_k)^2 + \frac{1}{\lambda} \cdot \left(1 + \frac{1}{\lambda^2} \cdot \sum_{j=1}^{m} (f_j)^2 \right) \cdot \left(\sum_{k=m+1}^{m+n} r_k \right)^2 . \tag{11c}$$

The first term on the right-hand side of (11c) is the total contribution to Q made by the neurons. The second term contains the contributions made by the combination of v_{m+n+1} and the input signal pathways. Vertex v_{m+n+1} in one sense "buffers" the input vertices from the neuron vertices in the graph inasmuch as the feedback to the input vertices is funneled through it. In terms of λ this is a kind of "bottlenecking" of information in the graph. It is interesting to note that (7) is not equivalent to a connection matrix of the form

$$\widetilde{T} = \left[\begin{array}{c|ccc} 0 & F & \cdots & F \\ \hline B & & C & \end{array}\right] \tag{12}$$

even though it might seem at first glance that (12) provides exactly the same feedback connectivity as (7). In general the maximum eigenvalue $\widetilde{\lambda}$ of $\widetilde{T}$ will be larger than λ, and $\widetilde{Q}$ for (12) will be larger than Q. The situation here is analogous to the effect of the minimum run length constraint on the capacity of run-length-limited codes where increasing the minimum run length decreases the capacity of the code by restricting the number of possible code sequences in a block [34], [40]. Nodes in the code graph that enforce the minimum run length constraint restrict the branching that is possible within the graph, and v_{m+n+1} plays a similar role here. The AOA capacity index for (12) is larger than for (7), and since (12) is indistinguishable from a multi-layer recurrent network with output layer feedback to the input layer, we learn from this comparison that information capacity in a recurrent network is decreased by "bottleneck neurons" forming a connection matrix of the form of (7) *ceteris paribus*.

Next we note that for the special case where F is an all-zero vector (12) is homomorphic to a connection matrix $\widetilde{H} = C$ whereas (7) is homomorphic to

$$H = \left[\begin{array}{ccc|c} & & & \vdots \\ & C & & 0 \\ & & & \vdots \\ \hline \cdots & 1 & \cdots & 0 \end{array}\right].$$

What is interesting here is that although they are quite different in terms of their eigenvectors, $\tilde{H}$ and H have identical maximum positive eigenvalues and therefore the same AOA capacity index. This is easily seen from

$$\left|\lambda I - H\right| = \lambda \cdot \left|\lambda I - C\right| = \lambda \cdot \left|\lambda I - \tilde{H}\right|.$$

Vertex v_{m+n+1} is not strongly connected in H (the other vertices cannot be reached from v_{m+n+1}) and therefore it does not make an independent contribution to the AOA capacity index in H. This provides us with a clue for how to interpret $q_{m+n+1} = r_{m+n+1}$ in the system described by (7). First, its lack of contribution to the AOA capacity index in the homomorphic image H suggests that its contribution $\lambda \cdot (r_{m+n+1})^2$ to Q in (11c) is to be regarded as a network average rather than an input pathway contribution. This part of its character is given emphasis by re-writing (11c) as

$$Q = \lambda \cdot \left[\sum_{k=m+1}^{m+n} (r_k)^2 + \frac{1}{\lambda^2} \cdot \left(\sum_{k=m+1}^{m+n} r_k\right)^2\right] + \frac{1}{\lambda^3} \cdot \left(\sum_{j=1}^{m} (f_j)^2\right) \cdot \left(\sum_{k=m+1}^{m+n} r_k\right)^2. \qquad (13)$$

Second, if we regard the vertex indices as being abstract representatives of information flow in a system stimulated by stochastic inputs, the form of v_{m+n+1}'s contribution to Q in (13) has more of the character of being related to a squared-mean rather than a mean-squared, whereas the leftmost term in the first bracketed term in (13) suggests more of the character of being related to a mean-squared (i.e., related to a covariance). This favors the idea of interpreting v_{m+n+1} as being representative of correlated behavior in the network whereas the terms $(r_k)^2$ would be more representative of the independent actions of individual neurons in their mapping functions P. Because v_{m+n+1} directly drives the input vertices of the graph, its interpretation in terms of squared-mean network activity implies that it is more closely related to (5) compared to the other terms in (13). This interpretation is favored by the observation that as the connectivity in C is reduced r_{m+n+1} tends to increase relative to the other r_k terms for a given B and non-zero F. This

increase is accompanied by a reduction in λ, which denotes a general decrease in the information capacity of the network. Both these effects would be expected in networks that were more selective in their responses to inputs.

Now consider the virtual feedback terms, represented by the right-most term on the right-hand side of (13). Referring again to (7), the terms in F distribute the information flow from v_{m+n+1} to the input vertices and B distributes from these to the neuronal vertices. The roles of F and B are not generally interchangeable in this distribution just as T and $\widetilde{T}$ were not generally interchangeable earlier. B represents the distribution of input signals into the neural network while F merely assigns some fraction f_k of the network activity measured by r_{m+n+1} to each input vertex according to (11b). Because v_{m+n+1} makes no discrimination among the neurons in the network, this virtual feedback is non-specific with regard to individual neurons and therefore the differences among the $q_i = r_i$, $i = 1, ..., m$, can be regarded as indices of differences in the relative information rates of the m inputs. F thus serves as an abstracted input activity pattern.

Typically the virtual feedback contribution to Q is smaller than the network contribution at lower values of the f_k terms in F. Multiplicative increases in F tend to raise the value of λ, although the increase is proportionally less than the increase in the terms of F, and increases the r_i input indices, principally at the expense of r_{m+n+1} which decreases significantly in value. The neuron indices r_k, $k = m + 1, ..., m + n$, tend to decrease as well, although not so dramatically as r_{m+n+1}, and the distribution among these values tends to become more uniform, although they maintain their relative ranking as F is uniformly increased. As the modulus of F increases a point is reached where the input indices exceed the neuron vertex indices, a situation that can be interpreted as saturation of the neural network's information-handling capability. This behavior is consistent with what is expected to occur as the rate of incoming information exceeds the information capacity of the network.

