

Evolution of the Neural Immunoglobulin Supergene Family  
and Functional Studies of One of its Members

Thesis by  
Robert P. Lane

In Partial Fulfillment of the Requirements for the  
Degree of Doctor of Philosophy

California Institute of Technology  
Pasadena, California  
1997  
(Defended September 25, 1996)





## Acknowledgments

I want to begin with an acknowledgment of two people who will never appear on my papers yet are “first authors” in my life and have made countless contributions to this endeavor. The first is my wife, Heather, who has shown more times than I deserve limitless patience, understanding, and support. I am continually reminded of just how far above my head I have married. The second is my dad whose name I share, who has been my most influential and inspirational teacher in most of life’s subjects. Especially during my Eastwood and Colgate years when my dad had every reason to be entirely focused on his own life, he remained an incredible source of love and wisdom.

My choice to consider a career in science can be traced primarily to two people. Ron Hoham, my mentor at Colgate University, showed me as an undergraduate how much fun it is to play with the scientific method. But despite these experiences, I would probably today be an entirely broke and desperate novelist or children’s story writer if it hadn’t been for my senior year high school English teacher (whose name I have since forgotten) who gave me my first “C” and politely exposed my prose woes. Even with this thankful wisdom, I remain entirely broke and desperate however.

But I should say that money is pretty unimportant to me, and it was finally my fast-track position in investment banking that convinced me to not play out the string until retirement in some high-paying and incredibly uninteresting job. I wondered at first whether Bill Dreyer’s interest in me as a prospective graduate student was related to this finance experience and the promise of portfolio management between gels while the other graduate students in the lab got to write grants and teach classes. As it turns out, Bill was even less interested in talking about money than I was, but I should say, he knows how to spend it, and a few scientifically slow winters were magically transformed by Devon Street Superbowl margaritas, dolphin and sun worshipping at the Sun Castle, or flights to Mammoth Lakes for the previous night’s powder.

But I want to acknowledge Bill for more than just the obvious lab perks and his ability to recognize when they are most needed during this long and sometimes lonely trek through Ph.D. work. There is no hiding the fact that Bill and I have had a lot of disagreement and difficulty over the years, both scientifically and philosophically,

but as I have grown to understand him better I have come to admire his ability to fly directly and confidently into turbulent winds. I have learned a great deal from Bill about the importance of allowing scientific data the chance to support alternative hypotheses, and to be a fearless interpreter of this data if for no other reason than to be immaculate with controls in subsequent experiments. Bill is a big thinker, sees the big picture, and likes to think and share big ideas. I think we like each other because we are both rather philosophers merely using the tools of science: more than anything else, I would like to acknowledge Bill for the many years of debate, engaging discussion, and thoughtful exchanges about how life works.

Finally, there are a several others that in a less verbose manner (but in many ways, no less deserving of full paragraphs) I would like to especially acknowledge. I wish to thank members of my thesis committee for their help and insight along the way: Pamela Bjorkman, Scott Fraser, Paul Sternberg, and Kai Zinn, several of whom provided helpful guidance and kind words regarding future postdoctoral opportunities. I am grateful to have worked with past and present members of the lab, including Faiz, Janet, Frank and Anita, all of whom have been providers of buffers, protocols, and good conversation. I would like to especially acknowledge Jost Vielmetter who has been a good friend and a constant advisor in the lab. I would like to thank members of the Benzer, Sternberg and Bjorkman labs for allowing me to so frequently wander into their territories, and for freely sharing their equipment and expertise over the years. I would also like to acknowledge Michael Frohlich who has taught me just about everything I know about phylogeny, and whose recent interest in my work with the Ig Superfamily has led to several stimulating conversations, not to mention what I hope will remain a good scientific friendship. Speaking of friendships, finally I wish to acknowledge many of my friends here at Caltech who have made life here nearly as nice as the weather, especially Dan, Mike, Dale, and Jeff, all the members of the hockey club, as well as all the guys on the football, softball and rotisserie teams. It has really been a good time.

### Abstract

The immunoglobulin supergene family is a diverse set of molecules that share in common the immunoglobulin (Ig) domain. In the nervous system, a large subfamily of these proteins has been characterized that contain, in addition to N-terminal Ig-like domains, numerous fibronectin (Fn) type III repeats. One group of these neural homologs has been well characterized and includes Neuroglian, Bravo/Nr-CAM, Neurofascin, L1, Ng-CAM, Contactin/ F11, Axonin-1/Tag-1, Big-1/Tag-like, and Big-2. Each of these proteins have six Ig-like domains and either four or five fibronectin type III repeats, and various developmental functions have been attributed to this group, including neurite outgrowth, fasciculation, cell adhesion and axon guidance.

Using structural modeling and cladistic analyses, the evolutionary relationships among these homologous neural Ig superfamily proteins were investigated. This study reinforces the idea that individual Ig-like and Fn domains are probably not distinct functional modules that can be shuffled in evolution, but rather that they may act in tandem. Patterns of conservation and divergence of specific residues along the various phylogenetic branches of the evolutionary tree suggest a model whereby important interactions may predominantly map between domains, with the "top" loops of one domain, the "bottom" loops of the adjacent domain, and the interdomain residues forming part of a ligand "pocket". The evolutionary analyses also permits an evaluation of the controversial identification of Ng-CAM and L1 as species orthologs, and in light of avian-mammalian speciation events, it appears these proteins are orthologous but perhaps not functionally identical.

A new member of this neural Ig subfamily has been cloned and identified as the human ortholog to the chicken Bravo/Nr-CAM protein. The complete coding sequence was determined, and like its

chicken homolog, it is composed of six V-like Ig-like domains, five fibronectin type III repeats, as well as a transmembrane and intracellular domain. Overall, the human protein is 82% identical to the chicken homolog, although the trans-membrane and intracellular domains are 100% conserved at the amino acid level. Independent cDNA's encoding four distinct isoforms were identified, all of which contain alternatively spliced variants around the fifth fibronectin repeat, including one isoform previously identified in chicken in which the entire 93 amino acid domain is spliced. Northern blot analysis reveals one mRNA species of approximately 7.0 kb in adult brain. Fluorescence *in situ* hybridization maps the human Bravo/Nr-CAM gene to human chromosome 7q31.1-31.2, a locus previously identified to contain a tumor suppressor gene.

Although the cell adhesion (CAM) nomenclature implies that Bravo/ Nr-CAM and its family members function merely as a sort of indiscriminate cell-cell "glue", evidence has mounted that these proteins participate in receptor-like intracellular signalling functions with cell behavioral consequences. Of particular interest with regard to the Bravo/Nr-CAM protein is the conserved alternative splicing of the membrane-near fibronectin domain, as well as the striking sequence conservation C-terminal to this alternative exon that extends through the membrane and inside the cell. To explore the function of these sequences, both the 93 amino acid alternative Fn5 exon and the 100% conserved intracellular domains of Bravo/Nr-CAM were separately produced in heterologous expression systems and purified by various biochemical techniques. Affinity chromatography and expression library screening were used in an attempt to identify putative ligands to these presumably important protein regions.

The significance of the fifth fibronectin alternative exon usage was also investigated by using the expressed domain to raise domain-specific monoclonal antibodies, and using the antibodies in a histological study of spatial and temporal regulation of these splicing

events. Using double labeling and confocal microscopy, as well as PCR analysis, in all tissues and across all stages of development, both the domain-containing and domain-lacking isoforms appear to be uniformly expressed in the same cells. Therefore, the developmental function of the complex array of alternatively spliced variants around the fifth fibronectin domain is subtle. A model is discussed whereby isoform diversity may provide a means to integrate multiple ligand-binding events involving the same protein on the same cell that interact with distinct ligands and co-receptors.

## Table of Contents

### **Chapter I    The Molecular Basis of Developmental Neurobiology.**

#### **The Developmental Problem of Nervous System Assembly**

<i>Complexity of the Problem: Parallels with the Immune System</i>	1
<i>Axon Guidance and Synaptic Targeting</i>	3
<i>Hard-wiring versus Plasticity</i>	4
<i>Gradients versus Compartments</i>	7
<b>Underlying Molecular Basis of Neuronal Connectivity</b>	
<i>Diverse Proteins that Function in Nervous System Assembly</i>	8
<i>Cadherins, Integrins and the Extracellular Matrix</i>	9
<i>N-CAM proteins</i>	10
<i>Neogenin and DCC</i>	12
<i>Kinases and Phosphatases</i>	12
<i>Address Molecules</i>	14
<i>Methods and Initial Experiments</i>	15
<b>Conclusions</b>	17
<b>References</b>	18
<b>Figures</b>	29

### **Chapter II    Cladistic and Evolutionary Analysis of the Six Ig-like Domain Subfamily of Neural Cell Adhesion Molecules**

<b>Introduction</b>	2
<b>Materials and Methods</b>	
<i>Sequence Alignments and Structural Modeling</i>	5
<i>Cladistics and Generation of Phylogenetic Trees</i>	7
<i>Examining the Tree Space</i>	9
<i>Simplification of Analysis by Collapsing Orthologous Sequences</i>	10
<i>Gene Trees Versus Domain Trees</i>	10
<i>Statistical Analysis of Conservation and Divergence</i>	11

**Results**

<i>Phylogenetic Analysis of the Six Ig -like Domain Subfamily</i>	12
<i>Avian-Mammalian Speciation Events</i>	14
<i>Evolutionary History of Individual Ig -like Domains</i>	14
<i>Evolution of the Six Ig -like Domain Prototypic Ancestral Molecule</i>	15
<i>Evolutionary History of Individual Fn Type III Repeats</i>	17
<i>Patterns of Conservation and Divergence Among Domains and Motifs</i>	18
<i>Tracing Functional Structural Motifs</i>	21

**Discussion**

<i>Phylogeny to Infer Function</i>	24
<i>Domain Trees Atop Gene Trees Atop Species Trees</i>	26
<i>Fibronectin Repeats Versus Ig-like Domains</i>	29
<i>Is Neuroglian Like the Ancestral Six Ig -like Domain Protein?</i>	30
<i>Is Ng-CAM the Chicken Ortholog to L1?</i>	32
<i>Paralogous Versus Orthologous Conservations</i>	34
<i>Do Important Ligand Interactions Map Between Domains?</i>	35
<b>Conclusions</b>	38
<b>Acknowledgment</b>	40
<b>References</b>	41
<b>Table and Figures</b>	45

### **Chapter III   The Human Ortholog to the Bravo/Nr-CAM Neural Ig Subfamily Protein**

<b>Introduction</b>	2
<b>Materials and Methods</b>	
<i>Generation of Human Fetal cDNA Library</i>	4
<i>Isolation and Sequencing of Human Bravo/Nr-CAM Clones</i>	5
<i>Northern Blot Analysis</i>	6
<i>Fluorescence in situ Hybridization (FISH)</i>	6
<b>Results</b>	
<i>Sequence Analysis of Human Bravo/Nr-CAM cDNA</i>	7

<i>Homology between the Chicken and Human Bravo/Nr-CAM Proteins</i>	8
<i>Northern Blot Analysis of Human Bravo/Nr-CAM</i>	9
<i>Human Bravo/Nr-CAM Maps to Human Chromosome 7q31.1-2</i>	10
<b>Discussion</b>	
<i>Structural Model and Sequence Features</i>	10
<i>The Remarkable Conservation of the Intracellular Domain</i>	11
<i>Kinase Motifs and Signal Transduction Considerations</i>	12
<i>Alternative Splicing of Human and Chicken Bravo/Nr-CAM RNA's</i>	13
<i>Could Bravo be a Tumor Suppressor Candidate?</i>	13
<b>Conclusions</b>	15
<b>References</b>	16
<b>Figures</b>	22

## **Chapter IV Are CAM's Really Receptors? Functional Studies Aimed at Elucidating Signal Transduction Events in Bravo/Nr-CAM.**

<b>Introduction</b>	
<i>The CAM Misnomer? Evidence of Signal Transduction</i>	2
<i>Intracellular Domain Sequences in the Bravo/Nr-CAM Subfamily</i>	2
<i>Fifth Fibronectin Repeats and Co-Receptor Function?</i>	3
<b>Materials and Methods</b>	
<i>Protein Expression Systems</i>	6
<i>Heterologous Expression of the Intracellular Domain</i>	8
<i>IMAC and Purification of the Expressed Intracellular Domain</i>	10
<i>Heterologous Expression of the Alternative Fn5 Exon</i>	11
<i>Inclusion Body Preparation of the Expressed Fn5 Domain</i>	12
<i>N-terminal Protein Sequencing of Expressed Domains</i>	13
<i>Preparation of a Domain-Specific Monoclonal Antibody</i>	14
<i>Histology to Investigate Regulation of Isoform Expression</i>	14
<i>PCR Analysis of Isoform Expression</i>	16
<i>Affinity Chromatography to Identify Putative Ligands</i>	16
<i>Affinity Chromatography Utilizing the His-Tail</i>	19



<i>Expression Library Screening to Identify Putative Ligands</i>	19
<b>Results</b>	
<i>Expression of the Intracellular and Fn5 Domains of Bravo/Nr-CAM</i>	21
<i>Generation of a Domain-Specific Monoclonal Antibody</i>	22
<i>Co-expression of Domain-Containing and Domain-Lacking Isoforms</i>	23
<i>Affinity Chromatography Identifies Putative 70 kD Ligand to Fn5</i>	24
<i>Intracellular Domain Ligand Results</i>	25
<b>Discussion</b>	
<i>Is this Alternative Splicing Event Functionally Significant?</i>	26
<i>Putative Ligands to the Fn5 and Intracellular Domains</i>	28
<b>Conclusions</b>	32
<b>References</b>	33
<b>Figures</b>	39

## **Appendix I -- Unpublished Commentary (1994)**

Perspectives on the Evolution, Structure and Function of Fibronectin Type III Repeats.  
(Fibronectin Domains are People Too... )

## **Appendix II -- Unpublished Commentary (1996)**

Do DNA rearrangements Play a Role in Development? Perspectives and Investigative  
Experiments that Explore the Role of DNA Changes in Development.

## Chapter I

### Introduction: The Molecular Basis of Developmental Neurobiology

#### **The Developmental Problem of Nervous System Assembly**

##### *Complexity of the Problem: Parallels with the Immune System*

It is probably no coincidence that the immune system and nervous system share several molecular characteristics. Perhaps no other developing systems have quite the same daunting task of sorting out the sheer number of required molecular contacts: immunity requires the recognition of enormous numbers of antigens and pathogens; the developing human nervous system must sort out approximately  $10^{15}$  synaptic connections. In this regard, both systems must have evolved a large and diverse set of cell surface proteins that permit this order of recognition, and although the homology between the two systems is at this point only loosely bridged by a few genes, including the Immunoglobulin Superfamily (IgSF), this common cell recognition problem may ultimately lead to other intriguing parallels.

If the immune system has solved its daunting developmental problem via complex mechanisms of DNA rearrangements that allow cells to express an enormously diverse range of immunoglobulin receptors, might similar systems exist to generate such diversity in neural development? It is perhaps tempting to presume so: DNA recombination is certainly an ancient developmental phenomenon that can be traced to cyanobacteria, some of the earliest life on our

planet (Haselkom, 1992). Similar mechanisms are evident in a diverse and dispersed set of organisms (for example: Gierl et al., 1989; Klar, 1990; Muller et al., 1991; Prescott, 1992), therefore it is perhaps parsimonious to presume that the immune system recombination machinery was not re-invented with the dawn of vertebrates and immunity. An intriguing transgenic mouse experimental result was reported recently (Matsuoka *et al.*, 1991) that might have indeed given compelling support of this notion. As it turns out, however, this experiment appears flawed and the result irreproducible (see Schatz and Chun, 1992), and there remains no solid evidence that immune system mechanisms were merely borrowed in evolution from older pre-existing neural or other developmental systems. Initial experimental inquiry and more extensive consideration of this subject of DNA changes outside the immune system is discussed in Appendix II.

There is one important difference between the two systems that might shed further light on this subject. In the immune system, adaptation is critical: immunoglobulins must keep pace with rapidly mutating viral strains and generally respond to antigens that the organism has never before seen in its evolutionary history. In this way, the full extent of immunity can not be selected for in previous generations, and survival is assured only by selection of maximally adaptive mechanisms. This is certainly in stark contrast with the cell-cell recognition problem in development of the nervous system. Synaptic connectivity, or for that matter any developmental cell-cell specificity problem, does not warrant nor indeed favor randomness. The receptors that guide developing neurons to make specific synaptic targets must be under exquisite control, and in this regard, it is unlikely that neurons will accomplish this task by producing novel recombinants. And so the underlying molecular mechanisms that account for the sheer number of distinct connections established in the developing nervous system remain enigmatic: an unparalleled cross

between the complexity of the immune system recognition problem and the preprogrammed specificity of the most complex electrical circuit board imaginable.

### *Axon Guidance and Synaptic Targeting*

If the emergent property of a functioning nervous system is the ability of the organism to integrate sensory or other inputs to generate concise behavioral outputs, it is likely the product of patterned synaptic connectivity. Differential synaptic strengths or weights among a patterned network of neurons may ultimately shape this output, and thus the precise connectivity in the neuronal web is probably the fundamental unit of neural activity, from reflex to creative thought. In this regard, the developmental mechanisms that underlie these patterned connections will likely provide an important part of the puzzle that demystifies the unfathomable: how the brain works.

Yet even the simplest, "instinctive" behaviors involve an astounding complexity of neuronal targeting and connectivity. One neural system that especially illustrates principles of complex choreography and elegant patterning is the vertebrate visual system, which also exemplifies a common aspect of the connectivity problem: neurons born in one geographical region must project to an appropriate set of target neurons in another region. Therefore in general, both mechanisms that guide extending axons to distant areas as well as specifically connect these axons to synaptic targets must be considered. In vertebrates, extensive information on both of these fronts has emerged in the last 50 years, and some of this data is discussed below in the context of one particularly well studied example: the retinotectal projection.

In the avian visual system, approximately one million retinal ganglion cells are born on the retinal dish in a developmental gradient, the most recently differentiated cells being born progressively more towards the periphery. These newly born ganglion cells extend axons

in a stereotyped fashion atop the fibers of their older siblings, fasciculating as they extend towards the optic fissure. The axons form one large fascicle as they arrive at the center of the retina, which exits the retina as the optic nerve. The projection to the visual center of the brain, the optic tectum, is entirely contralateral in chickens: the optic nerves from both eyes completely cross at the chiasm. As the axons extend past the chiasm they confront a large, glial cell population substrate, and perhaps not coincidentally, these axons begin to defasciculate as they pass this landmark and approach the target. At the optic tectum, discrete axonal bundles spread over the tectal surface and innervate in a highly stereotyped fashion.

In a set of pioneering experiments, Roger Sperry first described the fact that this visual system development was a topographical projection: retinal cells robustly maintain their neighborly relationships and connect with neighborly tectal cells. Anterior/nasal retina projects to posterior/caudal tectum; posterior/temporal retina to anterior/rostral tectum; dorsal to ventral; and ventral to dorsal; the result being that the visual image captured in the retina is rotated yet spatially preserved. Because this projection is reiterated in regenerative tissue of some vertebrates, and because in a series of remarkable ablation experiments, these same regenerating neurons will ignore unoccupied tectal targets and extend to appropriate posterior targets. Sperry concluded that specificity in the retinotectal projection is preprogrammed by "cytochemical identification tags" (Sperry, 1963). This notion of hard-wiring in the developing nervous system remains to this day a subject of debate, and the current data and perspectives on this subject are discussed below.

#### *Hard-wiring versus Plasticity*

In the retinotectal projection, the idea that retinal ganglion cells (RGC's) are born with a preprogrammed molecular code is one that is substantiated by a number of important studies spawned from Sperry's

landmark experiments. Perhaps the most compelling of these studies involve experimental manipulations that challenge future fates of cells when subject to inappropriate environmental cues. Early Stage 10-11 optic vesicles in chicken embryos, for example, when transplanted prior to RGC differentiation give rise to neurons that behave according to their original donor positions, rather than being re-specified by their new positions in the host eye (Dutting and Meyer, 1995). Transplantations of tectal targets (Yoon, 1971) and imposed misrouting of axons (Thanos, 1984) further reinforce the intrinsic and autonomous nature of the molecular code that specifies position in the retinotectal projection.

However, studies with several other vertebrates have progressively mounted evidence that demonstrate the plasticity of the system. Amphibians and fish, for example, are capable of regeneration and ongoing asymmetric neuronal growth in order to sustain the retinotopic map throughout life (Cline, 1991). In general, the ablation of a fraction of the retina ultimately does not result in a fraction of the tectum being naked and non-innervated. Conversely, the ablation of a fraction of the tectum ultimately does not exclude the corresponding retinal ganglion axons that would otherwise map to that missing target tissue (Gaze and Sharma, 1970; Fraser and Hunt, 1980; Hayes and Meyer, 1988). Indeed it appears as though the ultimate synaptic connections that can be made are quite flexible.

Retinotectal precision is at least in some cases, an activity-dependent process. Direct evidence that demonstrates the importance of neuronal activity in the final map comes from perturbation experiments, including tetrodotoxin (TTX) blockage of action potentials (Meyer, 1982; Meyer, 1983; Schmidt and Edwards, 1983; Fawcett and O'Leary, 1985; Reh and Constantin-Paton, 1985; Kobayashi *et al.*, 1990), stroboscopic imposition of synchronous action potentials (Schmidt and Eisele, 1985; Cook and Becker, 1990; Schmidt and Buzzard, 1990), and pharmacological manipulations that antagonize specific

transmitter/synaptic receptors (Cline and Constantine-Paton, 1989; Schmidt, 1990). In the retinal projection to the Superior Colliculus in rats, for example, early spontaneous firing patterns (Itaya *et al.*, 1995) and nitrous oxide mediated pre/post-synaptic activity-dependent mechanisms (Williams *et al.*, 1994; Wu *et al.*, 1994; Renteria and Constantine-Paton, 1996) have been implicated in the process of retinotopic modulation. As is the case with fish and frogs, partial rat retinal ablations result in an expansion of remaining retinal ganglion cell axons so that the full range of tectal targets are ultimately innervated, however unlike these lower vertebrates, the rat remodeling is remarkably unordered (Simon *et al.*, 1994). The implication of this finding is that competition and activity in at least higher vertebrates, may drive remodeling rather than intrinsic molecular code expressed originally on the target-seeking axons. The activity-dependent component generally appears to be a latent mechanism of refinement; even in TTX-blocked animals, a coarse retinotopic map is nevertheless formed, although the precision of this activity-independent map appears variable among specific vertebrates.

One of the most curious and inescapable conclusions is that evolution has arrived at dramatically different solutions for the retinotopic projection among a group of otherwise closely related vertebrate species (reviewed in Holt and Harris, 1993, for example). This seemingly unparsimonious result may generally reflect the at-large differences in visual dependency among the vertebrates. Except for the primates, whose erect posture has resulted in a nose that is off the ground and therefore a general decline in the importance of smell as the primary sensory modality, mammals are far less dependent on their visual system, especially at birth which for most, marks the beginning of a long period of parental care and protection. Perhaps not coincidentally, postnatal remodeling and activity-dependent refinement of the retinotopic map is especially prevalent in newborn mammals. Many avian species, on the other hand, rely on a fully



functional visual system at the moment of hatching in order to respond appropriately to a number of critical sign stimuli; consequently, activity-dependent remodeling is absent or at least far less prominent in postnatal chickens for example. In any case, the issue of hard-wiring and plasticity need not be mutually exclusive, and indeed the model that seems to best fit the seemingly contradictory data is one in which axon guidance and ultimate synaptic connectivity are separable, the former involving intrinsic molecular addressing and the latter involving activity-dependent refinement (reviewed in Fraser and Perkel, 1989). The differences among vertebrates, therefore, might be less related to distinct underlying molecular mechanisms, but rather merely to the relative contribution of the hard-wiring versus plasticity components of the system.

#### *Gradients versus Compartments*

If the Sperry-deduced affinity tag molecules of the retinotectal projection are largely responsible for axon navigation and establishment of the crude map of target connections, there are at least two underlying molecular mechanisms that could account for this hard-wired component. The first of these mechanistic models stems from the pioneering work of Bonhoeffer and his colleagues. In his stripe assay, in which retinal explants are given a growth choice between alternating stripes of anterior and posterior tectal tissue, temporal retinal axons demonstrably prefer anterior tectal substrates, and this preference is due to the graded distribution of an inhibitory/repulsive molecule (Walter *et al.*, 1987; Drescher *et al.*, 1995). Like lines of longitude and latitude on a globe, it is possible that two orthogonal gradients of cell surface protein expression on both retinal and tectal cells may specify crude target coordinates. Several candidate gradient molecules have been identified, including the TOP proteins (Trisler *et al.*, 1981), the Eph kinase family and their putative ligands (Holash *et al.*, 1995; Kenny *et al.*, 1995; Cheng *et al.*, 1995; Tessier-



Lavigne, 1995), and the homeobox-containing vertebrate *engrailed* homolog (Retaux *et al*, 1996; Itasaki and Nakamura, 1996), all of which have graded expression in both retina and tectum.

While gradient models are parsimonious and perhaps sufficient to account for the observed retinotopic map, a second mechanism that can not yet be excluded from consideration is compartmentalization. Compartmental boundaries of expression in some developmental systems define discrete geographies: like countries on a globe, cells recognize and respect the borders of a given segment or somite, for example. These compartments are defined by discrete patterns of gene expression, and may represent the smallest units of positional identity. In the retinal-tectal projection, the TRAP glycoprotein appears to be asymmetrically distributed, restricted in its expression to temporal retinal cells (Moskal *et al*, 1986). Perhaps TRAP is part of a combinatorial system, whereby the ordered expression of a specific array of proteins, like digits in a zip code, define that cell's destined navigational route.

## **Underlying Molecular Basis of Neuronal Connectivity**

### *Diverse Proteins that Function in Nervous System Assembly*

It is conceivable that the assembly of topographic maps does not require cell surface guidance molecules -- retinal axons, for example, might simply maintain neighborly relationships from the point of birth in the retina, through the optic nerve, ultimately filling the next available tectal synaptic space. Although in some species, maintenance of fiber order and selective fasciculation in this way, indeed is observed (reviewed in Kaprielian and Patterson, 1994; Holt and Harris, 1993), it is clear from various perturbation studies, that even when normal axon relationships are surgically disrupted, these "lost" axons nevertheless find their way to appropriate targets. What are the molecules that

might be involved with this programmed guidance and target-seeking function? Enormous progress has been made on this front, and several families of neural cell surface candidate molecules have been identified, some of which are discussed here.

*Cadherins, Integrins and the Extracellular Matrix*

Cell-cell interactions, in many systems, require  $\text{Ca}^{2+}$ . The cadherin and integrin neural cell surface proteins are two families that mediate  $\text{Ca}^{2+}$ -dependent adhesion. Cadherins play a diverse role in embryonic development, including compaction, cell adhesion, and neurite outgrowth (Blaschuk, 1990; Nose, 1990; Takeichi, 1985). Cadherin-mediated cell adhesion requires other components of the cellular architecture, as homophilic aggregation occurs in transfected cells, but not on synthetic substrates. The intracellular domains of cadherin proteins interact with the actin-associated catenins (Ozawa *et al.*, 1990), and this cytoskeletal-dependent cell adhesion component is part of the observed requirement *in vitro* for real cells versus microspheres.

Outside the cell, the integrin proteins primarily support interactions between neurons and the extracellular matrix (for example, see Albelda and Buck, 1990; Reichardt and Tomaselli, 1991). There is the potential for integrins to encode an expansive combinatorial system of ligand interactions: multiple alpha subunits combine non-covalently with multiple beta subunits in the complete heterodimeric protein (for example, Vogel *et al.*, 1990; Cheresch *et al.*, 1989). The extracellular matrix with which the integrins interact in neural development, is complex and contains a number of diverse proteins, including laminin, fibronectin and tenascin/cytotactin. Like the integrins, laminin is potentially modular, composed of multiple subunits that form specific tetramers (Sanes *et al.*, 1990); laminin is a natural substrate for axonal extension and is part of the signal that promotes neurite outgrowth (Liesi *et al.*, 1984). Fibronectin and

tenascin/cytotactin, although not exactly modular, might be described as “mosaic”, with a diverse set of functional modules within the full-length protein structure (for example, Hynes, 1990; Spring *et al.*, 1989); functionally diverse forms are further generated via developmentally-regulated alternative splicing of some of these distinct modules (for example, Weller *et al.*, 1991; Burton-Wurster *et al.*, 1989). In general, there is an enormous diversity of extracellular components that contribute to the substrate environment for target-seeking neurons, perhaps sufficient combinatorial power to account for much of the guidepost cues required for correct navigation.

#### *N-CAM Proteins*

The most extensively studied neural protein is the Neural Cell Adhesion Molecule (N-CAM). N-CAM is expressed early in the ontogeny of neurons and persists throughout development and life of the animal, and contrary to its name, is expressed in many cell types, including muscle, kidney, heart and various epithelial tissues (for example, Crossin *et al.*, 1985; Chuong and Edelman, 1985). The implication of this diverse realm of expression is that molecules involved with cell recognition may be re-used, and similar cellular “addressing” systems may underlie the assembly of many, if not all, tissues.

The “cell adhesion” concept implies that proteins like N-CAM are merely adhesive, a sort of indiscriminate “glue” on cells that causes them to stick together. Many of its functions include roles consistent with this concept, including homophilic axon-axon fasciculation and stabilization of synaptic junctions (Edelman, 1986), yet this narrow scope of N-CAM function clearly falls short of the mark. The N-CAM gene encodes a complex array of isoforms, the most significant difference among these alternative forms being the presence or extent of intracellular residues. While one of the deduced functions for the N-CAM intracellular alternative exons is to support interactions with

components of the cytoskeleton (Pollerberg *et al.*, 1987) and permit actin-dependent cell aggregation (Jaffe *et al.*, 1990), it is also probably true that the full range of N-CAM-mediated function involves classical signal transduction events that lead to developmental decisions within the cell. Perhaps the most significant evidence of this signalling capacity, is the co-receptor interaction between N-CAM and the fibroblast growth factor receptor (FGF-R), which in the absence of factor, leads to a kinase cascade and a trophic cellular response (Williams *et al.*, 1994). As it turns out, the amino acid sequences responsible for this interaction also support similar interactions with two other neural cell surface proteins, N-Cadherin and L1. Interestingly then, this common co-receptor-mediated signal transduction mechanism may underscore a widespread theme in neural development, as to how exactly, complex cell adhesion molecules exert their full range of effects.

Another common theme among neural cell surface proteins is the use of differential glycosylation, which has been particularly well-studied with the N-CAM protein. Sites for numerous carbohydrate moieties, including heparan sulfate proteoglycan and polysialic acid (PSA), have been identified in extracellular sequence motifs of virtually every neural cell surface protein identified. In general, glycosylation appears to play an inhibitory role: the carbohydrate moieties seem to simply mask amino acid functional residues until it is an appropriate developmental time to expose these residues and elicit corresponding functions. Embryonic N-CAM, for example, has high levels of PSA, and the shift from embryonic to the less-glycosylated adult N-CAM form is coincident with cessation of cell migration and axon outgrowth (Hekmat *et al.*, 1990). The degree of glycosylation also characterizes N-CAM differences proximal versus distal along retinal ganglion axons (Schlosshauer *et al.*, 1984), and in general, the conclusion is that the adhesive properties of the cell are decreased wherever the embryonic (high glycosylated) form is expressed (Hoffman *et al.*, 1982; Crossin *et al.*, 1984).

### *Neogenin and DCC*

Recently, a new and intriguing group of neural cell surface proteins has emerged which are homologous, although distant relatives to N-CAM. The neogenin and DCC proteins, like N-CAM, are members of the immunoglobulin superfamily, with extracellular immunoglobulin-like and fibronectin type III domains (Vielmetter et al., 1994). There is no relationship of the large and highly conserved intracellular domains of neogenin/DCC to any other sequence in the database, leaving open the question as to its signalling function. DCC is a putative tumor suppressor protein that is deleted in colorectal cancer; although similar roles in tumor progression have not yet been identified for the closely homologous neogenin protein, it is clear that related roles in neuronal differentiation are possible if not likely (Vielmetter et al, 1994). With so many neural cell adhesion proteins having been implicated in neurite outgrowth support (which is the initial differentiated phenotype of a neuron), and with expression profiles that correlate with the earliest point of differentiation in the ontogeny of specific cell types, it is not unreasonable to predict a general, wide-scoping role of this protein family in differentiation signalling pathways. If this is so, it is possible that CAM's may be important players in the problem of tumor progression (discussed in Johnson, 1991).

### *Kinases and Phosphatases*

If the "Holy Grail" in this field is to identify neural cell surface receptors that ultimately establish the patterned connectivity in the retinotopic map or elsewhere, then the recent identification of the Eph kinase family has understandably generated a rejuvenating excitement among those of us who were growing impatient with yet more candidates for supporting outgrowth and fasciculation. The Bonhoeffer assays and results suggesting that a repulsive molecule on posterior tectal cells was responsible for the retinal ganglion outgrowth

preferences, has led to the identification of RAGS (repulsive axon guidance signal) which indeed, is expressed predominantly in posterior tectum and causes collapse of temporal axons in culture (Walter *et al.*, 1987; Drescher *et al.*, 1995). RAGS is a member of a large ligand family, all of which interact in a promiscuous manner with Eph tyrosine kinase receptors (Holash *et al.*, 1995; Kenny *et al.*, 1995; Cheng *et al.*, 1995; Tessier-Lavigne, 1995). The ELF-1 ligand in the tectum, for example, is expressed in a counter-gradient to its Mek4 receptor in retinal ganglion cells; these matching gradients of receptor and ligand in axon and target are exactly what Sperry had predicted 50 years ago.

Where there are kinases, conventional wisdom would say, there are phosphatases too, and the recent identification of receptor-type protein tyrosine phosphatases (RPTP) in the developing nervous system adds another exciting piece to the puzzle. Tyrosine phosphorylation is now cited as important regulatory signal components of neural development (Chao, 1992; Zinn, 1993; Zipursky and Rubin, 1994); Ignelzi *et al.*, 1994; Umemori *et al.*, 1994), and is especially critical in general, for cellular responsiveness to proliferation and differentiation signals (Schlessinger and Ullrich, 1992). Receptor tyrosine phosphatases in the nervous system include RPTP- $\beta$ , RPTP- $\gamma$ , both of which contain a carbonic anhydrase (CAH) domain and fibronectin type III repeats (FNIII), and exist in multiple forms (membrane-bound versus secreted for example) with diverse post-translational modifications (isoform-specific glycosylation). The CAH domains are sufficient to bind the neural cell surface protein Contactin (Peles *et al.*, 1995); the phosphacan attachment in some isoforms interacts with Tenascin, N-CAM and Ng-CAM (Barnea *et al.*, 1994; Grumet *et al.*, 1993; Milev *et al.*, 1994); and the FNIII repeats are required for specific glial cell interactions (Peles *et al.*, 1995). So once again we are left with a complex, perhaps combinatorial array of interactions supported by a single receptor.

### *Address Molecules*

Perhaps the most compelling question is no longer “What are the underlying molecular components of neural assembly?”; to date, an astounding number of proteins have been identified, each complex and modular, and perhaps collectively, this large and diverse group is sufficient to account for the entire navigational code. Rather we are left with an even more difficult question for future neuroscience to tackle: “How is the multiplicity of these signals, inside the cell and out, integrated into the emergent property that is a functional web of neural connectivity?” In other words, how are these hundreds of puzzle pieces made sense of, so that a concert of axon growth is choreographed? Indeed, we remain seemingly distant from ultimately cracking the code. At a point in history when very few of the molecular components had been elucidated, our laboratory set out to crack this code from an admittedly different perspective: it was our vision that patterns of expression from as yet unidentified cell surface protein families would make obvious the various compartments, gradients and mapping principles that underlie this code.

If such a preprogrammed code exists, it might be appropriately described as an “address” or “zip code” molecular system. Just like a unique array of numbers in a given order can target an addressed letter to a unique mail box, so too might the unique array of molecules in a given temporal efficacy define the precise route to the ganglion cell target. Interestingly, some studies point to the fact that in the retinocollicular pathway in cats for example, nasal retinal cells that inappropriately project ipsilaterally nevertheless locate the appropriate target region, only on the wrong side of the brain (Chalupa et al, 1996). To the postal worker, this result may be familiar: a single erroneous zip number might cause routing to the wrong state, yet regardless, the remaining numbers still sufficiently encode its delivery to an otherwise appropriate destination. The implication of this observation is that ganglion cell axons may have numerous molecular digits in the



“zip code”, some which define ipsilateral versus contralateral choice, others which independently define the various additional pathfinding choices that ultimately complete the navigation to a specific compartmental destination in the tectum. As discussed previously here, this compartmental view of axon guidance is in contrast to gradient models; while the search continues in a number of laboratories to elucidate the role of gradient molecules in the retinotopic projection, we have rather initiated efforts to search for putative “address” proteins. These efforts are described in part, in the following section.

#### *Methods and Initial Experiments*

To investigate the role of molecular “addressing” in the retinotectal projection in chicken development, our laboratory has developed some new methods to identify cell surface proteins whose expression is compartmentalized, or otherwise consistent with Sperry’s original idea of cytochemical affinity tags. The method involves two components: first, the isolation of cell surface proteins from the far more abundant intracellular fractions, and second, the generation of monoclonal antibodies against presumed families of this enriched cell surface fraction. These methods are described in detail elsewhere (Kayyem et al, 1992b; also, see Figure 1), but briefly involve: extraction of intact retinal or tectal tissue and amino-labeling of biotin to exposed surface residues prior to lysis in a 5% detergent cocktail. Biotinylated cell surface protein is at least 1000-fold enriched from unlabelled intracellular protein via subsequent avidin chromatography, and this cell surface isolate was further fractionated by size via HPLC. The latter size fractionation step provides two advantages: 1) the reduction of complexity of the immunogen, thereby reducing the probability of immuno-dominant protein species being present in any given antigen mixture, and 2) the increased likelihood of any given immunogen being exclusively of a particular receptor family. On the former point,



previous efforts to raise monoclonal antibodies using a complex mixture of proteins has resulted in the majority of the mouse response being against a small number of abundant and immunodominant antigens, such as N-CAM; rarer, less dominant, and perhaps more interesting antigens have been therefore difficult to identify. On the latter point, two-dimensional gel electrophoresis of the cell surface fractions used as immunogens illustrate the potency of these methods: cell surface protein spots in an isoelectric array are evident at various sizes, indicating the possible enrichment and isolation of discrete protein families (see Figure 2).

The developmental screen for candidate "address" antigens was initially confined to the 120-130 kD immunogen (because at the time these studies were initiated, the 120 kD Ng-CAM protein was one of the only known neural cell surface antigen, and therefore could serve as a positive control for the methods). The HPLC-purified cell surface fraction corresponding to this initial inoculant is shown in Figure 1. The fusion was enormously successful, perhaps in part due fortuitously to the extra covalently-bound biotin group that was used in the purification step; the small biotin moiety may act as a hapten, and might have therefore greatly enhanced the mouse immune response. A large number of developmentally interesting expression patterns was identified, including antibodies against virtually every known cell surface protein of the Ng-CAM family. Western blot analysis of these and other as yet unidentified antigens illustrate in every case, disulfide-dependent epitopes, further suggesting that all antigens might be of a common family (Figure 3). Particular antigens were chosen for further histological and biochemical analysis; the monoclonal antibody itself was used to expression-clone the cDNA coding these antigens from embryonic brain libraries. From my own efforts on this front, the genes encoding chicken Bravo/Nr-CAM and Contactin were cloned (unpublished), both of which are members of the rapidly growing Immunoglobulin Supergene family.

## Conclusions

Many of the puzzle pieces are now on the table: the underlying molecular components so far identified are diverse and complex, mosaic and modular, and collectively generate sufficient combinatorial power to account for a significant fraction of the neural connectivity problem in development. The future of neuroscience is not unlike the future of many developmental systems. It is now apparent that complex developmental problems are not singular pathways involving a linear relationship among genes. Rather, it appears as though patterns of gene expression and overlapping weighted contributions ultimately are integrated by the cell in order to shape the ultimate response. Whether the question involves the field of signal transduction, and the multiplicity of receptor activations, or involves the field of gene transcription, and the multiplicity of nuclear factors, it is becoming increasingly clear that the interpretative power of the cell remains a mystery. The long road ahead is one that begins to elucidate how, exactly, a neuron integrates its complex, noisy environment.

Chapter I  
References

Albelda, S.M. and Buck, C.A. (1990). Integrins and other cell adhesion molecules. *FASEB J.* 4, 2868-2880.

Barnea, G., Silvennoinen, O., Shaanan, B., Honegger, A.M., Canoll, P.D., D'Eustachio, P., Morse, B., Levy, B., La Forgia, S., Huebner, K., Musacchio, J.M., Sap, J., and Schlessinger, J. (1993). Identification of a carbonic anhydrase-like domain in the extracellular region of RPTP-gamma defines a new subfamily of receptor tyrosine phosphatases. *Mol. Cell. Biol.* 13, 1497-1506.

Blashuk, O.W., Sullivan, R., David, S., and Pouliot, Y. (1990). Identification of a cadherin cell adhesion recognition sequence. *Dev. Biol.* 139, 227-229.

Burton-Wurster, N., Lust, G., and Wert, R. (1989). Expression of the ED B fibronectin isoform in adult human articular cartilage. *Biochem Biophys. Res. Comm.* 165, 782-787.