Information Weights and Entropy Index. The cardinality of the variety of responses a neuron can have to its various inputs is a measure of its information capacity, and its index q_k is a relative measure of its distinctive contribution to the overall network's information capacity. This index is given by (9b), and the $t_{k,j}$ terms in (9b) give the "connection weight" of the input pathways in the sense that each $t_{k,j}$ is a measure of the number of *significant* signals affecting neuron k from source vertex j. For neuron k we re-write (10) as

$$P_k : \Sigma_{k,1} \times \Sigma_{k,2} \times \cdots \times \Sigma_{k,m+n} \to \Sigma_k$$

and consider the $\Sigma_{k,j}$ terms in this expression. Each term represents a significant input message $\sigma_{k,j}$ and so $P_k(\sigma_{k,1},\ \sigma_{k,2},\ ...,\ \sigma_{k,m+n}) \mapsto \sigma \in \Sigma_k$ is the simplest general expression for the mapping performed by neuron k on its input messages. More generally, however, neuron k is also characterized by some internal state, $u_k \in U_k$, where U_k denotes a set of possible neuron states. (Such would be the case, e.g., in an integrate-and-fire neuron model, a BCM model or an Eckhorn neuron model). Therefore, the general model extends (10) to the coupled set of dynamical functions

$$\begin{aligned} P_k &: \Sigma_{k,1} \times \Sigma_{k,2} \times \cdots \times \Sigma_{k,m+n} \times U_k \to \Sigma_k \\ Z_k &: \Sigma_{k,1} \times \Sigma_{k,2} \times \cdots \times \Sigma_{k,m+n} \times U_k \to U_k \end{aligned} \tag{14}$$

where Z_k is the mapping function for neuron k's state transition function. The specific functions in (14) depend on the properties of the neuron model being used, and what concerns us at the more abstract level of AOA analysis is the cardinality of each of the sets $\Sigma_{k,j}$ in (14).

An input $\sigma_{k,j}$ can be significant in any of three ways: 1) it can affect the neuron's output σ, 2) it can affect the neuron's internal state u_k; or 3) it can affect both. Furthermore, whether or not a particular $\sigma_{k,j}$ is significant at any given time can depend on the other coincident input signals and the neuron's state at that time. Adding to this complication is the presence of recurrent connections in the network since feedback from neurons for which neuron k is a donator will generally set up a transient response in the network as a whole, which makes a crisp definition of a signal $\sigma_{k,j}$ at any particular moment in time somewhat problematic. A formal method is needed in order to resolve this ambiguity and permit evaluation of (14) and determination of the cardinalities of its various sets.

In establishing such a formal method the following definitions are useful.

<u>Def. 1</u>: A *closed cycle of activity* is any periodic activity pattern.

<u>Def. 2</u>: An *innovation* is any change from a closed cycle of activity to any other activity pattern.

<u>Def. 3</u>: A neuron is in *equilibrium* if its output activity is a closed cycle of activity.

<u>Def. 4</u>: An *assembly* is any defined set of neurons whose connectivity is described by a strongly-connected graph that does not include any vertices not corresponding to the members of the defined set of neurons.

Def. 5: An assembly is in equilibrium if all its member neurons are in equilibrium.

The utility of these definitions is owed to the following property of entropy rates. Let a_t be a symbol that describes a neuron's firing activity over some interval of time characterized by index t. For example, a_t might represent the elapsed time between an action potential and the neuron's previous action potential. Let the output activity of the neuron be represented by the sequence of symbols a_0, a_1, ..., a_{t-1}. The entropy rate of this sequence is defined as

$$h = \lim_{t\to\infty} \frac{H(a_0, a_1, \cdots, a_{t-1})}{t}$$

and by the chain rule for entropy this expression evaluates as

$$h = \lim_{t\to\infty} \frac{H(a_0) + \sum_{i=1}^{t} H(a_i | a_0, \cdots, a_{i-1})}{t}.$$

Now suppose that beyond some index $t = t_r$ the symbol sequence establishes a closed cycle. From this point on, the conditional entropies in the expression for h are zero and the numerator is finite. Therefore $h \to 0$ for any activity pattern that eventually forms a closed cycle of activity. From this we have the following lemma.

Lemma 1: The entropy rate of any neuron in equilibrium and any assembly in equilibrium is zero.

We may now define a time-limited message as any neuronal firing activity whose entropy rate is zero. Likewise, a time-limited input message set is any set $\{\sigma_{k,1}, \sigma_{k,2}, ..., \sigma_{k,m+n}\}$ whose *joint* entropy rate is zero. The donator set of a neuron can therefore be defined as a subset

$$D_k \subseteq \Sigma_{k,1} \times \Sigma_{k,2} \times \cdots \times \Sigma_{k,m+n} \tag{15}$$

such that every element of D_k is a time-limited input message set whose members are significant time-limited messages. Let $\chi_{k,j}$ represented the number of set-distinguishable signals that donator

j can transmit to neuron *k*. An upper bound on the cardinality of (15) is achieved if every set-distinguishable signal is also a time-limited input message, and therefore

$$\left|D_k\right| \le \prod_{j=1}^{m+n} \max\left(\chi_{k,j}, 1\right). \tag{16}$$

provided that neuron *k* has at least one donator. Otherwise $\left|D_k\right| = 0$ and cannot belong to any assembly.

The entropy $H(X_k)$ of the inputs to neuron *k* is upper bounded by

$$H\left(X_k\right) \le \log_2 \left|D_k\right|$$

with equality if and only if every element of Σ_k is equally probable. Applying (16) to this expression gives

$$H\left(X_k\right) \le \sum_{j=1}^{m+n} \log_2 \left[\max\left(\chi_{k,j}, 1\right)\right].$$

For each vertex r_j is a normalized index of information distributable from vertex v_j. It is therefore reasonable to assume that $\chi_{k,j}$ is proportional to the product of r_j and $t_{k,j}$. Let α a the constant of proportionality so that $\chi_{k,j} = \alpha \cdot t_{k,j} \cdot r_j$. Let d_k be the number of donator arcs terminating on neuron *k* such that $t_{k,j} \cdot r_j \neq 0$ and let $\{d_k\}$ denote the set of these arcs. Then

$$H\left(X_k\right) \le \sum_{\substack{j=1 \\ j \in \{d_k\}}}^{m+n} \log_2 \left(\alpha \cdot t_{k,j} \cdot r_j\right) = \sum_{\substack{j=1 \\ j \in \{d_k\}}}^{m+n} \log_2 \left(w_{k,j} \cdot r_j\right) \tag{17}$$

where $w_{k,j} = \alpha \cdot t_{k,j}$ will be called the information weight of arc (k, j).

Since the input entropy can never be negative but all r_j terms lie in the range $0 \le r_j \le 1$, (17) restricts the permissible values of the $w_{k,j}$ terms to those for which the sum of the logarithms in (17) is non-negative. (17) is therefore a constraint on the structure of the graph. Furthermore, since α has the interpretation given above, its value is constrained by the requirement that (17) must be satisfied at every v_k, $k \in [m+1, m+n+1]$, in the graph. Furthermore, since $t_{m+n+1,j} = 1$ for

all $j \in [m+1, m+n]$, $\alpha \geq 1/r_j$ for at least some subset of j in this range such that the constraint on the sum of logarithms in (17) is satisfied at $k = m+n+1$ as well. One consequence of these constraints is that any adaptation algorithm operating on the graph G will automatically be able to produce a minimum bound on α because these constraints must be satisfied at every step in the adaptation. It is therefore possible to establish a link, albeit only in terms of bounds, between the AOA indices and the entropies in the system. For this reason, α can be regarded as an entropy index for the system.