Chalupa, L.M., Snider, C.J., and Kirby, M.A. (1996). Topographic organization in the retinocollicular pathway of the fetal cat demonstrated by retrograde labeling of ganglion cells. *Journal of Comp. Neurology* 368, 295-303.

Chao, M.V. (1992). Neurotrophin receptors: a window into neuronal differentiation. *Neuron* 9, 583-593.

Cheng, H.J., Nakamoto, M., Bergemann, A.D., and Flanagan, J.G. (1995). Complementary gradients in expression and binding of ELF-1 and

Mek4 in development of the topographic retinotectal projection map. *Cell* 82, 371-381.

Cheresh, D.A., Smith, J.W., Cooper, H.M., and Quaranta, V. (1989). A novel vitronectin receptor integrin is responsible for distinct adhesive properties of carcinoma cells. *Cell* 57, 59-69.

Chuong, C.M., and Edelman, G.M. (1985). Expression of cell-adhesion molecules in embryonic induction II: morphogenesis of adult feathers. *J. Cell Biol.* 101, 1027-1043.

Cline, H.T. (1991). Activity-dependent plasticity in the visual systems of frogs and fish. *Trends Neurosci.* 14, 104-111.

Cline, H.T. and Constantine-Paton, M. (1989). NMDA receptor antagonists disrupt the retinotectal topographic map. *Neuron* 3, 413-426.

Cook, J.E., and Becker, D.L. (1990). Spontaneous activity as a determinant of axonal connections. *Eur. J. Neurosci.* 2, 162-169.

Crossin, K.L., Chuong, M.C., and Edelman, G.M. (1985). Expression sequences of cell adhesion molecules. *Proc. Natl. Acad. Sci. USA* 82, 6942-6946.

Drescher, U., Kremoser, C., Handwerker, C., Loschinger, J., Noda, M., and Bonhoeffer, F. (1995). *In vitro* guidance of retinal ganglion cell axons by RAGS, a 25 kDa tectal protein related to ligands for Eph receptor tyrosine kinases. *Cell* 82, 359-370.

Dutting, D. and Meyer, S.U. (1995). Transplantations of the chick eye anlage reveal an early determination of nasotemporal polarity. *International Journal of Developmental Biology* 39 (2), 252-259.

Edelman, G.M. (1986). Cell adhesion molecules in neural histogenesis. *Ann Rev. Physiol.* 48, 417-430.

Fawcett, J.W. and O'Leary, D.D.M. (1985). The role of electrical activity in the formation of topographic maps in the nervous system. *Trends Neurosci.* 8, 201-206.

Fraser, S.E. and Hunt, R.K. (1978). Neuroplasticity in *Xenopus*. *Biophys. J.* 21, 110a.

Fraser, S.E. and Hunt, R.K. (1980). Retinotectal specificity: models and experiments in search of a mapping function. *Ann. Rev. Neurosci.* 3, 319-352.

Fraser, S.E. and Perkel, D.H. (1989). Competitive and positional cues in the patterning of nerve connections. *Journal of Neurobiology*, 21, 51-72.

Gierl, A., Saldler, H., Peterson, P.A. (1989). Maize transposable elements. *Annu. Rev. Genet.* 23,, 71-85.

Grumet, M., Flaccus, A., and Margolis, R.U. (1993). Functional characterization of chondroitin sulfate proteoglycans of brain: interactions with neurons and neural cell adhesion molecules. *J. Cell Biol.* 120, 815-824.

Haselkorn, R. (1992). Developmentally regulated gene rearrangements in prokaryotes. *Annu. Rev. Genet.* 26, 113-130.

Hayes, W.P. and Meyer, R.L. (1988). Optic synapse number but not density is constrained during regeneration onto surgically-halved tectum in goldfish: HRP-EM evidence that optic fibers compete for fixed numbers of post-synaptic sites on the tectum. *J. Comp. Neurol.* 274, 539-559.

Hekmat, A.D., Bitter-Seurmann, D., and Schachner, M. (1990). Immunocytochemical localization of the highly polysialylated form of the neural cell adhesion molecule during development of the murine cerebellar cortex. *J. Comp. Neurol.* 291, 458-467.

Hoffman, S., Sorkin, B.C., White, P.C., Brackenbury, R., Mailhammer, R., Rutishauser, U., Cunningham, B.A., and Edelman, G.M. (1982). Chemical characterization of a neural cell adhesion molecule purified from embryonic brain membranes. *J. Biol. Chem.* 257, 7720-7729.

Holash, J.A., and Pasquale, E.B. (1995). Polarized expression of the receptor protein-tyrosine kinase *Cek5* in the developing avian visual system. *Developmental Biology* 172, 683-693.

Holt, C.E. and Harris, W.A. (1993). Position, guidance, and mapping in the developing visual system. *J. Neurobiol.* 24, 1400-1422.

Hynes, R.O., ed. 1990. *Fibronectins*. Springer-Verlag, New York.

Ignelzi, M.A.J., Miller, D.R., Soriano, P., and Maness, P. (1994). Impaired neurite outgrowth of *src*-minus cerebellar neurons on the cell adhesion molecule L1. *Neuron* 12, 873-884.

Itasaki, N. and Nakamura, H. (1996). A role for gradient *en* expression in positional specification on the optic tectum. *Neuron* 16 (1), 55-62.

Itaya, S.K., Fortin, S., and Molotchnikoff, S. (1995). Evolution of spontaneous activity in the developing rat superior colliculus. *Canadian Journal of Physiology and Pharmacology* 73, 1372-1377.

Jaffe, S.H., Friedlander, D.R., Matsuzaki, F., Crossin, K.L., Cunningham, B.A., and Edelman, G.M. (1990). Differential effects of the cytoplasmic domains of cell adhesion molecules on cell aggregation and sorting-out. *Proc. Natl. Acad. Sci. USA* 87, 3589-3593.

Johnson, J.P. (1991). Cell adhesion molecules of the immunoglobulin supergene family and their role in malignant transformation and progression to metastatic disease. *Cancer Metast. Rev.* 10, 11-22.

Kaprielian, Z. and Patterson, P.H. (1994). The molecular basis of retinotectal topography. *Bioessays* 16 (1), 1-10.

Kayyem, J.F., Roman, J.M., de la Rosa, E.J., Schwarz, U., and Dreyer, W.J. (1992). Bravo/Nr-CAM is closely related to the cell adhesion molecules L1 and Ng-CAM and has a similar heterodimer structure. *J. Cell Biol.* 118, 1259-1270.

Kayyem, J.F., Roman, J.M., Von Boxberg, Y., Schwarz, U., and Dreyer, W.J. (1992b). A method for the generation of monoclonal antibodies against rare cell-surface molecules. *Eur. J. Biochem* 208, 1-8.

Kenny, D., Bronner-Fraser, M., and Marcelle-C. (1995). The receptor tyrosine kinase Qek5 messenger RNA is expressed in a gradient within the neural retina and the tectum. *Developmental Biology* 172 (2), 708-716.

Klar, A.J.S. (1990). Regulation of fission yeast mating-type interconversion by chromosome imprinting. *Development (S)*, 3.

Kobayashi, T., Nakamura, H., and Yasuda, M. (1990). Disturbance of refinement of retinotectal projection in chick embryos by tetrodotoxin and grayanotoxin. *Dev. Brain Res.* 57, 29-35.

Liesi, P., Dahl, D., and Vaheri, A. (1984). Neurons cultured from developing rat brain attach and spread preferentially to laminin. *J. Neurosci. Res.* 11, 241-251.

Matsuoka, M., Nagawa, F., Okazaki, K., Kingsbury, L., Yoshida, K., Muller, U., Larue, D.T., Winer, J.A., Sakano, H. (1991). Detection of somatic recombination in the transgenic mouse brain. *Science* 254 (5028), 81-86.

Meyer, R.L. (1982). Tetrodotoxin blocks the formation of ocular dominance columns in goldfish. *Science* 218, 589-591.

Meyer, R.L. (1983). Tetrodotoxin inhibits the formation of refined retinotopography in goldfish. *Dev. Brain Res.* 6, 293-298.

Milev, P., Friedlander, D., Sakurai, T., Karthikeyan, L., Flad, M., Margolis, R.K., Grumet, M., and Margolis, R.U. (1994). Interactions of the chondroitin sulfate proteoglycan phosphacan, the extracellular domain of a receptor-type tyrosine phosphatase, with neurons, glia, and neural cell adhesion molecules. *J. Cell Biol.* 127, 1703-1715.

Moskal, J.R., Trisler, D., Schneider, M.D., and Nirenberg, M. (1986). Purification of a membrane protein distributed in a topographic gradient in chicken retina. *Proc. Nat. Acad. Sci. USA* 83, 4730-4733.



Reichardt, L.F., and Tomaselli, K.J. (1991). Extracellular matrix molecules and their receptors: functions in neural development. *Annu. Rev. Neurosci.* 14, 531-570.

Renteria, R.C. and Constantine-Paton, M. (1996). Exogenous nitric-oxide causes collapse of retinal ganglion cell axonal growth *in vitro*. *Journal of Neurobiology* 29 (4), 415-428.

Retaux, S., McNeill, L. and Harris, W.A. (1996). *Engrailed*, retinotectal targeting, and axonal patterning in the midbrain during *Xenopus* development - an antisense study. *Neuron* 16, 63-75.

Sanes, J.R., Engvall, E., Butkowski, R., and Hunter, D.D. (1990). Molecular heterogeneity of basal lamina: isoforms of laminin and collagen IV at the neuromuscular junction and elsewhere. *J. Cell Biol.* 111, 1685-1699.

Schatz, D.G., and Chun, J.J.M. (1992). V(D)J recombination and the transgenic blues. *New Biologist* 4 (3), 188-196.

Schlessinger, J., and Ullrich, Y. (1992). Growth factor signaling by receptor tyrosine kinases. *Neuron* 9, 383-391.

Schlosshauer, B., Schwarz, U., and Rutishauser, U. (1984). Topological distribution of different forms of neural cell adhesion molecule in the developing chick visual system. *Nature* 304.

Schmidt, J.T. (1990). Long-term potentiation and activity-dependent retinotopic sharpening in the regenerating retinotectal projection of goldfish: common sensitive period and sensitivity to NMDA blockers. *J. Neurosci.* 10, 233-246.

Schmidt, J.T. and Buzzard, D.L. (1990). Activity-driven sharpening of the retinotectal projection in goldfish: development under stroboscopic illumination prevents sharpening. *J. Neurosci.* 24, 384-399.

Schmidt, J.T. and Edwards, D.L. (1983). Activity sharpens the map during the regeneration of the retinotectal projection in goldfish. *Brain Res.* 209, 29-39.

Schmidt, J.T. and Eisele, L.E. (1985). Stroboscopic illumination and dark rearing block the sharpening of the regenerated retinotopic map in goldfish. *Neuroscience* 14, 535-546.

Simon, D.K., Roskies, A.L., and O'Leary, D.D.M. (1994). Plasticity in the development of topographic order in the mammalian retinocollicular projection. *Developmental Biology* 162, 384-393.

Sperry, R.W. (1963). Chemoaffinity in the orderly growth of nerve fiber patterns and connections. *Proc. Natl. Acad. Sci. USA* 50, 703-710.

Spring, J., Beck, K., and Chiquet-Ehrismann, R. (1989). Two contrary functions of tenascin: dissection of the active sites by recombinant tenascin fragments. *Cell* 59, 325-334.

Takeichi, M., Hatta, K., and Nagafuchi, A. (1985). Selective cell adhesion mechanisms: role of the calcium-dependent cell adhesion system. *In* *Molecular Determinants of Animal Form*, G.M. Edelman, ed. Alan R. Liss, New York. 223-233.

Thanos, S.F., Bonhoeffer, F., and Rutishauser, U. (1984). Fiber-fiber interaction and tectal cues influence the development of the chicken retinotectal projections. *Proc. Natl. Acad. Sci. USA* 81, 1906-1910.

Tessier-Lavigne, M. (1995). Eph receptor tyrosine kinases, axon repulsion, and the development of topographic maps. *Cell* 82, 1995.

Trisler, D.M., Schneider, M.D., and Nirenberg, M. (1981). A topographic gradient of molecules in retina can be used to identify cell position. *Proc. Natl. Acad. Sci. USA* 78, 2145-2149.

Umemori, H., Sato, S., Yagi, T., Aizawa, S., and Yamamoto, T. (1994). Initial events of myelination involve Fyn tyrosin kinase signaling. *Nature* 367, 572-576.

Vielmetter, J., Kayyem, J.F., Roman, J.M., and Dreyer, W.J. (1994). Neogenin, a cell surface protein expressed during terminal neuronal differentiation, is closely related to the human tumor suppressor molecule Deleted in Colorectal Cancer. *J. Cell Biol.* 127, 2009-2020.

Vogel, B.E., Tarone, G., Goamcotto, F.G., Gailit, J., and Ruoslahti, E. (1990). A novel fibronectin receptor with an unexpected subunit composition. *J. Biol. Chem.* 265, 5934-5937.

Walter, J.S., Henke-Fahle, S., and Bonhoeffer, F. (1987). Avoidance of posterior tectal membranes by temporal retinal axons. *Development* 101, 909-913.

Weller, A., Beck, S., and Ekblom, P. (1991). Amino acid sequence of mouse tenascin and differential expression of two tenascin isofors during embryogenesis. *J. Cell Biol.* 112, 355-362.

Williams, C.V., Nordquist, D., and McLoon, S.C. (1994). Correlation of nitric-oxide synthase expression with changing patterns of axonal projections in the developing visual system. *Journal of Neuroscience* 14 (3), 1746-1755.

Williams, E.J., Furness, J., Walsh, F.S., and Doherty, P. (1994). Activation of the FGF receptor underlies neurite outgrowth stimulated by L1, N-CAM, and N-Cadherin. *Neuron* 13, 583-594.

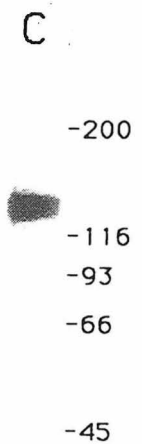
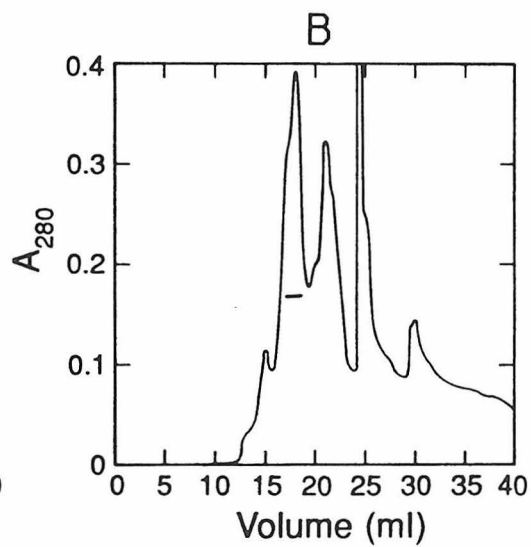
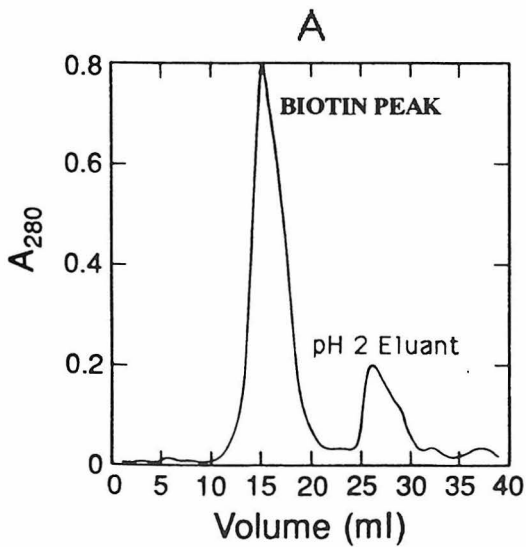
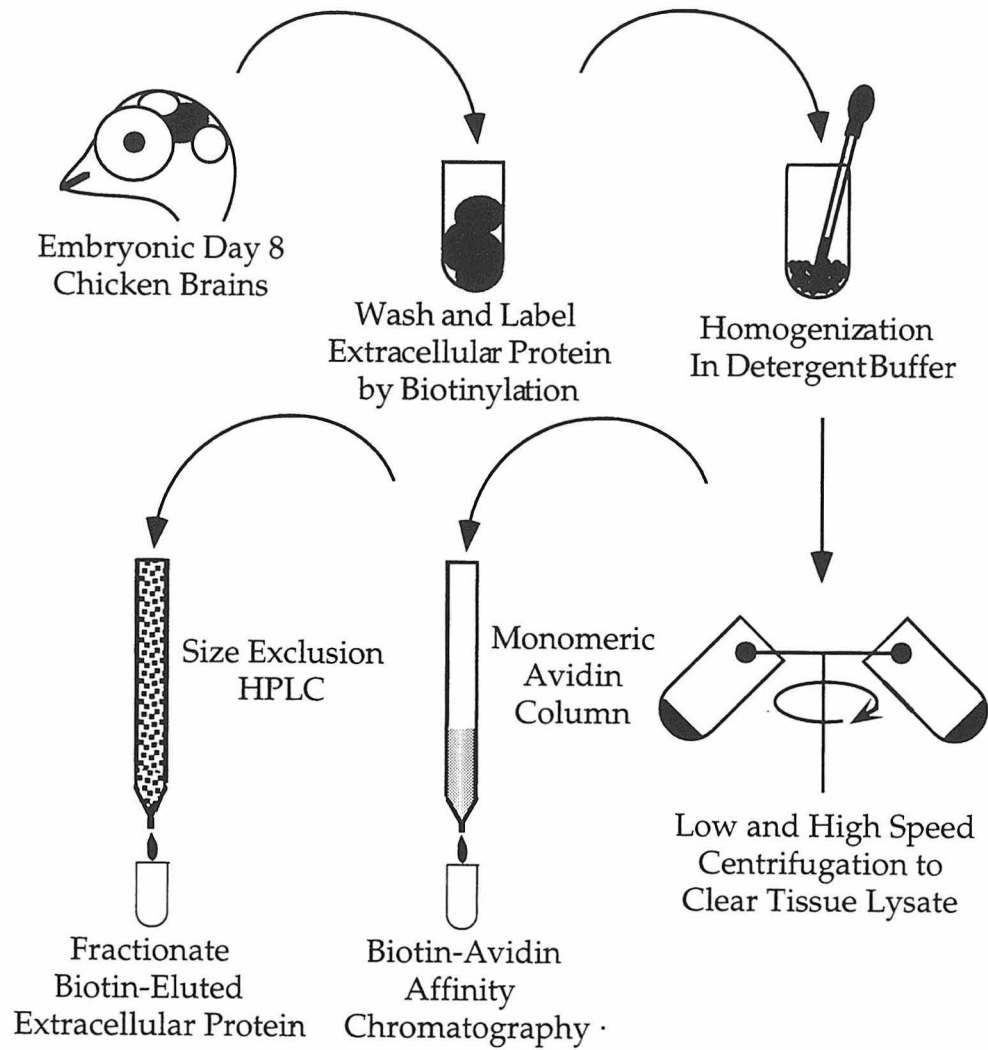
Wu, H.H., Williams, C.V., and McLoon, S.C. (1994). Involvement of nitric-oxide in the elimination of a transient retinotectal projection in development. *Science* 265, 1593-1596.

Yoon, M.G. (1971). Reorganization of retinotectal projection following surgical operations on the optic tectum in goldfish. *Expl. Neurol.* 33, 395-411.

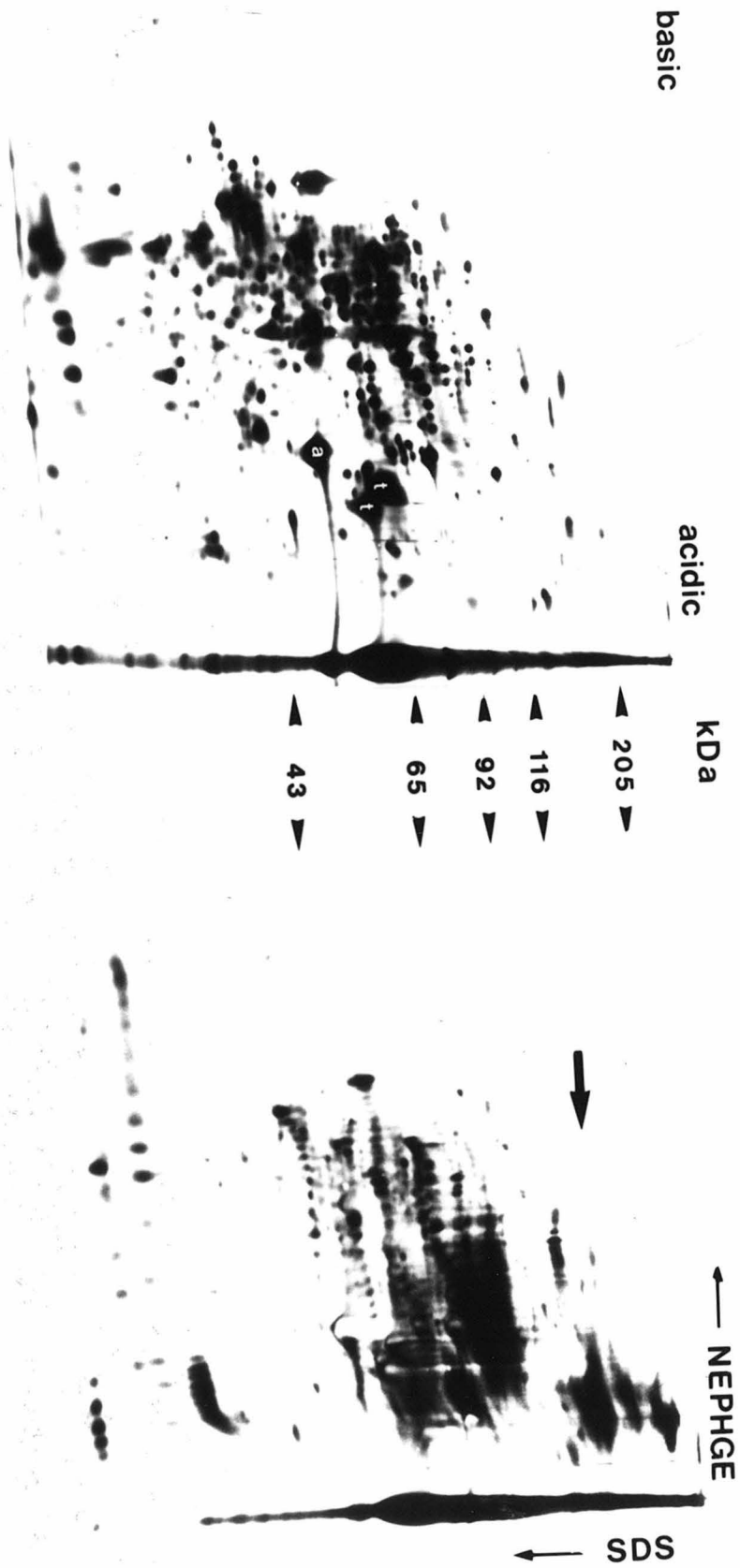
Zinn, K. (1993). *Drosophila* protein tyrosine phosphatases. *Semin. Cell Biol.* 4, 397-401.

Zipursky, S.L., and Rubin, G.M. (1994). Determination of neuronal cell fate: lessons from the R7 neuron of *Drosophila*. *Annu. Rev. Neurosci.* 17, 373-397.

**Fig. 1.**        *Method to Purify Cell Surface/Extracellular Protein.* As depicted in the cartoon, cell surface/extracellular protein was enriched at least 1000-fold (see Fig. 2) by surface biotinylation prior to lysis and subsequent purification of labeled protein by avidin chromatography. In order to raise monoclonal antibodies against putative receptor families, FPLC was used to separate cell surface protein into discrete size fractions; size fractionation also reduces the complexity of the immunogen which decreases the chance of especially abundant or immunodominant antigens dominating the immune response in any specific mouse. Below the cartoon, the optical density traces for both biotin (A) and HPLC (B) chromatography are shown (from Kayyem et al., 1992b). The peak corresponding to free-biotin (which preceded low-pH elution) illustrates that the majority of protein is specifically-eluted, and this peak was further fractionated by size via HPLC. (C) Blot stained with streptavidin alkaline phosphatase shows the avidin-purified/HPLC-fractionated 120-130 kD antigen mixture used for the initial immunogen/fusion (the corresponding HPLC peak is indicated in B).



**Fig. 2.**        *Two-dimensional gels..* E8 tectum was biotinylated and subjected to NEPHGE in the first dimension (right to left) and SDS PAGE in the second dimension (top to bottom). Left: Total protein pattern is blotted and stained with colloidal gold. The positions of actin (a) and tubulins (t) are indicated. Right: Autoradiograph of <sup>35</sup>S-streptavidin labeled blot reveals the patterns of biotinylated surface proteins. The arrow indicates the bulk of 120-140 kD proteins used in initial experiments. Note that most of the proteins in the total lysate are not visualized in (or would be subtracted away from) the cell surface fraction; it is estimated that the enrichment of surface proteins exceeds 1000-fold by the method described in Figure 1. This figure is from Kayyem et al., 1992b.





**Fig. 3.**        *Western blots.* In all lanes, approximately 1/10,000 of a brain's worth of P2 chicken total lysate was run on 10% SDS PAGE under non-reducing conditions. Blots are incubated with various monoclonal antibodies from the initial fusion, and visualized by using an AP-conjugated anti-mouse secondary antibody. Each lane is labeled (at the base) to indicate the specific antibody clone, and the antibodies that have been subsequently identified as Ng-CAM, Bravo, Neogenin, DM-Grasp, Contactin and Axonin-1 are noted. While there is some diversity in antigen apparent molecular weight (between 110-200 kD), each of these antigen epitopes is sensitive to reducing versus non-reducing conditions (no visible staining when disulfide bonds are compromised) which may indicate that they are of a common protein family. All antigens that have been characterized to date are indeed, members of the Ig superfamily whose component Ig-like domains require intact cysteine disulfide bonds for proper structural folding (and presumably, native epitope conformation).

NgCAM	
Bravo	
Axonin	
4-62G2	
DM-Grasp	
4-66A2	
4-60F1	
4-62G2	
Contactin	
4-64F6	
5-2B12	
5-1F3	
5-5-1C10	
4-63G4	
Neogenin	
4-66C1	

Chapter II

Evolutionary Analysis of the Six Immunoglobulin Domain Subfamily  
of Neural Cell Adhesion Molecules Suggests Important Ligand  
Interactions May Map Between Pairs of Domains.

Robert P. Lane, Jost Vielmetter, William J. Dreyer, and Michael W. Frohlich

## Introduction

The immunoglobulin supergene family (IgSF) is a large and diverse group of proteins that share in common the immunoglobulin (Ig) domain (Williams and Barclay, 1988). Ig domains are approximately 100 amino acids in length and form two opposing beta sheets. The overall structure is stabilized by characteristically spaced hydrophobic (cysteine and tryptophan) core residues (reviewed in Vaughn and Bjorkman, 1996). Although the domain is named for its presence in immunoglobulins in the immune system, a homologous domain structure has been identified in a number of proteins in various tissues including the nervous system of vertebrates and invertebrates. The nervous system subfamily of the IgSF, in addition to Ig-like domains, also often contain a number of more C-terminal fibronectin type III repeats (Fn domains) in the extracellular portion of the protein. Structurally, Fn domains closely resemble Ig-like domains: two stacked beta sheets, one with four beta strands, the other with three beta strands, connected by loops that organize a similar but not identical topology. Fn and Ig-like domains are stabilized by different hydrophobic core residues and therefore are almost certainly not evolutionarily homologous, but rather have arrived at this similar beta-barrel fold by means of convergent evolution. In every case where they have been examined, Ig-like and Fn domains are found on extracellular portions of proteins and have been selected for their capacity to carry out intermolecular interactions, and collectively, the neural subfamily of the IgSF are included in a large group of extracellular proteins that are referred to as Cell Adhesion Molecules (CAM's).

It is certainly no coincidence that the immune system and nervous system share many molecular characteristics. Perhaps no other developing systems have quite the same daunting task of sorting out the sheer number of required molecular contacts: immunity

requires the recognition of enormous numbers of antigens and pathogens; the developing human nervous system must sort out approximately  $10^{15}$  synaptic connections. In this regard, both systems must have evolved a large and diverse set of cell surface proteins that permit this order of specificity, and although the IgSF set of genes is at this point one of only a few bridges between the two systems, their common cell recognition problem may ultimately lead to other intriguing parallels. The nervous system subfamily of the IgSF is a rapidly growing family on its own account, and numerous developmental functions that pertain to the cell recognition problem of correct assembly of neuronal connectivity have already been attributed to members of the family, including the regulation of neurite extension, axon pathfinding, and synapse formation (reviewed in Goodman and Schatz, 1993).

In this study, we have focused on a neural subfamily of closely-related genes that include: Neuroglian (*Drosophila*), Bravo/Nr-CAM (chicken and human), Neurofascin (chicken and rat), L1 (mouse, rat, human), Ng-CAM (chicken), Contactin/F11 (chicken, mouse, human), Axonin-1/Tag-1 (chicken, rat, human), Tag-like/Big-1 (mouse, rat), and Big-2 (rat). Each of these genes encode six Ig-like domains at the N-terminus of the protein. About half of these proteins (Contactin/F-11, Tag-1/Axonin-1, Tag-like/Big-1 and Big-2) contain four Fn domains and are glycosylphosphatidylinositol (GPI)-linked to the cell membrane, while the other half (Neuroglian, Neurofascin, Bravo/Nr-CAM and L1/Ng-CAM) have five Fn domains and a well-conserved transmembrane and intracellular domain. In a few of these proteins, the intracellular domains interact with cytoskeletal components of the cell (Davis and Bennett, 1994; Otsuka *et al.*, 1995), and although little is known about other interactions inside the cell, there is abundant evidence that suggest these domains may be involved with various signal transduction pathways (see for example: Doherty *et al.*, 1991; Atashi *et al.*, 1992; Williams *et al.*, 1995; Klinz *et al.*, 1995; Goldman *et*

*al.*, 1996; Wong *et al.*, 1996). The extracellular Ig-like and Fn domains of several members of this subfamily are known to bind to each other, and both heterophilic and homophilic counter-receptor interactions have been characterized (see for example: Mauro *et al.*, 1992; Morales *et al.*, 1993; Suter *et al.*, 1995). Generally, it is likely that these large and complex proteins are able to carry out multiple functions involving a diverse set of ligands and signals both between cells and among proteins on the same cell.

In order to gain further insight into the functionally important residues, we have initiated studies on the evolutionary history and molecular diversification that has resulted in this large, closely-related gene family. The translated cDNA sequences for each of the IgSF homologs were obtained from Genbank databases. Cladistic software, such as Paup and MacClade utilized here, are powerful tools that reconstruct phylogenetic trees and in doing so, states of ancestral nodes within the tree. Cladistic analyses and phylogeny reconstruction is accomplished by the criterion of parsimony, i.e., by minimizing the total amount of change needed to account for modern character states. In this way, molecular evolutionary trees may be generated that reflect the probable phylogenetic history for families of genes.

The results of such analyses depends critically on the correct identification of evolutionarily homologous characters: in the specific case of molecular evolution, sequence alignment is the most important initial step because alignment specifies which amino acids in the protein sequence are considered homologous (i.e., share a common ancestor). In this study, we have taken advantage of an extensive structural databases on Ig-like and Fn domains to infer tertiary structure, with the tight constraints this places on alignments. This approach of identifying structurally homologous characters (i.e., specific residues that occupy the same place in similar structures) allows alignment with greater confidence than can be achieved by methods that rely on sequence alone.

With carefully determined amino acid alignments for this neural Ig subfamily, we utilized Paup and MacClade phylogenetic software to ask four evolutionary questions: 1) What is the phylogenetic relationship among these full-length proteins? 2) What are the evolutionary relationships among individual Ig-like and among Fn domains? 3) What do patterns of evolutionary conservation and divergence suggest about functionally important domains for specific proteins in the family? 4) What do patterns of evolutionary conservation and divergence suggest about particular beta strand and loop structural motifs within domains in terms of mapping function to domain topology? Our results suggest that the individual domains are probably not distinct and independent modules of function, but rather that they may act in tandem. In this regard, we propose a model whereby important interactions may map to regions between domains, with the "top" loops of one domain, the "bottom" loops of the adjacent, more N-terminal domain, and the interdomain residues forming a ligand "pocket". We also evaluate the controversial identification of Ng-CAM (chicken) and L1 (mammals) genes as functional orthologs in the light of avian-mammalian molecular diversification in this family. Finally, we analyze the evolutionary history of the fibronectin type III repeats and discuss the significance of these domains as distinguished from their better known Ig-like domain partners.

## Materials and Methods

### *Sequence Alignments and Structural Modeling*

Coding Sequences for members of the neural Ig subfamily of proteins that are predicted to contain six immunoglobulin-like domains were obtained from Genbank databases: *Drosophila* Neuroglian (Accession Code M28231); chicken and human Bravo/Nr-

CAM (Accession Codes L08960 and U55258 respectively); chicken and rat Neurofascin (Accession Codes X65224 and S76791 respectively); rat, mouse and human L1 (Accession Codes X59149, X12875 and M74387 respectively); chicken Ng-CAM (Accession Code X56969); chicken, mouse and human Contactin/F11 (Accession Codes X14877, X14943 and U07819 respectively); chicken, rat and human Tag-1/Axonin-1 (Accession Codes X63101, M31725 and X68274 respectively); mouse and rat Tag-like/Big-1 (Accession Codes L01991 and U11031); and rat Big-2 (Accession Code U35371). Auxiliary studies described in the Results section also included alignment of N-CAM sequences (Accession Codes M76710, X70342, X06564, Y00051, X55322).

In all studies, sequences were preliminarily aligned using GCG Pileup package software. The resulting alignment outputs were subsequently modified to reflect protein structure data for individual domains in the following ways: 1) known structural homologous core residues for various beta strands in Ig-like and fibronectin type III structures (reviewed in Vaughn and Bjorkman, 1996) were manually forced in alignment, 2) highest priority was given to alignment of alternating hydrophobic residues in the beta strand structures because it is most likely that these structurally homologous amino acids are also evolutionarily homologous (i.e., it is unlikely that the position of core residues has shifted in evolution), 3) the number of gaps was minimized under the presumption that elimination/creation of coding sequence is generally far less frequent/far more expensive a change than any given amino acid substitution, and 4) gaps in sequences were generally forced to exist in predicted loop structures as opposed to beta strands in order to reflect this empirical tendency in solved Ig-like and Fn domains (although 1-2 amino acid long beta-bulges in strands were permitted). All alignments used in this study are shown in Table 1 (at the end of this paper).

In order to further increase the confidence of alignments, all amino acid residues found outside structurally predictable landmarks



were excluded from the analysis: interdomain sequences align poorly and were therefore not included, nor were sequences beyond the highly conserved core residues (i.e., the cysteines in Ig-like domains; the tryptophans and tyrosines in Fn domains) for alignment of distantly related N-CAM proteins. In cases where proteins with a different number of domains (for example, N-CAM has one fewer Ig-like domain and two/three fewer Fn domains) were aligned, the Pileup output maximizing sequence similarity was used to determine which specific domains should be regarded as "missing" in the protein with fewer domains (i.e., to indicate where the domain-gaps should be placed). Finally, because interpretations are so tightly coupled to specific alignment choices, the analysis was performed with two different alignment sets, the alternative alignment (not shown) typically placed a higher priority on maximizing sequence similarity rather than minimizing the number of gaps. In this way, conclusions can be tested for how robust they are (i.e., how stable they are given other reasonable alignment choices).

#### *Cladistics and Generation of Phylogenetic Trees*

The above amino acid sequence alignments were analyzed by two phylogenetic software packages: Paup 3.1.1 and MacClade (Sinauer Associates, Inc). Evolutionary relationships among the 18 neural subfamily proteins that each contain six Ig-like domains were generated by a Paup heuristic search. MacClade was used to investigate properties of trees found by Paup, in particular, the cost to move specific branches in order to assess topological stability (see "Examining the Tree Space" below). All trees shown are "phylograms"; i.e., the length of horizontal branches is proportional to the amount of change along these branches.

All studies involved both an "Identity" and "Similarity" search for shortest phylogenetic trees. Identity searches score branch lengths in such a way that all amino acid changes are given the same cost of

one step. Similarity searches score branch lengths in such a way that specific amino acid changes are given different costs relative to others, and these searches therefore reflect a more highly resolved assessment of change. The relative costs utilized per amino acid change were obtained from a modified Blossum62 matrix, a similarity matrix commonly used to score sequence relationships (used in the GCG Blast and fasta algorithms, for example). For these studies, the matrix was altered to eliminate costs for identities (i.e., no relative penalty is assessed for unchanged amino acids along a phylogenetic tree branch). Gaps in sequences were considered as expensive as the least conservative amino acid substitution (i.e., assigned the maximum number of steps=15), although missing sequences (i.e., domain gaps such as the missing Fn 5 and intracellular domains in some proteins) were not penalized. All amino acid changes in the stepmatrix are scored between 8 and 15, the higher numbers reflecting the empirically more "expensive" nature of specific changes. This version of the Blossum62 stepmatrix is conservative and closely approximates (within 50%) what is otherwise obtainable by identity searches. While the conservative nature of this modified Blossum62 matrix minimizes the noise that an unmodified matrix is likely to contain (the matrix is statistical, and includes irrelevant data on gene families which have nothing to do with Ig-like and Fn domain structures -- see Discussion), the resolution in these similarity trees is nevertheless finer than can be expected for identity trees because information on the exact nature of substitutions is not ignored. The modified Blossum62 matrix showing costs for all amino acid changes is given in Table 1 (at the end of this paper).

The full Gene trees of Figure 1 were rooted by outgroup N-CAM orthologs (human, mouse, rat, bovine, chicken, and *Xenopus* orthologs were included in the analysis). Note that the N-CAM intracellular domain was not included (and treated as missing data) because it bears no sequence resemblance to the other intracellular

domains in the subfamily and is therefore probably not homologous. Also note that N-CAM has fewer Ig-like (5) and Fn (2) domains. For aligning these, N-terminus-to-C-terminus order on N-CAM was maintained, and the "missing" domains were assigned in such a way to maximize sequence similarity for the entire alignment (by using GCG Pileup output). In this way, the fourth Ig-like domain and the third, fourth and fifth Fn domains of the 6-Ig-like domain protein subfamily were specified as the "non-homologous" domains to N-CAM (i.e., the N-CAM sequence was treated as "missing" these specific domains for alignment purposes).

#### *Examining the Tree Space*

While only the shortest tree is reported and discussed in various analyses, the stability of these trees was examined extensively. First, in any given study, approximately 100 random starter trees were used to confirm that the heuristic minimi were reproducible from various entry points. Second, in all studies, at least the ten shortest trees were examined (usually within 5% total tree length of the shortest tree) to confirm that completely novel branch topologies (that would have made results uninterpretable) were not among this sample. Third, wherever possible, outgroup proteins (N-CAM orthologs) were included to confirm that relationships and topology remain unaffected when rooting the tree. Fourth, exhaustive searches were performed independently on individual domains (the intracellular domains, for example) and subclades (the Bravo/Neurofascin/L1/Ng-CAM clade, for example) to confirm that the resulting absolute minimum branch topology is superimposable on the corresponding portion of the heuristic tree in question. Fifth, within presumed orthologous groups, branch topology was compared to known evolutionary relationships in order to confirm that no phylogenetic violation was evident. And finally, using MacClade and the branch-swapping tool, the relative expense to change branch topology within specific clades was examined

in order to challenge particular drawn conclusions (for example, to ask: How expensive is it to move the Ng-CAM branch apart from L1? Or, is the placement of the Neuroglial branch nearer one subclade or another equally expensive?). By the above criteria, all minimal trees reported are robust, and any instabilities are acknowledged in the appropriate result descriptions.

*Simplification of Analysis by Collapsing Orthologous Sequences to Presumed Ancestral Character States.*

In some analyses, groups of two to four orthologous protein sequences were replaced by an artificial sequence exhibiting a reconstructed ancestral state, so as to reduce the number of Operational Taxonomic Units (OTU's) for tractability with phylogenetic software. Reconstructions were done by projecting character states down the branches that connect orthologous protein sequences using the tree topology of the known phylogenetic relationships of the organisms containing the genes. In this way, the two Bravo orthologs are collapsed to a single Bravo ancestral sequence, likewise the two Neurofascin orthologs, the three L1 orthologs, the three Contactin orthologs, the three Tag orthologs, and the two Big-1 orthologs, reducing the number of OTU's to nine (from eighteen). The six N-CAM orthologs were likewise collapsed to its presumed ancestral sequence. This method of reconstruction is identical to the "downward pass" described in Maddison *et al.* (1984). Analyses that utilize these reconstructions are indicated in figure legends (for example, were used to generate "Domain Trees", as discussed below).