Neuron Equivocation. We can also examine the consequences of (14) for the equivocation properties of a neuron. Since the output set Σ_k of neuron k is a function of its input set and state space, represented as the ordered pair of sets $\langle D_k, U_k \rangle$, it is meaningful to look at the mutual information

$$I\left(\langle D_k, U_k \rangle, \Sigma_k\right) = H\left(D_k\right) - H\left(D_k \middle| \Sigma_k\right) + H\left(U_k \middle| D_k\right) - H\left(U_k \middle| D_k, \Sigma_k\right).$$

Because mutual information is always non-negative, the upper bound on neuron k's equivocation is therefore

$$H\left(D_k \middle| \Sigma_k\right) \leq H\left(D_k\right) + H\left(U_k \middle| D_k\right) - H\left(U_k \middle| D_k, \Sigma_k\right). \tag{18}$$

$H(D_k) = H(X_k)$ and is upper-bounded by (17). The remaining two terms on the right-hand side of (18) depend on the properties of neuron k.

In system theory, a system is said to be *observable* if the state of the system, U_k, can be determined from observations of its inputs and outputs. Otherwise the system is said to be unobservable [41]. From this definition we have

Lemma 2: Neuron k is observable if and only if $H\left(U_k \middle| D_k, \Sigma_k\right) = 0$.

Complete observability means that every state variable making up the system's state can be uniquely determined from observation of the system's inputs and outputs. Unobservability does not necessarily mean that nothing can be known of the state variables of the system; it merely implies that *all* the individual state variables cannot be determined. There can therefore be *degrees* of unobservability according to what fraction of the state variables within the system

state are unobservable. $H(U_k|D_k,\Sigma_k)$ measures the degree of unobservability. In system engineering, and particularly in control system engineering, it is often the case that "observers" form part of the controlling mechanism of the system. This raises the interesting speculation that some neurons in a neural network might constitute an observer function within the network, particularly if the function of that network is served by maximizing the equivocation (18).

In system theory a system is said to be *controllable* if the state of the system can be set through the application of particular input signals [41]. Otherwise it is said to be uncontrollable. The term $H(U_k|D_k)$ is a measure of the uncertainty in state U_k given inputs D_k. $H(U_k|D_k) = 0$ implies that the neuron's state is completely and uniquely determinable by the inputs regardless of the initial state of the neuron. From this we have

Lemma 3: Neuron k is controllable if and only if $H(U_k|D_k) = 0$.

Complete controllability means that every state variable in the system state can be uniquely set independently of the other state variables. Here, too, uncontrollability is a matter of degree. If two or more state variables cannot be set independently of each other, while others can be so set, the system is only partially uncontrollable. $H(U_k|D_k)$ is therefore a measure of uncontrollability for neuron k. In system engineering, and particularly in control system engineering, a "controller" forms part of the mechanism of the system, and again we have the interesting speculation that some neurons in a network might constitute a controller function.

The equivocation for neuron k, given a fixed set of inputs, is maximized by maximizing the difference $H(U_k|D_k) - H(U_k|D_k,\Sigma_k)$. If neuron k is part of a cell assembly, (18) is easily extended to describe the cell assembly merely by defining the state of the assembly in terms of the states of all its member neurons and restricting the assembly's donator set to only those inputs coming into the assembly from without. If that cell assembly is a feature-representing network (a feature MCU), its selectivity is maximized by maximizing the assembly's equivocation, as discussed earlier. Formally, let assembly A be a set of neurons denoted by the set K of vertex numbers. Let U_A denote the state of the assembly,

$$U_A = \bigcup_{k \in K} U_k \ .$$

Let D_A denote the assembly's inputs

$$D_A \subset \bigcup_{k \in K} D_k$$

and let Σ_A denote the assembly's outputs

$$\Sigma_A \subset \bigcup_{k \in K} \Sigma_k \ .$$

The equivocation of the assembly is then given from (18) as

$$H\left(D_A \middle| \Sigma_A\right) \leq H\left(D_A\right) + H\left(U_A \middle| D_A\right) - H\left(U_A \middle| D_A, \Sigma_A\right). \tag{19}$$

The task of any adaptation algorithm for producing a feature-detecting assembly is the maximization of (19) through restrictions on D_A and Σ_A and formation of an intra-assembly connectivity (i.e. a graph G_A). It follows from what was said above that this process quite probably involves the specialization of some neurons in the assembly to "controller" and "observer" tasks within the assembly. Furthermore, it can be expected that some elements of D_A, entering the assembly from other assemblies, will constitute a representation of context states S as discussed previously.

Seen in this way, e.g. as the problem of maximizing (19) for each assembly, this description of the adaptation task suggests an approach for dealing with one obvious computational issue that attends applying AOA to the DLA structure identification problem. That issue is the computational complexity of evaluating the eigenvalues and eigenvectors of a large graph G. But because AOA indices are always *relative* rather than *absolute* indices of information, the task represented by (19) suggests that a large neural network graph can be regarded at a higher level of abstraction in terms of an equivalent *assembly graph* in which the vertices represent cell assemblies and input *tracts* rather than individual neurons and data inputs. The first phase of adaptation then involves establishing inter-assembly connectivities and tract routing, to be followed in the next phase by treatment of individual large assemblies in terms of their constitutive smaller assemblies. The basic idea here is one of *top-down decomposition* of the overall problem, and is made possible by the relative nature of AOA indices. To realize this ability, the adaptation algorithm used must produce results in lower-level network graphs that provide for inter-assembly connectivity at the next higher level of analysis.

Adaptation of the Connection Matrix. By introducing the idea that elements $t_{i,j}$ in B and C are an abstract representation of the number of donator messages that are significant to the receptor, we are no longer bound to regard the $t_{i,j}$ as integers (although they must remain non-negative). In the original AOA concept, the design indices λ and r_k were to be used to design dynamic soft re-linking of decentralized systems. However, how this is to be done was not yet established in [35].

In the context of feature-detecting network formation there are two properties of the AOA indices that reflect whether an adaptation of the B and C submatrices of T is serving the objective functions stated earlier. First, because $Q = \lambda$ is a global measure of network capacity (for a given F vector), adaptation should maintain or increase Q. Increasing Q corresponds to increasing the information capacity of the network by elimination of information-lossy interconnects. Second, the adaptation should decrease r_{m+n+1} (again for a fixed F). This is because this index represents redundancy within the network, and therefore decreasing it while increasing Q implies that regions of the network are becoming specialized in their response to input patterns F. Preliminary numerical experiments indicate that both these properties, as well as the production of inter-assembly context links, are satisfied by the adaptation method presented in this section.