*Gene Trees Versus Domain Trees.*

In various analyses, both "Gene Trees" and "Domain Trees" are discussed. A Gene Tree is a phylogenetic tree that relates full length sequences of the protein subfamily; a Domain Tree is a phylogenetic tree that relates the individual domains of these proteins (i.e., the

individual Ig-like or Fn domains were treated as their own distinct OTU's). The latter permits analyses of the evolutionary history of specific domain shuffling, origins, duplications and relatedness that ultimately has culminated in the current six Ig-like /five fibronectin domain organization of the various full-length proteins in the subfamily. To generate Domain trees, distinct domain sequences for collapsed orthologous ancestors (see above) are defined as independent OTU's and aligned, which in the entire subfamily, results in 54 Ig-like domain alignments (9 ancestral proteins  $\times$  6 Ig-like domains each) and 40 fibronectin domain alignments (4 ancestral proteins  $\times$  5 Fn domains + 5 ancestral proteins  $\times$  4 Fn domains). Because Ig-like and Fn domains are not alignable (i.e., not homologous), the group of 54 Ig-like domains and the group of 40 fibronectin repeats were analyzed separately.

#### *Statistical Analysis of Conservation and Divergence*

Patterns of conservation and divergence were determined for particular domains or sequence subsets along specific phylogenetic branches of the Gene tree. Within the Gene tree of Figure 1, the amount of character state change for each branch was analyzed for each domain and structural motif. This was done in Paup using the "Inclusion"/"Exclusion" option to analyze particular parts of the genes in isolation within the context of the established topology. An analysis of this type identifies segments of the gene and localized regions of the tree that exhibit unusual conservation or divergence. Figure 5, for example, shows overlapping phylograms for individual domains (each branch length normalized per 100 amino acid length) in order to illustrate relative change along each branch in the Gene tree at a domain-specific resolution.

Figures 6, 7 and 8 summarize overall conservation and divergence for each beta strand and loop predicted structure for each domain over the entire phylogenetic tree. The total change for any

given strand or loop is the sum of Blossum62 "costs": every amino acid substitution along every branch in the tree was scored a given cost according to the modified Blossum62 matrix of Table 1 (at the back of this paper). These cost numbers were normalized for length (i.e., per 10 amino acid sequence length) in order to more accurately reflect the degree of change for each domain and structural motif. Both the mean and standard deviation of total change was calculated separately for all Ig-like and fibronectin structural elements. In this way, specific strands or loops can be identified as more significantly conserved (smaller number of steps along a branch) or more diverged (greater number of steps along a branch) for particular domains.

## Results

### *Phylogenetic Analysis of the Six Ig-like Domain Subfamily of Neural Cell Adhesion Molecules.*

The 6-Ig-like domain neural subfamily of proteins includes Neuroglian (*Drosophila*), Bravo (chicken and human orthologs), Neurofascin (chicken and rat orthologs), L1 (mouse, rat and human orthologs), Ng-CAM (chicken), Contactin/F11 (chicken, mouse and human orthologs), Tag-1/Axonin-1 (chicken, rat and human orthologs), Big-1/Tag-like (mouse and rat orthologs), and Big-2 (rat). These proteins are grouped in a subfamily of the Ig Superfamily based on sequence similarity and on structural grounds: they are each predicted to contain six Ig-like domains at the N-terminus and either four or five fibronectin type III repeats C-terminal to these. In order to determine the probable evolutionary relationships among this subfamily of proteins, a cladistic analysis was performed using Paup 3.1.1 phylogenetic software (Sinauer Associates, Inc). Using the sequence alignments shown in Table 1, heuristic searches based on the identities or similarities step matrices generated trees of identical

topology. With the identities stepmatrix, in which all amino acid substitutions are of equal cost, the shortest tree was 4844 substitutions long. With the similarities stepmatrix, in which all amino acid substitutions cost between 8 to 15 steps as derived from the Blossum62 matrix, the shortest tree was 54692 steps long. Exploration of the tree space did not reveal any near shortest trees of grossly differing topology, with only the placement of the Neuroglial branch exhibiting noteworthy instability (see below). The shortest "similarity" tree (rooted by N-CAM proteins) is shown in Figure 1.

As predicted, the tree divides itself into two discrete clades: the first clade that contains Bravo, Neurofascin, L1 and Ng-CAM, and the second clade that contains Contactin, Tag-1, Big-1 and Big-2. The former proteins grouped in the upper clade of the trees in Figure 1 all contain five fibronectin type III repeats and an intracellular domain; the latter group in the lower clade all contain four fibronectin type III repeats and are linked to the membrane via a PI-anchor, and therefore on these grounds alone, the division into these two clades was anticipated. The placement of the *Drosophila* Neuroglial protein, often called the L1-homolog, into one of these two clades is ambiguous. On the grounds of predicted structure, Neuroglial fits with the Bravo/Neurofascin/L1 clade with its five fibronectin domains and intracellular domain, however in the 20 shortest Gene trees, Neuroglial is not topologically closer to the proteins in this clade than to the Contactin/Tag/Big group. Of these 20 shortest trees, three tree topologies are consistently and significantly shorter than the others (within 0.2% total tree length of the shortest tree): one in which the Neuroglial branch is at the root of the upper Bravo/Neurofascin/L1 clade (length 4844/54692, the tree in Fig.1), one in which the Neuroglial branch is at the root of the lower Contactin/Tag/Big clade (length 4847/54781, not shown), and one in which the Neuroglial branch is at the root of both clades (length 4850/54753, not shown).



### *Avian-Mammalian Speciation Events*

One of the curious features of this phylogenetic analysis is the placement of the chicken Ng-CAM protein branch in the Gene tree. The tree suggests that Ng-CAM is the avian ortholog to L1, however its distance (i.e., total branch length) from L1 far exceeds other mammalian-avian divergences in the tree. In order to explore this L1-Ng-CAM question further, chicken-human orthologs for the protein family, along with a more distant chicken and human N-CAM orthologous pair, were aligned and analyzed. Figure 2A illustrates the shortest phylogenetic tree that results from a heuristic similarities search. In each case, approximately 20% sequence divergence defines the rate of the apparent "molecular clock" since the avian-mammalian split, with the notable exception of Ng-CAM-L1 which are approximately three-times more diverged than the others. It is possible that specific individual domains of Ng-CAM/L1 account for most of this exceptional divergence, and other domains may indeed have a degree of conservation that reflect a similar rate of divergence (20%) as other avian-mammalian orthologs in the subfamily. To explore this, individual Ig-like domains for a subset of chicken/human proteins were aligned and analyzed. Figure 2B illustrates the shortest domain-by-domain phylogenetic tree for Ig-like domains of the L1/Ng-CAM clade. Indeed it appears that specific Ng-CAM-L1 domain branch topologies are highly variable, the extremes being the sixth Ig-like domain in which these two are not even monophyletic and greatly diverged (515 steps apart), and the second Ig-like domain which approaches a level of similarity that is observed for other avian-mammalian protein domain branch lengths (205 steps apart).

### *Evolutionary History of Individual Immunoglobulin Domains*

It is not uncommon to think of the individual Ig-like and Fn domains as distinct modules that impart distinct functions or ligand interactions. In order to explore the evolutionary history of these



“modules” for the neural Ig subfamily studied here, the domains were analyzed independently. For each of the 9 orthologs, a common ancestral state was deduced using the known phylogenetic relationships among vertebrate species (see Methods); in this way, only a single Bravo, Neurofascin, L1, Contactin, Tag-1 and Big-1 amino acid sequence is considered, and this reduction in OTU’s makes the analysis tractable. For each of the 9 presumed ancestral states of the various proteins, a cladistic analysis was performed on the 54 individual Ig-like domain ancestral sequence modules in order to elucidate their evolutionary history. The five N-CAM Ig-like domains were also included in order to root the tree. A single shortest tree suggesting the evolutionary relationship of the individual Ig-like domains across the entire subfamily is illustrated in Figure 3.

The first and most important feature of this tree is the fact that generally, domains that occupy the same place on a given homolog (i.e., their placement from N-terminus to C-terminus in the complete protein) are grouped together. That is, all of the first Ig-like domains are closely related and monophyletic, all of the second Ig-likes are similarly grouped in the tree, and all of the thirds, fourths, fifths and sixths are likewise grouped together. For each individual Ig-like domain clade, not only is the *sub*-subfamily division (i.e., Bravo/Neurofascin/L1 versus Contactin/Tag/Big division) maintained across all domains, but in addition, the specific branch topology established in the Gene tree is recapitulated from domain-to-domain (compare the Gene tree to the Ig I clade, for example). Again, the exception is the Neuroglial branch, which like the shortest Gene tree is part of the upper Bravo/Neurofascin/L1 division in the Ig I and Ig IV clades, yet part of the lower Contactin/Tag/Big division in the Ig II, Ig V and Ig VI clades, and at the root of both divisions in the Ig III clade.

*Evolution of the Six Ig -like Domain Prototypic Ancestral Molecule*

The global phylogenetic tree generated in Figure 1 predicts that a common six Ig-like domain ancestral molecule existed that ultimately gave rise to the entire subfamily of proteins. It is probable that the evolutionary history prior to this six Ig-like domain state involved domain recruitment or duplication, ultimately derived from a single domain ancestor. The first and second Ig-like domains, for example, are together a monophyletic clade in Figure 3, and this may reflect a duplication event whereby one of these domains directly came from the other (although it is also entirely possible that the close phylogenetic relationship may rather reflect convergence). The phylogenetic Domain tree in Figure 3 includes the five N-CAM Ig-like domains. N-CAM is likely to have a common ancestral state with the six-Ig-like domain subfamily for two primary reasons: 1) like the six-Ig-like subfamily, N-CAM is expressed on the cell surface in the developing nervous system with related functions, and is a member of the Ig supergene family, and 2) the N-CAM extracellular protein is entirely composed of Ig-like and Fn domains, and has only one fewer Ig-like domain than the six-Ig-like subfamily. It is possible, if not likely therefore, that either an Ig-like domain duplication led to the arrival of the six-Ig-like subfamily from N-CAM precursors, or an Ig-like domain deletion led to the arrival of N-CAM from six-Ig-like subfamily precursors.

Figure 3 does not, however, support such a simple model. Interestingly, the second Ig-like domain of N-CAM roots the Ig II clade, while all other N-CAM Ig-like domains are grouped at the root of the Ig III clade. This suggests that in at least four of the five N-CAM Ig-like domains, individual domains are closer to other domains on N-CAM than they are to the corresponding domains on any of the other proteins in the tree. Although it is possible that this branch topology is an artifact of poor alignment (N-CAM sequences are highly diverged from the subfamily) or convergence noise, the model most firmly substantiated by such a finding is one in which the common ancestor

between N-CAM and the six-Ig-like subfamily contained only two Ig-like domains (Ig II and Ig III). In this model, the remaining N-CAM Ig-like domains would have arisen from duplications (and subsequent divergences) of the Ig III ancestor after the split from the six Ig-like subfamily. In this way, these newer N-CAM domains would be expected to share greater homology with the ancestral Ig III domain than any of the other six Ig-like domains (and vice-versa) because they arose separately, diverged to distinct functions, and have been subject to presumably unrelated selective pressures.

*Evolutionary History of Individual Fibronectin Type III Repeats.*

In order to explore the evolutionary history of the fibronectin type III repeat "modules", these discrete amino acid sequences were also analyzed independently. Again, for each of the 9 orthologous groups, a common ancestral state was deduced and a cladistic analysis was performed on these 40 individual ancestral fibronectin sequences. A single shortest tree showing the probable relationships of the individual Fn domains across the entire subfamily is illustrated in Figure 4.

As was the case for the Ig-like domain-by-domain phylogenetic tree, the groupings of individual fibronectin repeats reflect their order on the full-length protein, and within these domain groupings, Gene tree topology is generally recapitulated. The first fibronectin type III repeat subclade, for example, is both monophyletic and has a branch pattern that superimposes perfectly with the global Gene tree of Figure 1. The Fn 4 domains are an exception to this orderly grouping, and yet these domains are grouped in two monophyletic clades that reflect the predicted division of these proteins into the two *sub*-subfamilies (i.e., there is a monophyletic Fn 4 clade that contains the Bravo/Neurofascin/L1 *sub*-subfamily; and a second monophyletic Fn 4 clade that contains the Contactin/Tag-1/Big *sub*-subfamily), and these two Fn 4 clades are topologically near in the tree. With regard to the

Neuroglian branch placement, again its grouping with either *sub*-subfamily is ambiguous: it is at the root of the Bravo/Neurofascin/L1 group in the Fn 1 clade, at the root of the Contactin/Tag/Big group in the Fn 3 and Fn 4 clades, and below the root of both groups in the Fn 2 clade.

The most interesting fibronectin type III repeat is the fifth domain: it is alternatively spliced in Bravo, diverged to the point of non-recognition in Neurofascin, and altogether lost/missing in the Contactin/Tag/Big *sub*-subfamily. In the fibronectin domain-by-domain tree in Figure 4, the Fn5 domains are grouped in a monophyletic clade, connected by atypically long branches. This certainly reflects the uniqueness and diversity of the amino acid sequences of the various fifth domains in the family, although it should be pointed out that among orthologs (chicken and human Bravo for example) the level of conservation of this domain is not unusually low. It may be that this domain is "on its way in", with recently derived important functions that delineate specific roles among the various proteins; or, on the other hand, this domain may be "on its way out", with the various proteins demonstrating differing stages of diverged sequences and lost function. These alternative perspectives are elaborated on in the Discussion section.

*Patterns of Conservation and Divergence Of Specific Domains and Motifs Along Specific Branches in Evolutionary History.*

Assuming the six Ig-like domain subfamily of neural proteins share a common ancestral state, there is only one correct evolutionary history that links them. The cladistic analysis on full length amino acid sequences that gives rise to the phylogenetic tree in Figure 1 is the most probable of possible evolutionary histories (see Discussion). If this tree and the ancestral nodal states reconstructed by parsimony are assumed to be an accurate depiction of the actual molecular events in evolution, an amino acid by amino acid mutational analysis is possible.

That is, it is possible to examine the patterns of sequence change along specific phylogenetic branches for highly localized regions of the protein, such as specific domains or structural motifs. In practice, the Paup software permits the inclusion/exclusion of specific characters which, while maintaining fixed branch patterning, calculates the degree of divergence along each branch exclusively for these "included" characters. In this way, phylograms illustrating divergence of specified segments of the whole sequence can be isolated in the context of the ancestral history defined by the full length protein tree. Figure 5a and 5b illustrate the superimposition of conservation/divergence patterns for individual domains analyzed in this way. Individual domains with significant levels of conservation or divergence along any specific branch are noted.

Curiously, the two domains that define the most obvious division of the six Ig-like domain subfamily into two discrete *sub*-subfamilies represent the extreme cases in this analysis. The Bravo/Neurofascin/L1 *sub*-subfamily is considered distinct from the Contactin/Tag/Big *sub*-subfamily primarily on structural grounds: the former group contains both an extra fibronectin type III repeat (Fn5) and intracellular domain. The superimposed phylograms in Figure 5b illustrate that the intracellular domain is by far the most conserved domain in the overall tree, having changed the least of all domains along every branch in the clade (extrapolated to 3902 total steps). This domain, quite remarkably, is 100% conserved since avian-mammalian divergence in the Bravo proteins. At the other extreme is the fifth fibronectin type III repeat, whose degree of divergence is inferred as significant on greater than half the phylogenetic branches, with a total divergence extrapolated to over 8000 total steps. The Fn5 branches in the Neurofascin clade are difficult to assess: the actual sequences contain short elements that resemble the Fn5 sequences in Bravo, so it is likely that divergence has taken place from an ancestor that contained a more recognizable fibronectin domain, yet there are long,

repeating intervening sequences that make alignments almost arbitrary. Nevertheless, even if this clade is ignored, the Fn5 phylograms remain by far the most diverged branches in L1 and Ng-CAM, and the second-most diverged in Bravo.

In general, the trend among fibronectin domains is clear: the degree of divergence in the tree is correlated with domain number (i.e., more C-terminal domains are increasingly more diverged). The most C-terminal Fn 4 domains of Contactin/Tag/Big proteins, are significantly diverged to an extreme similar to the Fn5 domains of Bravo/Neurofascin/L1 proteins. It is possible that the more C-terminal domains have fewer functions which may have resulted in reduced selective pressure in ancestral molecules. This seems reasonable given that these more C-terminal domains are increasingly less accessible to the extracellular environment. Most of this increased divergence in C-terminal Fn domains, however, takes place in internal branches; branches connecting orthologs are generally not exceptionally long. It is also possible therefore, that these domains encode mostly newer functions which may have evolved since paralogous duplications and thus define functional differences among orthologous sets.

Overall, fibronectin type III repeats are approximately 25% more diverged than the Ig-like domains (Fig 5b). Ig II and Ig IV are especially well conserved in the tree; most of the conservation of Ig II is localized within the L1/Ng-CAM branches, while Ig IV is highly conserved among Neurofascin and Bravo branches. The Ig III clade in Figure 3 implies that this particular domain may be responsible for some ancestral/ancient functions shared between the N-CAM and six-Ig-like domain subfamily, which is indeed the case in one known example in which FGF-receptor binding to L1 and N-CAM has been mapped to the C-terminal region of Ig III (Doherty *et al.*, 1995; Williams *et al.*, 1994). Ig III, like Ig II, is highly conserved within the L1 branch, yet unlike Ig II, its conservation does not extend out the Ng-CAM branch. Ig I, like Ig IV, is highly conserved among the Bravo and Neurofascin branches



despite an overall divergence of this domain elsewhere in the tree. Similarly, Ig V is the most conserved domain among the Contactin and Tag branches despite an overall significant level of divergence of this domain elsewhere in the tree. It appears as if these patterns of conservation and divergence delineate specific domains as having uniquely high levels of selective pressure among specific protein sets: Ig I and Ig IV for Bravo and Neurofascin, Ig II and Ig III for L1 and Ng-CAM, and Ig V for Contactin and Tag. Only Ig VI seems to be uniformly diverged along all branches in the tree.

#### *Tracing Functional Structural Motifs*

Both immunoglobulin-like domains and fibronectin type III repeats are structurally very similar, consisting of two opposing beta sheets with similar strand topology. These similar structures are almost certainly an example of convergent evolution because they do not share common conserved core residues that are required for proper domain folding. In every case where they have been studied, Ig-like and Fn domains play a role in ligand interactions or cell adhesion. Partial structural solutions suggest that these Ig-like and Fn domains may be structurally-separated modules, linked yet erectly extending out from the cell membrane in such a way that would minimize contacts with other domains and maximize contacts in the surrounding environment. Each of these separated domain modules along the full length of the protein would then be able to adopt a similar and repetitive two-dimensional orientation, with the beta sheets facing out, perpendicular to the line of intervening sequence that connects the domains. There are six conceivable "structural landscapes" that may serve as ligand interfaces for individual domains: the hydrophilic outer surfaces of each of the two beta sheets (the BED sheet and the AGFC sheet of Ig-like domains), the outer side of the "barrel" opposite to domain-connecting intervening sequence (C strand-CD loop-D strand of Ig-like domains), the inner side of the barrel

nearest the domain-connecting intervening sequence (A strand-AB loop-B strand of Ig-like domains), the three upper loops that arc away from the cell membrane (i.e., in the N-terminal direction, BC loop+DE loop+FG loop of Ig-like domains), and the three bottom loops that arc towards the cell membrane (i.e., in the C-terminal direction, AB loop+CD loop+EF loop of Ig-like domains). In order to gain insight into possible functional significance for particular structural motifs, patterns of amino acid conservation/divergence were calculated and contrasted for each beta strand and loop for all domains. These data are summarized in Figures 6 (Ig-like domains) and 7 (Fn domains).

Although there is significant domain-by-domain variation, if the overall change is summarized for all domains and all branches of the tree, some dramatic trends emerge. In both Ig-like and Fn domains, there are specific beta strands that predictably have undergone very little change over the course of evolution: the B, C and F strands in particular are highly conserved because they contain characteristic core residues required to maintain domain structure. Interestingly, for both kinds of domains, the greatest divergence is found at the outer side of the structure: the CD-D-DE elements of Ig-like domains, and the CC'-C'-C'E-E elements of Fn domains. It is interesting to note that at least in one specific case, the CD loop of V-CAM is responsible for supporting an interaction with its integrin ligand (Jones *et al.*, 1995). The enormous divergence of this particular loop for a wide variety of Ig domains has led to the proposal that the C-CD-D side of the beta barrel may be a general region for intermolecular contacts (Jones *et al.*, 1995), and that the CD loop in particular may provide specificity to these interactions (hence, its divergence from protein to protein).

Another trend involves the loop regions. In both cases, the bottom loops (AB+CD+EF) are significantly diverged, while a single, highly conserved top loop (BC loop of Ig-like domains; FG loop of fibronectin domains) is outstanding. It is possible, for example, that the highly conserved upper loop of one domain and the highly diverged



lower loops of the adjacent domain might intermingle in the interdomain space, the conserved loop forming a sort of common ligand-binding platform, while the diverged loops providing the ligand-binding specificity, functioning like toggles in a combination lock among various proteins. To investigate the possible functional significance of these general structural conservation/divergence trends, specific strand-loop conservation was examined for individual domains of various orthologs which are presumed to be functionally equivalent in their respective species. The first and most confident assumption that can be made about functionally important residues is that these specific residues should be conserved among proteins that have the same function; i.e., among orthologous proteins. Figure 8a (Ig-like domains) and 8b (Fn domains) indicate specific beta strands and loops that are 100% conserved among the orthologs.

On the previously mentioned proposal that the beta barrel side (C-CD-D strands/loops of Ig-like domains) may be a general ligand-binding specificity region, it is perhaps noteworthy that despite the overall divergence elsewhere in the tree, the D-DE region is 100% conserved in Ig I of Contactin; Ig II of Bravo, Neurofascin and Contactin; Ig IV of Neurofascin; Ig V of L1; Ig VI of Neurofascin and L1; Fn 2 of Bravo; and Fn 3 of Tag-1 (C' strand only). The CD loop to which the V-CAM-integrin interaction has been mapped, however, remains highly diverged even among orthologs, and therefore this loop's particular divergence may represent diminished selective pressure as opposed to functional significance. The notable exception may be Ig IV of Contactin, which despite extreme divergence elsewhere in the tree, has a 100% conserved CD loop (although in this particular case, the CD loop is predicted to be quite short).

There are other examples where particular loops are significantly diverged overall in the phylogenetic history, yet absolutely immutable among orthologs. It is perhaps functionally significant that in many of these cases, the opposing bottom loops of one domain and top loops of

the adjacent domain exemplify this same striking pattern of overall divergence yet orthologous conservation (examples include loops between Bravo Ig I-II, Neurofascin Ig I-II and Ig III-IV; L1 Ig I-II and Fn 1-2). Because this striking pattern is often a characteristic of opposing loops on adjacent domains, an intriguing model emerges: it is possible these bottom and top loop conservations may indicate that important ligand interactions span or map to interdomain regions (see Discussion below).

## Discussion

### *Phylogeny to Infer Function*

The recent and ongoing genome sequencing efforts in various species has presented an opportunity to study evolution unlike any time since the work of Darwin. The most frequent use of this sequence data in evolutionary studies is to obtain evidence on the phylogenetic history of the organisms from which these sequences come. In this particular study, the phylogenetic relationships of the organisms are well established, and we are rather interested in what phylogeny can imply about the history and function of the genes themselves. Our questions are aimed at sorting out the molecular duplication and divergence events that has resulted in this gene family: Which are the orthologs? Where have the paralogous duplications occurred? Is there domain shuffling which would impart common functions to proteins that recruit specific modules? Are there discrete regions of the protein that have outstanding levels of conservation or divergence? Each of these questions is aimed at gleaning further insight into the functional history of this gene family which may be revealed by such molecular evolutionary studies.

It is worth noting that we have used amino acids as the character states in phylogenetic reconstruction, rather than nucleotides, as has

been the more common practice. Nucleotide sequences suffer from the bimodal degree of significance of changes at the third codon position: some third position changes result in amino acid substitutions, while identical changes in other codons do not. The former represent important components of the informational signal, while the latter are mostly noise, especially among distant proteins, in which synonymous changes in the third codon position are probably saturated. The latter may be even worse than ordinary noise, as differences in codon bias may prevent the noise component from being random. At the nucleotide level, these two levels of significance can not be distinguished, as contrasted to the amino acid level, where synonymous codon substitutions are ignored, and signal is exclusively derived from changes that result in amino acid substitutions. It is only these amino acid coding changes that have functional consequence for the protein; furthermore, specific amino acid changes can be weighted (i.e., by use of a stepmatrix) to reflect the presumed relative similarity in structure/function of the amino acids.

Perhaps the most thorough use of data would incorporate both codon usage and amino acid content: a 64-by-64 stepmatrix could be designed in which the entire triplet codon defines the character state (with corresponding modification of the phylogenetic software to accommodate such a matrix and its three-letter character state codes), with low cost steps between synonymous codons (weighted for transversions vs. transitions, and for codon usage), and high cost steps between non-synonymous codons (weighted for specific amino acid changes as per the Blossum62 matrix utilized here). Although this approach would be ideal because all of the genetic information would be used, a caveat due to codon bias must also be considered: cross-species codon comparisons could introduce a systematic/artificial similarity and lead to significant non-random noise in the data. Furthermore, no information is available as to whether codon bias has oscillated at various places in history, and therefore it is perhaps

dangerous to simply subtract away the noise generated by the present-day bias.

When considering amino acid sequence data, it should be pointed out that homologous residues are not always identified by standard alignment programs. If evolutionary change from one residue to another is the measure of phylogenetic relatedness, then deciding which residues are homologous character states is the most critical step in the analytic process. Especially in long sequences with gaps that disrupt regions of alignment, the results of alignment programs are not always compatible with homology as determined by amino acid tolerance in solved structures. The reason for this is probably that amino acid similarity matrices are entirely empirical and statistical, integrating the probability of specific substitutions in consideration of a diverse and often unrelated set of proteins. Among the Ig-like and Fn domains of the Ig Superfamily, for example, specific amino acid changes which are frequent and tolerated in the domain structure (such as phenylalanine to tyrosine), are nevertheless considered rarer and more expensive when the entire scope of the protein world is considered.

Standard programs treat all residues as equally subject to change, yet structural alignments show that some residues are very unlikely to change without major disruption to the protein structure. These highly conserved residues can be identified and can serve as a guide to alignments, as in this case, even if only a few proteins of a family have fully solved structures. Such structurally refined alignments are likely to be significantly more reliable than those done without structural information.

#### *Domain Trees Atop Gene Trees Atop Species Trees*

The Gene tree in Figure 1 that relates a subfamily of proteins in the Ig Supergene family shows which of the homologs are likely orthologous, and where the paralogous duplications occurred. The

distinct subclades in which the relationships among proteins follows the known phylogenetic relationships among their organisms indicate the orthologous branches in the tree (i.e., indicate which proteins are orthologs). In the Ig subfamily studied here, the orthologous branches appear to be the terminal branches, strongly implying that paralogous duplications occurred more internally in the tree and probably prior to vertebrate speciation. This principle of superimposing a gene tree over a species tree, and by virtue of shared branch patterning, inferring where the orthologous and paralogous branching occurs, is an important tool in studying molecular evolution and gene families.

The entire Ig Supergene family is a very diverse group of proteins, and the number of Ig or Fn domains among different subfamilies is highly variable. It is not uncommon to regard individual Ig and Fn domains as distinct "functional modules" that can be shuffled in evolution. In this context, individual Ig and Fn modules can be described in terms of their independence: they can fold independently (i.e., adopt the correct structural fold regardless of adjacent flanking sequences or domains) and in some cases, represent independent exons in the genome that are alternatively spliced. Given that there are 10-11 of these domains in the extracellular portions of the subfamily considered herein, the issue as to whether or not these modules have indeed been inherited as a unit on the time scale represented in the Gene tree, remains a serious threat to the premise that a common ancestor with 6 Ig-like domains even existed. To explore this issue, the individual "Domain Trees" were generated (Figures 3 and 4), which to first approximation, demonstrate that shuffling has probably not occurred (nor gene conversion or significant convergence), and not only has domain number been conserved in evolution, but so to has domain order. Therefore, the Domain trees strongly suggest that a 6 Ig-like domain ancestor indeed existed, because domain duplications have apparently not occurred since paralogous diversification.

The process of placing a Domain tree on a Gene tree is comparable to that of placing a Gene tree onto a Species tree. In both cases, the units analyzed in the first tree (i.e., domains or genes) are inherited within the units analyzed in the larger tree (i.e., genes or species). Any events which violate this internal inheritance of smaller units, such as gene conversion or hybridization between species, are likely to result in such inexplicable branch patterns that such events will be revealed. Furthermore, if the trees are fully reliable and rooted, the pattern of superimposition is highly constrained topologically and indicates where paralogous duplications are likely to have occurred (i.e., the history of domain diversification or gene diversification).

The expectation is that in addition to Gene trees being superimposable upon Species trees, so to will Domain trees be superimposable upon Gene trees. Within the individual domain clades of Figures 3 and 4, the Gene tree topology is in most cases indeed recapitulated: the first Fn domain clade of Figure 4, for example, has identical topology to the overall Gene tree of Figure 1. Yet individual domain clades in Figures 3 and 4 in some cases, are topologically different than the Gene tree. While on the one hand this may likely be due to local "noise" in the data (domain sequences are naturally shorter than full protein sequences, and the reduced number of informative residues may fail to correctly or definitively resolve branch patterns), we believe that these local violations indicate that local convergence events may be contaminating the analysis. The sixth Ig-like domain of Contactin, for example, is topologically closer to the Big proteins than to Tag-1, and the second Fn domains of Neurofascin and L1 are likewise unexpectedly similar. If we are to believe that the overall Gene tree is the true phylogenetic history (and there is no gene conversion), then this exceptional similarity between what ought to be more greatly diverged protein sequences, might represent convergent domain-specific functions. It is intriguing that many of the proteins in this subfamily seem to share common ligands and functions, and it is



possible therefore that these domain-specific branching anomalies may indicate where overlapping functions may reside in otherwise more distant proteins.

#### *Fibronectin Repeats Versus Immunoglobulin-like Domains*

The neural subfamily of the IgSF discussed here (and indeed, numerous other cell surface proteins expressed within and outside the nervous system) have an extracellular multi-domain mosaic composed of both Ig-like and Fn domains. These convergent beta-sheet structural modules were first identified in immunoglobulins of the immune system and the fibronectin molecule of the extracellular matrix respectively. It is clear that in general, Fn domains of the 6-Ig-like domain subfamily have diverged far more than Ig domains (Fig. 5). There are many possible reasons why this might be so, among them: 1) Fn domains have significantly fewer critical structural core residues than V-like Ig-like domains, and therefore overall, there is probably less tight constraint on sequence content required for proper folding, 2) Fn domains are more C-terminal, closer to the membrane, and probably less accessible for interactions with ligands, and therefore, may have fewer functions and less overall selective pressure associated with these domains, 3) Ig-like domains may impart more of the relic functions, while Fn domains may provide some of the newer, specific functional differences within the family, and therefore the Fn domains may have been evolving functions recently while Ig-like domains may have reached more of an equilibrium status prior to the time scope of the tree, 4) Fn domains and Ig-like domains may have completely distinct roles, and the molecules with which they interact may themselves be subject to differing selective pressures: Ig-like domains may, for example, interact primarily in *trans* with a large set of extracellular matrix and cell membrane spanning proteins that collectively may reside in a relatively constant environment, while Fn domains may interact primarily in *cis* with a completely different set of

co-receptors in possibly a more dynamic environment encompassing a diverse and wide-scoping range of signal transduction machinery.

On the latter point, it is interesting to note that in the IgSF subfamily discussed here, all of the proteins that contain the highly diverged Fn5 module, also contain a highly conserved intracellular domain; the converse is also true, and all subfamily members that are missing this Fn5 domain are PI-linked to the membrane. It is possible, therefore, that the Fn5 domain and intracellular domain may be subject to co-evolution, the function of one dependent on the function of the other. In the Bravo/Nr-CAM protein, for example, the Fn5 domain is alternatively spliced, and the entire amino acid sequence adjacent and C-terminal to this domain exon is 100% conserved between chicken and humans. This remarkable conservation includes the entire transmembrane and intracellular domains, and especially considering the absence of change within the cell membrane (where one might expect only the requirement for hydrophobicity and otherwise little selective pressure), it is perhaps likely that *cis* co-receptor interactions are taking place, possibly spanning the alternative Fn5 exon, transmembrane and intracellular sequences. This particular fibronectin type III repeat, therefore, may play a critical role in modulating signal transduction events by virtue of providing co-receptor dimerization specificity. This model of *cis* co-receptor mediated function is conceptually similar to what is observed in the L1-fibroblast growth factor (FGF) receptor coupling and resulting kinase signaling event (Doherty *et al.*, 1995; Williams *et al.*, 1994).

#### *Is Neuroglian Like the Ancestral Six Ig-like Domain Protein?*

Although *Drosophila* Neuroglian is often referred to as the invertebrate ortholog to L1, the phylogenetic tree of Figure 1 does not suggest it is evolutionarily closer to any of the orthologous groups. In fact, when the alignments are rooted by the outgroup N-CAM proteins, the shortest trees are equally likely to place Neuroglian in either of the



two *sub*-subfamilies (i.e., with either the Bravo/Neurofascin/L1 *sub*-subfamily or the Contactin/Tag/Big *sub*-subfamily) or at the root below both *sub*-subfamilies, which, given this ambiguity, may be the correct placement of Neuroglian. It should also be pointed out that in all the genetic characterization in *Drosophila*, as well as the extensive genome and sequence tag work in this species, no other 6-Ig-like domain subfamily member has been identified. In this regard, it is perhaps unlikely that there is a *Drosophila* Bravo, Neurofascin, L1, Contactin, Tag, Big-1 or Big-2 homolog, especially considering that most of the RNA's encoding these proteins are abundant and well-represented in vertebrate cloning libraries. It is therefore likely that the paralogous duplication events that have led to the full complement of vertebrate homologs may have occurred after deuterostome divergence (i.e., after the *Drosophila*/invertebrate split) but prior to the reptilian radiation (more than 200 million years ago).

If this model is correct, then in terms of the full range of neurodevelopmental functions covered by this subfamily, the Neuroglian protein may be the lone and sufficient invertebrate family member. Therefore, two possible scenarios regarding function might be considered: either the full range of vertebrate homologs are redundant and accomplish in development what a single protein in *Drosophila* manages, or, the increased complexity of the vertebrate nervous system has obligated paralogous duplication and divergence, resulting in an increased range of molecules and functions. The fact that null mutations with some of the vertebrate homologs results in subtle or even no phenotypes, and the fact that several of these homologs are promiscuous with each other, bind common ligands and support very similar functions *in vitro*, the notion of redundancy (or at least functional overlap *in vivo*) must be especially considered.

If Neuroglian is regarded as sister to the ancestral molecule of the rest of the six Ig-like domain subfamily, then the parent molecule that gave rise to the entire subfamily, like Neuroglian, was composed

of five fibronectin type III repeats as well as an intracellular domain. It follows therefore, that evolution in this group of proteins would then be moving in the direction of streamlining and reduced complexity; i.e., by eliminating domains. That is, if the ancestral molecule was a Neuroglian-like protein, then at the point where the Contactin/Tag/Big *sub*-subfamily divergence occurred, presumably the fifth fibronectin repeat and intracellular domain were lost as compared to the ancestral state. In this regard, the alternative splicing of Fn5 in Bravo and the loss of coherent fibronectin-like sequence in Neurofascin, may be further indication of a domain "on its way out".

*Is Ng-CAM the Chicken Ortholog to L1?*

Avian Ng-CAM is the sister to the mammalian L1 proteins, but it is so far diverged from the L1 proteins that in terms of pairwise scoring by percent similarities, it is nearly equal in distance from the Bravo proteins as to the L1 proteins. It had even been argued (prior to the identification of the human Bravo protein) that perhaps Ng-CAM and Bravo were exclusively avian molecules that each partially accomplish the mammalian L1 function, and this redundancy has reduced the selective pressure on the two chicken proteins (Kayyem, 1992). The inclusion of the Neurofascin proteins in the current cladistic analysis clearly establishes Ng-CAM as evolutionarily closer to L1 than Bravo, so the question remains: Is Ng-CAM the L1 of chicken?

It is unlikely that a new set of chicken L1s or mammalian Ng-CAMs will be identified: Ng-CAM/L1 cDNAs are some of the most abundant clones in brain libraries, and random sequence tag databases routinely and frequently catalogue cDNA sequences from each of the other, less abundant proteins in the family. In this regard, Ng-CAM is indeed very likely the ortholog to L1, but the significant sequence divergence between the proteins may indicate that they are not functionally equivalent. It is possible, for example, that L1 and Ng-CAM diverged because they are functioning in species-specific

environments or tissues. Unlike Bravo for example, Ng-CAM/L1 proteins are expressed in non-nervous system tissue and it is possible that the range of tissue expression might include environments that are different in birds and mammals (feathers versus hair, for example). Even in the nervous system, the molecular mechanisms that underlie developmental assembly are expected to have species differences. In the projection of retinal ganglion cells to the visual centers of the brain, for example, the orchestration is distinct between birds and mammals, involving quite different organizing principles (contralateral versus ipsilateral projections; autonomous hardwiring versus synaptic remodeling/activity-dependent processes (for example, reviewed in Holt and Harris, 1993). In this way, differences between Ng-CAM and L1 might be part of the distinct underlying molecular machinery that manifests these kinds of uniquely avian or mammalian neuronal phenotypes. On the other hand, given that there appears to be a degree of functional redundancy among the subfamily proteins (see above section), it is also possible that the Ng-CAM/L1 higher-than-expected divergence rather than indicating uniqueness in function, may indicate maximum overlap of function -- if Ng-CAM/L1 has a range of functions that are mostly redundant (i.e., otherwise accounted for by other members of the subfamily), then it would follow that there might be less selective pressure on these more dispensable homologs.

In any case, whether the low sequence similarity between chicken Ng-CAM and mammalian L1 is due to redundancy or species-specific issues, it is nevertheless conceivable that individual domains may account for the bulk of this higher-than-expected divergence. A domain-by-domain analysis of avian-mammalian orthologs was performed to investigate possible localized levels of conservation that approached the same, low degree of divergence (approximately 20% sequence divergence) observed for the rest of the family (Figure 2B). The second Ig-like domain of L1/Ng-CAM fits this criterion and appears to be as conserved in the phylogenetic tree at a similar level as

other avian-mammalian proteins (this conservation of Ig II is also illustrated in the overlapping domain phylograms of Fig. 5). Accordingly, Figure 8 illustrates structural regions that are significantly conserved between L1 and Ng-CAM, and by far the majority of this conservation is again isolated in Ig II. It is possible, therefore, that if redundancy is the reason for sequence divergence between Ng-CAM and L1, then the second Ig-like domain has exceptionally high selective pressure and may represent the most indispensable (and non-redundant) domain in the protein.

#### *Paralogous Divergence Versus Orthologous Conservations*

It is interesting to note in Figure 8 the frequency with which loops and strands that are highly diverged elsewhere in the tree, are absolutely unchanged in orthologous proteins, probably representing more than 200 million years of immutability (since avian-mammalian divergence). In Ig I, the highly diverged DE loop is conserved in Bravo, Neurofascin and L1 orthologs. In Ig II, this DE loop is also highly diverged yet conserved in Bravo, Neurofascin, L1, Contactin and Tag-1 orthologs; and the most significantly diverged FG loop of Ig II is conserved in Bravo and Neurofascin orthologs. In Ig III, the highly diverged DE loop is conserved in Neurofascin, L1, Contactin and Tag-1 orthologs, and the highly diverged E strand is conserved in Contactin. In Ig IV, the highly diverged CD loop is conserved in Contactin, and the highly diverged D strand is conserved in Neurofascin. In Ig V, the AB loop is conserved in Contactin, and the D strand is conserved in L1. In Ig VI, the D strand and DE loop is conserved in Neurofascin and L1. In Fn 1, the highly diverged AB loop is conserved in Tag-1. In Fn 2, the diverged CC' loop is conserved in Bravo and L1, and the diverged C' strand is conserved in Bravo. In Fn 3, the diverged C' strand is conserved in Tag-1. In Fn 4, the diverged AB loop is conserved in Contactin, and the C' strand is conserved in Bravo. And in Fn 5, the diverged EF loop and A strand is conserved in L1. It appears, therefore,

that this global divergence yet orthologous conservation is more the norm than exception. For each of these examples, a high degree of divergence, like changing the toggles in a lock, occurs in paralogous branches, yet it would follow that once a function is arrived at in a given protein, it would then be necessarily conserved in orthologous branches in order maintain this function across species. In this regard, many of the newer, ortholog-specific functions may map to these specific loops or strands.

*Do Important Ligand Interactions Take Place Between Domains?*

The domain-by-domain analysis depicted in Figures 3 and 4 illustrate that domains that occupy the same N-terminal-to-C-terminal place in the protein (all the first Ig-like domains, for example), group together. Any given domain, therefore, shares greater similarity to the same domain on other proteins than it does to the other domains on the same protein. This indicates that not only is domain number been conserved for hundreds of millions of years in evolutionary history, but so to has domain order. The individual Ig-like domains and fibronectin type III repeats are therefore (in this case) not distinct modules of function that get shuffled in evolution, rather, it appears the functions of the domains are linked to their specific placement on the protein with respect to their neighboring domains. There are a number of models that one could consider to relate function to domain order. For example, perhaps there is a sequential ratcheting down the protein to progressively tighten an intermolecular interaction with other proteins of the subfamily. This model would predict that the first Ig-like domain may bind with a certain affinity to the first Fn or Ig-like domain of the counter-ligand as an initial intermolecular contact, but would bind with increasingly higher affinity to more C-terminal domains as it ratcheted down the protein. Ultimately, the most stable, high-affinity binding would result when opposing domains are maximally overlapped. This simple, mechanistically-based model

makes specific predictions for domain to domain binding affinities, and would certainly be one mechanism which would account for conservation of domain order in evolution.

Another model that would account for such stringent preservation of domain order would involve specific interactions that depended on more than one domain. That is, if the domains are acting in concert, important ligand interactions may span domain boundaries, and therefore selective pressure would maintain appropriate neighborly relationships. If this is so, one prediction is that the bottom loops of one domain and the top loops of the adjacent domain might be critical in such between-domain ligand interactions, especially considering that the interdomain sequences are short and would probably permit these loops to reside in spatial proximity. As discussed in the previous section, it is already predicted that many ortholog-specific functions may map to a number of specific loops across all Ig-like and Fn domains, and so in general, it is anticipated that although highly diverged, loop structures will emerge as important functional components of domain function. The only known ligand interaction that has been mapped to a specific, structurally-identifiable domain sequence among the neural Ig Superfamily of proteins discussed here, is the RGD motif that is sufficient to support specific integrin-fibronectin domain interactions. This RGD motif maps to the upper FG loop, and perhaps significantly, this fibronectin FG loop is generally by far the most well conserved (Fn 5 excepted) of the six loop structures in the family tree (Figure 7). Hypervariable loops in antibody Ig domains are also known to be critical for specific interactions in the immune system, and these loops are structurally homologous to the upper loops of the V-like Ig-like domains in the nervous system, including the highly conserved BC loop in these proteins (Figure 6). So, among functionally equivalent proteins (i.e., orthologs), is there further evidence for the importance of loops, and in particular, for the loops of adjacent domains acting in tandem? There are numerous



examples illustrated in Figure 8 where there is 100% conservation of loops between domains, often including three or more opposing loops that are identical among orthologs (Ig I-II in Bravo and Neurofascin; Ig II-III in L1; Ig III-IV in Bravo, Neurofascin and L1; Ig IV-V in Neurofascin and L1; Ig V-VI in Neurofascin and L1; Ig VI-Fn 1 in L1; Fn 1-2 in L1). Many of these highly conserved opposing loops include specific loops that are otherwise diverged significantly elsewhere in the tree (indicated by red and orange color codes in the figure).

The "beads on a string" description sometimes used to describe these multi-domain proteins implies that the domains themselves are the important protein regions, and the intervening sequences merely the backbone that connects domains. However, if one examines the Bravo protein sequence for example, it is clear that this so-called "string" sequence that falls between domains is almost certainly functionally important. First, three alternatively spliced exons have been characterized as Bravo isoforms, and they each map to sequences between domains (the large Fn 5 alternative exon excepted): between Ig II and Ig III, between Ig VI and Fn 1, and between Fn 4 and Fn 5 (Kayyem *et al.*, 1992). It is presumed that alternative splicing events, especially if they are conserved in evolution (for example, the 12 amino acid alternative exon between Fn 4 and Fn 5 conserved in Bravo orthologs), have functional importance, perhaps to modulate a ligand interaction that maps near or at the spliced exon. Second, there are several intervening domain sequences that are absolutely conserved between chicken and human Bravo (Lane *et al.*, 1996): between Ig I and Ig II, between Ig III and Ig IV, between Ig IV and Ig V, and between Fn 2 and Fn 3, all of which in addition, were exemplified above as having remarkable conservation of opposing loops that presumably mingle in this interdomain space (i.e., opposing loop and interdomain conservations are correlated). Third, a recently identified interaction between L1 and the fibroblast growth factor receptor (FGF-R) has been mapped to an AAPYW sequence that falls between Ig III and Ig IV, and

this same AAPYW sequence is found between these same domains in Bravo. And finally, while most of the best illustrative examples cited here especially pertain to Ig-like domain sequences, it is apparent that fibronectin domains may likewise be subject to similar selective pressure: a specific tenascin-contactin interaction, for example, has been shown to be sensitive to the spacing of particular fibronectin domains (Zisch et al., 1992), which may further indicate the importance of multiple domains making specific ligand contacts over a discrete and well-defined space.

Any of these arguments in isolation are probably not especially intriguing, yet given that domain order/spacing is apparently important to function and has probably not been altered since at least invertebrate-vertebrate divergence, that striking patterns of conservation among opposing loops of specific domain pairs is evident among orthologs, that intervening sequences connecting these same domain pairs in the case of Bravo are also conserved, and that alternative exons and ligand binding motifs in specific cases map between domains or onto near-by loops, the evidence, although circumstantial, is at least compelling. Strict domain order, therefore, may be important because individual domains may be acting in concert, with highly conserved sequences forming ligand-binding pockets in between.

## Conclusions

A cladistic analysis is the generation of a phylogenetic tree that shows the relationships among organisms by the principle of parsimony; i.e., by minimizing the number of steps or branch lengths in the tree. Cladistic tools were developed to infer evolutionary history among various species, and these relationships traditionally have been established by gross morphological or behavioral character states. With the progress in DNA technology over the past couple decades, the use



of orthologous gene sequences as character states provides a wealth of high-resolution data that can be used to infer phylogenetic relationships.

So, why is any of this of any consequence to molecular biology? Why is the study of evolution relevant to our quest to understand how modern genes function in modern organisms? Historians defend their academic pursuits with the argument that the past teaches us something about the present, and similarly, so might the history of genes, their duplications, divergences, and patterns of conservation likewise provide important insights into their current functions. The key to this premise is the fact that cladistics, in addition to simply generating a phylogenetic tree, also reconstruct internal ancestral states at the various branch nodes. Thus, by examining concise patterns of conservation and divergence of gene sequences along specific internal branches, one is permitted a glimpse of exactly where in the gene (i.e., at the resolution of functional motifs or domains) there is unusual levels of selective pressure at particular points in history.

The idea of inferring "functional hot spots" from evolutionary patterns of conservation and divergence is a novel and as yet untested application of phylogenetics. This chapter explores the evolutionary history of a family of neural cell surface proteins, and the results of these phylogenetic experiments make specific, testable functional predictions which may even be useful in the design of future experiments. Whether or not this is the case, it is certainly true that an increasingly important question has emerged in the current climate of various genome initiatives: What do all these sequences mean? In this regard, it is my belief that a marriage between those who think about genes and the protein structures they encode, and those who think about genome evolution and molecular phylogeny is not only inevitable, but perhaps will prove enormously fruitful for those who wish to eventually sort out functions among gene families.

**Acknowledgment**

We thank Daniel Vaughn for critical reading of the manuscript as well as valuable comments with regard to crystal structure data and assessment of structurally homologous alignments.

## Chapter II

### References

Atashi, J.R., Klinz, S.G., Ingraham, C.A., Matten, W.T., Schachner, M., and Maness, P.F. (1992). Neural cell adhesion molecules modulate tyrosine phosphorylation of tublin in nerve growth cone membranes. *Neuron* 8 (5), 831-842.

Davis, J.Q., and Bennett, V. (1994). Ankryin binding activity shared by the neurofascin/L1/Nr-Cam family of nervous system cell adhesion molecules. *J. Biol. Chem.* 269, 27163-27166.

Doherty, P., Ashton, S.V., and Moore, S.E. (1991). Morphoregulatory activities of N-CAM and N-Cadherin can be accounted for by G-protein-dependent activations of L-type and N-type neuronal calcium channels. *Cell* 67 (1), 21-33.

Doherty, P., Williams, E., and Walsh, F.S. (1995). A soluble chimeric form of the L1 glycoprotein stimulates neurite outgrowth. *Neuron* 14, 57-66.

Goldman, S.A., Williams, S., Borami, K., Lemmon, V., and Nedergaard, M. (1996). Transient coupling of Ng-CAM expression to Ng-CAM-dependent calcium signaling during migration of new neurons in the adult songbird brain. *Molecular Cellular Neuroscience* 7 (1), 29-45.

Goodman, C.S., and Schatz, C.J. (1993). Developmental mechanisms that generate precise patterns of neuronal connectivity. *Cell* 72 (Suppl), 77-98.

Holt, C.E. and Harris, W.A. (1993). Position, guidance, and mapping in the developing visual system. *J. Neurobiology* 24 (10), 1400-1422.

Jones, E.Y., Harlos, K., Bottomley, M.J., Robinson, R.C., Driscoll, P.C., Edwards, R.M., Clements, J.M., Dudgeon, T.J., and Stuart, D.I. (1995). Crystal structure of an integrin-binding fragment of vascular cell adhesion molecule-1 at 1.8 Å resolution. *Nature* 373, 539-544.

Kayyem, J.F. (1992). Thesis: Bravo, a new immunoglobulin superfamily member in the developing nervous system, is identified using a new method. California Institute of Technology.

Klinz, S.G., Schachner, M., and Moness, P.F. (1995). L1 and N-CAM antibodies trigger protein phosphorylation activity in growth cone enriched membranes. *Journal of Neurochemistry* 65, 84-95.

Lane, R.P., Chen, X-N., Yamakawa, K., Vielmetter, J., Korenberg, J.R., and Dreyer, W.J. (1996). Characterization of a highly conserved human homolog to the chicken neural cell surface protein Bravo/Nr-CAM that maps to chromosome band 7q31.

Maddison, W.P., Donoghue, M.J., and Maddison, D.R. (1984). Outgroup analysis and parsimony. *Syst Zool* 33 (1), 83-103.

Mauro, V.P., Krushel, L.A., Cunningham, B.A., and Edelman, G.M. (1992). Homophilic and heterophilic binding activities of Nr-CAM, a nervous system cell adhesion molecule. *J. Cell Biol.* 119, 191-202.

Morales, G., Hubert, M., Brummendorf, T., Treubert, U., Tarnok, A., Schwarz, U., and Rathjen, F.G. (1993). Induction of axonal growth by heterophilic interactions between the cell surface recognition protein f11 and protein Nr-CAM/Bravo. *Neuron* 11 (6), 1113-1122.

Otsuka, A.J., Franco, R., Yang, B., Shim, K.H., Tang, L.Z., Zhang, Y.Y., Boontrakulpoontawee, P., Jeyaprakash, A., and Hedgecock, E. (1995). An ankryin-related gene (*unc-44*) is necessary for proper axonal guidance in *Caenorhabditis elegans*. *J. Cell Biol.* 129 (4), 1081-1092.

Suter, D.M., Pollerberg, G.E., Buchstaller, A., Giger, R.J., Dreyer, W.J., and Sonderegger, P. (1995). Binding between the neural cell adhesion molecules axonin-1 and Nr-CAM/Bravo is involved in neuron-glia interaction. *J. Cell Biol.* 131, 1067-1081.

Vaughn, D.E., and Bjorkman, P.J. (1996). The (Greek) key to structures of neural adhesion molecules. *Neuron* 16, 261-273.

Williams, A.F., and Barclay, A.N. (1988). The immunoglobulin superfamily -- domains for cell surface recognition. *Annu. Rev. Immunol.* 6. 381-405.

Williams, E.J., Furness, J., Walsh, F.S., and Doherty, P. (1994). Activation of the FGF receptor underlies neurite outgrowth stimulated by L1, N-CAM, and N-Cadherin. *Neuron* 13, 583-594.

Williams, E.J., Mittal, B., Walsh, F.S., Doherty, P. (1995). A calcium/calmodulin kinase inhibitor, KN-62, inhibits neurite outgrowth stimulated by CAMs and FGF. *Molecular and Cellular Neurosciences* 6 (1), 69-79.

Wong, E.V., Schaefer, A.W., Landreth, G., and Lemmon, V. (1996). Involvement of P90 (RSK) in neurite outgrowth mediated by the cell adhesion molecule L1. *Journal of Biological Chemistry* 271 (30), 18217-18223.

Zisch, A.H., Dalessandri, L., Ranscht, B., Falchetto, R., Winterhalter, K.H., and Vaughan, L. (1992). Neuronal cell adhesion molecule contactin/F11 binds to tenascin via its Immunoglobulin-like domains. *Journal of Cell Biology* 119, 203-213.

**Table 1**      *Alignments and Blossum62 Stepmatrix*

All sequences used in this study are aligned and separated by predicted domain. Sequences shown are orthologous reconstructions of presumed ancestral states (see Methods) as utilized for Domain tree searches. Protein names are prefixed with the domain-number (1-6 for Ig-like domains, 1-5 for Fn domains); intracellular sequences are prefixed with "I-" in the bottom five rows. Alignments of specific species sequences are inferred: for example, the "I/V" at character 6 in "1Bravo" indicates that either amino acid occupied this position in the ancestral Ig I domain of Bravo, and that the isoleucine at this position in human Bravo and the valine at this position in chicken Bravo were aligned when generating Gene trees (i.e., in those studies in which reconstructions were not used). Structural components of each domain are indicated at the top of the table for the Ig-like domains, and again above the first alignments for Fn domains in row 66. The modified Blossum62 stepmatrix used for similarity searches follows these sequence alignments. Stars (\*) indicate gaps, and were given the maximum cost of 15 (as expensive as a leucine-to-glycine substitution, for example).

[illegible]

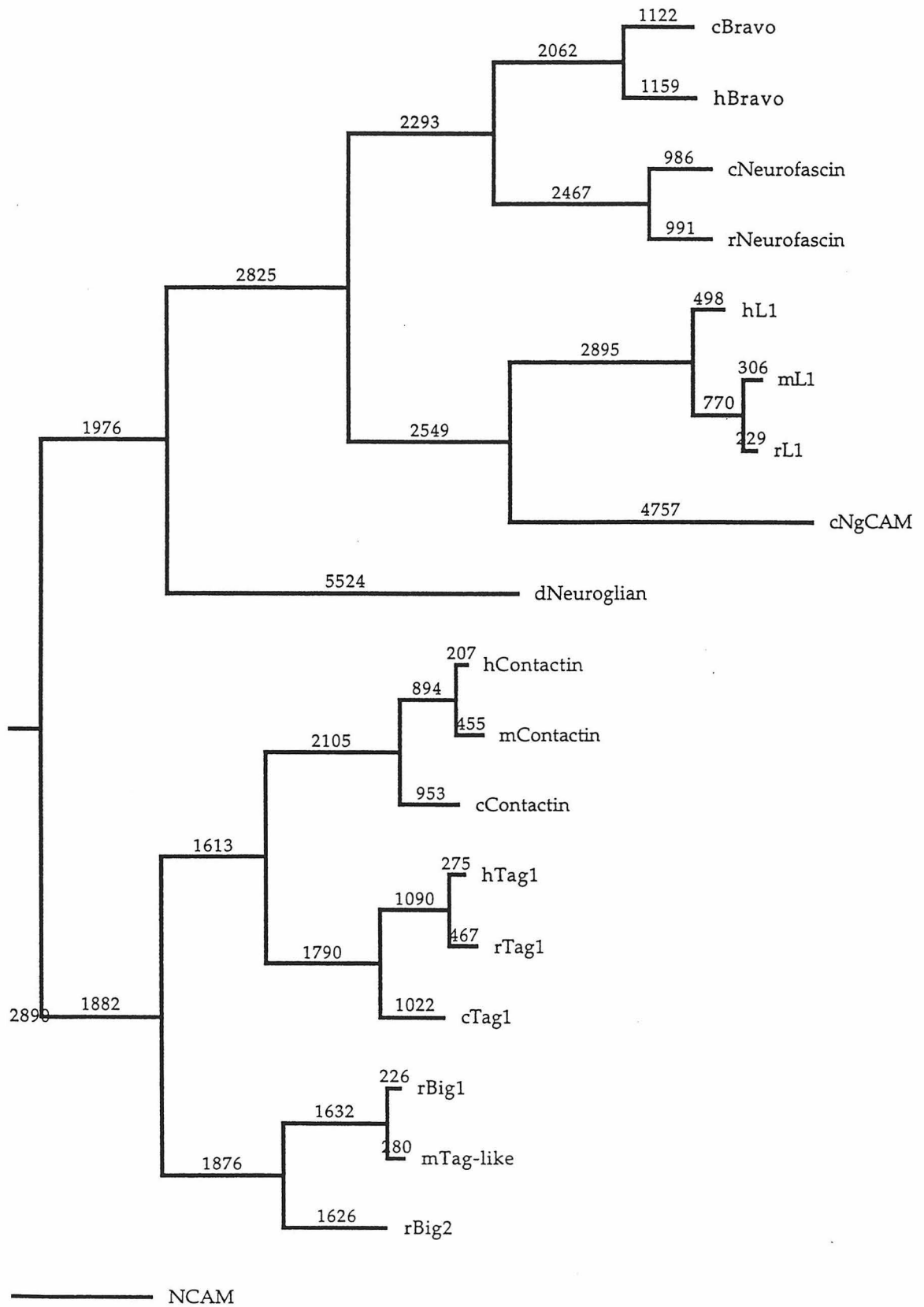




### Blossum62 Stepmatrix

**Figure 1**      *Gene Tree*

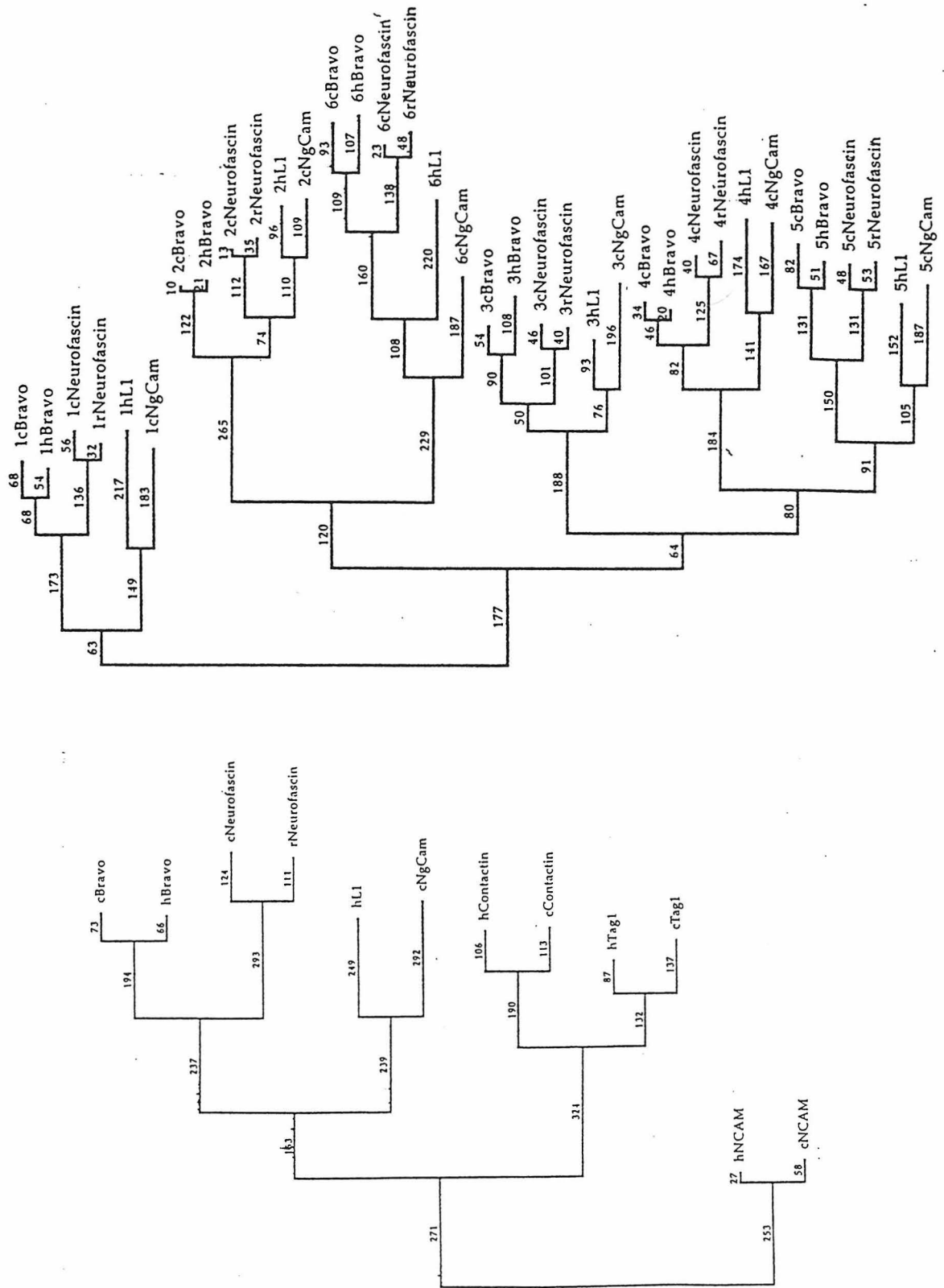
Phylogram illustrating the shortest similarity tree (rooted by an N-CAM reconstructed ancestral protein sequence) relating the entire subfamily of six-Ig-like domain cell adhesion proteins. The tree was generated by a heuristic Paup3.1.1 search and its total length is 54692 steps (scored by Blossum62 stepmatrix, see Table 1). Branch lengths are indicated, and the species from which the various molecules come are indicated by a small-case prefix letter: d=*Drosophila*, c=chicken, h=human, r=rat, m=mouse. An identical shortest tree is generated by an unordered/identities search and with an alternative alignment set (data not shown).



**Figure 2**     *Avian-Mammalian Trees*

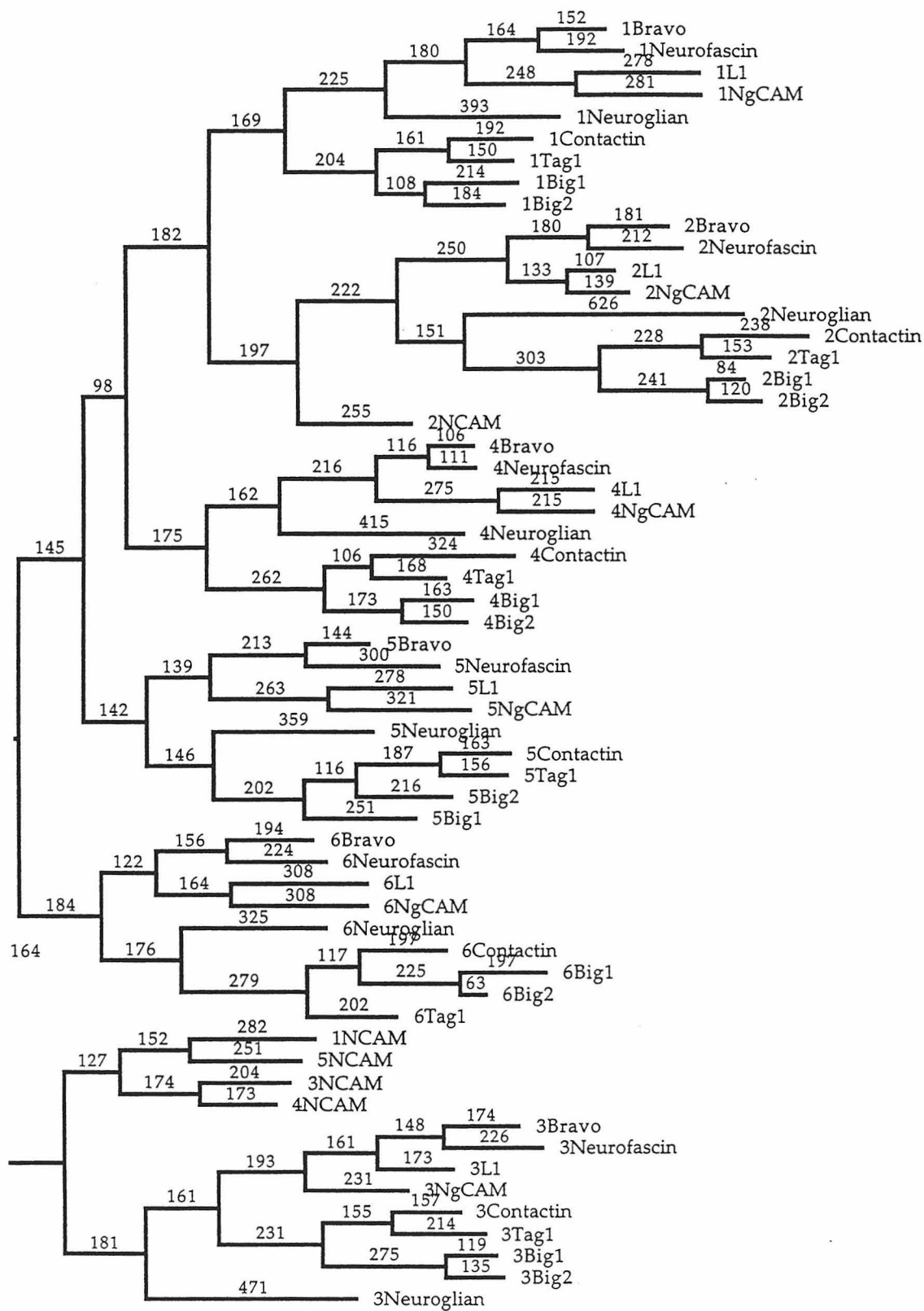
Left: Phylogram illustrating the shortest similarity (Blossum62) tree relating various avian and mammalian orthologs in the subfamily, including a pair of outgroup N-CAM orthologs. The tree was generated by a heuristic Paup3.1.1 search. Branch lengths are indicated, and the avian molecules in all cases are from chicken (indicated with the prefix "c"); the mammalian molecules are from human (indicated with the prefix "h") except for Neurofascin, which is from rat (indicated with the prefix "r").

Right: Phylogram illustrating the shortest similarity (Bossum62) Domain tree relating the individual Ig-like domains for the avian-mammalian orthologs of the Bravo, Neurofascin, L1 and Ng-CAM *sub*-subfamily. The tree was generated by a heuristic Paup3.1.1 search. Branch lengths are shown above branches, and the domain-number (1-6) is indicated as a prefix in the OTU labels. The species is also indicated in the OTU labels (c=chicken, r=rat, h=human), which immediately precedes the name of the molecule.



**Figure 3**     *Ig-like Domain Tree*

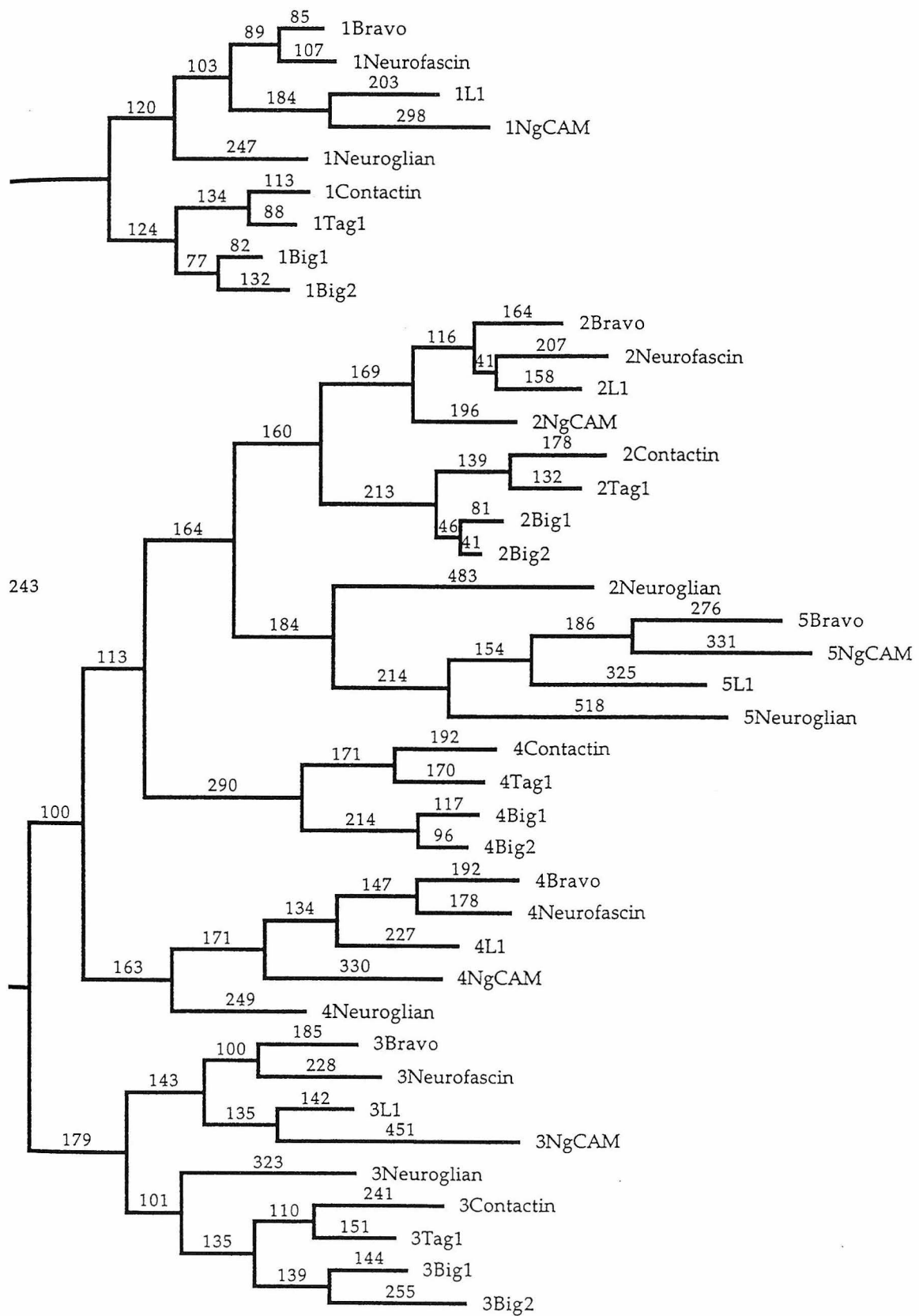
Phylogram illustrating the shortest similarity (Blossum62) Domain tree relating the individual Ig-like domains for the entire subfamily, as well as N-CAM outgroup domains. Orthologs were collapsed to derived ancestral states (see Methods) and the tree was generated by a heuristic Paup3.1.1 search. Individual branch lengths are indicated along branches and domain number (1-6) is indicated as a prefix to each ortholog name.





**Figure 4**      *Fibronectin Domain Tree*

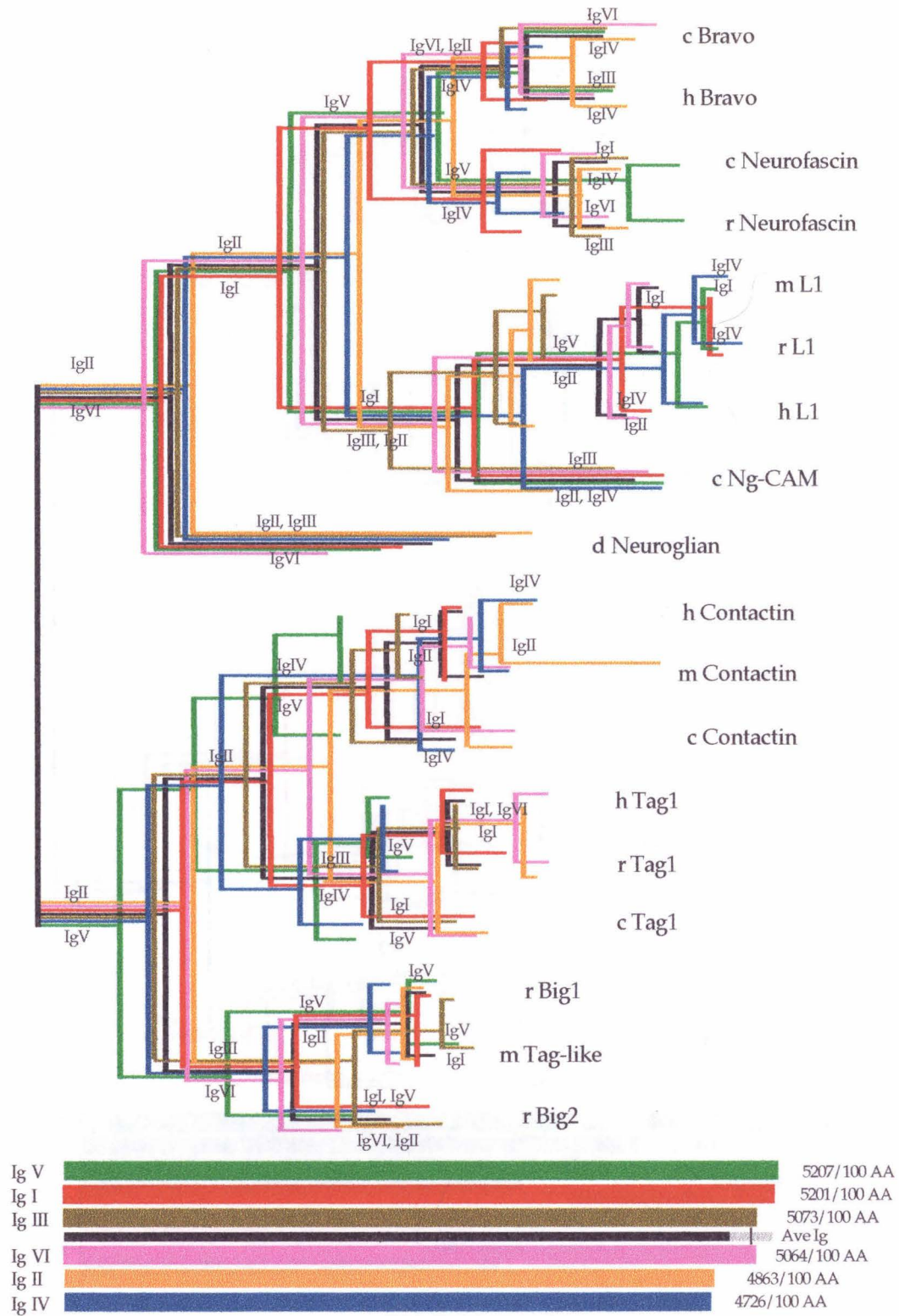
Phylogram illustrating the shortest similarity (Blossum62) Domain tree relating the individual Fn domains for the entire subfamily. Orthologs were collapsed to derived ancestral states (see Methods) and the tree was generated by a heuristic Paup3.1.1 search. Individual branch lengths are indicated along branches and domain number (1-5) is indicated as a prefix to each ortholog name.

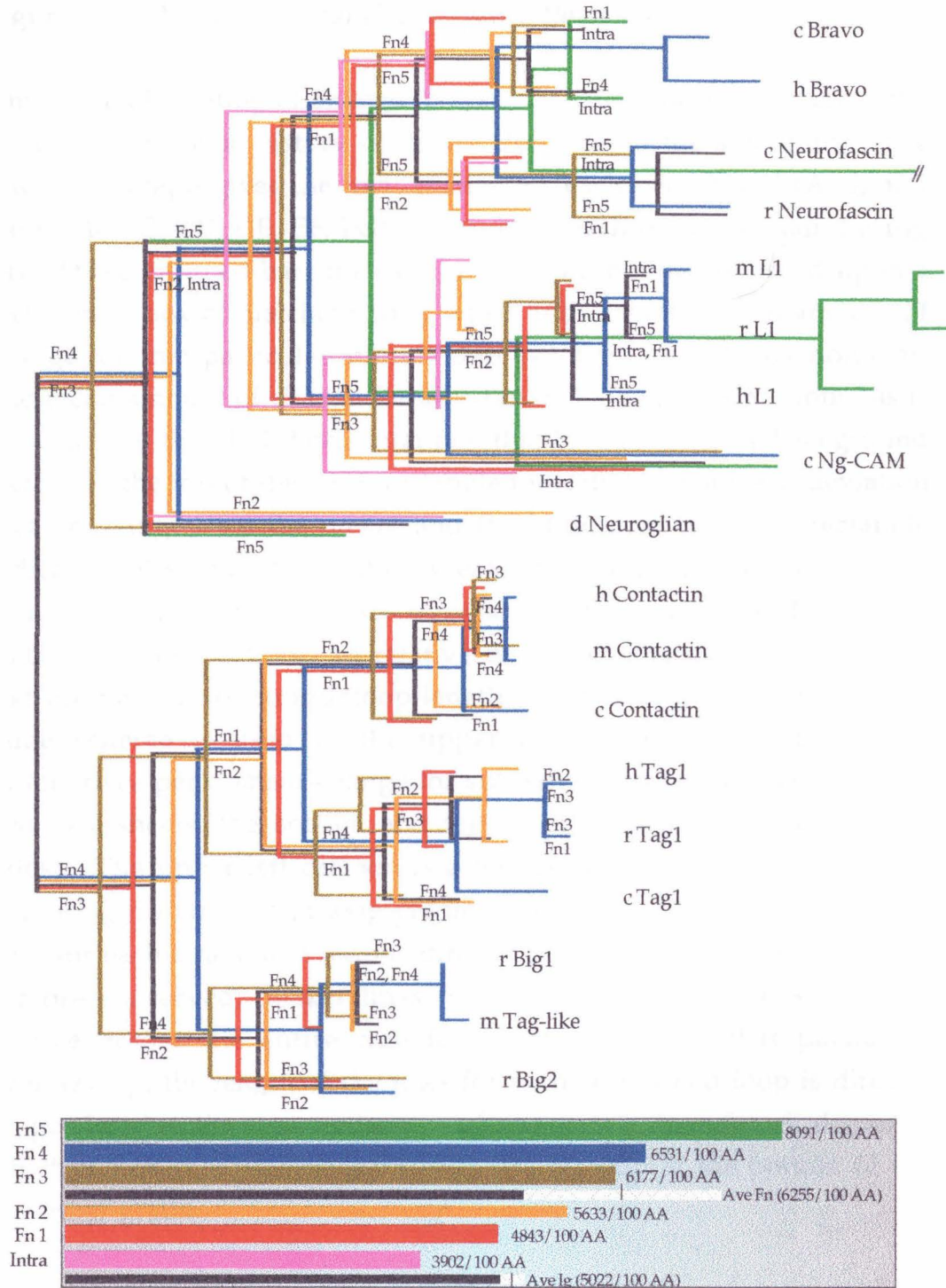


**Figure 5**      *Overlapping Domain Phylograms*

a) Superimposed (rooted) phylograms for each Ig-like domain. The relative lengths of the horizontal branches are proportional to the number of steps (i.e., amount of divergence) for each domain along each branch in the phylogenetic tree. Each domain is color-coded as indicated beneath the phylograms; the length of the color-coded bars is proportional to the total amount of amino acid change for each domain across the entire tree (normalized per 100 AA length). Above each set of branches in the trees, the domain(s) most significantly diverged (above the branches) or conserved (below the branches) is/are indicated. Along any particular branch, the top-to-bottom orientation of the colored lines is organized to illustrate most diverged-to-conserved domain along that specific branch.

b) Superimposed (rooted) phylograms for each Fn domain, as well as the intracellular domain. All other features of this figure are as described above. The bar graph below the trees extrapolates total divergence for the Fn 5 and Intracellular domains as if they were present on all branches: averaged for all Fn domains, the branches in which Fn5 is present and measured represents 58.6% of total divergence; averaged for all domains, the branches in which the intracellular domain is present and measured represents 56.5% of total divergence.