The squared eigenvector element $(r_k)^2$ is the fraction of the total information index Q concentrated at vertex k for a given input pattern defined by F. The virtual term $(r_{m+n+1})^2$ is a non-specific global index describing the network as a whole, and we have

$$1 = \sum_{k=1}^{m+n} (r_k)^2 + r_{m+n+1}^2 .$$

Therefore the relative local information contribution to the *network distribution* by vertex k, $k < m+n+1$, can be described as

$$\rho_k^2 \stackrel{\Delta}{=} \frac{r_k^2}{1 - r_{m+n+1}^2} .$$

The square root of the denominator of this term fills the role of a normalizing factor during changes to the network connectivity or changes in the input distribution F.

Now, regardless of how many messages a donator can emit that might be significant for a particular receptor, a particular input path or neuron output emits one firing pattern at a time. Network adaptation must consider how the information carried in this signal is distributed through the network. Define the *direct distribution matrix* as the $n \times m+n$ matrix

$$D \overset{\Delta}{=} [B \mid C] \begin{bmatrix} r_1 & & & \\ & r_2 & & 0 \\ & & \ddots & \\ & 0 & & r_{m+n} \end{bmatrix} \cdot \frac{1}{\lambda \cdot \sqrt{1 - r_{m+n+1}^2}} \overset{\Delta}{=} \begin{bmatrix} \frac{D_1}{D_2} \\ \vdots \\ \frac{}{D_n} \end{bmatrix}. \tag{20a}$$

The elements of D are AOA indices of how donator information is distributed among that donator's receptors. The columns of D correspond to the individual donators; the rows of D correspond to the individual receptors. Columns 1 through m correspond to input donators, and the remaining columns correspond to neurons in the network. These indices are abstract measures of the distribution of donator information among the receptors. Matrix $[B \mid C]$ is, of course, the submatrix of T consisting of its rows m+1 to m+n and columns 1 to m+n.

Under Hebb's hypothesis, adaptation of synapses depends on the product of input activity and the responding activity of the postsynaptic cell. Similarly, AOA indices are defined by the matrix

$$\delta \overset{\Delta}{=} \begin{bmatrix} \frac{r_{m+1} \cdot D_1}{} \\ \frac{r_{m+2} \cdot D_2}{} \\ \vdots \\ \frac{}{r_{m+n} \cdot D_n} \end{bmatrix}. \tag{20b}$$

Additionally, Hebbian-like adaptation requires some threshold function that determines if the connectivity will be increased or decreased. In our present context, this threshold corresponds to a determination of the degree to which a donator can be said to be "participating" in the responses of its receptors. We make the hypothesis that a donator j is "participating" with receptor i when its distribution element $\delta_{i,j}$ in (20b) exceeds a weighted row average taken over some subset of donators, the weighting being a function of the corresponding elements $t_{i,j}$ in a subset $j_z \in \{j_0, \cdots, j_f\}$ of row i. We define sub-row averages for each such partition of row i as

$$\overline{\delta}(i, z) \overset{\Delta}{=} \frac{\sum_{j \in j_z} \delta_{i,j}}{\sum_{j \in j_z} \Theta\left(t_{i+m,j}\right)} \tag{20c}$$

where z is called the "zone number." $\Theta(x) = 1$ for $x > 0$ and 0 otherwise. The set of indices j_z defines a "competitive zone." This concept is discussed below. Each receptor has at least one competitive zone, and the union of all a receptor's competitive zones covers all its donators. From the competitive zone definitions we obtain from the set of (20c) a threshold matrix

$$\overline{\overline{\delta}} \overset{\Delta}{=} \begin{bmatrix} \overline{\overline{\delta}}(1, z_{1,1}) & \overline{\overline{\delta}}(1, z_{1,2}) & \cdots & \overline{\overline{\delta}}(1, z_{1,m+n}) \\ \overline{\overline{\delta}}(2, z_{2,1}) & \overline{\overline{\delta}}(2, z_{2,2}) & \cdots & \overline{\overline{\delta}}(2, z_{2,m+n}) \\ \vdots & & & \\ \overline{\overline{\delta}}(n, z_{n,1}) & \overline{\overline{\delta}}(n, z_{n,2}) & \cdots & \overline{\overline{\delta}}(n, z_{n,m+n}) \end{bmatrix}. \tag{20d}$$

Each $z_{i,j}$ in (20d) specifies a particular competitive zone of the sub-row average. Competitive zones are defined row-by-row for each individual receptor and (20c) is computed for each zone.

The connection update within each competitive zone is provided by some function, called the adaptation rule, g,

$$\Delta t_{i+m,j} = g\left(\delta_{i,j} - \overline{\delta}_{i,j}\right), \quad i = 1, \cdots, n;\ j = 1, \cdots, m+n \ . \tag{21a}$$

The corresponding new element in T becomes

$$t_{i+m,j} \leftarrow \max\left(0,\ t_{i+m,j} + \eta \cdot \Delta t_{i+m,j}\right) \tag{21b}$$

where η is an adaptation rate constant.

The adaptation rule is governed by the following considerations: 1) if $\delta_{i,j} >> \overline{\delta}_{i,j}$ this implies that donator v_j is providing relatively high message traffic to receptor v_i in comparison to its other donators. However, one information tract within a neural network is typically not sufficient by itself to provide a level of excitation to the receptor consistent with firing responses or the induction of plastic changes in the synaptic connections. Therefore, the number of messages that actually are in participation with messages from other donators is going to be merely a fraction of the traffic from v_j. Accordingly, Δt should be kept relatively small; 2) if $\delta_{i,j} << \overline{\delta}_{i,j}$ this implies that v_j is either not very active or that most of its messages are not participating at receptor v_i. Its connection weight should therefore be reduced (under Hebb's hypothesis). However, if a large Δt is permitted, this is tantamount to assuming that either v_j makes many synaptic connections to v_i

or that it is providing many significant messages to the receptor. Neither of these suppositions is consistent with the condition $\delta_{i,j} << \bar{\delta}_{i,j}$. Therefore, Δt should again be kept relatively small; 3) if $\delta_{i,j} = \bar{\delta}_{i,j}$, on the average v_j will be a participant half the time and a non-participant the other half. Its gains and losses in connection weight will therefore tend to cancel out, implying $\Delta t = 0$; 4) finally, considerations 1 – 3 above imply there is some region $\delta_{i,j} - \bar{\delta}_{i,j}$ in which $|\Delta t|$ is maximal. A function that satisfies all four of these criteria is

$$g(x) = x \cdot \exp[-\sigma \cdot |x|] \tag{22}$$

where σ is a constant controlling the rate of change in Δt. In the more general case, σ should be made dependent on the sign of x, producing a faster exponential decay for $x < 0$ to account for the fact that $t_{i,j}$ is lower-bounded by zero.

Adaptation under equations (21) results in non-integer values in the connection matrix. These are interpreted as indices of the relative participation of the various interconnection pathways in the activity of the network. η is selected such that changes in the connection weights vary by no more than a small amount during any one adaptation cycle.