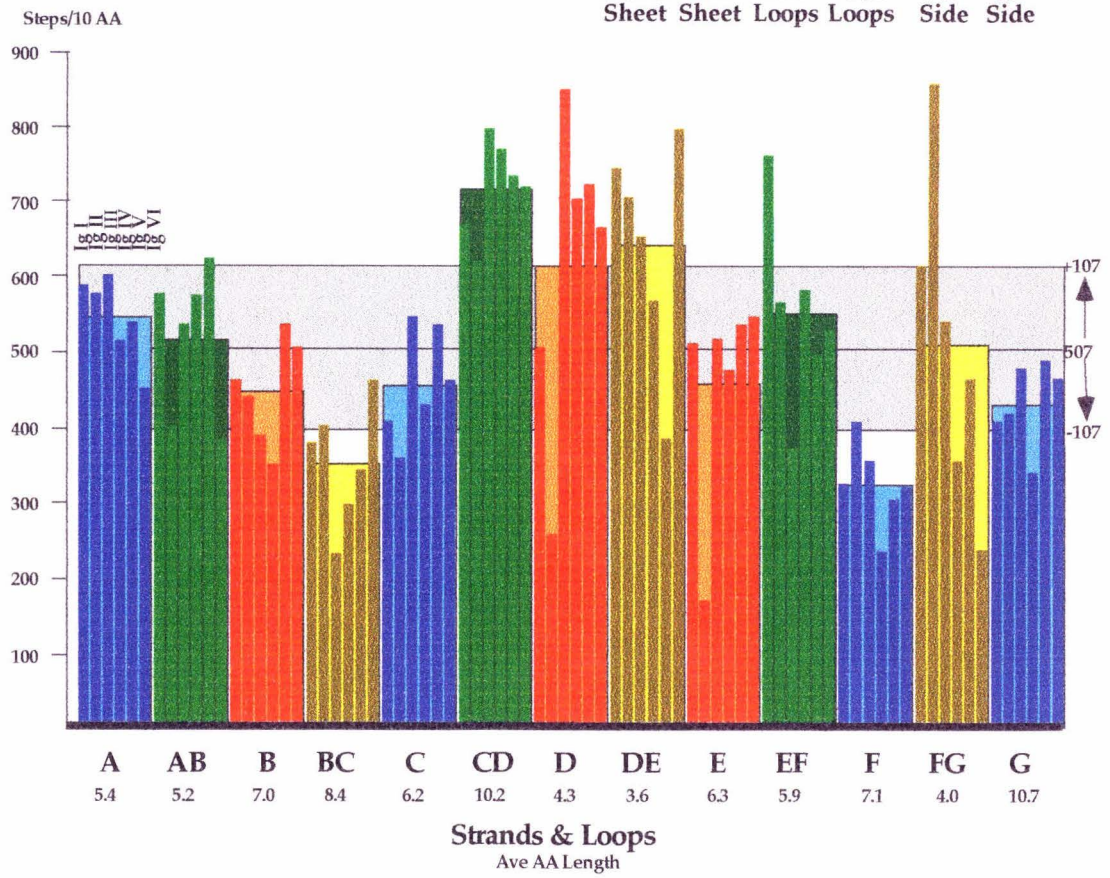
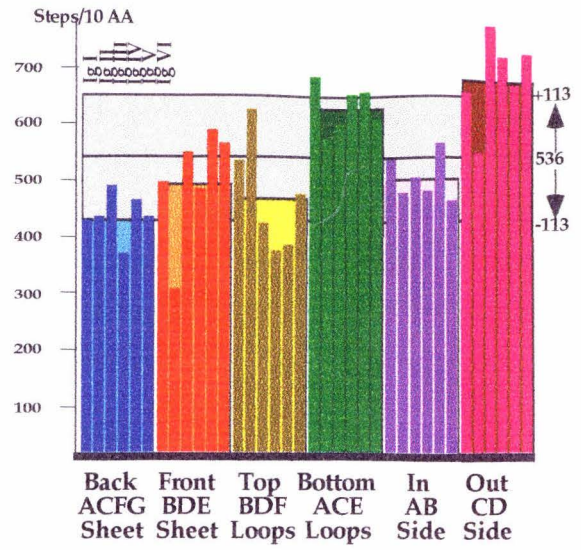
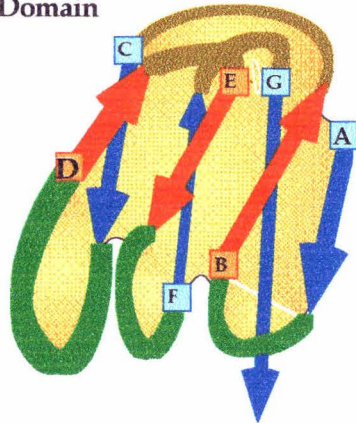




**Figure 6**     *Ig-like Structural Divergence Patterns*

Amount and location of change for specific structural motifs across the entire phylogenetic history of Ig-like domains in the subfamily. The lower bar graph gives the total change for each beta-strand (A-G) and loop (AB, BC, CD, DE, EF, FG) structure for each Ig-like domain (Ig I-Ig VI). These grouped bars for each strand and loop are overlaid upon a rectangular box of matching color which indicates the mean amount of change for that particular strand/loop averaged over all six domains. The mean amount of change for all strands/loops across all domains is indicated by the black horizontal line that bisects the gray background rectangle; the top of this gray rectangle indicates 1.0 standard deviation greater change than the mean, and the bottom of this gray rectangle indicates 1.0 standard deviation greater conservation (or less change) than the mean. The units are steps as determined by the Blossum62 matrix (i.e., proportional to relative degree of change or divergence) and normalized for strand/loop length (i.e., per 10 amino acids). This same scheme is used in the upper right inset bar graph which summarizes per domain change for six "structural landscapes" (i.e., the two beta-sheets, the bottom and top loops, and the inner and outer sides). The upper left cartoon is a model of a V-like Ig-like domain, indicating the strand and loop predicted topology. The thickness of the lines for each strand and loop is directly proportional (i.e., thinner lines = more conserved; thicker lines = more diverged) to the amount of change across the entire tree for all domains for that particular strand/loop; the length of the lines for each strand and loop is directly proportional to the average lengths of that strand/loop for all domains, which is also indicated beneath each strand/loop at the bottom of the lower bar graph.

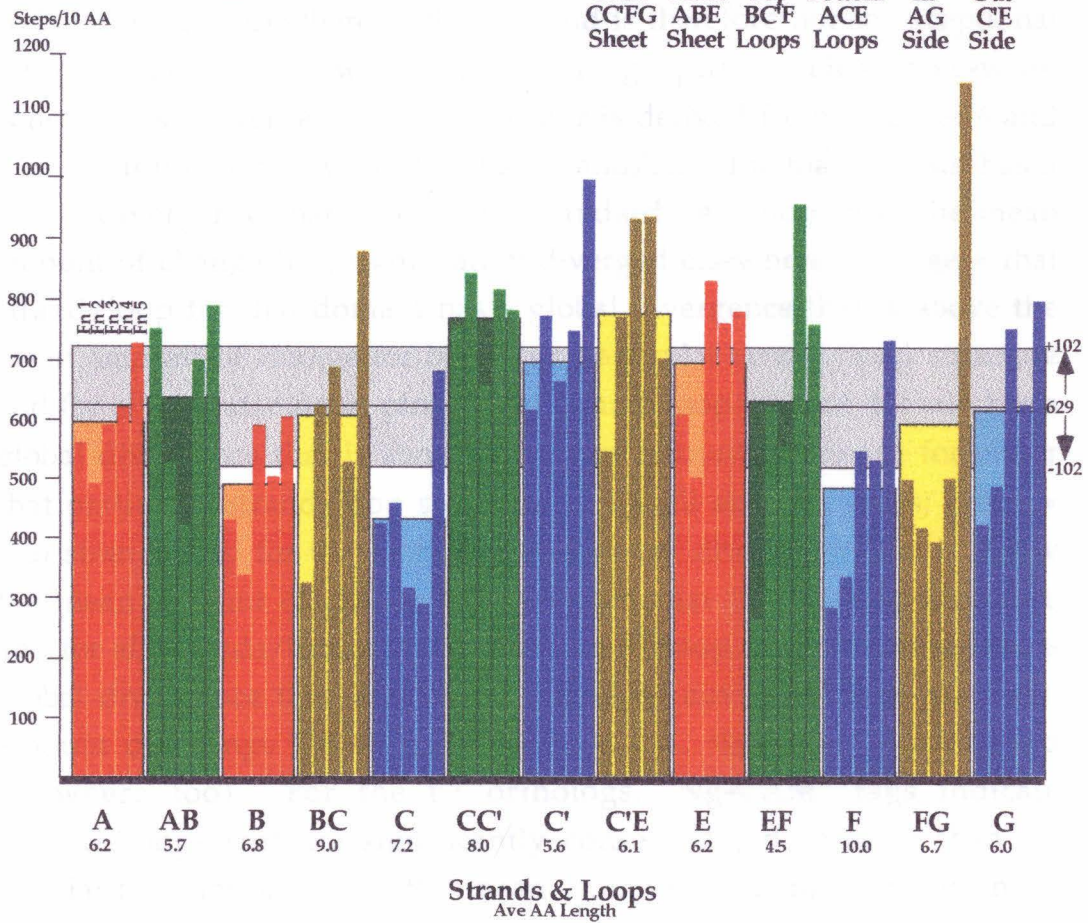
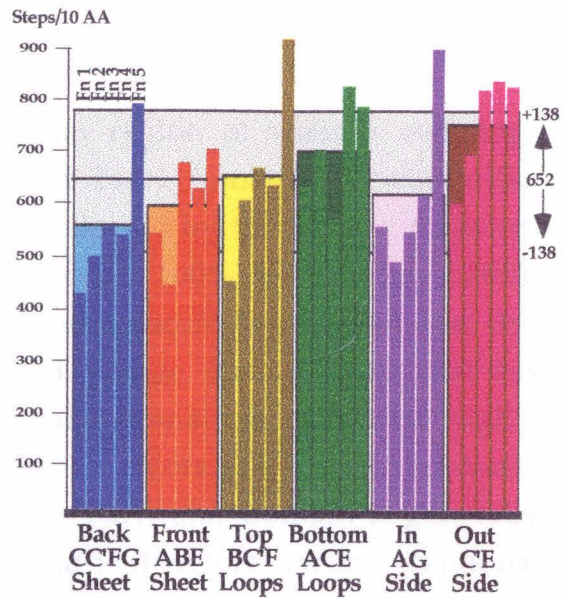
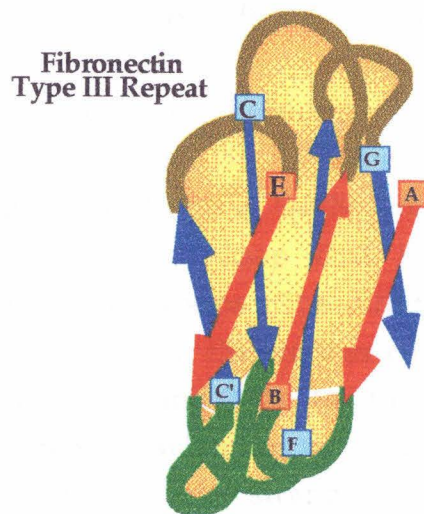
V-like Ig Domain



**Figure 7**      *Fn Domain Structural Divergence Patterns*

Amount and location of change for specific structural motifs across the entire phylogenetic history of Fn domains in the subfamily. The lower bar graph gives the total change for each beta-strand (A-G) and loop (AB, BC, CC', C'E, EF, FG) structure for each Fn domain (Fn 1-5). All other features of this figure are otherwise exactly as described for Figure 6. Note that the short EF loop, given the strand topology, places constraints on possible ways this cartoon can be drawn, which may likewise restrict/enforce tighter conformation *in vivo*.





**Figure 8**      *Orthologous Conservation Patterns*

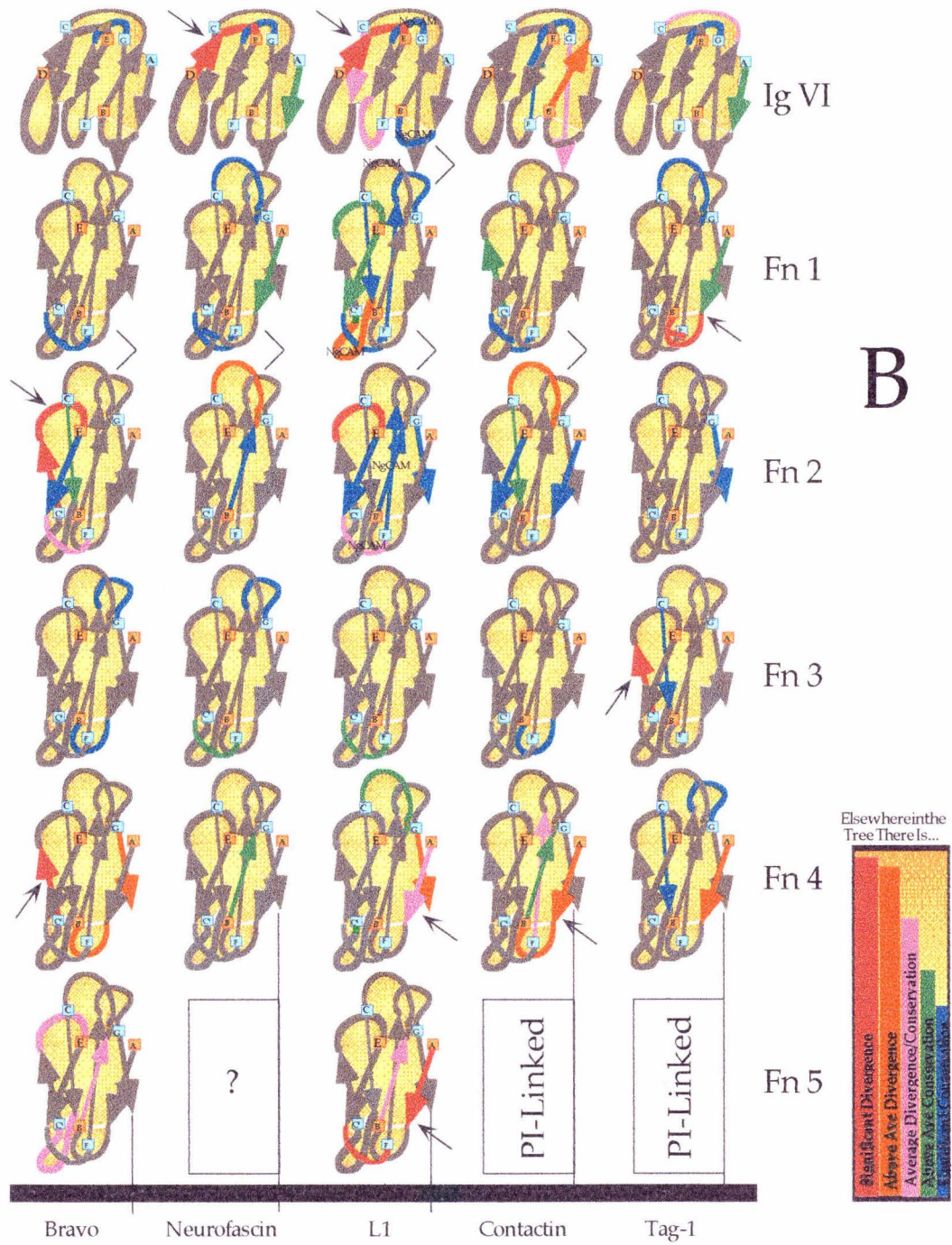
A) Models of the entire extracellular region of various orthologs (Bravo, Neurofascin, L1, Contactin and Tag-1) showing all six Ig-like domains (Ig I=top at the N-terminus) followed by the first Fn domains. Any strand or loop that is colored is 100% conserved within orthologous groups. The differing colors indicate how exceptional this conservation is: the warmer colors (red, orange, pink) indicate that this particular strand/loop for this particular domain is otherwise highly diverged elsewhere in the tree; the cooler colors (blue, green) indicate that this particular strand/loop for this particular domain is generally well conserved elsewhere in the tree and is therefore not so exceptional (summarized in lower right bar graph). This tree-wide conservation/divergence color indicator is derived from Figures 6 and 7 in the following way: red = that strand/loop for that domain has a global divergence that exceeds 1.0 standard deviation from the mean amount of change (i.e., significantly diverged elsewhere); orange = that strand/loop for that domain has a global divergence that is above the mean amount of change for both that particular strand/loop structure and for all strands/loops; pink = that strand/loop for that domain has a global divergence that is above the mean amount of change for either that particular strand/loop structure or for all strands/loops; green = that strand/loop for that domain has a global divergence that is below the mean amount of change for that particular strand/loop structure and for all strands/loops; blue = that strand/loop for that domain has a global divergence that is below 1.0 standard deviation from the mean amount of change for all strands/loops (i.e., significantly conserved elsewhere too). For the L1 orthologs, "Ng-CAM" tags indicate strands/loops that are significantly conserved (i.e., greater than 1.0 standard deviation below the mean amount of change) between L1 and Ng-CAM orthologs. Arrows indicate strands or sides that may be ligand-binding vicinities (see Discussion); left brackets (>) indicates

opposing loops that may indicate ligand-binding interdomain pockets (see Discussion).

B) Models of the entire extracellular region of various orthologs (Bravo, Neurofascin, L1, Contactin and Tag-1) showing the sixth Ig-like domains followed by the four or five Fn domains. All other features in this figure are exactly as described above. The box containing a question mark (?) for Neurofascin is meant to indicate that the sequence corresponding to the fifth Fn domain for this pair of orthologs is unrecognizable as a domain structure.







Chapter III

The Human Ortholog to the Bravo/Nr-CAM  
Neural Ig Subfamily Protein

Robert P. Lane, Xiao-Ning Chen, Kazuhiro Yamakawa, Jost Vielmetter,  
Julie R. Korenberg, and William J. Dreyer



### Introduction

The immunoglobulin superfamily (IgSF) is a large and diverse group of proteins that share in common the immunoglobulin (Ig) domain (Williams and Barclay, 1988). Ig domains are approximately 100 amino acids in length, and form two stacked beta-sheets, a topology also known as a beta-barrel structure. This structure is stabilized by characteristically spaced cysteine and tryptophan core residues (reviewed in Vaughn and Bjorkman, 1996). Although the domain is so-named because of its significance in numerous molecules of the immune system including the immunoglobulins, this same domain structure has been identified in a number of proteins in various tissues including the nervous systems of vertebrates and invertebrates. Collectively this neural subfamily of IgSF proteins are often referred to as cell adhesion molecules (CAMs). Neural cell adhesion molecules generally are cell surface proteins that consist of multiple Ig-like domains at the N-terminus, followed by multiple fibronectin type III repeats nearer the membrane, and either a transmembrane-spanning intracellular domain or a glyco-phosphatidylinositol (GPI)-linked membrane anchor at the C-terminus (Goodman and Schatz, 1993, Sonderegger and Rathjen, 1992).

The chicken Bravo/Nr-CAM molecule is a neural IgSF protein that is subgrouped with the evolutionarily related CAMs L1/Ng-CAM and neurofascin, each sharing a common domain structure of six N-terminal V-like Ig-like domains, five fibronectin type III repeats, a transmembrane region and intracellular domain (Grumet *et al.*, 1991; Kayyem *et al.*, 1992). Numerous studies on chicken embryos suggest Bravo plays an important role in the development of the vertebrate nervous system. In the retinotectal projection of retinal ganglion cell axons to their targets in the optic tectum, Bravo expression correlates with the onset of neuronal differentiation. Bravo is expressed on

growing neurites as well as on radial cells including their ventricular endfeet which are believed to provide a substrate for ganglion cell axons as they grow en route to the optic fissure and optic nerve exit (Kayyem *et al.*, 1992; de la Rosa *et al.*, 1990). As the axons leave the optic chiasm and project contralaterally, Bravo is down-regulated and little or no expression is observed on the distal portions of these axons. This is coincident with changes in nerve fiber order that occur as these axons approach their target sites in the tectum (Scholes, 1981). It is possible that the establishment of the retinotectal projection involves both homophilic (Mauro *et al.*, 1992) and heterophilic (Suter *et al.*, 1995; Morales *et al.*, 1993) interactions of Bravo with other neural cell surface proteins that allow retinal ganglion axons to be correctly guided and/or appropriately fasciculated. In the developing spinal cord, an interaction between Bravo and another IgSF molecule, axonin-1, is necessary to guide commissural axons along correct pathways when transversing the floorplate in the spinal cord (Stoeckli *et al.*, 1995). These studies and others on closely related cell adhesion molecules suggest that this family of proteins play diverse roles in the assembly of the nervous system, including the regulation of neurite extension, pathfinding, and synapse formation (reviewed in Goodman and Schatz, 1993).

The cell adhesion nomenclature implies that this large and complex group of proteins merely play an adhesive role between cells. It is becoming increasingly clear, however, that many cell adhesion molecules function as receptors whose interactions result in intracellular signalling. One intriguing example is the trophic/growth-stimulating effect of one of Bravo's closest IgSF relative, L1, whose interaction with the fibroblast growth factor receptor (FGF-R) is sufficient to bring about neurite extension via FGF-R phosphorylation events (Doherty *et al.*, 1995; Williams *et al.*, 1994). Core sequence motifs responsible for this interaction, as well as interactions between the FGF-R and N-CAM, another IgSF protein, and N-Cadherin, a cell



surface protein mediating  $\text{Ca}^{2+}$ -dependent cell adhesion, are also partly conserved in Bravo leaving open the possibility that similar interactions with growth factor receptors may be widespread among cell adhesion molecules. IgSF proteins may also bring about intracellular changes mediated by direct interactions of their cytoplasmic domains. L1, Bravo and neurofascin, for example, may bring about specific growth and guidance effects in growth cones by virtue of direct interaction with intracellular cytostructure, as the intracellular domains of all three molecules interact with the spectrin-binding ankyrin protein (Davis and Bennett, 1994; Otsuka *et al.*, 1995).

We initiated studies to characterize the human Bravo homolog. Similar studies with L1 have been fruitful in linking the molecule with specific human disorders by virtue of mapping the gene to a chromosomal locus associated with a genetic defect (Jouet *et al.*, 1994). We report the cloning and sequencing of the human Bravo molecule and identify its genetic locus within human chromosome 7q31.1-31.2. To date, six genes have been assigned to this chromosomal region, including four sequenced genes that are clearly not Bravo, and two genes not yet sequenced and whose identity can not at this time be distinguished from Bravo. The latter two are a known tumor suppressor gene (Zenklusen *et al.*, 1994) and the gene strongly implicated in the connective tissue disorder *cutis laxa with marfanoid phenotype* (Bonneau *et al.*, 1991).

## Materials and Methods

### *Generation of Human Fetal cDNA Library.*

A 14 week trisomy 21 fetal brain cDNA library was constructed using the ZAP-cDNA synthesis kit (Stratagene) which is designed for construction of a unidirectional cDNA library (Yamakawa *et al.*, 1995). Briefly, double stranded cDNA was synthesized from 5ug trisomy 21

fetal brain poly(A)+ RNA using a hybrid oligo(dT)-XhoI linker primer with 5-methyl dCTP. EcoRI linkers were attached to the 5'-ends and products were digested with EcoRI and XhoI, and cloned into the UNI-ZAP XR vector. The library was packaged using the GigapackII Gold packaging extract (Stratagene). The number of independent clones was  $1.1 \times 10^6$ . The library was amplified once, and a blue-white color assay indicated that 99% of the clones had inserts with an average size of 1.9kb (calculated from 14 clones).

#### *Isolation and Sequence Analysis of Human Bravo cDNA Clones.*

Using the chicken Bravo sequence, oligonucleotides were synthesized (by the Caltech microchemical facility; sense oligonucleotide 5' CAAATCCACCGCTTGACTTG, and antisense oligonucleotide 5' TGGCCATTTACACCTTCTCC) and utilized in low stringency PCR using the human fetal library as template. A 1.6 kb PCR product (nucleotides 1939-3794 with missing alternative exon sequence) was cloned into the pCR<sup>TM</sup>II vector using the TA cloning site (Invitrogen). Preliminary DNA sequencing on the plasmid DNA confirmed the isolation of the human Bravo homolog. The 1.6 kb human Bravo PCR product was labeled with [ $\alpha$ -<sup>32</sup>P]dCTP (New England Nuclear) incorporation by random priming (Boehringer Mannheim), and used to screen the human fetal library. The library was plated at a titer of approximately  $10^5$  pfu/plate ( $10^6$  total plaques). Replica filters (Hybond-N, Amersham) were lifted and hybridized at high stringency (50% formamide, 4x SSPE, 1% SDS, 0.5% Carnation Non-Fat Dry Milk, 0.1% 37°C shaking overnight); the filters were washed at increasing stringency (highest stringency: 0.1x SSC, 0.1% SDS, 65°C for 10 min). Ten positive cDNA clones were isolated, and all were transformed into pBluescript plasmids by *in vivo* excision (Stratagene). One of the ten clones was determined to contain the complete coding sequence of the human Bravo cDNA, and this template was selected for sequence analysis. Both strands were

sequenced using synthetic oligonucleotide primers and dideoxy chain termination (Sanger *et al.*, 1977; USB Sequenase Version 2.0). Three of the other isolated cDNA clones were partly sequenced by these same methods in order to investigate alternative splicing. All sequences were analyzed and compared using the GCG system software package.

#### *Northern Blot Analysis.*

An adult human RNA blot was obtained from Clontech Laboratories, Inc., which contained blotted RNA populations from various brain tissues: amygdala, caudate nucleus, corpus callosum, hippocampus, hypothalamus, substantia nigra, subthalamic nucleus and thalamus. A 1.1 kb PCR product was generated using synthetic oligonucleotide primers complementary to N-terminal immunoglobulin domain sequences (between nucleotides 488 and 1596). This PCR product was labeled with [ $\alpha$ - $^{32}$ P]dCTP (New England Nuclear) incorporation by random priming (Boehringer Mannheim kit) and used as a probe to hybridize at high stringency (50% formamide, 5x SSPE, 2% SDS, 10x Denhardt's solution, 100  $\mu$ g/ml sheared salmon sperm DNA); the blot was washed at increasing stringency (0.1x SSC, 0.1% SDS, 65°C for 10 min). The labelled autoradiograph was exposed for 28 h.

#### *Fluorescence in situ Hybridization (FISH).*

The originally isolated 1.6 kb human Bravo clone (PCR generated between nucleotides 1939-3974 excluding alternative FNIII-5 exon sequence) was used as a probe to map the gene to human chromosomes by *fluorescence in situ hybridization* (FISH). The probe was labeled with biotin-14-dATP (GIBCO, BRL) using nick-translation and was subsequently hybridized to metaphase chromosomes prepared from normal male peripheral blood lymphocytes using the bromodeoxy-uridine synchronization method (Korenberg and Chen, 1996). FISH was performed essentially according to the method

described in Korenberg and Chen (1996). The hybridization solution contained 200 ng probe DNA, 5 µg Cot 1 DNA and 5 µg sonicated salmon sperm DNA per 10 µl hybridization solution (50% formamide, 10% dextran sulfate, 2x Sodium Saline Citrate (SSC)). The DNA was pre-annealed at 37°C for 15 minutes and then applied to denatured chromosome slides. Post-hybridization washes (4x) were performed at 44°C for 5 minutes each in 2xSSC and 50% formamide, followed by (4x) at 55°C for 5 minutes in 1xSSC. Hybridized DNAs were detected by using avidin-conjugated fluorescein isothiocyanate (Vector Labs). One amplification was carried out using biotinylated-anti-avidin (Vector Labs). To generate clear reverse bands, metaphase chromosomes were counterstained with chromomycin A3 followed by distamycin A (Korenberg and Chen, 1996). The image was captured by using the Photometrics Cooled-CCD camera (CH250) and the BDS image analysis system (Oncor Imaging, Gaithersburg, MD).

## Results

### *Sequence Analysis of Human Bravo /Nr-CAM cDNA*

Ten independent cDNA clones encoding four isoforms of human Bravo/Nr-CAM were isolated, including a full-length cDNA which was selected for sequencing (Fig 1). This full-length clone contains an open reading frame (ORF) with two possible ATG translation initiation codons at the 5' end and a stop codon 3912bp downstream of the first more upstream ATG. The predicted amino acid sequence begins with a hydrophobic stretch that shares significant identity (58%) with the signal peptide of chicken Bravo/Nr-CAM, including an eight amino acid sequence just upstream of the cleavage site that is identical to the chicken signal peptide. When the signal peptides of the two homologs are aligned, the second more 3' ATG start codon aligns with the chicken Bravo/Nr-CAM ATG start codon; the first more upstream

chicken Bravo/Nr-CAM ATG start codon; the first more upstream ATG start codon if used would result in a signal peptide that is five amino acids longer. No Kozak consensus translation start signal sequence (Kozak, 1989) is present, however a stop codon is encoded only 8 codons upstream of the first ATG and we therefore conclude that one of these two start codons is the 5' end of the mRNA. The N-terminus of the mature protein is predicted by homology (Fig 2) to the chicken Bravo N-terminal protein sequence which has been determined previously (de la Rosa *et al.*, 1990).

The amino acid sequence of human Bravo includes several motifs and predicted structural domains. There are six predicted N-terminal immunoglobulin-like (Ig-like) domains based on characteristically spaced cysteine and tryptophan core residues that are required for the properly folded beta-sheet domain structure. These Ig-like domains are followed by five fibronectin type III repeats (FNIII), which are predicted based on characteristically spaced tryptophans and tyrosines that are also structurally critical core residues (Kayyem *et al.*, 1992; reviewed in Vaughn and Bjorkman, 1996). A 23 amino acid hydrophobic stretch predicts a transmembrane spanning region followed by a C-terminal 114 amino acid intracellular domain. The extracellular Ig-like and FNIII domains contain 17 motifs for asparagine-linked glycosylation (Marshall, 1972) and the intracellular domain contains 7 potential serine/threonine phosphorylation sites (Woodgett *et al.*, 1986).

#### *Homology between the Chicken and Human Bravo /Nr-CAM Proteins*

Overall, the amino acid sequence of the human Bravo/Nr-CAM protein is 82% identical to chicken Bravo/Nr-CAM, the highest degree of identity between the human molecule and any amino acid sequence in the databases (Figure 2). The structural topology of six Ig-like domains, five fibronectin type III repeats, transmembrane and intracellular domain is identical to chicken Bravo, as well as to L1, Ng-CAM and neurofascin, and together, these proteins are subgrouped as

evolutionarily close relatives in the Ig Superfamily of cell adhesion molecules (Kayyem *et al.*, 1992). Individual extracellular domains of human Bravo range from 66% identical (Ig VI) to 93% identical (Ig IV) to its chicken homolog. The transmembrane and intracellular domains of the human Bravo protein are 100% identical between the orthologs.

Several alternatively spliced isoforms have been identified for chicken Bravo, including a 19 amino acid alternative exon (AE19) between Ig II and Ig III, a 10 amino acid alternative exon (AE10) between Ig VI and FNIII-1, a 12 amino acid alternative exon (AE12) just 5' to FNIII-5, and a 93 amino acid alternative exon (AE93) representing the entire fifth fibronectin repeat. The 19 amino acid sequence homologous to chicken AE19, and the 10 amino acid sequence homologous to chicken AE10 are both present in all 10 human Bravo cDNA's that were isolated, and therefore there is no evidence at this point of alternative splicing of these exons in human Bravo/Nr-CAM. Alternative splicing around the fifth fibronectin type III repeat (FNIII-5) however, is diverse and at least partly conserved between the orthologs. In both species, an isoform that is missing AE93 (the entire fifth fibronectin type III repeat) has been identified. In addition, there are two chicken isoforms and three human isoforms that have not yet been identified across species, and each of these variants is illustrated in Figure 3.

#### *Northern Blot Analysis of Human Bravo/Nr-CAM*

A 1.1kb probe (between nucleotides 488 and 1596) was generated by PCR and used to probe a Northern blot (obtained from Clontech Laboratories) of RNA from various brain tissues: amygdala, caudate nucleus, corpus callosum, hippocampus, hypothalamus, substantia nigra, subthalamic nucleus and thalamus. A 7.0 kb band was detected in all tissues examined, and one of these brain tissue blots is shown in Figure 4.

*Human Bravo /Nr-CAM Maps to Human Chromosome 7q31.1-2.*

The PCR-generated human Bravo/Nr-CAM fragment that was used as a probe to screen the human embryonic library for a full-length cDNA was also used for fluorescence *in situ* hybridization on metaphase chromosomes prepared from normal male peripheral blood lymphocytes. The human Bravo gene was mapped to chromosome band 7q31.1-31.2 (Fig. 5). Two independent experiments were performed, and over 100 metaphase cells were evaluated. Signals were clearly detected on two chromatids of at least one chromosome band 7q31.1-31.2 in 45% of cells. No other chromosomal sites with consistent signals were detected in greater than 1.0% of cells. To date, six genes have been mapped to the q31 band of chromosome 7, including four genes whose sequence is known and identity is distinct from Bravo: the met proto-oncogene (MET hepatocyte growth factor receptor), wingless-type MMTV integration site 2, human homolog (WNT2), sperm adhesion molecule 1 (SPAM1), and human CAP Z (CAPZA2). Two other genes have been assigned to this locus whose sequences have not yet been determined and therefore can not be distinguished from Bravo. The first gene has been mapped between 7q31-q32 and is strongly implicated in the human condition *cutis laxa with marfanoid phenotype* which is a lethal connective tissue disorder (Bonneau *et al.*, 1991). There is evidence that this gene is one of the subunits of the extracellular laminin protein, and therefore Bravo/Nr-CAM is not a likely candidate. The second gene is a presumed tumor suppressor gene that is mapped to the 7q31.1-31.2 chromosomal region by loss of heterozygosity (LOH) analysis (Zenklusen *et al.*, 1994).

## Discussion

### *Structural Model and Sequence Features*

We have identified and report the cDNA sequence of the



human Bravo/Nr-CAM molecule, and have mapped its genetic locus to human chromosome 7q31.1-31.2. Overall there is 82% amino acid sequence identity between the chicken and human homologs. Remarkably, the C-terminal 154 amino acids that include the transmembrane and intracellular domains are **100%** conserved. The two Bravo homologs are subgrouped in the Ig Superfamily and are structurally related to Ng-CAM, L1 and Neurofascin. Each consists of six immunoglobulin-like (Ig-like) domains, five fibronectin type III repeats (FNIII), a transmembrane spanning domain, and a cytoplasmic intracellular domain. Evolutionary relationships to the human Bravo protein are illustrated in the family gene tree in Figure 6. It is interesting to note that for the Bravo homologs and in fact for nearly every molecule of the neural Ig Superfamily where sequences are known, approximately 20% amino acid divergence has been tolerated since the furcation of birds and mammals. The notable exception to this trend are the L1 and Ng-CAM homologs which indeed may not be functionally equivalent (see Chapter II).

#### *The Remarkable Conservation of the Intracellular Domain*

Each of the proteins in the subfamily that include human Bravo and its closest evolutionary relatives (Bravo, Neurofascin, L1, Ng-CAM) has an intracellular domain whose high degree of sequence conservation suggests functional significance. In particular, the intracellular domains of the mammalian and avian Bravo homologs have not a single amino acid substitution over more than 200 million years of evolution. If the coding sequence were permitted to accumulate mutations at a random rate over this period of time, every nucleotide in the sequence would have undergone mutation at least once (Wilson *et al.*, 1977; Jukes, 1980). Although part of this stringent conservation between the two Bravo homologs, as well as the Neurofascin and L1 intracellular domains, is accounted for by the common requirement of sequences to bind the spectrin-binding



cytoskeletal ankryin molecule (Davis and Bennett, 1994), the expectation is that there are other important intracellular ligands not yet identified.

#### *Kinase Motifs and Signal Transduction Considerations*

At least in the case of the mammalian and avian Bravo homologs, there are putative phosphorylation recognition motifs that suggest that the intracellular domains may be involved in several intermolecular interactions. Eight potential serine/threonine phosphorylation motifs have been conserved in the two proteins which are putative recognition sequences for casein kinase II, cAMP-dependent kinase, and protein kinase C, each of which may be involved with regulating a signal in a Bravo-mediated transduction pathway or interactions with other membrane spanning co-receptors. It is interesting to note that the transmembrane spanning sequence has also not changed over this period of evolution. This may further imply that membrane interactions such as receptor dimerization or construction of transduction machinery, may depend on transmembrane-spanning residues.

Recently it has been demonstrated that one of Bravo's closest evolutionary relatives, L1, manifests some of its trophic effects on cultured neurons via an interaction with the fibroblast growth factor receptor (FGF-R), a receptor tyrosine kinase. This receptor interaction evokes FGF-R kinase activity that ultimately leads to neurite outgrowth (Doherty *et al.*, 1995; Williams *et al.*, 1994). These studies demonstrated that an "AAPYW" motif, shared by the FGF-R and its co-receptors L1, N-CAM, and N-Cadherin (all of which bind FGF-R and elicit the trophic effect), was part of the minimal sequence responsible for these interactions. Interestingly, both the chicken and human Bravo homologs also have this "AAPYW" motif conserved, leaving open the possibility that Bravo may participate in similar growth factor receptor interactions.

### *Alternative Splicing of Human and Chick Bravo/Nr-CAM RNA's*

Alternative splicing of mRNA precursors is an important means to develop programmed functional diversity among many proteins including some of Bravo's closest evolutionary relatives (Doherty *et al.*, 1992, for example). The two Bravo homologs each have diverse isoform expression around the fifth fibronectin type III repeat, and including this study, six unique variants have been identified as well as a common isoform shared between the two species. The sequence identity between the chicken and human alternative exon AE93 is among the highest of any of the eleven extracellular domains (82% identity, see Figure 2). To date, no specific function has been attributed to this domain, although its membrane proximity might further suggest isoform-specific co-receptor coupling. It is interesting to note numerous examples so far identified where the alternative inclusion or exclusion of fibronectin type III repeats in extracellular proteins are observed, including tenascin (Dorries and Schachner, 1994; Carnemolla *et al.*, 1992; Tucker *et al.*, 1994) and fibronectin (Kaczmarak *et al.*, 1994).

### *Could Bravo be a Tumor Suppressor Candidate?*

Within the resolution of the FISH technique, the human Bravo gene maps to the chromosomal locus of a known tumor suppressor gene (Zenklusen *et al.*, 1994), raising an interesting question: Is it possible that Bravo might indeed be a tumor suppressor molecule? While the 7q31.1-31.2 chromosomal band might contain as many as 75 genes (Fields *et al.*, 1994), there are some considerations that lead us to believe that Bravo should not be ruled out as a candidate. First, Bravo is expressed early on the cell surface of neural precursor cells at a stage when they cease to divide and begin to differentiate into post-mitotic neurons (Kayyem *et al.*, 1992). This is compatible with the notion that Bravo may be involved in cell-cell interactions that remove stem cells from the cell cycle and trigger differentiation events, including axonal outgrowth in neuronal cells. Second, an Ig Superfamily molecule

Deleted in Colorectal Cancer (DCC), which itself may be a tumor suppressor gene, is expressed in the nervous system in a developmental pattern similar to that of Bravo (Cho and Fearon, 1995; Vielmetter *et al.*, 1994). DCC may be one of the very few known examples of a membrane-spanning tumor suppressor gene, and like Bravo, is composed of immunoglobulin-like domains and fibronectin type III repeats. Finally, while we are accustomed to thinking about members of this family of cell surface receptors in the context of specialized functions of the developing nervous system, including neurite outgrowth, fasciculation, pathfinding and synapse formation, most may function in related ways in many tissues outside the nervous system. It is our expectation that important cell-cell recognition and signalling events in the development of numerous cell types and tissues will involve some of these same proteins (Fazeli *et al.*, 1995; Cho and Fearon, 1995). This recognition process may include signals that prevent abnormal cell growth or tumor development (Fearon *et al.*, 1990; Johnson, 1991; Marshall, 1991).

Although the exact molecular functions of Bravo remain to be elucidated, the striking sequence conservation of especially the membrane-spanning and intracellular domains between chickens and humans argues for critical functions involving a molecular pathway that likewise has been stringently conserved. Future studies that are directed at identifying ligands in these pathways, both outside and inside the cell, will likely provide important insights as to role of Bravo/Nr-CAM in development.

## Conclusions

When the chicken Bravo/Nr-CAM full coding sequence was first published (Grumet et al., 1991; Kayyem et al., 1992), it was identified as being a close homolog to the chicken Ng-CAM and mammalian L1 proteins. This trio of genes were approximately equally distant from each other, and sequence similarity scores between L1 and either of the chicken proteins were considerably lower than what would be expected for orthologous proteins (i.e., species or functional homologs). This prompted one of the authors of the chicken Bravo/Nr-CAM studies to propose that perhaps the single L1 protein in mammals accomplishes the full range of developmental functions that both avian Bravo/Nr-CAM and Ng-CAM proteins accomplish (Kayyem et al., 1992). In this way, functional redundancy between the two chicken proteins might explain both the equidistance of both to L1, as well as the higher than expected divergence among these orthologs. This hypothesis predicts that no avian L1 homolog exists (which is so far true) and that no mammalian Ng-CAM nor Bravo/Nr-CAM homolog exists (which is so far true in the former case). The current identification of the human ortholog to Bravo/Nr-CAM challenges this model, as does the recent identification of avian and mammalian Neurofascin proteins which further distinguish Bravo/Nr-CAM from L1/Ng-CAM in evolution. The question as to whether or not L1 and Ng-CAM are functionally equivalent and orthologous remains open, and this issue was discussed extensively in Chapter II.

Chapter III

References

Bonneau, D., Huret, J. L., Godeau, G., Couet, D., Putterman, M., Tanzer, J., Babin, P., and Larregue, M. (1991). Recurrent ctb(7)(q31.3) and possible laminin involvement in a neonatal cutis laxa with a marfanoid phenotype. *Human Genetics* **87**: 317-319.

Borsi, L., Carnemolla, B., Nicolo, G., Spina, B., Tanara, G., and Zardi, L. (1992). Expression of different tenascin isoforms in normal, hyperplastic and neoplastic human breast tissues. *International Journal of Cancer* **52** (5): 688-692.

Cho, K. R., and Fearon, E. R. (1995). DCC: linking tumor suppressor genes and altered cell surface interactions in cancer? *European Journal of Cancer* **31A** (7/8): 1055-1060.

Davis, J. Q., and Bennett, V. (1994). Ankyrin binding-activity shared by the Neurofascin/L1/Nr-Cam family of nervous-system cell-adhesion molecules. *J. Biol. Chem.* **269** (44): 27163-27166.

de la Rosa, E. J., Kayyem, J. F., Roman, J. M., Stierhof, D., Dreyer, W. J., and Schwarz, U. (1990). Topologically restricted appearance in the developing chick retinotectal system of Bravo, a neural surface protein - experimental modulation by environmental cues. *J. Cell Biol* **111**: 3087-3096.

Doherty, P., Moolenaar, C. E. C. K., Ashton, S. V., Michalides, R. J. A. M., and Walsh, F. S. (1992). The vase exon down-regulates the neurite growth-promoting activity of N-CAM-140. *Nature* **356**: 791-793.

Doherty, P., Williams, E., and Walsh, F. S. (1995). A soluble chimeric form of the L1 glycoprotein stimulates neurite outgrowth. *Neuron* **14**: 57-66.

Dorries, U., and Schachner, M. (1994). Tenascin messenger-RNA isoforms in the developing mouse-brain. *Journal of Neuroscience Research* **37** (3): 336-347.

Fazeli, M. S., Hobbs, C., Tonge, D., Wells, D. J., and Walsh, F. S. (1995). Ectopic expression of the Neural Cell-Adhesion Molecule in skeletal-muscle of transgenic mice leads to altered neuromuscular development. *J. Neurochem.* **65** (Suppl.): S78.

Fearon, E. R., Cho, R. K., Nigro, M. J., Kern, E. S., Simons, W. J., Ruppert, M. J., Hamilton, R. S., Preisinger, C. A., Thomas, G., Kinzler, W. K., and Vogelstein, B. (1990). Identification of a chromosome 18q gene that is altered in colorectal cancers. *Science* **247**: 49-56.