Competitive Zones. The hypothesis that synaptic plasticity involves competition among synapses at the target cell is a commonly employed modeling assumption in a number of important adaptation rules [42] and is an assumption employed in the correlation theory of brain function [43]. Most adaptation rules employing this hypothesis, either explicitly or implicitly, treat synaptic competition at the receptor cell in terms of cell-wide (i.e. single compartment model) competition. However, the biological plausibility of treating synaptic competition on a the whole-cell level is somewhat strained for at least those neurons having extensive dendritic arbors. There is evidence favoring the hypothesis that axons may have preferences for specific portions of a target cell's surface [44]. Such a preference, combined with the localization of properties often exhibited by dendritic compartments in multi-compartment neuron models, suggests that at least some synaptic competition may be "regionalized" to specific portions of the receptor cell. Furthermore, at a higher level of modeling where one vertex represents a cell ensemble, there is even more justification for making the assumption that competition among inputs may be localized. A region of localized competition is what is meant here by "competitive zone."

Def. 6: A set of connections $T_{i,z} = \{t_{i,j_0}, \cdots, t_{i,j_f}\} \subset \{t_{i,1}, t_{i,2}, \cdots, t_{i,m+n}\},\ i \in [m+1, m+n]$, is a competitive zone in receptor *i* iff all $t_{i,j} \in T_{i,z}$ are adapted using the same threshold $\bar{\delta}(i, z)$ defined by equation (20c).

Different sets $T_{i,z}$ defined on receptor *i* are disjoint, and the union of all such sets defined on receptor *i* constitutes row B_iC_i within *T*. At higher levels of model abstraction, the simplest partitioning for competitive zones is to partition zones between *B* and *C* in *T*. This can be an appropriate partitioning when one vertex represents a cell ensemble since the number of input fibers in the input tract is typically much smaller than the number of neurons in the receiving network, most of which are interneurons [45]. Functionally it is reasonable to suppose that competition between inputs in this case is effectively compartmentalized by both interactions among "input layer neurons" and the distribution of synaptic connections made by different neurons in the local circuit [46]-[47].

At a lower level of abstraction the determination of competitive zone partitions bears some resemblance to the specification of clusters in neural networks based on higher-order neuron models [12]. Although cluster-learning algorithms that make no *a priori* assumptions about specific cluster internal structure have been reported, clustering models must nonetheless specify the number of clusters themselves, i.e. must specify the gross structure (how many clusters are present, what their input sources are, etc.). AOA analysis offers a possible approach to this initial structural definition. In our preliminary experiments so far conducted, it has been observed that the final graph structure obtained from applying the adaptation method above tends to eliminate connections from donator vertices having the smallest eigenvector component in the initial configuration of the network. This is a consequence of using row-averages $\bar{\delta}_{i,j}$ in (21a). Averaging is an inherently information-lossy process, and (21a) can be regarded as analogous to using a binary threshold quantizer in a communication receiver. It should be recognized that this information loss mechanism is one that the *modeling* introduces into the network structure, much as the introduction of the quantization characteristics of an analog-to-digital converter introduces information loss in a receiver [48]. In order to reduce the degree of this modeling loss, competitive zones can be introduced into a receptor based on grouping donators with like-valued eigenvector elements into the same competitive zone. This tends to prevent vertices with initially high eigenvector elements in the starting configuration from unduly dominating those with lower values during the crucial initial direction taken by the adaptation of the structure. Note that each

receptor can have its own competitive zone definition, which allows for overlapping the competition of specific donators at different receptors.

Synopsis of Preliminary Results. The AOA approach was publicly introduced outside of Japan in the first week of November, 2003. In the time that has elapsed since then the method has been extended to apply to the neural network structure problem in our laboratory with the following initial findings. These findings are based on relatively small graphs with varying initial configurations, and so represent preliminary results.

1) Stability. The adaptation algorithm (20)-(22) converges to stable final configurations for input patterns F consisting of both constant inputs and small sets of different input patterns, $\{F\}$ applied sequentially. For simple competitive zone partitioning BC, submatrix C typically converges first to a steady-state solution. With repetitive patterns, submatrix B usually exhibits a small, bounded limit cycle behavior as the input connections $t_{i,j}$ are slightly modulated by F. Within this limit cycle the B matrix $t_{i,j}$ do converge to steady-state values for each particular input pattern. "Stability" of the B submatrix is here bounded-input bounded-output (BIBO) stability in the Lyapunov sense. Depending on the number of patterns and the size of the network, convergence to steady-state typically occurs for small networks within 40 to 100 iterations, one iteration per input pattern application. The C submatrix also converges to a fixed steady-state configuration under zero-input conditions ($F = 0$) as well.

2) Capacity Index. The AOA index $Q = \lambda$ increases from its initial value to its final value, adjusted for variations due to different input patterns. For almost all steps in the iteration λ increases at each step. In a small number cases, typically one or two iterations out of the total, λ may decrease momentarily when a donator column in C goes to all-zero values, which means that this vertex no longer feeds back into the network, i.e. the vertex has been relegated to the role of being merely a "relay neuron." (Note that no vertex is ever allowed to lose its connection to the virtual feedback vertex v_{m+n+1}). This transient decrease in Q appears to be due to discretization effects related to the adaptation rate η since the system recovers a higher Q value at the next step. Hence, this appears to be merely a numerical artifact rather than being indicative of a violation of Bellman's Principle.

When simple BC competitive zone partitioning is used, no vertex has been observed to lose all of its donator connections. The virtual feedback index r_{m+n+1} likewise decreases at almost every step, adjusted for variations due to different input patterns, indicating more specific information concentration at vertices within the network, i.e. improved input pattern selectivity. Among the other eigenvector indices r_k, $k = m+1, ..., m+n$, the largest final values tend to be found at vertices

that have been configured as "output" (i.e. "relay") vertices driven by vertices that remained strongly connected in the final configuration. Furthermore, these "output vertices" have eigenvector indices that respond strongly to the input pattern *F*, which indicates selective routing of the input information and good separation of input patterns.

3) Reduction of Network Pathways. When exposed to a small suite of input patterns, the final network configuration tends to produce a final configuration in which the strongly-connected final subnetwork contains many fewer vertices than the initial configuration. Define a recurrent multi-layer network as a network in which some of the vertices in the initial configuration are not given direct paths from the input vertices but do have recurrent feedback to vertices that do have such connections. (Such vertices will here be called "strict interneurons"). The adaptation tends to reduce the number of layers in such a network in the final configuration in the sense that it tends to abolish the recurrent connections of strict interneurons back into the remaining strongly-connected subnetwork whenever possible. Thus, these vertices tend to be relegated to the role of "secondary relay neurons" and their eigenvector indices tend to be smaller than average in response to "familiar" input patterns. Provisionally, it may be the case that these secondary relay vertices may be well-suited to function as "observer vertices" reporting on the degree of "recognition" of the input pattern *F* since they do tend to have strong connections $t_{i,j}$ to some, but not all, donators in the strongly-connected subnetwork. In our preliminary studies, the network configuration seems to prefer lateral connections to multi-layer feedback connections, although not enough cases have been tested yet to allow us to form any firm hypothesis in this regard. In summary, the system shows a high degree of partitioning of the network in response to a small suite of correlated input patterns *F*.