Fields, C., Adams, M. D., White, O., and Venter, J. C. (1994). How many genes in the human genome. *Nature Genetics* **7**: 345-346.

Goodman, C. S., and Schatz, C. J. (1993). Developmental mechanisms that generate precise patterns of neuronal connectivity. *Cell* **72** (Suppl.): 77-98.

Grumet, M. V., Mauro, M. P., Burgoon, G. M., Edelman, G. M., and Cunningham, B. A. (1991). Structure of a new nervous system

glycoprotein, Nr-CAM, and its relationship to subgroups of neural cell adhesion molecules. *J. Cell Biol.* **113**: 1399-1412.

Johnson, J. P. (1991). Cell adhesion molecules of the immunoglobulin supergene family and their role in malignant transformation and progression to metastatic disease. *Cancer Metastasis Rev.* **10**: 11-22.

Jouet, M., Rosenthal, A., Armstrong, G., MacFarlane, J., Stevenson, R., Paterson, J., Metzenberg, A., Ionasescu, V., Temple, K., and Kenwrick, S. (1994). X-linked spastic paraplegia (SPG1), MASA syndrome and X-linked hydrocephalus result from mutations in the L1 gene. *Nature Genetics* **7**: 402-407.

Jukes, T. H. (1980). Silent nucleotide substitutions and the molecular evolutionary clock. *Science* **210**: 973-978.

Kaczmarek, J., Castellani, P., Nicolo, G., Spina, B., Allemanni, G., and Zardi, L. (1994). Distribution of oncofetal fibronectin isoforms in normal, hyperplastic and neoplastic human breast tissues. *International Journal of Cancer* **59** (1): 11-16.

Kayyem, J. F., Roman, J. M., de la Rosa, E. J., Schwarz, U., and Dreyer, W. J. (1992). Bravo/Nr-CAM is closely related to the cell adhesion molecules L1 and Ng-CAM and has a similar heterodimer structure. *J. Cell Biol.* **118**: 1259-1270.

Korenberg, J.R. and Chen, X-N. (1996). Human cDNA mapping using a high resolution R-banding technique and fluorescence *in situ* hybridization. *Cytogenetic Cell* **69**: 196-200.

Kozak, M. (1989). The scanning model for translation. *J. Cell Biol.* **108**: 229-241.

- Marshall, C.J. (1991). Tumor Suppressor Genes. *Cell* **64**: 313-326.
- Marshall, R. D. (1972). Glycoproteins. *Annu. Rev. Biochem.* **41**: 673-707.
- Mauro, V. P., Krushel, L. A., Cunningham, B. A., and Edelman, G. M. (1992). Homophilic and heterophilic binding activities of Nr-CAM, a nervous-system cell-adhesion molecule. *J. Cell Biol.* **119**: 191-202.
- Morales, G., Hubert, M., Brummendorf, T., Treubert, U., Tarnok, A., Schwarz, U., and Rathjen, F. G. (1993). Induction of axonal growth by heterophilic interactions between the cell-surface recognition protein-f11 and protein-Nr-CAM/Bravo. *Neuron* **11** (6): 1113-1122.
- Otsuka, A. J., Franco, R., Yang, B., Shim, K. H., Tang, L. Z., Zhang, Y. Y., Boontrakulpoontawee, P., Jeyaparakash, A., and Hedgecock, E. (1995). An ankyrin-related gene (unc-44) is necessary for proper axonal guidance in *Caenorhabditis-elegans*. *J. Cell Biol.* **129** (4): 1081-1092.
- Scholes, J. H. (1981). Ribbon optic nerves and axonal growth patterns in the retinal projection to the tectum. In: "Development in the Nervous System" (D. R. Garrot, and J. D. Feldman, Eds), pp181-214, Cambridge University Press, Cambridge.
- Sonderegger, P., and Rathjen, F. G. (1992). Regulation of axonal growth in the vertebrate nervous-system by interactions between glycoproteins belonging to 2 subgroups of the immunoglobulin superfamily. *J. Cell Biol.* **119**: 1387-1394.



Stoeckli, E. T., and Landmesser, L. T. (1995). Axonin-1, Nr-CAM, and Ng-CAM play different roles in the in-vivo guidance of chick commissural neurons. *Neuron* **14** (6): 1165-1179.

Suter, D. M., Pollerberg, G. E., Buchstaller, A., Giger, R. J., Dreyer W. J., and Sonderegger, P. (1995). Binding between the neural cell-adhesion molecules axonin-1 and Nr-CAM/Bravo is involved in neuron-glia interaction. *J. Cell Biol.* **131**: 1067-1081.

Tucker, R. P., Spring, J., Baumgartner, S., Martin, D., Hagios, C., Poss, P. M., and Chiquetehrisman, R. (1994). Novel tenascin variants with a distinctive pattern of expression in the avian embryo. *Development* **120** (3): 637-647.

Vaughn, D. E. and Bjorkman, P. J. (1996). The (Greek) key to structures of neural adhesion molecules. *Neuron* **16**: 261-273.

Vielmetter, J., Kayyem, J. F., Roman, J. M., and Dreyer, W. J. (1994). Neogenin, an cell surface protein expressed during terminal neuronal differentiation, is closely related to the human tumor suppressor molecule Deleted in Colorectal Cancer. *J. Cell Biol.* **127**: 2009-2020.

Williams, A. F., and Barclay, A. N. (1988). The immunoglobulin superfamily - domains for cell surface recognition. *Annu. Rev. Immunol.* **6**: 381-405.

Williams, E. J., Furness, J., Walsh, F. S., and Doherty, P. (1994). Activation of the FGF receptor underlies neurite outgrowth stimulated by L1, N-CAM, and N-Cadherin. *Neuron* **13**: 583-594.

Wilson, A. C., Carlson, S. S., and White, T. J. (1977). Biochemical evolution. *Annu. Rev. Biochem.* **46**: 573-639.

Woodgett, J. R., Gould, K. L., and Hunter, T. (1986). Substrate-specificity of protein kinase C: use of synthetic peptides corresponding to physiological sites as probes for substrate recognition requirements. *Eur. J. Biochem* **161**: 177-184.

Yamakawa, K., Mitchell S., Hubert R., Chen X-N., Colbern S., Huo Y-K., Gadomski C., Kim U-J., and Korenberg J. R. (1995). Isolation and characterization of a candidate gene for progressive myoclonus epilepsy on 21q22.3. *Human Molecular Genetics* **4**:709-716.

Zenklusen, J. C., Bieche, I., Lidereau, R., and Conti, C. J. (1994). (c-a)(n) Microsatellite repeat d7s522 is the most commonly deleted region in human primary breast-cancer. *Proc. Natl. Acad. Sci. USA* **91**: 12155-12158.

**Figure 1. Nucleotide sequence and deduced amino acid sequence of human Bravo.** The longest open reading frame consists of 1,304 amino acids. The hydrophobic signal peptide (-24/-29 to -1) and transmembrane region (1,127 to 1,149) are underlined. A second potential start codon that would result in a longer signal peptide (five additional amino acids) is bracketed (); the two possible start codons are indicated by > (see discussion in text). The Ig-like domains are indicated IgI to IgVI over the conserved tryptophans and cysteines; the fibronectin type III repeats are indicated Fn1 to Fn5 over the conserved tryptophans and tyrosines. Potential phosphorylation sites are indicated by asterisks; potential sites of asparagine-linked glycosylation are indicated by plus (+) signs. The alternative exon AE12 is bracketed [] because its sequence was not included in the full-length cDNA.

[illegible]

**Figure 2.** *Amino acid alignments of chicken and human Bravo proteins.* Sequences are separated by predicted domain structure. For each domain, the amino acid identities are indicated. N-terminal signal peptide and hydrophobic transmembrane region are underlined and in italics. Alternatively spliced exons are enclosed by square brackets, with amino acid length indicated (AE19, AE10, AE12, AE93). The amino acid identity for the entire protein is 82%.

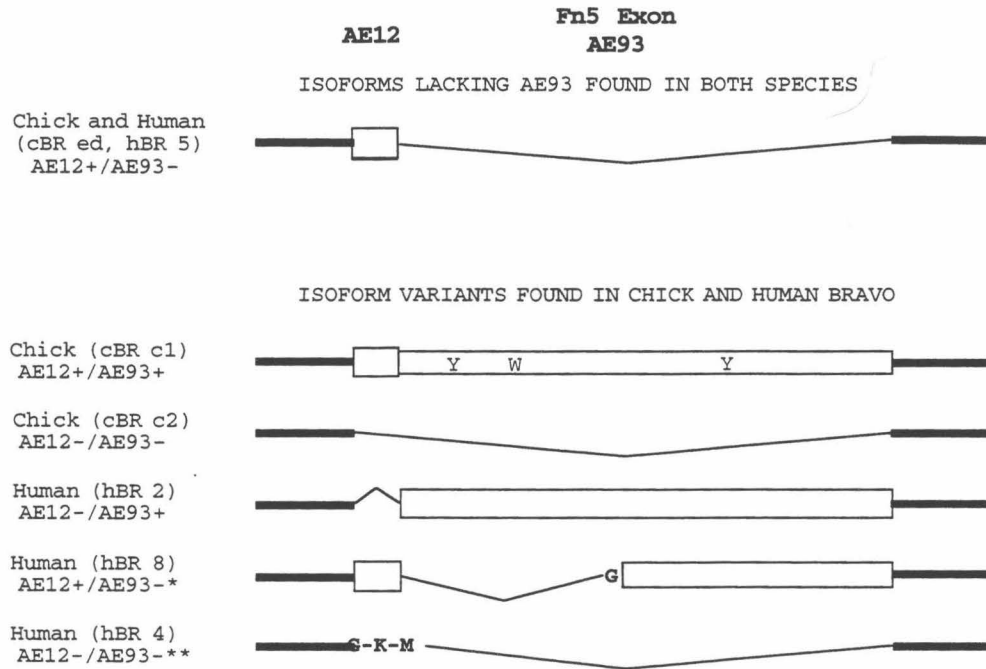
-24 WKKPKTSISKASIVFELCOMTSALDVLDSKLLLELSOPPTITQGS chick BRAVO  
 Signal Peptide/N-terminus  
 -24 WKKPKTSISKASIVFELCOMTSALDVLDSKLLLELSOPPTITQGS human BRAVO  
 24 PKDYIVDPRENIQCEAKGKPPPSFWTRNGTHFDIDKDAQVTMKPNSGTLVWNIHMGVKAAYEGVYQCTARNERGAASNNIVIR chick BRAVO  
 24 PKDYIIDPRENIQCEAKGKPPPSFWTRNGTHFDIDKDLPLVTHKPGTGTLLIINHSKGAEITYEGVYQCTARNERGAASNNIVIR human BRAVO  
 112 PSRSPLMTKEKLE 124 chick BRAVO  
 112 PSRSPLMTKEKLE 124 Igi-II: 100% Identity human BRAVO  
 125 PRHVREGDSLVLAKRPPVGLPPPIIFWMDNAPQLPQSERVSGQLNGDLYFSWQPEDTREDYICVARNHTQTICQKQPSIVKVS chick BRAVO  
 125 PITLQSQSLVLCRPPVGLPPPIIFWMDNAPQLPQSERVSGQLNGDLYFSWQPEDTREDYICVARNHTQTICQKQPSIVKVS human BRAVO  
 212 [MDSINDTIAANLSDTYGA] KPVTERRPVLIT 243 chick BRAVO  
 212 [MDSINDTIAANLSDTYGA] KPVTERRPVLIT 243 Igi-III (AE19): 72% Identity human BRAVO  
 244 PMGSTSNKVRGNVLLLECAAGLPPIVIRWIKEGGELPANRTFFENFKTLKIIDVSEADSGNYKCTARNLTGSTHNVISVTV chick BRAVO  
 244 PMGSTSNKVRGNVLLLECAAGLPPIVIRWIKEGGELPANRTFFENFKTLKIIDVSEADSGNYKCTARNLTGSTHNVISVTV human BRAVO  
 329 KAAPFWITA 338 chick BRAVO  
 329 KAAPFWITA 338 Igi-IV (AAPW): 100% Identity human BRAVO  
 339 PRNLVLSPOEGDTLIRANGNPKPSISWLTNGVPIAIPEDPSRKVDGDTIIFSAVQERSAVYQCNASNEYGLLANAFNV chick BRAVO  
 339 PRNLVLSPOEGDTLIRANGNPKPSISWLTNGVPIAIPEDPSRKVDGDTIIFSAVQERSAVYQCNASNEYGLLANAFNV human BRAVO  
 421 LAEPRIILT 429 chick BRAVO  
 421 LAEPRIILT 429 IgiV-V: 100% Identity human BRAVO  
 430 PANKLVQIADSPALIDCAFYCSKPEIEMFRGVKGSILRNEYVHNGTLEIPVAQKDGSTGYTCVARNKLGKTCNEVLEV chick BRAVO  
 430 PANKLVQIADSPALIDCAFYCSKPEIEMFRGVKGSILRNEYVHNGTLEIPVAQKDGSTGYTCVARNKLGKTCNEVLEV human BRAVO  
 514 KDPTMIKQ 522 chick BRAVO  
 514 KDPTMIKQ 522 IgiV-VI: 67% Identity human BRAVO  
 523 POYKVIQSAQSFECVFKHDPITLPTVWLKNNELPDERFLVGNLNTIMNVTDKDDGTYTCVANTLDSVSASAVLTV chick BRAVO  
 523 POYKVIQSAQSFECVFKHDPITLPTVWLKNNELPDERFLVGNLNTIMNVTDKDDGTYTCVANTLDSVSASAVLTV human BRAVO  
 605 VA(APPTPAIYA)RNP 620 chick BRAVO  
 605 VA(APPTPAIYA)RNP 620 IgiV-Fn1 (AE10): 63% Identity human BRAVO  
 621 PLDELTDQLSERSIELSWPGEENNSPITNFVIEYEDGLHEPGVMHYQTEVPGSQTTVQLKSPVNYSPVIAVNEIGRSQSEPSQ chick BRAVO  
 621 PLDELTDQLSERSIELSWPGEENNSPITNFVIEYEDGLHEPGVMHYQTEVPGSQTTVQLKSPVNYSPVIAVNEIGRSQSEPSQ human BRAVO  
 710 YLTASNPDEN 720 chick BRAVO  
 710 YLTASNPDEN 720 Fn1-2: 64% Identity human BRAVO  
 721 PSNVQIGSEPDNLITWESLKGPSQNGPGLQYKVSWRQKDDDBWTSVVANVSKYIVSGTPTFPVYKIQALNDGYPAPSEVIG chick BRAVO  
 721 PSNVQIGSEPDNLITWESLKGPSQNGPGLQYKVSWRQKDDDBWTSVVANVSKYIVSGTPTFPVYKIQALNDGYPAPSEVIG human BRAVO  
 810 HSGEDLPWA 819 chick BRAVO  
 810 HSGEDLPWA 819 Fn2-3: 100% Identity human BRAVO  
 820 PGNVQVRIINSTLAKVHWDVPLKSVRHLQCYKYVYWKVSLRRSKRVHVEKILTFRGNKTFQMLPGLEPYSSYKLVVNVNGKGEPPASDVKV chick BRAVO  
 820 PGNVQVRIINSTLAKVHWDVPLKSVRHLQCYKYVYWKVSLRRSKRVHVEKILTFRGNKTFQMLPGLEPYSSYKLVVNVNGKGEPPASDVKV human BRAVO  
 916 FKTPEGVSP 925 chick BRAVO  
 916 FKTPEGVSP 925 Fn3-4: 80% Identity human BRAVO  
 926 PSFLKINPTLDSLTLWQSPHNPVLTSTYILKFPINNTNELGPIVEIRIPANESSLILKNLNYSTRYKFFYNAQTSVGSQSQITEZAV chick BRAVO  
 926 PSFLKINPTLDSLTLWQSPHNPVLTSTYILKFPINNTNELGPIVEIRIPANESSLILKNLNYSTRYKFFYNAQTSVGSQSQITEZAV human BRAVO  
 1017 TIMDE[AGILRPVAGAK] 1033 chick BRAVO  
 1017 TIMDE[AGILRPVAGAK] 1033 Fn4-5 (AE12): 76% Identity human BRAVO  
 1034 [VQPLPRIRNVTAAAEYANISWEYEGPDHANFYVEYGVAGSKEDWKEIVNGSRSFVFLKGLTPOTAYKVRVGAEGLSGFRSSEDLFETGP] chick BRAVO  
 1034 [VQPLPRIRNVTAAAEYANISWEYEGPDHANFYVEYGVAGSKEDWKEIVNGSRSFVFLKGLTPOTAYKVRVGAEGLSGFRSSEDLFETGP] human BRAVO  
 1127 AMASROVDIATQGFELMCVAALLLILALVCEI 1161 chick BRAVO  
 1127 AMASROVDIATQGFELMCVAALLLILALVCEI 1161 Fn5-Intra/Transmembrane: 100% Identity human BRAVO  
 1162 RRNGGKYVPKEKEDAHADPEIQPKEDDGTGSEYDAEDHKLKKGSRTPSDRTVYKEDSDSLVDYGEVNGQPNEDGSLQYSGKKEKPAEGNESSEAPSPVNAHNSFV chick BRAVO  
 1162 RRNGGKYVPKEKEDAHADPEIQPKEDDGTGSEYDAEDHKLKKGSRTPSDRTVYKEDSDSLVDYGEVNGQPNEDGSLQYSGKKEKPAEGNESSEAPSPVNAHNSFV human BRAVO



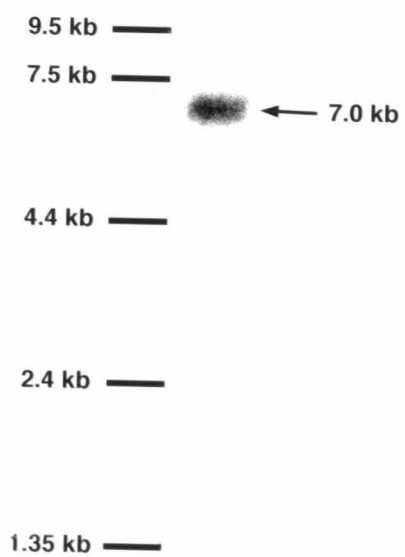
**Figure 3. *Alternative splicing around the 5th fibronectin type III repeat of the chicken and human Bravo/Nr-CAM molecules.*** Specific cDNA clones are indicated and coded as follows: "cBR-nr" indicates a chicken Bravo cDNA clone isolated in Grumet *et al.*, 1991; "cBR-c1/c2" indicates chicken Bravo cDNA clones #c1 and #c2 isolated in Kayyem *et al.*, 1992; "hBR-5/8/4" indicates human Bravo cDNA clones #5, #8, and #4 isolated here. \*Note in hBR-8, the isoform is missing the N-terminal half of AE93 between K-1021 and K-1058 with an additional glycine inserted. \*\*Note in hBR-4, the isoform is missing both AE19 and AE93 and in their place is the tripeptide glycine-lysine-methionine.



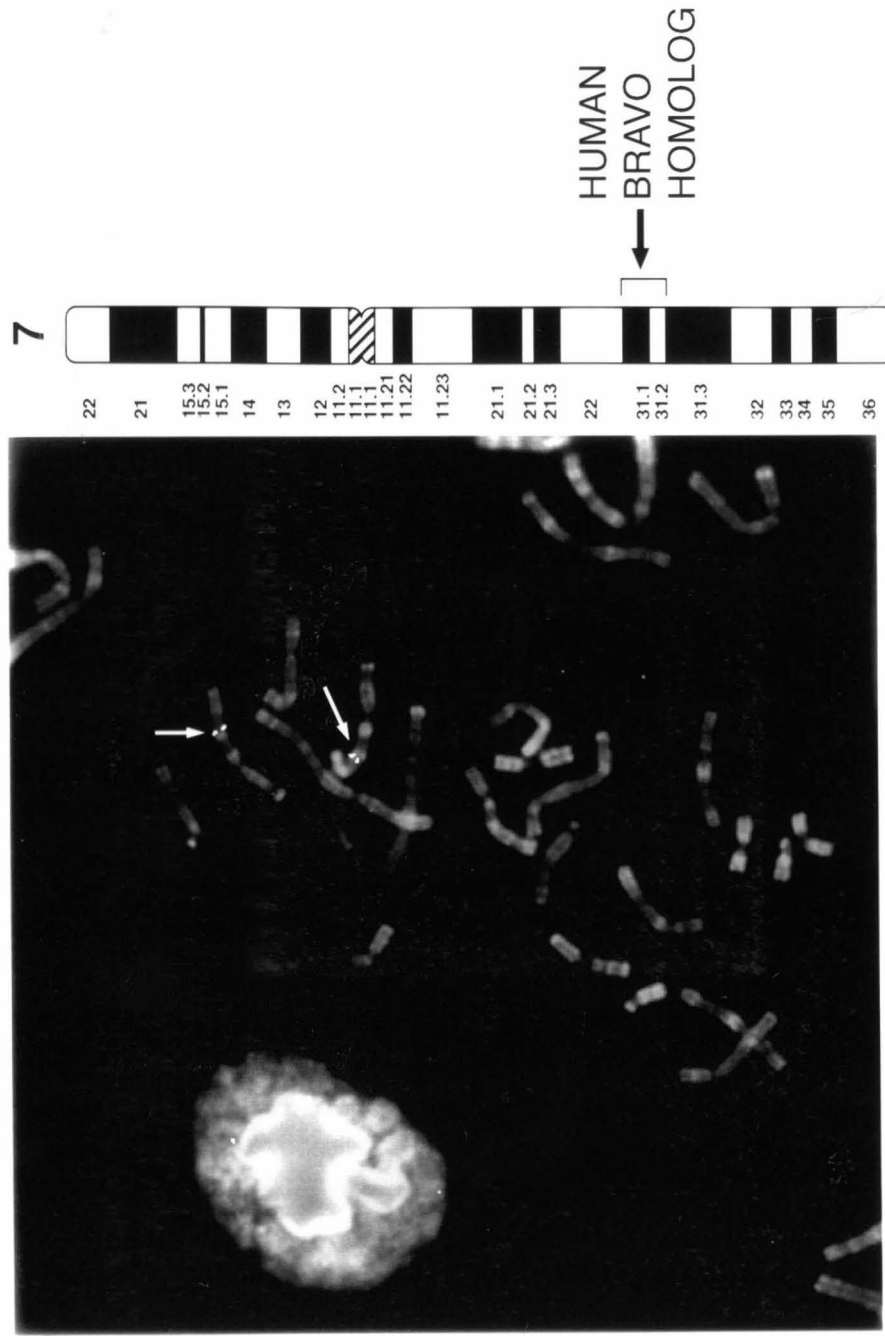
**ALTERNATIVE SPLICING AROUND THE Fn5 DOMAIN**



**Figure 4.** *Northern blot.* Human RNA isolated from adult brain tissue probed with P<sup>32</sup>-labelled human Bravo PCR fragment. The hybridized RNA species shown is 7.0 kb.



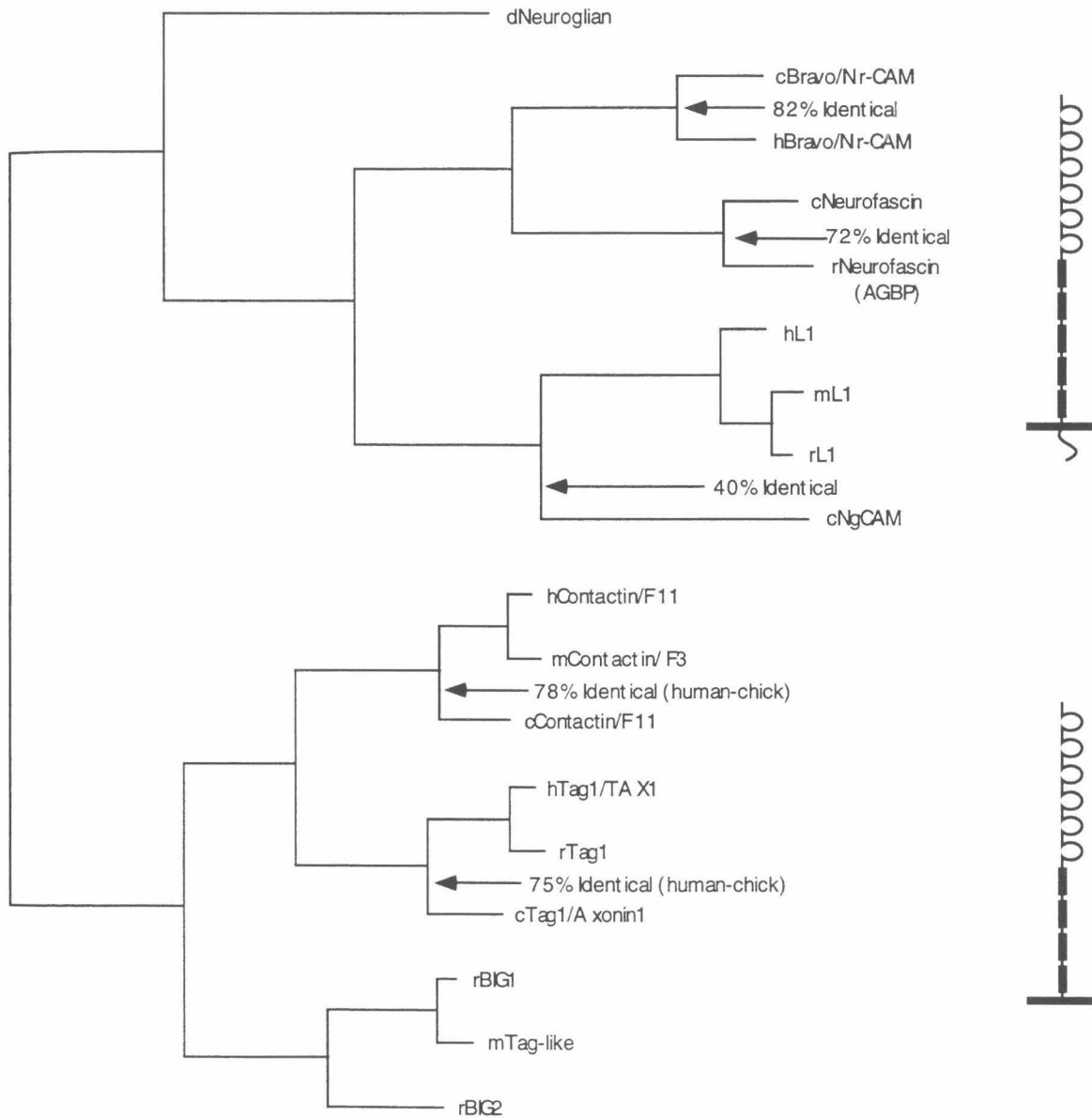
**Figure 5.** *Fluorescent in situ hybridization (FISH).* A 1.6 kb human Bravo PCR fragment hybridizes to human chromosome 7, band q31.1-q31.2.



**Fluorescent In Situ Hybridization  
Maps hBravo to Chromosome 7q31.1-2**

**Figure 6.** *Shortest phylogenetic tree of two closely related subfamilies of the Ig Superfamily.* The phylogenetic analysis was accomplished using Paup 3.1.1 system software of aligned sequences. Horizontal lengths of tree branches are proportional to the amount of sequence divergence from the common ancestral node. Mammalian-Avian identity scores are indicated for species homologs. Not shown here are members of the N-CAM and neogenin neural Ig subfamilies, whose mammalian-avian sequence identities are 80% and 87% respectively.

Shortest Phylogenetic Tree of Two Families of the IG Superfamily





Chapter IV

Are CAM's Really Receptors? A Functional Study Aimed at  
Elucidating Putative Signal Transduction Events in Bravo/Nr-CAM

Robert P. Lane, Jost Vielmetter, Frank Miskevich and William J. Dreyer

## Introduction

### *The CAM Misnomer? Evidence of Signal Transduction*

The neural Ig Superfamily, of which Bravo/Nr-CAM is a member, is a diverse group of proteins that contain multiple Ig-like and fibronectin domains. These proteins, along with several other extracellular proteins, are collectively referred to as Cell Adhesion Molecules (CAMs). While it is true that in many cases so far examined, CAMs indeed have significant functions relating to cell-cell interactions and adhesion (for example, see Edelman, 1986), the name implies that these proteins are a sort of indiscriminate "glue" that merely provides a molecular basis for certain cell types to stick to each other. In this regard, until recently, the Bravo/Nr-CAM protein and its relatives were rarely thought of as true receptors (i.e., part of a developmental pathway), with the capacity to transduce extracellular signals into an appropriate intracellular response. Lately, however, an abundance of evidence has emerged that suggests CAMs indeed are linked to classical signal transduction pathways, including calcium fluxes (Doherty *et al.*, 1991; Williams *et al.*, 1995; Goldman *et al.*, 1996) and kinase cascades (Atashi *et al.*, 1992; Klinz *et al.*, 1995; Wong *et al.*, 1996). Even PI-linked proteins, like Contactin/F11 and Axonin/Tag-1, may trigger 2nd messenger systems via receptor coupling, which as discussed below, may represent a general mechanism for signaling among CAMs.

### *Intracellular Domain Sequences in the Bravo/Nr-CAM subfamily*

The intracellular domain of Bravo/Nr-CAM is a highly conserved 114 amino acid sequence. The domain is 100% conserved since avian-mammalian divergence (i.e., between the chicken and human orthologs), a period of evolution that, if this sequence were permitted to accumulate mutations at a random rate, would have resulted in every amino acid changing at least once (Wilson *et al.*, 1977;

Jukes, 1980). Furthermore, comparing Bravo/Nr-CAM's closest IgSF relatives (L1, Neurofascin, Ng-CAM and Neuroglian), the intracellular domains are by far the most conserved of all 12 domains among these proteins (see Chapter II). Although part of this stringent conservation among these intracellular domains is accounted for by the common interaction with the spectrin-binding cytoskeletal ankryin molecule (Davis and Bennett, 1994), the expectation is that there are other important intracellular ligands not yet identified. Among these putative interactions, might be numerous kinases: in Bravo/Nr-CAM, eight potential serine/threonine phosphorylation motifs have been conserved including putative recognition sequences for casein kinase II, cAMP-dependent kinase, and protein kinase C (see Chapter III). While the domain itself does not share homology with other receptor-type cytoplasmic signal transduction domains, and therefore may not directly elicit an established 2nd messenger system, it is possible (if not likely) that these phosphorylation motif sequences may regulate interactions with other transducers or co-receptors which themselves may consequently generate a signal. There are numerous examples in which receptor dimerization results in signal transduction events, including the interaction between L1 (of the IgSF) and the FGF-receptor which leads to a kinase cascade and neuronal trophic effects (Williams *et al.*, 1994; Doherty *et al.*, 1995). In principle, dimerization or *cis* interactions of this type may be widespread among CAM's and generally provide a mechanism for even PI-linked proteins to relay various extracellular signals into appropriate cellular responses.

#### *Fifth Fibronectin Repeats and Co-Receptor Function?*

The Bravo/Nr-CAM cDNA has been cloned and sequenced in two vertebrate species (chicken, Kayyem *et al.*, 1992; and human, Lane *et al.*, 1996; see Chapter III), and in both, a diverse set of alternatively spliced isoforms has been identified. Among these isoforms are at least six distinct variants that exhibit alternative splicing around the fifth

and most membrane-near fibronectin type III repeat (Fn5), including an isoform in both species that completely eliminates the domain (93 amino acids). It is also interesting to note that the amino acid sequence that follows this alternative exon is 100% conserved between the chicken and human proteins (more than 150 amino acids at the C-terminus), including the entire transmembrane and intracellular domains. This enormous degree of sequence immutability that has survived over 200 million years of evolution (since avian-mammalian divergence) suggests an important membrane-spanning function that might be linked to the adjacent Fn5 alternative splicing events.

Fibronectin type III repeats are abundant, found on a wide range of proteins on the cell membrane and extracellular matrix (see Bork and Doolittle, 1992, for example). There are a number of examples among these, in which an entire repeat is alternatively spliced, and in many cases tissue-specific or temporal regulation of these splicing events has elucidated important developmental significance to isoform modulation. The fibronectin and tenascin proteins, for example, each have fibronectin type III alternative exons some of which may play a role in cell cycle/post-mitotic differentiation events: the inclusion of the domain is correlated with tumor growth and development, while the domain-lacking form is exclusively expressed in post-developmental cells of adult tissue (Burton-Wurster *et al.*, 1989; Weller *et al.*, 1991; Dorries and Schachner, 1994; Kaczmarek *et al.*, 1994; Tucker *et al.*, 1994). Alternative splicing in various proteins of the Ig Superfamily has illustrated spatial, temporal and subcellular mechanisms of splicing regulation (for example: Persohn and Schachner, 1987; Pollerberg *et al.*, 1987; Huhtala *et al.*, 1990; Suzuki and Naitoh, 1990), and in this regard, isoform variation is a potent means to diversify functions of these already large and complex proteins.

In this Chapter, we have examined the isoform expression pattern of the alternatively spliced fifth fibronectin type III repeat in Bravo/Nr-CAM. The 93 amino acid alternative domain was produced

in a heterologous prokaryotic expression system, purified, and used to raise an exon-specific monoclonal antibody. This antibody was subsequently utilized in a developmental histological study (double labeling and confocal microscopy) in order to explore spatial and temporal regulation of the splicing events. These studies, as well as supplemental PCR analyses, suggest that both domain-containing and domain-lacking Bravo/Nr-CAM isoforms are expressed at similar levels on all cells and tissues throughout development. Because these alternative splicing events have been conserved since avian-mammalian divergence, it is presumed to have functional consequence, and we propose a model that might account for the observed co-expression while concurrently, providing an important and distinguishable role for both variants.

Also in this Chapter, we have initiated experiments aimed at deciphering the possible functions of both this alternatively spliced Fn5 exon, as well as the highly conserved intracellular sequences C-terminal to this domain. Because the Fn5 and intracellular domain appear to have co-evolved among proteins in the IgSF subfamily that contain them (see Chapter II), it is possible that the functions of both domains are coupled. The most likely way that an extracellular and intracellular domain may share a common function is by virtue of supporting a *cis* co-receptor dimerization (with interacting residues that span both sides of the cell membrane), a possibility further supported by the fact that the transmembrane sequence connecting these two domains is also 100% conserved. Co-receptor interactions may, for example, modulate signal transduction events, as evident with the L1-FGF-receptor example cited previously. In order to explore this possibility, we have expressed both domains in a heterologous expression system, and used the domains in two kinds of approaches to identify ligands: affinity chromatography and expression library screening.

Affinity chromatography is a method used routinely in many laboratories to isolate specific proteins based on their interaction with a ligand covalently bound to a solid phase support (reviewed in Phizicky and Fields, 1995). This method has been used successfully, for example, to identify a specific tenascin-contactin/F11 interaction (Zisch *et al.*, 1992). In the experiments described in the following sections, both cell surface and total chicken brain lysates were used as a putative ligand source; in the case of the intracellular domain, it would also be possible to use bovine lysate (one brain contains more than 1000 chicken brains worth of protein) because this particular domain is 100% conserved between avian-mammalian species. Identified ligand bands would be typically identified by protein sequencing, which is only possible if it is not N-terminally blocked and if the yield approaches  $\mu\text{g}/\text{nm}$  levels. Expression library screening, on the other hand, has the advantage of directly cloning cDNAs that express the putative ligand. This method has been used successfully previously (Young and Davis, 1983; Sikela and Hahn, 1987; MacGregor *et al.*, 1990; Blackwood and Eisenman, 1991; Ayer *et al.*, 1993), and is widely used for cloning antigens using antibody probes (for example, Kayyem *et al.*, 1992; Vielmetter *et al.*, 1994). There are two possible pitfalls to this approach however: 1) the binding constant must be at least in the  $\mu\text{M}$  range, and preferably close to that of antibody-antigen nM  $K_D$ s, and 2) the expressed protein must present native epitopes in a foreign bacterial environment.

## Materials and Methods

### *Protein Expression Systems*

In order to raise domain-specific monoclonal antibodies, and to generate copious amounts of protein for functional studies, sequences corresponding to various extracellular and intracellular portions of Bravo/Nr-CAM were expressed in two different expression systems: *E.*

*coli* and *Pichia pastoris*. The former system utilizes a powerful T7 RNA polymerase promoter that drives expression of heterologous protein to levels exceeding 100 mg/l culture. This overexpression causes the bacteria to produce inclusion bodies, the preparation of which provides a one-step method of purification that yields greater than 95% purity. The latter system maintains some of the simplicity of prokaryotic bacterial expression systems, while providing some of the presumably important eukaryotic post-translational events. Especially when expressing extracellular protein sequences, the ability to utilize secretion signal peptides not only significantly reduces the complexity of subsequent purification steps, but also mimics appropriate folding conditions while the expressed protein is escorted and inserted through the cell membrane. Like bacterial systems, expression in yeast is robust and at high quantities (at least 10-50 mg/l culture), and products containing a 6-histidine tag can be purified from supernatants to greater than 99% in a single IMAC chromatographic step.

Although for some purposes the use of mammalian expression systems may provide the best approximation to *in vivo* conditions, and therefore represent the most reliable source of properly folded and modified product, both the bacterial and yeast expression systems provide a rapid, inexpensive means to huge quantities of desired product. Furthermore, in several wide-scoping expression experiments, protein produced by these means has been functional and therefore appropriate for further study (e.g., see Cregg et al., 1993). For our purposes, while no explicit functional assay exists, the expressed proteins are at least presenting native epitopes as assayed by various monoclonal antibody reactivities. One important assay for structural suitability is whether or not the proteins are able to perform appropriate functions, including binding appropriate ligands, and this application is stringently put to test by an elaborate set of biochemical experiments described subsequently in this Chapter.



*Heterologous Expression of the Intracellular Domain*

A previously isolated chicken Bravo/Nr-CAM clone (Kayyem *et al.*, 1992) was used as a template for PCR in order to amplify the 342 bp nucleotide sequence corresponding to the intracellular domain. All cloning primers used in the experiments described in this Chapter were synthesized by the Caltech oligonucleotide facility. Primers were designed to contain extra flanking sequences, including 5' and 3' restriction sites in frame with the pPIC9K *Pichia* expression vector, as well as a 3' six-His tail and stop codon following the intracellular coding sequence. The specific primer sequences used were:

5' GAATTCAGGAGGAATAAAGGTGGCA 3' (sense)

5' GCGGCCGCTTAATGGTGATGGTGGTGATGCACAAATGAACTCATGGCA 3'

Note that the sense primer begins with an EcoRI overhang (underlined); the antisense primer has a NotI (underlined), Stop Codon (italics) and six histidine (small font) overhang. Also note that the antisense primer introduces a silent mutation (bold, large font) that eliminates a naturally occurring EcoRI site in order to facilitate cloning.

PCR was carried out under standard reaction conditions, using Vent (exo-) enzyme (New England Biolabs), 60°C annealing temperature and 30 cycles. The resulting product was visualized on a 1% agarose TAE gel, and the product cloned using the TA cloning kit (Invitrogen). Positive transformants (blue-white color selection) were further selected by identical PCR conditions, plasmid prepared from overnight cultures, and sequences were confirmed using automated dye terminator sequencing in the Caltech sequencing facility. A single clone (with no reading frame mutations or PCR artifacts) was selected for expression, and using the engineered EcoRI and NotI restriction sites, the intracellular coding sequence was isolated in preparative 1% agarose-TAE gels, and ligated (T4 ligase, Boehringer Mannheim) to dephosphorylated (calf intestinal alkaline phosphatase, Promega)

*Pichia* pPIC9K expression vector at 12°C for 10 h. The ligation reaction was heat-terminated (65°C for 10 min), precipitated with glycogen, prior to retransformation by electroporation in XL1-blue *E. coli* bacteria (Stratagene). The pPIC9K plasmid was prepared using Wizard miniprep kits (Promega), eluting with TE buffer pH 7.5. Approximately 5 µg plasmid DNA was linearized using 10 units Bgl II restriction enzyme for 2h at 37°C, phenol-chloroform extracted, and precipitated with 40 mg glycogen. Linear pPIC9K plasmid was washed thoroughly to remove salts and re-suspended in 5 µL (1.0 mg/ml) dH<sub>2</sub>O to be used for transformation into the *Pichia* yeast expression cells.