4) Time Course of Network Connections. Although in many cases the $t_{i,j}$ tend to monotonically increase or decrease from their initial value, this is not true of every connection. In some cases, particularly those involving donators to vertices destined to become relay vertices, $t_{i,j}$ values may initially rise, only to fall later. This has an interesting possible and unanticipated implication for the structure of the network. Increasing the value of a $t_{i,j}$ is interpreted as indicative of the establishment of "messages" that are "significant" to the receptor. Biologically, there is a strong temptation to regard this as indicative of "stabilizing" synaptic connections. If, therefore, such a $t_{i,j}$ is later greatly reduced or driven to zero, the proper interpretation is not so much that a previously established synaptic connection has become disestablished; rather, it seems more likely that "significant messages" are being blocked by inhibitory actions.

If this provisional interpretation holds up under additional testing and research, the implication is that changes in direction in the time-course of the $t_{i,j}$ can be used to construct a second graph to

be overlaid upon the first. This second graph would represent *inhibitory connections* within the network, such inhibitory connections serving to optimize the objective function in a neuron-level model of the structure defined by *T*. Each vertex in this *inhibitory graph*, *S*, is interpreted as representing an inhibitory interneuron or an ensemble of interneurons with inhibitory outputs. Excitatory donator connections from graph *T* to graph *S* are easy to identify from the $\delta_{i,j}$ in (21a), and inhibitory output connections would extend to the receptors v_i undergoing a direction change in one or more of their $t_{i,j}$. If some but not all these $t_{i,j}$ values change direction, this tends to imply presynaptic inhibition of these pathways. If all these $t_{i,j}$ undergoing changes in direction, this tends to imply inhibition applied to v_i itself. Linkage between *T* and *S* can in principle be effected as an inhibitory modulation

$$t_{i,j} = \left(t_{i,j}\right)_{\max} \cdot \rho_{k,i,j} \tag{23}$$

where $\rho_{k,i,j}$ is the connection strength of the inhibition running from vertex *k* in *S* to arc $t_{i,j}$ in *T*. Parameter $(t_{i,j})_{\max}$ represents the peak value attained by $t_{i,j}$ during the adaptation prior to its later reduction.

5) Dynamic Links. We have observed that competitive zone partitions confined to within the *C* submatrix tend to converge to stable configurations before those involving the *B* submatrix. Furthermore, the *B* submatrix tends to exhibit BIBO limit cycle responses to periodically-applied input pattern suites $\{F\}$. This of course means that the $t_{i,j}$ connections running to "neuron" vertices v_k, $k \in [m+1, m+n]$, have input-dependent optimum values. Convergence of the *C* submatrix is an optimization over the input suite, and once *C* has converged it undergoes no more changes provided that it contains no competitive zones overlapping into *B*. This property can be exploited to obtain input-dependent *dynamic links* in the input pathway.

This can be achieved as follows. Let the input suite consist of *L* input patterns

$$\mathfrak{F} = \{F_1, F_2, \ldots, F_L\}.$$

Following convergence of the *C* submatrix, continued application of each pattern F_ℓ drives the *B* submatrix to convergence to input-dependent steady-state values associated with each F_ℓ. The limit cycle behavior of *B* is a consequence of changing input patterns. The resulting B_ℓ for each element of $\mathfrak{F}$ represents an optimal value conditioned on a pattern-by-pattern basis. Because the

adaptation produces a maximal pattern-dependent value of λ for each separate case, this gives us a pattern-dependent optimal set

$$\mathfrak{B} = \{B_1, B_2, \ldots, B_L\}.$$

For an arbitrary input pattern F compute

$$M = F'\left[F_1 \mid F_2 \mid \cdots \mid F_L\right]$$

where prime denotes transpose. The maximum element in row vector M identifies the element of $\mathfrak{F}$ that F most closely matches. The corresponding element of $\mathfrak{B}$ then corresponds to the best match of connections $t_{i,j}$ for this input in submatrix B. Obviously this simple scheme can be refined, and the possibility is open that better and more sophisticated methods of dynamic link modulation of connections await discovery. What must be borne in mind is that the AOA information indices $t_{i,j}$ do not themselves directly represent synaptic strength. Their application to lower-level, more physiologically-based neural network models calls for examination of the signal modulation properties of the neurons. The $t_{i,j}$ should not merely be regarded as synaptic weights.

Damasio Networks. Extension of the graph network (7) to general structures in the form of Damasio's multi-level network architecture [2] is straightforward. It involves merely a partitioning of the F vector, B matrix, and C matrix in (7). Suppose we have a three-MCU system composed of two feature-representing networks at the first level projecting to a single convergence zone network at the second level. Further assume that the convergence zone network receives no direct sensory input but only outputs from the first-level networks. Let $F^{(1)}$ and $B^{(1)}$ be the pattern vector and input distribution matrix for feature network 1, and let $F^{(2)}$ and $B^{(2)}$ be the pattern vector and input distribution matrix for feature network 2. The total number of inputs to the system is $m = m_1 + m_2$ and the total number of neurons is $n = n_1 + n_2 + n_3$. $F^{(1)}$ is $m_1 \times 1$, $F^{(2)}$ is $m_2 \times 1$, and the input matrices are $n_1 \times m_1$ and $n_2 \times m_2$, respectively.

The C matrix is partitioned into local network submatrices, $C^{(i,i)}$, and cross-connection submatrices, $C^{(i,j)}$. Each $C^{(i,i)}$ is $n_i \times n_i$, and each $C^{(i,j)}$ is $n_i \times n_j$. (7) then becomes

$$T = \left[\begin{array}{cc|ccc|c} \multicolumn{2}{c|}{0} & \multicolumn{3}{c|}{0} & \begin{matrix} F^{(1)} \\ \hdashline F^{(2)} \end{matrix} \\ \hline \begin{matrix} B^{(1)} & 0 \\ \hdashline 0 & B^{(2)} \end{matrix} & \begin{matrix} 0 \\ \hdashline 0 \end{matrix} & \begin{matrix} C^{(1,1)} & C^{(1,2)} & C^{(1,3)} \\ \hdashline C^{(2,1)} & C^{(2,2)} & C^{(2,3)} \end{matrix} & \begin{matrix} C^{(3,1)} & C^{(3,2)} & C^{(3,3)} \end{matrix} & \begin{matrix} \\ \hdashline 0 \end{matrix} \\ \hline \multicolumn{2}{c|}{\cdots 0 \cdots} & \multicolumn{3}{c|}{\cdots 1 \cdots} & 0 \end{array}\right]. \qquad (24)$$

As before, the system contains a single virtual feedback node, v_{m+n+1} that monitors all the "neuron vertices" in the system and supplies excitation to the pattern vector F. T is $(m+n+1) \times (m+n+1)$. Generalization of the form of T to any number of levels and any number of MCUs is straightforward.

The adaptation behavior of (24) will be determined in significant part by the definitions of competitive zones in each of rows $m + 1$ to $m + n$. Preliminary results to date indicate that for at least some competitive zone definitions different sub-graphs within C will converge to a steady-state configuration prior to other sub-graphs within C. For example, this is expected if competitive zone boundaries align going down the columns of C. This raises one experimental research question and also one implication for dynamic link architectures.

The research question is: What is the correct determination of competitive zone definitions for models of actual neurological networks? This is a difficult question because the cellular-level dynamics of synaptic competition are not presently completely understood. For example, do synapses made on distal dendrites undergo competition with those made on proximal dendrites? Do they do so in some cell types but not in others? These and many other questions have no clear cut answers at the present state of knowledge. In this arena, the model proposed here has value for helping to identify and design experiments. Definitions chosen for the competitive zones based on morphological studies of real neural networks will affect the final structure of the system, and at a sufficiently low level of abstraction will make imputations about the nature of synaptic competition in biological neurons. It follows that in those cases where a particular competitive zone definition produces structures at odds with experimental findings, it is unlikely that synaptic competition in the biological system has the competitive domain established by the competitive zone definition of the model. On the other hand, if experimental findings agree with the structural characteristics produced by the model, this lends support to the hypothesis that the

mechanics of synaptic competition in the experimental case follow the competitive domain structure assumed in the model.

The implication for dynamic link architectures is a generalization of the approach to modeling dynamic links discussed in the previous section. Subgraphs in C that stabilize before the system as a whole due to their competitive zone definitions are unaffected by subsequent variations in the rest of the system under the application of a suite of input patterns $\mathfrak{F}$. On the other hand, cross-linking terms $C^{(i,j)}$ are physically more likely to have competitive zones that overlap either with inputs or with competitive zones $C^{(k,k)}$ in the target network. Likewise, within any particular MCU $C^{(k,k)}$ a given donator v_j may compete with different sets of donators on different receptors v_i. BIBO limit cycles are more likely to occur for these types of competitive zone definitions. Consequently, by including all subgraphs in (24) exhibiting BIBO limit cycles in the method proposed earlier, the dynamic link mapping can be made to include these terms, thus defining pattern-dependent optimal sets $\mathfrak{C}^{(i,j)}$ to accompany those in $\mathfrak{B}$ described previously. As was the case in the discussion above, further research can be expected to lead to additional refinements in this simple scheme.

Summary. This paper has presented a new approach to the problem of neural network structure identification. It is based on an extension of the recently introduced AOA method. It has been established that AOA indices are related to information-theoretic measures of optimization, and that consequently AOA indices can serve as a basis for an objective function for optimization of information-theoretic properties in the structure of a network. A simple algorithm has been found that produces a number of preliminary results consistent with what one would expect from optimization of an information-theoretic objective function. The algorithm does not require direct measurements of entropies or probability distributions, which is advantageous because such direct measures are notoriously difficult to obtain in practice. AOA treats the network at an abstract level and is suitable for a hierarchical approach to large neural network systems. Preliminary work has evidenced significant promise for this approach, but the research is still in its early stages and clearly much work remains to be carried out.

Acknowledgement. This work is supported by the NSF-Idaho EPSCoR Program and by the National Science Foundation under award number EPS-0132626.

References

1. A.R. Damasio, “Time-locked multiregional retroactivation: A systems-level proposal for the

neural substrates of recall and recognition," *Cognition*, **33** (1989) 25-62.

2. A.R. Damasio, "The brain binds entities and events by multiregional activation from convergence zones," *Neural Computation*, vol. 1, 1989, pp. 123-132.
3. E.R. Kandel, T.M. Jessell, and J.R. Sanes, "Sensory experience and the fine-tuning of synaptic connections," in *Principles of Neural Science*, 4th ed., E.R. Kandel, J.H. Schwartz, and T.M. Jessell (eds.),

   NY: McGraw-Hill, 2000 , pp. 1115-1129.
4. C.S. Goodman, C.J. Shatz, "Developmental mechanisms that generate precise patterns of neuronal connectivity, *Cell* **72:** 77-98, 1993.
5. L.C. Katz and C.J. Shatz, "Synaptic activity and the construction of cortical circuits, *Science* **274:** 1133-1138, 1996.
6. T.M. Jessell and J.R. Sanes, "The generation and survival of nerve cells," in *Principles of Neural Science*, 4th ed., E.R. Kandel, J.H. Schwartz, and T.M. Jessell (eds.), NY: McGraw-Hill, 2000 , pp. 1041-1062.
7. E.L. Bienenstock, L.N. Cooper, and P.W. Munro, "Theory for the development of neuron selectivity: Orientation specificity and binocular interaction in visual cortex," *J. Neurosci.*, vol. 2, no. 1, pp. 32-48, Jan., 1982.
8. S.-I. Amari, "Mathematical foundations of neurocomputing," *Proc. IEEE*, vol. 78, no. 9, Sept. 1990, pp. 1443-1463.
9. Ch. von der Malsburg and W. Schneider, "A neural cocktail-party processor," *Biol. Cybern.* **54**, 29-40, 1986.
10. R. Eckhorn, H.J. Reitboeck, M. Arndt, and P. Dicke, "Feature linking via synchronization among distributed assemblies: Simulations of results from cat visual cortex," *Neural Comput*. **2**, 293-307, 1990.
11. R.B. Wells, "Modulation channels in biomimic artificial neurons," *Proc. 28th An. Conf. IEEE Indus. Electron. Soc. IECON'02*, Sevilla, Spain, Nov. 5-8, 2002, pp. 3209-3214.
12. M.W. Spratling and G.M. Hayes, "Learning synaptic clusters for nonlinear dendritic processing," *Neural Processing Letters* **11:** 17-27, 2000.
13. W. Gerstner and W. Kistler, *Spiking Neuron Models*, Cambridge, UK: Cambridge University Press, 2002, pp. 380-383.
14. C. Aizenman, P.B. Manis, and D.J. Linden, "Polarity of long-term synaptic gain change is related to postsynaptic spike firing at a cerebellar inhibitory synapse," *Neuron* **21:** 827-835, 1998.
15. A. Marty and I. Llano, "Modulation of inhibitory synapses in the mammalian brain," *Curr.*

Opin. Neurobiol. **5:** 335-341, 1995.