A single colony of *Pichia pastoris* (strain GS115, *his4* genotype) was grown in 25 ml MD-his broth (1% glycerol, 400 µg/l biotin, 13.4 g/l yeast nitrogen base w/o amino acids) at 30°C, shaking to an O.D.<sub>600</sub>=1.0. Cells were washed three times with 25 ml ice cold dH<sub>2</sub>O, once with 25 ml ice cold 1.0 M sorbitol, and resuspended in 250 µl 1.0 M sorbitol. 40 µl of these electrocompetent *Pichia* cells were placed in a 0.1 cm electroporation cuvette along with 1 µl (1 µg) of the linearized pPIC9K plasmid containing insert, and transformations were carried out at 400 ohms, 1.25 kV, 25 µF (time constant generally between 7.0 - 9.0). Immediately after the electroporation, 900 µl of ice cold 1.0 M sorbitol was added to the suspension, the cells incubated for 15 min with gentle shaking at room temperature, and 100 µl were spread on an MD plate (MD broth plus 15 g/l agar). Plates were incubated for 3 days at 30°C, and independent yeast colonies were picked into 150 µl dH<sub>2</sub>O, and 1-2 µl of this dilution was spotted onto an identical quadrant of both an MD and MM (same as MD plate except glycerol is replaced by 0.5% methanol) plate. Transformed colonies with the insert appropriately targeted within the methanol-inducible AOX-1 gene of the pPIC9K vector are unable to grow robustly in media with methanol as the only carbon source, and therefore, positive colonies were selected by virtue of their slow growth on MM plates after a 24-48 h incubation at 30°C. Twelve slow-growers (see Fig. 1) were selected to further test for high

levels of expression: colonies were picked and grown in 10 ml BMGY broth (20 g/l peptone, 10 g/l yeast extract, 1% glycerol, 400 µg/ml biotin, 13.4 g/l yeast nitrogen base, buffered with 100 mM potassium phosphate, pH 6.0), shaking vigorously at 30°C for 48 h. The yeast cultures were spun for 10 min at 5000g, resuspended in 2 ml BMMY broth (same as BMGY except the glycerol is replaced with 0.5 % methanol), and continued shaking growth (with cheese cloth cover in order to permit maximum aeration) at 30°C for another 48 h. After the first 24 h in BMMY, the methanol lost by evaporation was restored by adding an additional 10 µl methanol to the culture. Yeast cells were spun and supernatants collected for analysis of secreted expression product by coomassie blue-stained SDS PAGE. High-expressing clones were subsequently re-grown in 10 ml BMGY to O.D.<sub>600</sub>=0.5 and frozen at -70°C in 15 % glycerol.

#### *IMAC and Purification of the Expressed Intracellular Domain*

Full expression cultures using a single, high-expressing clone identified by the above procedures, were typically grown in 1.0 litre batches. 10 ml starter cultures were seeded by scraping from frozen colonies into MD broth and grown at 30°C shaking for 2 days. Growth and expression conditions were identical to the test expression cultures described above, taking care to maintain approximately the same surface area and aeration conditions in these scale-up procedures. The resulting 200 ml BMMY supernatant (from 1.0 litre BMGY culture) was spun twice (4000g for 15 min, 10,000g for 30 min) at 4°C, and all remaining yeast cells removed by 0.2 µm filtration. Protease inhibitors (1 µl/ml aprotinin, 1 mM PMSF) were added to the filtrate, and the expressed protein was concentrated using Amicon 3 kD cutoff membrane approximately 10-fold (20-30 ml from 200 ml BMMY supernatant). The concentrated supernatant was subsequently washed twice with 200 ml each of 0.1 M ammonium bicarbonate, concentrating back to 20-30 ml between washes; after the second wash, the

supernatant was once again concentrated back to 20-30 ml. The supernatant was prepared for purification by adding an equal volume of 2X running buffer (2.0 M sodium chloride, 100 mM sodium phosphate, pH 8.0).

The expressed protein was purified using IMAC chromatography and the engineered six-His tail, that is encoded at the C-terminus of the protein. The IMAC column was prepared by taking 1 ml IDA-agarose (Pharmacia), washing with 10 volumes of dH<sub>2</sub>O, and then charged with 3 volumes 10% nickel chloride (in dH<sub>2</sub>O). The charged column was washed again with 10 volumes dH<sub>2</sub>O, then equilibrated with 10 volumes running buffer (1.0 M sodium chloride, 50 mM sodium phosphate, pH 8.0). Concentrated protein supernatant in 0.05 M ammonium bicarbonate/1X running buffer (see above) was loaded onto the charged and equilibrated column at a flow rate of 1 ml/min. The column was washed with 20 volumes running buffer, and weakly bound contaminate proteins removed by washing with 10-20 column volumes of running buffer supplemented with 10 mM imidazole. Expressed protein containing the His-tail was eluted from the column using running buffer plus 200 mM imidazole, collecting 10 fractions of one-half column volume each. Fractions were analyzed by coomassie-blue stained SDS-PAGE, and those that contain expressed protein were pooled and dialyzed versus PBS. Figure 1 summarizes the entire *Pichia pastoris* expression system and protein purification methodology.

#### *Heterologous Expression of the Alternative Fn5 Exon*

Using the chicken Bravo/Nr-CAM sequence (Genbank Accession Code L08960), a Polymerase Chain Reaction (PCR) experiment was designed to clone the alternatively spliced 93 amino acid fifth fibronectin exon. Sense (5' CATATGGTGCAACCACTTTAT CCA 3') and antisense (5' GGATCCTTA TGGACCTGTCTCAAACAG 3') primers were designed to include addition cloning sequences on both the 5' and 3' end of the amplified domain: the 5' end included an *ndel*

restriction site (underlined) which contains an ATG start codon, and the 3' end included a stop codon (*italics*) followed by BamHI restriction site (underlined). The PCR was executed on a C-terminal chicken Bravo/Nr-CAM plasmid clone previously characterized in our laboratory (Kayyem et al, 1992) under the following conditions: 30 cycles of 92°C denaturation for 30s, 55°C annealing for 1 min, 72°C extension for 3 min. The resulting 290 bp band was excised from a 1% agarose gel containing 1 µg/ml ethidium bromide, and cloned into the pCR vector according to the TA cloning kit protocol (Invitrogen). A single positive transformed colony was selected for sequencing in order to confirm that the complete expected PCR product had been cloned, and that no reading frame or other mutations had been introduced by the Vent enzyme (New England Biolabs). The insert was re-cloned into the pET3A vector (Studier et al., 1990) using the *ndeI* and *bamHI* sites that had been introduced by the PCR primers immediately adjacent to coding sequence.

Expression of the alternatively spliced 93 amino acid fifth fibronectin domain was accomplished in PLysS strain of *E. coli* bacteria according to previously established protocols (Studier et al., 1990). Briefly, a pET3A transformant was selected (Ampicillin) and grown at 37°C until the culture density is OD<sub>600</sub>=0.5. At this time, expression was driven by the addition of IPTG to 0.4 mM, and the culture was grown at 37°C for another 3 hours. Typically, IPTG drives the M13 promoter of pET3A to express the heterologous protein at levels greater than 100 mg/l, which is sufficient to cause the PLysS cells to produce inclusion bodies. The inclusion bodies containing the overexpressed heterologous protein are prepared as follows (according to Nagai and Thogersen, 1987).

#### *Inclusion Body Preparation of the Expressed Fn5 Domain*

The bacteria were pelleted at 5000g for 5 min at 4°C and resuspended in 2 ml lysis buffer (50 mM Tris pH 8.0, 25% sucrose, 1

mM EDTA, 0.5 mM PMSF, 2.5 mg/ml lysozyme) per 250 ml culture. Cells were lysed for 30 min on ice, at which time DNaseI is added to 20 µg/ml, along with 25 µl 1.0 M MgCl<sub>2</sub> and 2.5 µl 1.0 M MnCl<sub>2</sub>, and the lysate was incubated at room temperature for 30 min. Two volumes of detergent buffer (0.2 M NaCl, 20 mM Tris pH 7.5, 2 mM EDTA, 1% NP-40, 1% deoxycholic acid, 0.5 mM PMSF) were added, mixed vigorously, and the inclusion bodies were pelleted at 5000g for 10 min at 4°C. The inclusion body pellet was washed five times by resuspension in wash buffer (1 mM EDTA pH 8.0, 0.5% Triton X-100, 0.5 mM PMSF) and re-centrifugation.

The inclusion bodies were solubilized by 6M guanidine hydrochloride in 50 mM Tris pH 8. Subsequent dialysis against PBS for at least six hours causes reprecipitation of insoluble material, however the expressed 9.5 kD fifth fibronectin exon remained in solution and was isolated from precipitates by centrifugation. The purity of the expressed domain in the resulting supernatant was approximately 95% as confirmed by capillary electrophoresis. Figure 2 summarizes the bacterial expression strategy.

#### *N-terminal Protein Sequencing of Expressed Domains*

At least 10 µg of the expressed and purified Fn5 and intracellular domains were run on a 15% SDS PAGE gel, and electroblotted to methanol pretreated Pro-blot (Bio-Rad) PDVF for 12 h at 150 mA in Towbin's transfer buffer (39 mM glycine, 48 mM tris base, 0.074% SDS, 20% methanol). The protein-blotted PDVF was rinsed in dH<sub>2</sub>O, then methanol, and submerged for 1 min in Coomassie stain (0.1 % Coomassie-blue in 40% methanol/1% acetic acid) and de-stained twice for several hours each in 50% methanol. Blotted, visualized bands corresponding to expressed domains were excised and thoroughly rinsed in dH<sub>2</sub>O before sequencing. Proteins were sequenced in the Caltech Protein/Peptide Microanalytical Laboratory using an ABI 373A

protein sequencer; ten N-terminal residues were definitively determined for both expressed domains.

#### *Preparation of a Domain-Specific Monoclonal Antibody*

The bacterially-expressed 93 amino acid alternative fifth fibronectin exon was prepared as an immunogen according to protocols established in Hoare and Koshland (1967). 10 µg of KLH-coupled antigen emulsified with Freund's Complete Adjuvant (Sigma) was injected subcutaneously on the back and tail of anesthetized Robertsonian 8.12 mice (Jackson Labs). Three subsequent booster immunizations were given with 10 µg uncoupled antigen emulsified in Freund's Incomplete Adjuvant (Sigma) in the same locations at 3 month intervals. A final boost was injected in the spleen three days prior to the fusion, which was performed with the cell line and method described by Taggart and Samloff (1983). The supernatants from cultures were screened for Bravo/Nr-CAM positive reactivity on tissue sections according to methods described below. Positives were subcloned, and ascites fluid was prepared for one of these (6D2).

#### *Histology to Investigate Regulation of Isoform Expression*

Chicken embryos at the following stages were studied: E3 (embryonic day 3), E4, E5, E6, E6.5, E7, E7.5, E8, E10, E13, E18. Embryos at these stages were isolated and fixed for 5-8 hours in 4% paraformaldehyde in 0.1 M sodium phosphate buffer (pH 7.0), and soaked for two days in 25% sucrose. Spinal cord, cerebellar and retinotectal sections for each of these developmental stages were prepared using a freezing microtome. Two monoclonal antibodies were used in the histology: 6D2 anti-Fn5 (against the alternatively spliced fifth fibronectin domain of Bravo/Nr-CAM) and 2B3 (against an immunoglobulin domain of Bravo/Nr-CAM, an epitope present in all isoforms so far identified). Preliminary studies included side-by-side comparison of the two monoclonal antibody staining patterns on



serial sections of the various tissues and stages. For these experiments, sections were pre-blocked in PBS containing 10% fetal calf serum (FCS) for 10 min, incubated in a 1:1000 dilution of monoclonal antibody ascites fluid in PBS/10% FCS for 1 hour, washed in PBS three times for 5 min, incubated in a 1:200 dilution of goat anti-mouse FITC or Rhodamine secondary antibody in PBS/10% FCS for 1 hour, washed in PBS three times for 5 min, and mounted in glycerol and examined using fluorescent microscopy.

For subsequent double-labeling experiments, attempts to label either monoclonal antibody with DIG or biotin at various multiplicity rendered the antibodies non-functional on tissue sections, and therefore distinct secondary antibody recognition of the two monoclonals was not readily possible. In order to explore whether specific cells exclusively express the Bravo/Nr-CAM isoform that is missing the fifth fibronectin domain, a method of sequential primary antibody incubations was used. In the first incubation, the exon-specific 6D2 monoclonal was permitted to react with all Bravo/Nr-CAM antigen that contains the alternatively spliced Fn5 domain. This initial primary incubation was followed by a specific anti-mouse secondary antibody incubation (both Rhodamine and FITC was used in various experiments). This specific secondary antibody was washed completely from the section, and the 6D2-Secondary complex was fixed with a 1 min soak in 4% PFA/PBS. The PFA was thoroughly rinsed from the section with PBS and the second 2B3 monoclonal antibody (against all Bravo/Nr-CAM isoforms) was incubated atop this fixed 6D2-Secondary complex. A differently labeled anti-mouse secondary antibody incubation followed. This second secondary incubation will of course label the Ig-bound 2B3, plus any unoccupied Fn5-bound 6D2 epitopes, which with both reactivities, labels the full range of Bravo/Nr-CAM antigen distribution. Thus, this sequential method permits examination of specific cells that are positive for Bravo/Nr-CAM staining, but do not contain the alternatively spliced Fn5 epitopes

(i.e., in order to establish which subset of Bravo/Nr-CAM positive cells express the spliced fifth fibronectin domain). All sections were examined at the full range of tissue depth (12  $\mu\text{m}$ ) using confocal microscopy. Control sections in which tissue was pre-incubated with 4% PFA prior to a single 2B3 staining indicated that this fixation step does not destroy Bravo 2B3 epitopes.

#### *PCR Analysis of Isoform Expression.*

In order to investigate the presence of the Bravo/Nr-CAM isoform that is lacking the alternatively spliced fifth fibronectin domain, a series of PCR reactions was performed across a wide range of developmental stages in retina, tectum and cerebellum. RNA and subsequent cDNA was prepared according to Maniatis et al. (1982). Sense (5' ATGAATTAGGTCCCTTGG 3') and antisense (5' TAGCAATGTCTACCTGCCG 3') primers outside the alternative exon were used under the following PCR conditions: Taq polymerase for 25 cycles of 92°C denaturation for 30 s, 55°C annealing for 1 min, and 72°C extension for 3 min. Resulting products were analyzed on 1% agarose gels containing 1  $\mu\text{g}/\text{ml}$  ethidium bromide.

#### *Affinity Chromatography to Identify Putative Ligands*

The heterologously expressed Bravo/Nr-CAM Fn5 and intracellular domains were covalently coupled to agarose solid support in order to identify putative ligands via affinity chromatography. In either case, 1.0 mg expressed domain in PBS was incubated with prewashed 1.0 ml affi-10 agarose gel (Biorad) for 1 h at room temperature under mild shaking conditions according to the manufacturer's protocol. The remainder of reactive sites was quenched by adding 10 mg glycine and continuing the incubation for another hour before dialyzing versus PBS. The domain-coupled agarose was loaded into a column and suspended in 30% glycerol/PBS prior to storage at -70°C.

An embryonic day-8 chicken brain lysate was prepared as a putative ligand source as follows: brains were teased apart from remaining epithelial and body tissue and placed in buffered Hanks solution on ice (10 brains per 10 ml), tissue was spun at low speed (500g) and washed three times with Label buffer (140 mM NaCl, 5 mM KCl, 5 mM glucose, 7 mM NaHCO<sub>3</sub>, 1.5 mM MgSO<sub>4</sub>, 1.5 mM CaCl<sub>2</sub> and supplemented with protease inhibitors: 2 mg/ml iodoacetiminide, 0.2 mg/ml PMSF, 50 µg/ml soybean trypsin inhibitor) by centrifugation. After the final wash, the tissue was gently dispersed (but not lysed or homogenized), and the supernatant was replaced with Label buffer (1 ml per brain) supplemented with 100 µl/10 ml of freshly prepared Biotin-X-NHS (Calbiochem) stock solution (100 mg/ml in DMSO). The mixture was shaken gently for 15 min at room temperature in order to label extracellular residues with biotin via the linked succinimide ester. After labeling, the reaction was terminated and quenched by washing the tissue three times in DMEM (Gibco). A total protein lysate was then produced by replacing the DMEM with a high-detergent lysis buffer (10 mM Hepes pH 7.5, 140 mM NaCl, 4 mM EDTA, 2.5% NP40, 2.5% Zwittergent, 0.02% azide, 0.5 µl/ml aprotinin, 2 mg/ml iodoacetiminide, 0.2 mg/ml PMSF, 50 µg/ml soybean trypsin inhibitor) at approximately 2.5 ml per brain, and shaken vigorously for 15 min at room temperature. Cell debris was centrifuged away at both low (15 min at 3000g, 4°C) and high (30 min at 50,000g, 4°C) speeds. The total lysate supernatants were prepared further by two separate procedures. In one set of experiments, the lysis buffer was replaced with detergent-free physiological buffer (0.06% KH<sub>2</sub>PO<sub>4</sub>, 0.04% KCl, 0.8% NaCl, 0.005% Na<sub>2</sub>HPO<sub>4</sub>, 0.02% CaCl<sub>2</sub>, 0.02% MgSO<sub>4</sub>) by three Amicon 3 kD concentration runs in order to remove excess detergent and low molecular weight materials. This protein lysate was separated into 50 brain aliquots (approximately 100 µl each), shell-frozen in alcohol-dry ice, and stored at -70°C. In a second set of experiments, the biotinylated cell surface protein was extracted from the total lysate using avidin-

agarose (Sigma) chromatography: 100 brains-worth (200 ml) of total lysate was loaded onto a 1.0 ml avidin-agarose column, the column was washed with 50 ml Wash buffer (0.1 M sodium phosphate pH 7.0 with 0.15% Zwittergent), and the biotin-labeled cell surface protein was competitively eluted with Wash buffer supplemented with 1 mg/ml D-biotin. Fractions containing biotinylated protein were pooled (approximately 1 mg yield from 100 brains eluted in 10 ml buffer) and dialyzed versus the above physiological buffer for 2 h at room temperature, and concentrated 10-fold (final concentration approximately 1 mg/ml). By both of these methods, both total and biotinylated cell surface enriched lysates were produced, the former to be used as a ligand source for the intracellular domain, the latter to be used as a ligand source for the extracellular Fn5 domain.

Affinity chromatography was performed by equilibrating the domain-coupled columns with Running buffer (the above physiological buffer containing 1% detergent @ 0.5% Zwittergent, 0.5% NP-40). The columns were pre-blocked with 5 column volumes of a 1 mg/ml cytochrome C solution in this Running buffer. Protein aliquots were quick-thawed in a 37°C water bath, and ultracentrifuged in a SW-27 rotor for 12 hours at 20000 rpm in a Beckman L5-50 ultracentrifuge; any remaining macroscopic lipids were removed by 0.45 µm filtration. The brain lysate supernatant was loaded onto the column at a gravity-driven 1 ml/min flow rate, followed by a 10 column volume wash with Running buffer. Two different elution conditions were used: in the first, putative ligands were eluted competitively using the expressed domain itself (3 column volumes of 1 mg/ml free domain); in the second, two elution conditions were used sequentially, including 5-10 column volumes of high-salt Running buffer (50 mM Tris pH 8.0, 500 mM NaCl, 0.5% NP-40) followed by 5-10 column volumes of high pH buffer (150 mM NaCl, 0.1% NP-40, 50 mM triethanolamine pH 11.5); high pH eluted fractions were immediately neutralized with one-tenth volume of 1 M Tris-HCl, pH 7.0. The column was re-

equilibrated with Running buffer, washed with several volumes PBS, and stored in 30% glycerol/PBS; columns were used no more than three times before disposal. All elution fractions were analyzed by SDS-PAGE, and in the Fn5 experiments in which the putative ligand source was biotinylated, elution fractions were analyzed by Western blots (stained with AP-streptavidin) as per protocols described above (see *N-terminal Protein Sequencing of Expressed Domains*).

#### *Affinity Chromatography Utilizing the His-Tail*

The six-histidine tail expressed at the C-terminus of the intracellular domain was utilized in one additional affinity chromatography experiment. The purified domain was loaded onto the IMAC column in Physiological buffer (see above), and a 50-brain total lysate dialyzed versus this same Physiological buffer was run through the column at 1.0 ml/min. After a 20 column volume wash in Physiological buffer, the domain along with any putative ligands was eluted with 100 mM imidazole in Physiological Buffer. This method of affinity chromatography affords two important advantages to the previously described methods: 1) the expressed domain is not coupled directly to agarose, and therefore the protein has greater degrees of freedom and higher probability of proper conformation/folding, 2) the specific elution of the domain-ligand complex by imidazole does not bring down (as is the case with high salt or pH) non-specific agarose-interacting species.

#### *Expression Library Screening to Identify Putative Ligands*

While putative ligands to extracellular domains are presumed to be themselves, extracellular with appropriate hydrophobic secretion-permissive sequences and possibly involving chaperones and/or post-translational modifications in order to fold properly, in contrast, putative ligands to intracellular domains in general are predicted to fold accurately more readily in foreign environments, such as in a

bacterial expression library. For this reason, putative ligands to the Bravo/Nr-CAM intracellular domain were searched for by labeling the domain and using it as a probe to directly identify a plaque clone that expresses an interacting protein coding sequence. An E8 chicken brain lambda library was plated at 25,000 pfu/NZY agar plate (20 plates or  $5.0 \times 10^5$  total plaques) by inoculating phage with 600  $\mu\text{L}$  per plate of O.D.<sub>600</sub>=0.5 XL1-Blue *E.coli* (in 10 mM  $\text{MgSO}_4$ ) at 37°C for 15 min, mixing the inoculate with 3.5 ml melted top agar (10 g/l NZ Amine, 5 g/l NaCl, 2 g/l  $\text{MgSO}_4$ , 5 g/l Bacto-yeast extract, 7 g/l agarose) cooled to 55°C, and swirling the *E.coli*-phage-top agar while pouring evenly atop pre-warmed plates. The plates were placed at 42°C for 3.5 hours until tiny plaques were just visible, at which time nitrocellulose filters (Hybond-C, Amersham) were treated with 10 mM IPTG in  $\text{dH}_2\text{O}$  (filters were wetted slowly by wicking) and set to dry for one-half hour. After this half-hour drying, the filters were placed carefully, avoiding air bubbles, atop the plaques on the plates, and the plates were incubated at 37°C for an additional 3 hours. Duplicate IPTG-pretreated filters were applied for an additional 3 hour incubation. Using a needle dipped in waterproof ink, the orientation of the nitrocellulose filters in relation to the plates was marked by piercing the filter and agar in several places. Filters were lifted from the plates and blocked for at least 2 hours (occasionally overnight) in 5% Carnation Instant Milk in PBS.

Protein was labeled by iodination. One iodobead (Pierce) was washed in iodination buffer (0.1 M sodium phosphate buffer, pH 6.5) and blotted dry on a kim-wipe. To 100  $\mu\text{L}$  of iodination buffer, 2 mCi of  $\text{Na}^{125}\text{I}$  (I.C.N) was added and incubated for 5 min at room temperature. 250  $\mu\text{g}$  purified protein was added to this mixture and the reaction was allowed to occur for 10 min at room temperature. The iodobead was then removed from the reaction vessel, and free  $^{125}\text{I}$  was removed by gel filtration (2 ml Sephacryl S-200 column). The void volume was collected and incorporation was measured by blotting dilutions of iodinated protein onto nitrocellulose filters, washing four times in TGI



buffer (25 mM Tris pH 8.3, 192 mM glycine, 20% methanol, 10 mM sodium iodide), and measuring gamma radiation with a scintillation counter. Typical incorporation was  $10^8$  cpm for the 250  $\mu$ g protein. After filter blocking was completed, the labeled protein was added directly to the blocking solution at a concentration of 100  $\mu$ g/100 ml (more than  $10^6$  cpm per filter), and filters were incubated for 1-2 hours at 8°C. The filters were washed 4-5 times with PBS containing 0.05 % Tween-20 (PBST), and autorads developed for 24 hours.

## Results

### *Heterologous Expression of the Intracellular and Fn5 Domains of Bravo/Nr-CAM*

The *Pichia pastoris* pPIC-9K expression vector uses the methanol inducible promoter of the AOX-1 gene to drive expression of the downstream coding sequence. Protein is directed to be secreted from the yeast using the *cerevisiae* alpha factor mating sequence. The intracellular domain of Bravo/Nr-CAM was expressed and secreted at high levels (greater than 10 mg/L) in 100% (12 of 12) of slow-growing AOX-1-inserted clones (typically, high level expression is detectable in approximately 20% of slow-growers). Two major protein bands were detectable, a higher molecular weight form that does not IMAC-purify, and a lower molecular weight form that was eluted at 200 mM imidazole from the IDA-agarose column. For all biochemical studies, only the IMAC-purifiable form was used. The form that does not IMAC-purify may adopt a non-native conformation that in addition to rendering the His+ tail inaccessible or ineffective to the IDA-nickel complex, also renders the intracellular domain far less reactive with the anti-Bravo polyclonal serum as compared to the purifiable form.

The Bravo/Nr-CAM fifth fibronectin type III repeat (Fn5 domain) under the control of the bacterial T7 promoter was



overproduced in BL21-pLYS-S *E. coli* following induction with IPTG. pLYS-S refers to an independent plasmid in the BL21 bacterial strain which produces low levels of T7 lysozyme, an natural inhibitor of T7 RNA polymerase, and therefore activity at this T7 promoter is reduced in the absence of the potent IPTG inducer (lowers toxicity and generally permits denser cultures and higher eventual expression levels). The overproduction of the Fn5 domain resulted in the vast majority of it being found in inclusion bodies at levels exceeding 100 mg/l culture (Fig. 3).

#### *Generation of a Domain-Specific Monoclonal Antibody*

In order to investigate the spatial and temporal regulation of alternative splicing of the fifth fibronectin type III repeat of Bravo/Nr-CAM, we have raised a domain-specific monoclonal antibody (6D2). The domain was PCR-cloned with engineered flanking sequences in the primers that permitted directional cloning into a bacterial expression vector. One of the main concerns with heterologous protein expression, especially in prokaryotic systems, is the issue of proper folding and tertiary structure. On this front, three positive statements may be made on the expression of the Fn5 domain: 1) it remains in solution at high concentrations (at least 50 mg/ml), 2) an anti-Bravo polyclonal serum recognizes native epitopes of the expressed domain, and 3) the expressed domain successfully generated monoclonal antibodies that subsequently recognize the native epitope on tissue sections and Western blots. It should also be noted that the resulting anti-Fn5 monoclonal antibody used in this study is highly specific: it does not recognize other extracellular domains of the Bravo/Nr-CAM protein, nor does it cross-react with domains of closely related homologs (Ng-CAM, Neurofascin, Axonin-1 or Contactin).

On Western blots, the 6D2 monoclonal antibody raised against the alternatively spliced Fn5 exon, reacts with a single, broad band between 80-90 kD and this reactivity is identical when both

immunopurified Bravo/Nr-CAM and total protein lysate is blotted (Fig. 4). The histological staining pattern in all tissues examined also precisely mirrors Bravo/Nr-CAM antigen distribution and this immunoreactivity *in situ* as well as the aforementioned Western blot results indicate that this monoclonal antibody indeed recognizes the intended antigen. The broad 80-90 kD band corresponds to the beta chain of the Bravo/Nr-CAM protein: the membrane-spanning subunit of Bravo/Nr-CAM that has been previously identified to contain the intracellular domain and the two C-terminal fibronectin type III repeats (including the targeted alternative Fn5 exon). The intact Bravo protein is thought to be cleaved after insertion into the membrane, and the resulting alpha and beta chains are associated non-covalently although can be readily co-immunopurified (Kayyem *et al.*, 1992). Of the entire family of closely related proteins, including chicken Neurofascin, Axonin-1, F11 and Ng-CAM, only the latter is known to have this Bravo/Nr-CAM-like heterodimer structure and protein subunits in the 80-90 kD range. Therefore, on the basis of Western blot non-reactivity in the size range (120-200 kD) expected for these other proteins (Ng-CAM excepted), and histological staining that clearly is distinguishable from each of these other proteins (Ng-CAM included), it is inferred that the 6D2 anti-Bravo/Nr-CAM Fn5 monoclonal antibody is highly specific and exhibits no evidence of cross-reactivity.

#### *Co-expression of Domain-Containing and Domain-Lacking Isoforms*

Serial sections were used to investigate staining of the anti-Fn5 alternatively spliced exon of Bravo/Nr-CAM as compared to overall Bravo/Nr-CAM staining. Monoclonal antibody staining (6D2 MAb against Fn5 versus 2B3 MAb against all Bravo/Nr-CAM isoforms) was studied across a wide set of developmental stages (between E3 and E18 of the chicken embryo) in all tissues so far identified as Bravo/Nr-CAM positive. These tissues include retina, optic nerve, optic tectum, cerebellum, spinal cord, and peripheral ganglia. In all tissues at all

stages, the expression of the alternative Fn5 exon was indistinguishable from the full extent of Bravo/Nr-CAM expression. To further investigate this question, double-labeling on the same histological section was performed, and confocal microscopy used to examine isoform expression patterns at various cellular depths in the selected tissues and developmental stages. Here again, the Fn5-containing isoform appears to be expressed in all cells that otherwise express Bravo/Nr-CAM throughout development (Fig 5).

PCR experiments confirm that in all tissues and stages examined, both isoforms are expressed at similar levels, with the possible exception of cerebellum (Fig 6). Given these PCR results and that no gradient or differential expression of the Fn5 isoform was evident *in situ*, it is possible that this alternative splicing event is neither regulated spatially, temporally nor quantitatively in development. No reliable quantitative measure can be made, however, as to the differential levels of isoform-lacking isoforms, which is a difficult if not impossible question to ask directly given that *in situ* hybridization probes designed to recognize isoform-lacking sequences would entirely contain flanking sequences to Fn5 that are found in all isoforms.

#### *Affinity Chromatography Identifies a Putative 70 kD Ligand to the Fn5 Domain*

An affinity chromatography experiment was performed using 1 ml agarose solid support coupled to 1.0 mg heterologously expressed Fn5 domain. Biotinylated cell surface E8 chicken lysate (100 brains in 1.0 ml Physiological Buffer containing residual detergent after dialysis, see Methods) was run through the column, and the column subsequently washed with 10 column volumes of Physiological Buffer containing 0.5% NP40/0.5% Zwittergent. Putative bound ligands retained in the column after this wash were competitively eluted with 1 ml of Physiological Buffer/0.5% NP40/0.5% Zwittergent containing 1

mg expressed Fn5 domain. Half-column-volume fractions were collected for the wash and elution steps, run on a 10% SDS PAGE gel, electroblotted, and biotinylated cell surface putative ligands stained using AP-streptavidin. The wash fractions contained several high molecular weight bands, and the Fn5 competitive elution fractions contained in addition, a single dominant 70 kD putative ligand (Fig. 7). The yield of this putative ligand is probably in the sub-nanogram range, as these fractions in their entirety are barely silver-stainable; at least 5000 chicken brains may be required in order to consider N-terminal sequencing of this provocative protein band.

#### *Intracellular Domain Results*

Two separate kinds of experiments were performed in an effort to search for putative ligands to the intracellular domain of Bravo/Nr-CAM. In the first set of experiments, affinity chromatographic methods were used in which the heterologously expressed domain was attached to solid agarose support either directly by covalent coupling or indirectly by non-covalent attachment of the histidine tail to Ni<sup>2+</sup>-activated IDA-agarose. In both cases, E8 and P2 chicken brain protein lysates were applied to the column as a putative ligand source. With an exhaustive effort, under a wide range of ionic strength, detergent and divalent cation conditions, no silver-stainable amount of specifically-bound proteins were identified as compared to negative controls. Figure 7 illustrates experimental versus control results of specific elution fractions. In the second set of experiments, expression library screening was used in which the heterologously expressed domain was labeled by radioactive iodination (Fig 7 shows a blot of the labeled domain) in order to identify library plaques that express a putative ligand. In two separate efforts, no (double-lift) positive plaques were identified.

## Discussion

### *Is this Alternative Splicing Event Functionally Significant?*

With the data in hand, we conclude that all cells or brain regions express both Fn5-containing and Fn5-lacking isoforms without an obvious spatial, temporal or quantitative regulatory preference. At least six distinct isoforms around the Fn5 domain have been identified in Bravo/Nr-CAM proteins, and specific alternative splicing events have been conserved since avian-mammalian divergence at least 200 million years ago. The Fn5 domain itself is among the most conserved extracellular domain in Bravo/Nr-CAM (>80% identical), and therefore is likely to have functional importance. Furthermore, the entire remainder of C-terminal sequences adjacent to the alternative Fn5 exon that includes the transmembrane and intracellular domain (approximately 150 amino acids in length) is 100% conserved between the chicken and human proteins (Lane *et al.*, 1996).

It is possible that the Fn5 alternative splicing events may be functionally coupled to these remarkably conserved adjacent sequences, perhaps modulating *cis* co-receptor interactions within and on both sides of the cell membrane. Consistent with this notion, it is interesting to note that among the closely related six-Ig-like domain neural cell adhesion proteins so far identified, all those that do not have transmembrane and intracellular domains (the PI-linked Contactin/F11, Axonin-1/Tag-1, Big-1/Tag-like and Big-2 proteins), also do not have an evolutionary related fifth fibronectin type III repeat. This extracellular domain, therefore, may be subject to co-evolution with the intracellular regions of the protein, consistent with the notion that their functions are coupled.

One of Bravo/Nr-CAM's closest Ig Superfamily relative, L1, along with the neural N-CAM and N-Cadherin proteins, all manifest important intracellular kinase cascades by virtue of a co-receptor interaction with the fibroblast growth factor receptor (Williams *et al.*,

1994; Doherty *et al.*, 1995). Given that numerous neural cell surface proteins are PI-linked to the cell membrane, and many others have well-conserved yet short cytoplasmic domains, it is possible that such co-receptor coupling to elicit important intracellular responses may be widespread. We propose a model, therefore, where the function of the alternatively spliced fifth fibronectin type III repeat may be intimately involved in signal transduction events that are manifested via co-receptor coupling. Putative transmembrane-spanning interactions may be modulated or altogether eliminated when the Fn5 exon is excluded in specific isoforms.

Given that extracellular domains of the entire neural cell adhesion protein family are apparently quite promiscuous (several heterophilic combinations of interactions have been identified), it is likely that some integrative mechanism may be present in order to allow the neuron to distinguish among these various specific ligand interactions along the full length of the axon. For example, if Bravo/Nr-CAM is capable of interacting with both Axonin-1/Tag-1 (Suter *et al.*, 1995) and F11/Contactin (Morales *et al.*, 1993), and these different interactions are able to elicit a different intracellular response, then some mechanism might be in place to permit a particular signal transduction pathway in the one case but not the other. One model whereby three different cellular responses are distinguished for two distinct extracellular signals by virtue of Fn5 alternative splicing events is illustrated in Figure 8B. In this way, a uniform expression of both isoforms along the extending axon might permit distinct intracellular contributions for multiple extracellular ligands, the overall emergent effect being an integrated, singular cellular behavior, or as proposed in the figure, localized cellular behaviors at various places along the axon.

Alternatively and more simply, it is possible that the co-expression of both Bravo/Nr-CAM variants permit the formation of isoform heterodimers: the domain-containing and domain-lacking forms may associate to form a staggered co-receptor Bravo/Nr-CAM

structure. Bravo/Nr-CAM may be able to interact with itself (Mauro *et al.*, 1992), and if this interaction can occur *cis* (i.e., Bravo dimerization), then the isoform homodimer versus heterodimer variations might have *trans* ligand-binding consequence (Fig. 8A). It should be reiterated, however, that it is unlikely that the Fn5 exon merely plays a structural role of this nature considering the high degree of sequence conservation in the domain. And so while this model of heterodimerization is possible, the full consequence of this alternative splicing event is probably beyond such a simple scenario, perhaps involving both *trans* ligand-binding and *cis* signal transduction modulations.

#### *Putative Ligands to the Fn5 and Intracellular Domains*

This Chapter describes a significant effort that has spanned more than three years in which expressed domains of Bravo/Nr-CAM were used to search for ligands. Using two different domains (Fn5 and intracellular), three different methods (affinity chromatography, co-immunoprecipitation and expression library screening) and a wide range of experimental conditions (variables include ionic strength, detergent concentrations, divalent cations, coupling methods and elution conditions), these efforts were not successful in so far as obtaining tractable amounts of putative ligands (i.e., at least  $\mu\text{g}$  or super pmole quantities). While the putative 70 kD ligand to the Fn5 domain remains provocative (in terms of reproducibility, specificity and presumed binding constant), the result does illustrate a major problem with these kinds of affinity chromatographic experiments: unless the experiment is designed to assay a specific, known ligand (using an antibody for example), the quantitation is less than desirable in terms of ultimately identifying the interacting species. Consider that the yield of total cell surface protein in 100 chicken brains is approximately 1.0 mg, and therefore in order to capture 1  $\mu\text{g}$  of ligand (near the minimum in order to obtain N-terminal protein sequence data), the



ligand must represent at least 0.1% of this total protein; this supposition also assumes the best-case scenario in which the binding constant is sufficient to capture 100% of available ligand.

For both expression library screening and affinity chromatographic methods, the result can only be as good as the proteins in the assay. In this case, both the structure/folding of the expressed domains and the putative ligand source must be considered. This seems to be an especially critical point considering that in the library screening experiment, clones were not even identified for ankryin cDNA, which is a known binding partner to the Bravo/Nr-CAM intracellular domain. While there is no functional assay for proper folding of Bravo/Nr-CAM nor any of its specific domains, a few positive comments may be made with regard to the folding of the expressed Fn5 and intracellular domains: 1) anti-Bravo polyclonal serum recognizes epitopes on both domains, 2) both domains elute in a tight FPLC peak, and 3) both domains remain in solution at high concentrations (i.e., at least 10 mg/ml). On the other hand, the intracellular domain runs larger than its predicted molecular weight on SDS-PAGE, even when it is de-glycosylated, and this larger-than-expected size may indicate non-native conformational problems.

Even if the domain is perfectly folded (or its structure adequately presents functional epitopes), the nature of these specific experiments also is accompanied by an important caveat. In the case of affinity chromatography, our preliminary studies indicated that satisfactory yield of rare cell surface receptors was only obtainable in the presence of a relatively high 5% detergent cocktail (and not the 0.5% detergent concentrations reported elsewhere). While in some experiments, attempts were made to reduce this detergent concentration, this amount of even mild detergent may render proteins unable to interact with their *in vivo* partners. In addition, preparation and column running times were on the order of several hours, and even in the presence of potent protease inhibitors, there remains an ample

opportunity for proteolysis to further sabotage the experiment. In the case of expression library screening, many of these issues are irrelevant, however perhaps an even more serious threat emerges: the issue as to whether functional epitopes are conformationally compromised in a bacterial phage lysate. In my own experience with monoclonal antibody library screening, less than 20% of our own catalog of high-affinity IgG monoclonal antibodies that otherwise react potently on Western blots, immunohistological tissue sections, and with immunopurification experiments, nevertheless fail to recognize their antigen cDNA's expressed in a phage plaque. This is almost certainly due to prokaryotic incompetence in folding heterologous, eukaryotic protein.

Finally, it should be pointed out that these experiments are demanding even in the absence of receptor and ligand folding problems. Because the putative ligand source, in the case of affinity chromatography, can not easily be concentrated to levels that would allow its volume to be 10% or less of a column volume (i.e.,  $> 1$  brain/ $\mu$ l), these experiments were not designed to detect chromatographing species. In this way, only high affinity interactions (i.e., ligands that would be retained in the column) were investigated. Likewise for expression screening, the five high-volume (10 ml/filter) filter washes in detergent severely challenge low-affinity interactions. With both methods, therefore, it is possible that more transiently interacting ligands may have escaped detection.

Especially in the case of the intracellular domain, the yeast two-hybrid system (Fields and Song, 1989; Chien *et al.*, 1991; Fields and Sternglanz, 1994) provides an attractive solution to many of these problems. First, both the domain and the ligand source are expressed under similar conditions, in a eukaryotic cell, with none of the detergent/quantitation/coupling/proteolysis pitfalls discussed here. Assuming that both domain and ligand can fold properly in the cytoplasm and both are able to be targeted to the nucleus, this method

has proven to be an exciting and productive way to identify binding partners. In terms of biochemical methods that might improve the lot of ligand-expeditions in the case of extracellular domains not readily targeted to the nucleus (as presumed to be the case for the Fn5 domain), both Biacore and Mass Spectrometry technology may present a bright future. In both cases, the most damning issue of quantitation is alleviated, which furthermore might permit milder (and more native) preparative methods. The immediate goal in ligand experiments such as those described here should begin to shift from N-terminal sequencing for probe design (i.e., to screen a library), to database matching (i.e., to screen Genbank). In this way, the requirement for nanomole quantitation should wane as more sensitive methods are developed that are not encumbered by the requirement for contiguous sequence information. For example, Mass Spectrometry applications aimed at amino acid content (i.e., total mass of a given amino acid) might be sufficient input for search algorithms, especially if the search can be narrowed by species, cell type or subcellular localization. Mass Spectrometry (as well as Biacore technology) also permits extremely sensitive deciphering of bound versus unbound ligand moieties, and in this way, the affinity experiment, detection and identification process could be combined in one, extremely powerful tool. Obviously, not soon enough for me.

## Conclusions

This Chapter explores the function of Bravo/Nr-CAM by using heterologously expressed protein and *in vitro* biochemistry. In general, advances in protein expression make possible this sort of powerful alternative to genetic epistasis/suppressor experiments as a means to elucidate molecular pathways. With many vertebrate gene knockout experiments, including some of Bravo/Nr-CAM's close relatives, occasionally yielding minimal or no phenotypic clues, biochemical approaches emerge as an important route to glean further functional insights. Furthermore, with various genome initiatives, the quest for sequence is no longer the issue; rather, experimentalists aim to decipher functions with sequence in hand. Manufacturing proteins to order from existing sequences, raising antibodies against the expressed proteins, and investigating function via *in vivo* perturbations and *in vitro* assays is an exciting frontier that should accelerate molecular-level understanding. Perhaps even more importantly, one critical aspect of molecular biology that remains almost untouched is the detailed biochemistry within the cell. For example, while numerous receptors seem to converge on a small number of common signal transduction messenger systems, little is understood how the cell integrates this overlapping information. The key to this integrative capacity may reside in the biochemical details: stoichiometry, molar ratios, localized ionic strength or pH, extent of phosphorylation, protein half-life and binding constants, etc. If so, predicting exactly how a cell is going to respond to a given broad set of extracellular signals generated in both the present and recent past, may require more than identifying the linear pathway, rather, it may require extensive system-level information on the full range of cellular components and chemistry. In this regard, *in vitro* (and *in cyber*) cell modeling using expressed proteins and old fashioned biochemistry may become a focal point in the next era of biological science.