16. M.F. Bear and D.J. Linden, "The mechanisms and meaning of long term synaptic depression in the mammalian brain," in *Synapses*, W.M. Cowan, T.C. Südhof, and C.F. Stevens (eds.), Baltimore, MD: Johns Hopkins University Press, 2001, pp. 455-517.
17. R.C. Malenka and S.A. Siegelbaum, "Synaptic plasticity: Diverse targets and mechanisms for regulating synaptic efficacy," in *Synapses*, W.M. Cowan, T.C. Südhof, and C.F. Stevens (eds.), Baltimore, MD: Johns Hopkins University Press, 2001, pp. 393-453.
18. B. Hammer, "Compositionality in neural systems," in *The Handbook of Brain Theory and Neural Networks*, M.A. Arbib (ed.), 2nd ed., Cambridge, MA: The MIT Press, 2003, pp. 244-248.
19. Ch. von der Malsburg, "Dynamic link architecture," in *The Handbook of Brain Theory and Neural Networks*, M.A. Arbib (ed.), 2nd ed., Cambridge, MA: The MIT Press, 2003, pp. 365-368.
20. C.E. Shannon and W. Weaver, *The Mathematical Theory of Communication*, Urbana, IL: The University of Illinois Press, 1964.
21. S. Becker and R.S. Zemel, "Unsupervised learning with global objective functions," in *The Handbook of Brain Theory and Neural Networks*, M.A. Arbib (ed.), 2nd ed., Cambridge, MA: The MIT Press, 2003, pp. 1183-1187.
22. A.J. Bell, "Independent Component Analysis," in *The Handbook of Brain Theory and Neural Networks*, M.A. Arbib (ed.), 2nd ed., Cambridge, MA: The MIT Press, 2003, pp. 569-575.
23. S.-I. Amari, "Natural gradient works efficiently in learning," *Neural Comput.* **10:** 251-276, 1997.
24. R. Linsker, "Local synaptic learning rules suffice to maximize mutual information," *Neural Comput.* **9:** 1661-1665, 1997.
25. A.J. Bell and T.J. Sejnowski, "An information maximization approach to blind separation and blind deconvolution," *Neural Comput.* **7:** 1129-1159, 1995.
26. J. Feng, Y. Sun, H. Buxton, and G. Wei, "Training integrate-and-fire neurons with the Informax Principle II," *IEEE Trans. Neural Networks*, vol. 14, no. 2, Mar. 2003, pp. 326-336.
27. S. Becker and M. Plumbley, "Unsupervised neural network learning procedures for feature extraction and classification," *Int. J. App. Intell.* **6**(3): 185-205, 1996.
28. H.B. Barlow, "Unsupervised learning," *Neural Comput.* **1:** 295-311, 1989.
29. S. Becker, "Mutual information maximization: Models of cortical self-organization," *Netw. Computat. Neural Syst.* **7:** 7-31, 1996.

30. J.-F. Cardoso and B.H. Laheld, "Equivariant adaptive source separation," *IEEE Trans. Signal Process.*, vol. 44, pp. 3017-3030, 1996.

31. L. Zhaoping, "Optimal sensory encoding," in *The Handbook of Brain Theory and Neural Networks*, M.A. Arbib (ed.), 2nd ed., Cambridge, MA: The MIT Press, 2003, pp. 815-819.

32. J. Hertz and S. Panzeri, "Sensory coding and information transmission," in *The Handbook of Brain Theory and Neural Networks*, M.A. Arbib (ed.), 2nd ed., Cambridge, MA: The MIT Press, 2003, pp. 1023-1026.

33. A.V.M. Herz, "Temporal sequences: Learning and global analysis," in *The Handbook of Brain Theory and Neural Networks*, M.A. Arbib (ed.), 2nd ed., Cambridge, MA: The MIT Press, 2003, pp. 1167-1171.

34. R.B. Wells, *Applied Coding and Information Theory for Engineers*, Upper Saddle River, NJ: Prentice-Hall, 1999.

35. K. Akuzawa and K. Ohnishi, "Design indices for information connection in decentralized systems," *Proc. 29th An. Conf. IEEE Indus. Electron. Soc. IECON'03*, Roanoke, VA, Nov. 2-6, 2003, pp. 2417-2422.

36. F.P. Preparata and R.T. Yeh, *Introduction to Discrete Structures*, Menlo Park, CA: Addison-Wesley, 1974.

37. K.A.S. Immink, "Run-length limited sequences," *Proc. IEEE*, vol. 78, no. 11, pp. 1745-1749, 1990.

38. G. Ungerboeck, "Channel coding with multilevel/phase signals," *IEEE Trans. Inform. Th.*, vol. IT-28, no. 1, 1982, pp. 55-67.

39. R.B. Wells, "Application of set-membership techniques to symbol-by-symbol decoding for binary data transmission systems," *IEEE Trans. Inform. Th.*, vol. 42, no. 4, 1996, pp. 1285-1290.

40. K.A.S. Immink, "Runlength-limited sequences," *Proc. IEEE*, vol. 78, no. 11, 1990, pp. 1745-1759.

41. C.T. Chen, *Introduction to Linear System Theory*, NY: Holt, Rinehart and Winston, 1970.

42. Y. Frégnac, "Hebbian synaptic plasticity," in *The Handbook of Brain Theory and Neural Networks*, M.A. Arbib (ed.), 2nd ed., Cambridge, MA: The MIT Press, 2003, pp. 515-522.

43. Ch. von der Malsburg, "The correlation theory of brain function," in *Models of Neural Networks II*, E. Domany, J.L. van Hemmen, & K. Schulten (eds.), NY: Springer-Verlag, 1994, pp. 95-119.

44. J.R. Sanes and T.M. Jessell, "The formation and regeneration of synapses," in *Principles of*

Neural Science, 4th ed., E.R. Kandel, J.H. Schwartz, and T.M. Jessell (eds.), NY: McGraw-Hill, 2000 , pp. 1087-1114.

45. A. Schüz, "Neuroanatomy in computational perspective," in *The Handbook of Brain Theory and Neural Networks*, M.A. Arbib (ed.), 2nd ed., Cambridge, MA: The MIT Press, 2003, pp. 733-737.
46. R. Douglas and K. Martin, "Neocortex," in *Synaptic Organization of the Brain*, 4th ed., G.M. Shepherd (ed.), NY: Oxford University Press, 1998, pp. 459-509.
47. M. Toledo-Rodriguez, A. Gupta, Y. Wang, C.Z. Wu and H. Markram, "Neocortex: Basic neuron types," in *The Handbook of Brain Theory and Neural Networks*, M.A. Arbib (ed.), 2nd ed., Cambridge, MA: The MIT Press, 2003, pp. 719-725.
48. R.B. Wells and G.L. Bartles, "Simplified calculation of likelihood metrics for Viterbi decoding in partial response systems," *IEEE Trans. Magn.*, vol. 32, no. 5, 1996, pp. 5226-5237.
49. B. Widrow and M.A. Lehr, "30 years of adaptive neural networks: Perceptron, Madaline, and backpropagation," *Proc. IEEE*, vol. 78, no. 9, pp. 1415-1442, 1990.