Chapter IV

References

Atashi, J.R., Klinz, S.G., Ingraham, C.A., Matten, W.T., Schachner, M., and Maness, P.F. (1992). Neural cell adhesion molecules modulate tyrosine phosphorylation of tublin in nerve growth cone membranes. *Neuron* 8 (5), 831-842.

Ayer, D.E., Kretzner, L., and Eisenmann, R.N. (1993). Mad: a heterodimeric partner for Max that antagonizes Myc transcriptional activity. *Cell* 72, 211-222.

Blackwood, E.M., and Eisenman, R.N. (1991). Max: a helix-loop-helix zipper protein that forms a sequence-specific DNA-binding complex with myc. *Science* 251, 1211-1217.

Bork, P., and Doolittle, R.F. (1992). Proposed acquisition of an animal protein domain by bacteria. *Proc. Natl. Acad. Sci. USA* 89, 8990-8994.

Burton-Wurster, N., Lust, G., and Wert, R. (1989). Expression of the ED B fibronectin isoform in adult human articular cartilage. *Biochem Biophys. Res. Comm.* 165, 782-787.

Chien, C-T., Bartel, P.L., Sternglanz, R., and Fields, S. (1991). The two hybrid system: a method to identify and clone genes for proteins that interact with a protein of interest. *Proc. Natl. Acad. Sci. USA* 88, 9578-9582.

Cregg, J.M., Vedvick, T.S., and Raschke, W.C. (1993). Recent advances in the expression of foreign gene in *Pichia pastoris*. *Bio/Technology* 11. 905-910.

Davis, J. Q., and Bennett, V. (1994). Ankyrin binding-activity shared by the Neurofascin/L1/Nr-Cam family of nervous-system cell-adhesion molecules. *J. Biol. Chem.* 269 (44), 27163-27166.

Doherty, P., Ashton, S.V., and Moore, S.E. (1991). Morphoregulatory activities of N-CAM and N-Cadherin can be accounted for by G-protein-dependent activations of L-type and N-type neuronal calcium channels. *Cell* 67 (1), 21-33.

Doherty, P., Moolenaar, C. E. C. K., Ashton, S. V., Michalides, R. J. A. M., and Walsh, F. S. (1992). The vase exon down-regulates the neurite growth-promoting activity of N-CAM-140. *Nature* 356, 791-793.

Doherty, P., Williams, E., and Walsh, F.S. (1995). A soluble chimeric form of the L1 glycoprotein stimulates neurite outgrowth. *Neuron* 14, 57-66.

Dorries, U., and Schachner, M. (1994). Tenascin messenger-RNA isoforms in the developing mouse-brain. *Journal of Neuroscience Research* 37 (3): 336-347.

Edelman, G.M. (1986). Cell adhesion molecules in neural histogenesis. *Ann Rev. Physiol.* 48, 417-430.

Fields, S., and Song, O-K. (1989). A novel genetic system to detect protein-protein interactions. *Nature* 340, 245-246.

Fields, S., and Sternglanz, R. (1994). The two-hybrid system: an assay for protein-protein interactions. *Trends Genet.* 10, 286-292.

Goldman, S.A., Williams, S., Borami, K., Lemmon, V., and Nedergaard, M. (1996). Transient coupling of Ng-CAM expression to Ng-CAM-dependent calcium signaling during migration of new neurons in the adult songbird brain. *Molecular Cellular Neuroscience* 7 (1), 29-45.

Hoare, D.G., and Koshland, D.E. (1967). A method for the quantitative modification and estimation of carboxylic acid groups of proteins. *J. Biol. Chem.* 242, 2447-2453.

Huhtala, P., Chow, L.T., and Tryggvason, K. (1990). Structure of the human type IV collagenase gene. *Biol. Chem.* 265, 11077-11082.

Jukes, T. H. (1980). Silent nucleotide substitutions and the molecular evolutionary clock. *Science* 210:, 973-978.

Kaczmarek, J., Castellani, P., Nicolo, G., Spina, B., Allemanni, G., and Zardi, L. (1994). Distribution of oncofetal fibronectin isoforms in normal, hyperplastic and neoplastic human breast tissues. *International Journal of Cancer* 59 (1), 11-16.

Kayyem, J.F., Roman, J.M., de la Rosa, E.J., Schwarz, U., and Dreyer, W.J. (1992). Bravo/Nr-CAM is closely related to the cell adhesion molecules L1 and Ng-CAM and has a similar heterodimer structure. *J. Cell Biol.* 118, 1259-1270.

Klinz, S.G., Schachner, M., and Moness, P.F. (1995). L1 and N-CAM antibodies trigger protein phosphorylation activity in growth cone enriched membranes. *Journal of Neurochemistry* 65, 84-95.



Lane, R.P., Chen, X-N., Yamakawa, K., Vielmetter, J., Korenberg, J.R., and Dreyer, W.J. (1996). Characterization of a highly conserved human homolog to the chicken neural cell surface protein Bravo/Nr-CAM that maps to chromosome band 7q31.

MacGregor, P.F., Abate, C., and Curran, T. (1990). Direct cloning of leucine zipper proteins: jun binds cooperatively to the CRE with CRE-BP1. *Oncogene* 5, 451-458.

Maniatis, T., Fritsch, E.F., and Sambrook, J. (1982). *Molecular Cloning*, Cold Spring Harbor, New York.

Mauro, V.P., Krushel, L.A., Cunningham, B.A., and Edelman, G.M. (1992). Homophilic and heterophilic binding activities of Nr-CAM, a nervous system cell adhesion molecule. *J. Cell. Biol.* 119, 191-202.

Morales, G., Hubert, M., Brummendorf, T., Treubert, U., Tarnok, A., Schwarz, U., and Rathjen, F.G. (1993). Induction of axonal growth by heterophilic interactions between the cell-surface recognition protein f11 and protein Nr-CAM/Bravo. *Neuron* 11 (6), 1113-1122.

Persohn, E., and Schacher, M. (1987). Immunoelectron microscopic localization of the neural cell adhesion molecules L1 and N-CAM during postnatal development of the mouse cerebellum. *J. Cell. Biol.* 105, 569-576.

Phizicky, E.M., and Fields, S. (1995). Protein-protein interactions: methods for detection and analysis. *Microbiological Reviews* 59 (1), 94-123.

Pollerberg, E., Burridge, K., Krebs, K., Goodman, S., and Schachner, M. (1987). The 180 kD component of the neural cell adhesion molecule N-CAM is involved in cell-cell contacts and cytoskeleton-membrane interactions. *Cell Tissue Res.* 250, 227-236.

Silkela, J.M., and Hahn, W.E. (1987). Screening an expression library with a ligand probe: isolation and sequence of a cDNA corresponding to a brain calmodulin-binding protein. *Proc. Natl. Acad. Sci. USA* 84, 3038-3042.

Suter, D.M., Pollerberg, G.E., Buchstaller, A., Giger, R.J., Dreyer, W.J., and Sonderegger, P. (1995). Binding between the neural cell-adhesion molecules axonin-1 and Nr-CAM/Bravo is involved in neuron-glia interaction. *J. Cell. Biol.* 131, 1067-1081.

Suzuki, S., and Naitoh, Y. (1990). Amino acid sequence of a novel integrin beta four subunit and primary expression of the mRNA in epithelial cells. *EMBO J.* 9, 757-763.

Taggart, R.T., and Samloff, I.M. (1982). Stable antibody-producing murine hybridomas. *Science* 219, 1228-1230.

Tucker, R. P., Spring, J., Baumgartner, S., Martin, D., Hagios, C., Poss, P. M., and Chiquetehrisman, R. (1994). Novel tenascin variants with a distinctive pattern of expression in the avian embryo. *Development* 120 (3), 637-647.

Vielmetter, J., Kayyem, J.F., Roman, J.M., and Dreyer, W.J. (1994). Neogenin, a cell surface protein expressed during terminal neuronal differentiation, is closely related to the human tumor suppressor molecule Deleted in Colorectal Cancer. *J. Cell Biol.* 127, 2009-2020.

Weller, A., Beck, S., and Ekblom, P. (1991). Amino acid sequence of mouse tenascin and differential expression of two tenascin isoforms during embryogenesis. *J. Cell Biol.* 112, 355-362.

Williams, E.J., Mittal, B., Walsh, F.S., Doherty, P. (1995). A calcium/calmodulin kinase inhibitor, KN-62, inhibits neurite outgrowth stimulated by CAMs and FGF. *Molecular and Cellular Neurosciences* 6 (1), 69-79.

Williams, E.J., Furness, J., Walsh, F.S., and Doherty, P. (1994). Activation of the FGF receptor underlies neurite outgrowth stimulated by L1, N-CAM, and N-Cadherin. *Neuron* 13, 583-594.

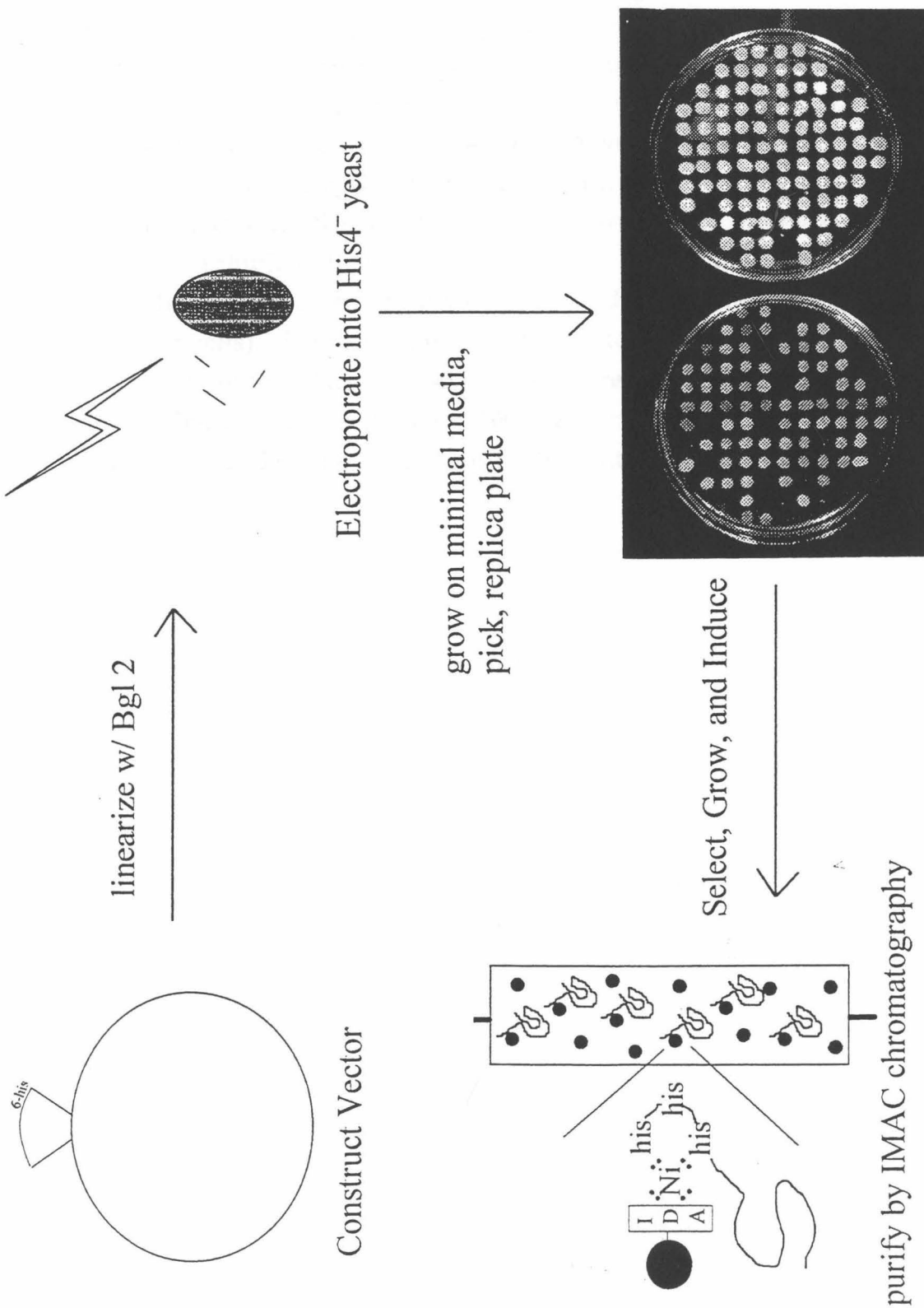
Wilson, A. C., Carlson, S. S., and White, T. J. (1977). Biochemical evolution. *Annu. Rev. Biochem.* 46, 573-639.

Wong, E.V., Schaefer, A.W., Landreth, G., and Lemmon, V. (1996). Involvement of P90 (RSK) in neurite outgrowth mediated by the cell adhesion molecule L1. *Journal of Biological Chemistry* 271 (30), 18217-18223.

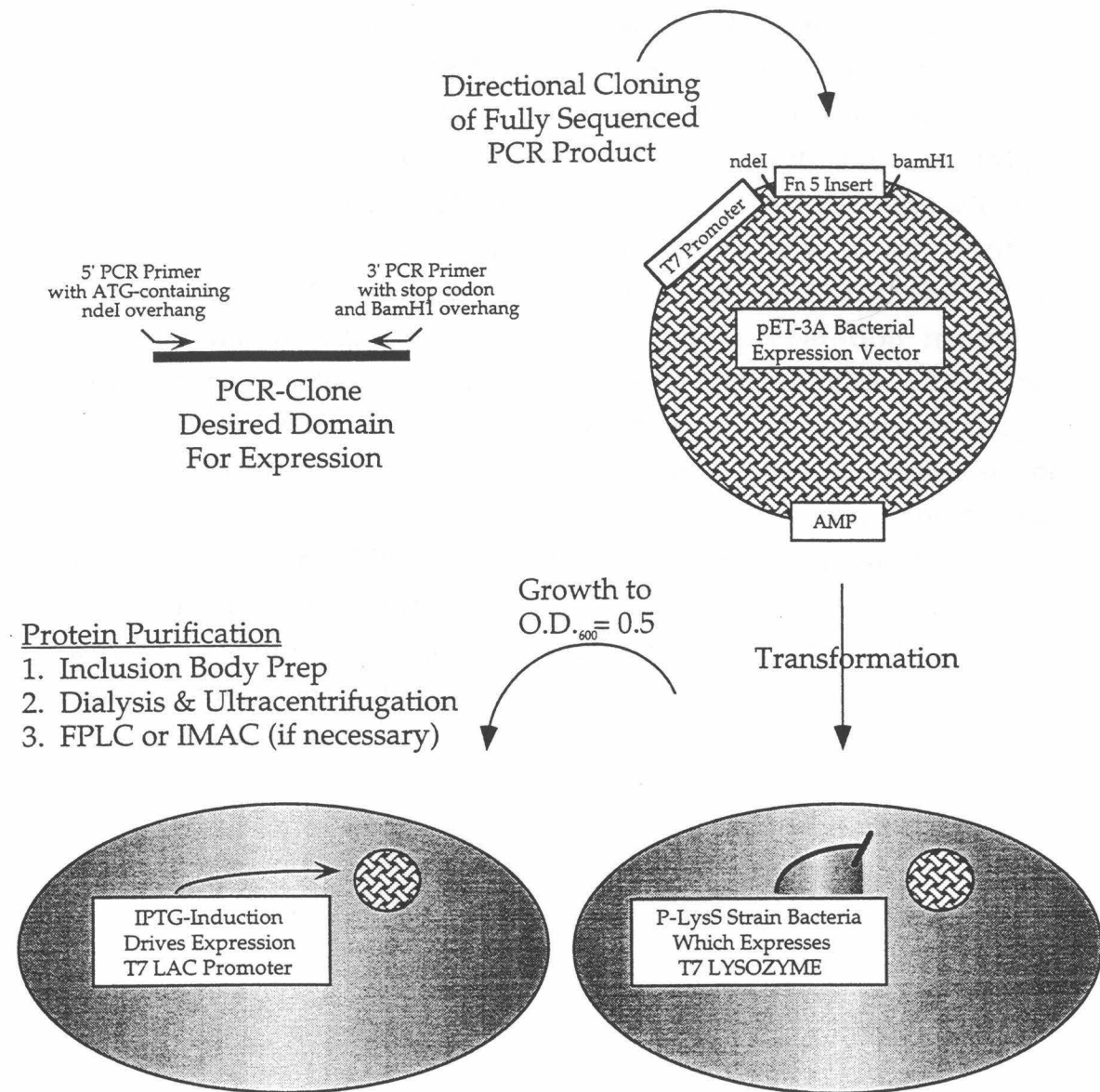
Young, R., and Davis, R. (1983). Yeast RNA polymerase II genes: isolation with antibody probes. *Science* 222, 778-782.

Zisch, A.H., Dalessandri, L., Ranscht, B., Falchetto, R., Winterhalter, K.H., and Vaughan, L. (1992). Neuronal cell adhesion molecule contactin/F11 binds to tenascin via its Immunoglobulin-like domains. *Journal of Cell Biology* 119, 203-213.

**Fig. 1.**      *Expression and purification of heterologous proteins using Pichia (yeast).* The pPIC9K vector was used for expression, and the insert cloned into the AOX-1 locus which has two advantages: 1) selection by slow methanol growth (AOX-1 gene is used for methanol metabolism when there are no other carbon sources available), and 2) high expression levels (the methanol-induced AOX-1 promoter drives as much as 50% of total cellular protein content). The vector also contains an alpha-factor signal sequence to drive secretion of the expression product. The photograph (lower right) shows transformed *Pichia* colonies grown on carbohydrate-based MD plates versus methanol-based MM plates, and illustrates the dramatic phenotype used for selection. IMAC chromatography was used for purifying expressed proteins from yeast culture supernatants utilizing the six-histidine tail that chelates nickel immobilized on the column. High level (10 mg/L) expression of the intracellular domain of Bravo was successful in 12 of 12 methanol slow-growing clones.



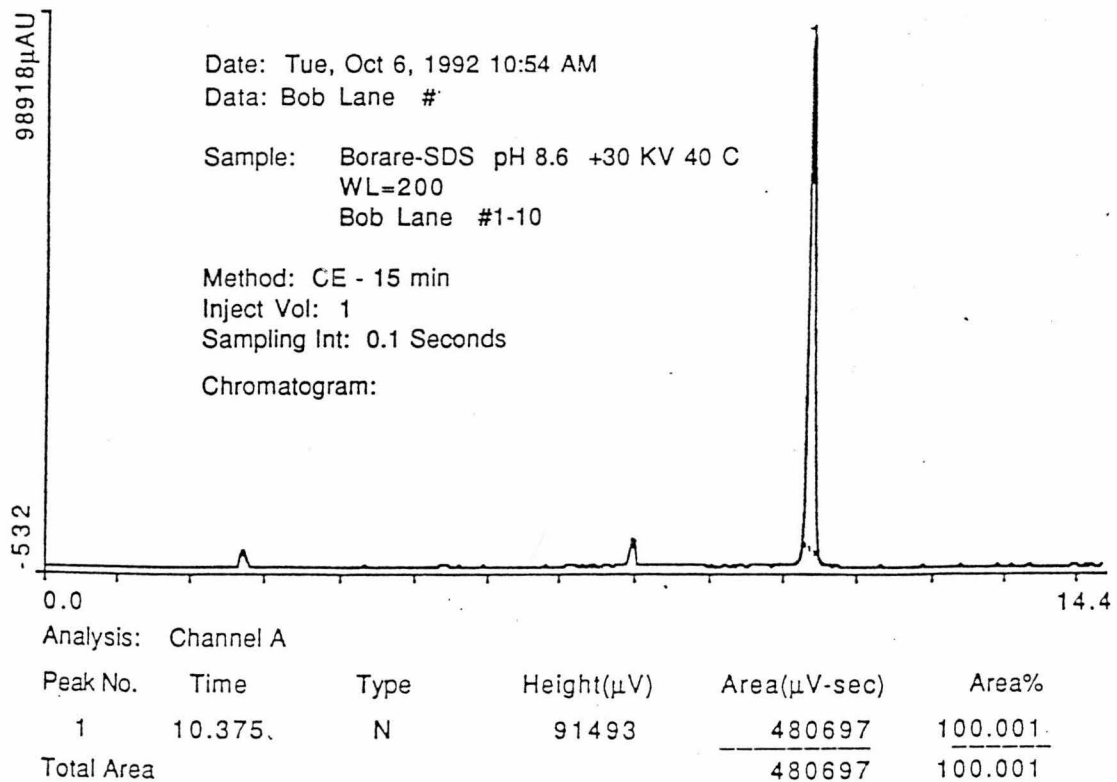
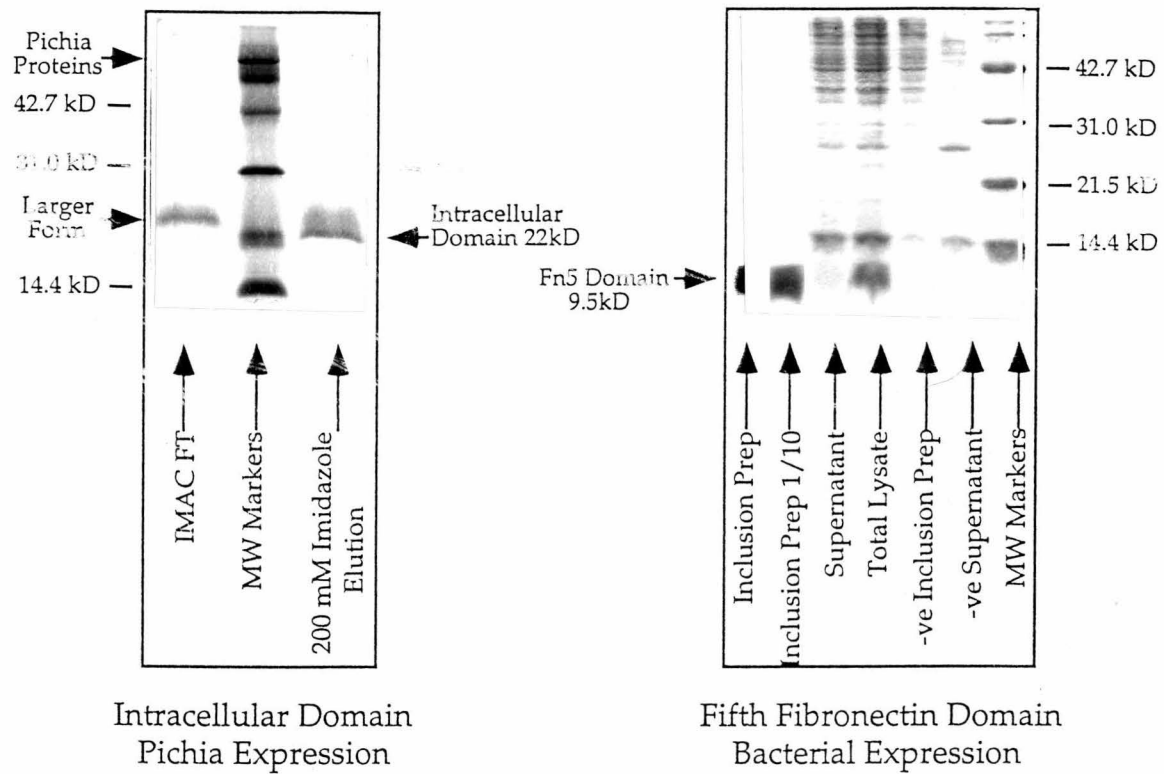
**Fig. 2.**      *Expression and purification strategy in bacteria.* Cartoon illustrating the general protein expression protocol and strategy in bacteria, including targeting specific domains by PCR which also permits primers designed with appropriate restriction sites, stop codons, and purification tags. The pET-3A vector is selectable using ampicillin, and cloning is directed upstream of a powerful T7-LAC promoter. The P-LysS host cells maintain low levels of lysozyme, a polymerase inhibitor that keeps expression levels low during the growth phase (high heterologous expression during growth is usually toxic to the cells). Expression is driven by addition of IPTG to grown cultures (not yet log-phase, OD=0.5). For expression of the Fn5 domain of Bravo/Nr-CAM, levels exceeded 100 mg/L culture which was almost exclusively included (i.e., inclusion body content was 95% expressed domain).



Yield = Approx. 100 mg/L



**Fig. 3.**        *SDS PAGE and capillary electrophoresis indicating yield and purity of expressed fifth fibronectin type III repeat and intracellular domains of Bravo/Nr-CAM.* Upper gels illustrate expression levels and purity of both domains: the intracellular gel (left) shows the IMAC non-retained yeast supernatant (indicated by "IMAC FT") and 200 mM imidazole-eluted fraction containing approximately 10 mg purified domain (from 1.0 L original culture); the Fn5 gel (right) shown the various components of the inclusion body preparation including inclusion body solubilized in 6 M guanidine hydrochloride. Note that the 9.5 kD Fn5 domain is perhaps between 10-50% of the total lysate, yet almost entirely pure in the inclusion body. After solubilization of inclusion bodies in buffered 6 M guanidine hydrochloride, the protein was dialyzed versus PBS which resulted in re-precipitation of insoluble material. The ultracentrifuged supernatant contained 99% pure Fn5 domain in solution as assayed by capillary electrophoresis (below gels; running conditions indicated in the inset).

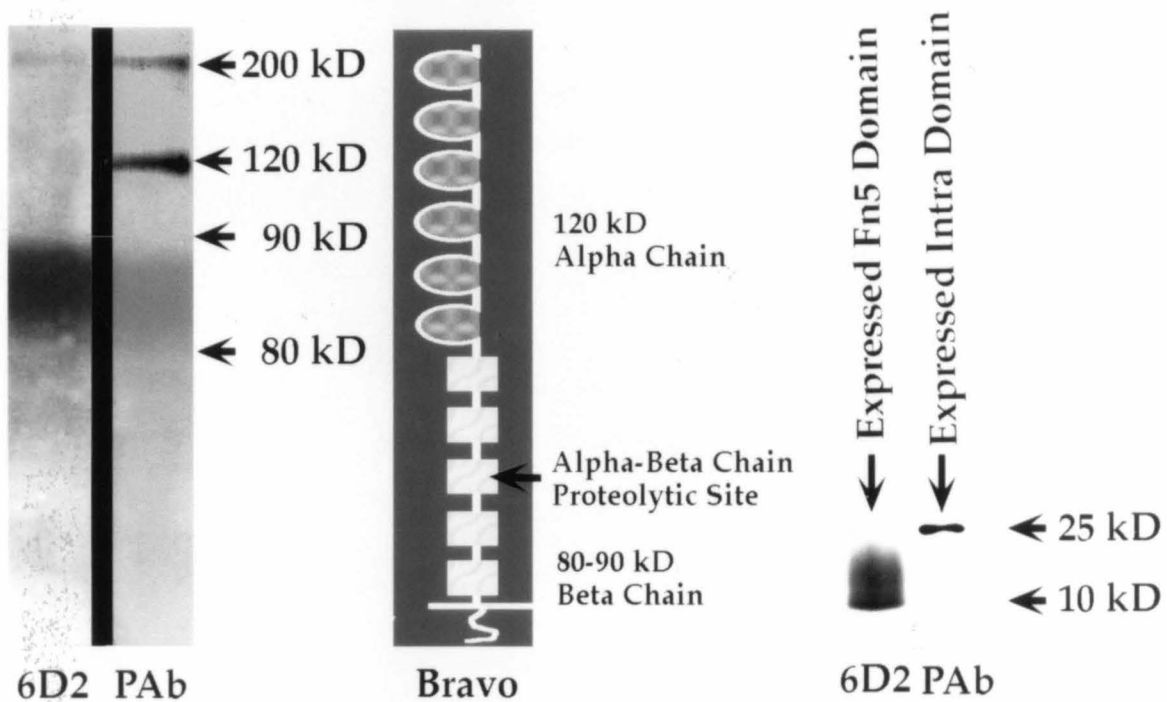


**Fig. 4.**        *Western blots using antibodies and heterologously-expressed antigens.*

Left: P2 chick brain lysate was blotted in order to demonstrate specific reactivity of the anti-Fn5 monoclonal antibody (6D2). Bravo polyclonal antiserum (PAb) was included as a positive control. The anti-Fn5 monoclonal antibody recognizes a diffuse 80-90 kD set of protein bands, as well as faintly recognizing a 200 kD band. The anti-Bravo polyclonal recognizes proteins at three locations in the blot: it stains the same 200 kD and broad 80-90 kD set that the domain-specific antibody recognizes, and in addition, it stains a 120 kD band. The cartoon of the Bravo structure to the right of this blot illustrates that these staining patterns are exactly as expected: the 200 kD intact Bravo form, and the 80-90 kD beta chain are the two Bravo proteins in the lysate expected to contain the Fn5 domain; the 120 kD band recognized in addition by the polyclonal serum corresponds to the alpha chain which does not contain the Fn5 domain. Note that the anti-Fn5 monoclonal antibody is quite cleanly negative for this 120 kD range, indicating that it does not cross-react with fibronectin domains from many of Bravo's closest Ig Superfamily relatives, nor does it cross-react with any of Bravo's Ig-like domains or any of its first three fibronectin domains.

Right: Antibody reactivity with the heterologously expressed Bravo domains. The expressed Fn5 and intracellular domains were run on 12% SDS PAGE, blotted, and stained with the anti-Fn5 and anti-Bravo polyclonal serum respectively. This indicates that antibodies, which otherwise recognize Bravo epitopes in P2 lysate and on tissue sections, also recognize the expressed Bravo protein fragments.

## Western Blot Analysis

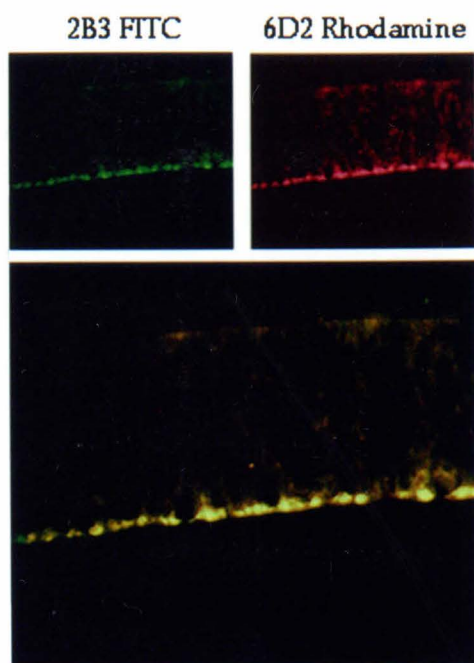


Total P2 Brain Lysate  
Stained with Anti-Fn5  
and Anti-Bravo polyclonal

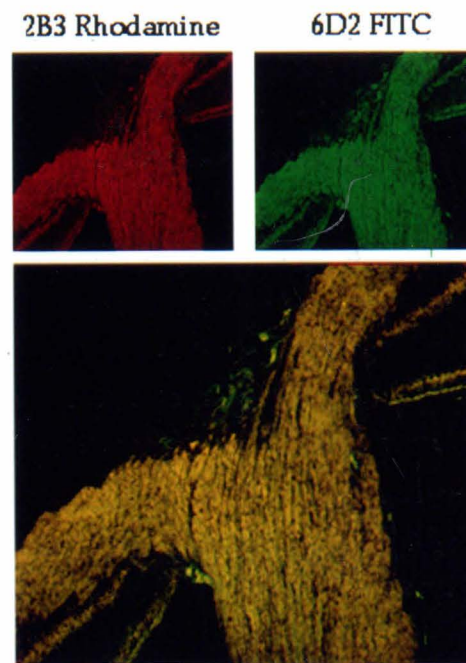
Fn5 and Intra. Domains  
Stained With Anti-Fn5  
and Anti-Bravo polyclonal

**Fig. 5.**        *Confocal microscopy and double-labeling using the 6D2 (anti-Fn5) and 2B3 (anti-Bravo) monoclonal antibodies in order to examine isoform expression patterns of Bravo/Nr-CAM.* Four photographic sets shown as representative samples: E5 retina, E8.5 optic nerve exit, E5 tectum, E5 spinal cord. All sections were labeled initially with 6D2 monoclonal antibody which specifically recognizes the alternatively spliced isoform of Bravo, followed by 2B3 monoclonal antibody which recognizes all Bravo isoforms. Confocal microscopy was used to investigate expression profiles, and overlapping digital images were captured separately using rhodamine and fitc-specific filters (the double-label overlap is displayed in the lower picture of each set). In all cases (including those not shown at various other stages), the alternatively spliced Fn5 domain is expressed in all cells that otherwise express Bravo, which implies that the domain-containing form is not developmentally regulated.

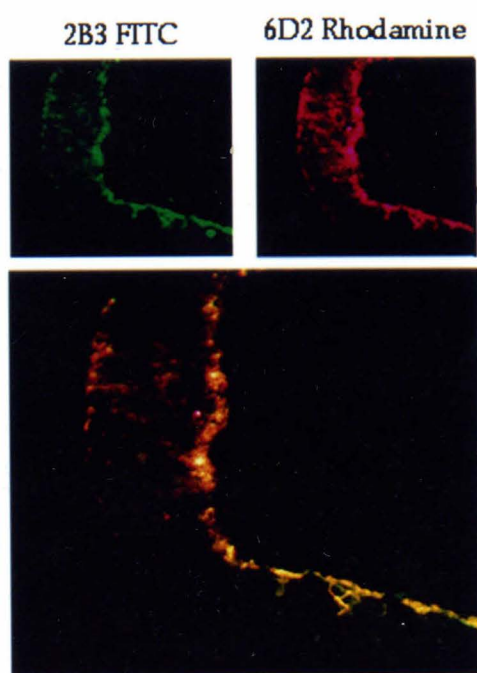
### E6 Retina



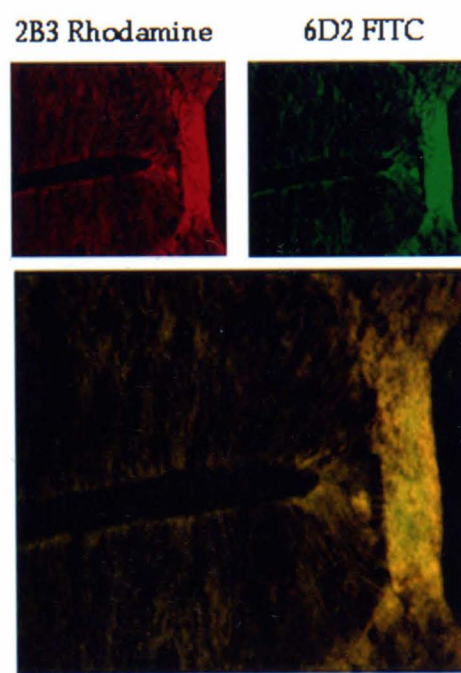
### E 8.5 Nerve Exit



### E5 Tectum

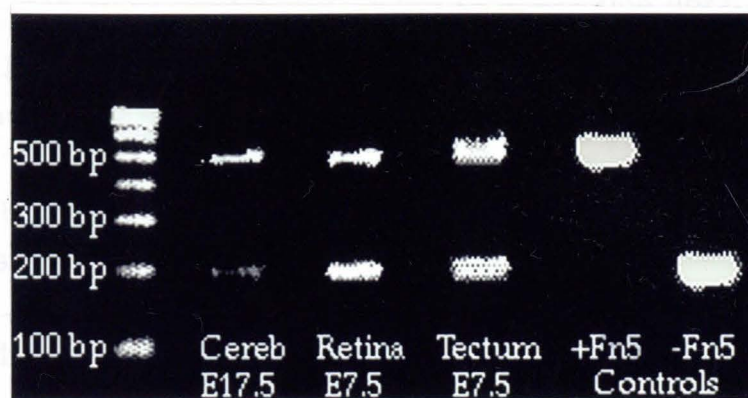


### E5 Spinal Cord



**Fig. 6.**        *RT-PCR*. Reverse-transcriptase PCR on RNA prepared from retina, tectum and cerebellum to investigate isoform expression of Bravo/Nr-CAM. RNA was isolated from retina, tecta, and cerebella at various stages and investigated for Fn5 isoform distribution by PCR using primers surrounding the alternatively spliced domain. Domain-containing and domain-lacking control plasmids were included in the PCR reaction and these corresponding products are shown (Fn5+ and Fn5-). In all samples, both isoforms are detectable at essentially uniform quantities (one representative sample is shown for each tissue type), although the low level of Fn5-lacking isoform product in cerebellar tissues may be significant.





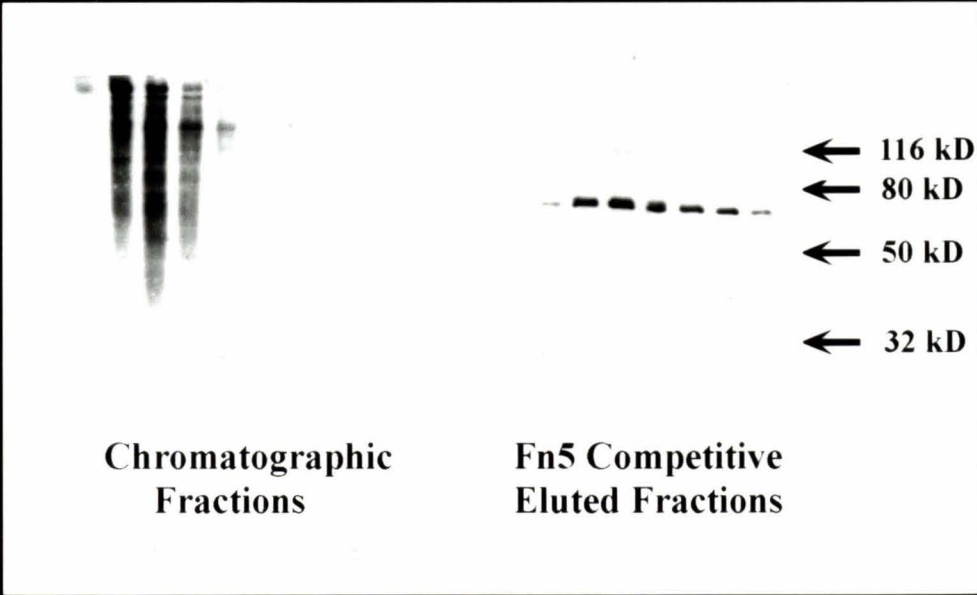
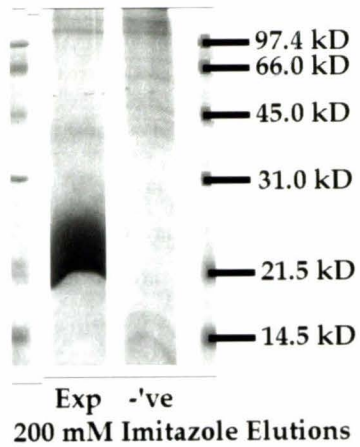
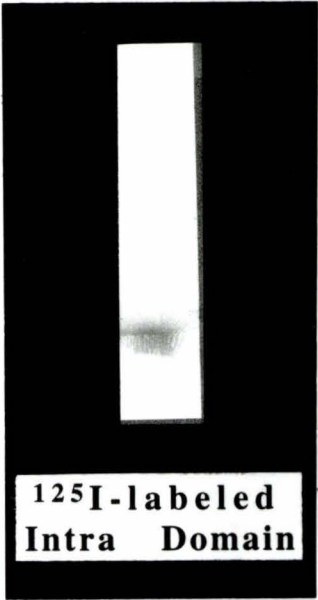
**Fig. 7.**        *Methods to identify putative ligands to the alternatively-spliced fifth fibronectin repeat and intracellular domains of Bravo/Nr-CAM.*

Upper Left: Autoradiograph showing the cleanly  $^{125}\text{I}$ -labeled intracellular domain used for screening an expression library. Of a 1:10 dilution, 1.0  $\mu\text{L}$  of labeled protein was run on 12% SDS PAGE, and the gel was autoradiographed for 1 hour. No contaminant protein label is evident, and the apparent molecular weight of the intracellular domain did not decrease nor broaden indicating that significant proteolysis did not take place during the freeze/thaw or labeling process. Numerous false positives but no double-filter positives were identified in the subsequent library screen.

Upper Right: affinity chromatography by non-covalently anchoring the expressed intracellular domain to the nickel-IDA-agarose solid support via its C-terminal six-histidine tail (left). E8 chicken brain lysate in low detergent (see Methods) was pre-absorbed by multiple runs through an IMAC column which nevertheless, did not remove all of the chelating protein in the lysate (right lane in the gel shows imidazole-eluted proteins that chelate nickel in the negative control). The experimental lane (left) shows the eluted intracellular domain excess, plus a number of protein bands all of which exactly correspond to bands found in the negative control (as assayed by 2D gel electrophoresis; data not shown). Therefore, no novel intracellular-domain specific interactions were retained in the affinity column.

Lower: Affinity chromatography Western blot showing putative ligands to the Fn5 domain. Using a column containing 1 mg recombinant fifth fibronectin type III repeat (Fn5) of Bravo/Nr-CAM. Biotinylated cell surface protein from E8 chicken brain lysate was run through the column and collected in the first two fractions. Ten

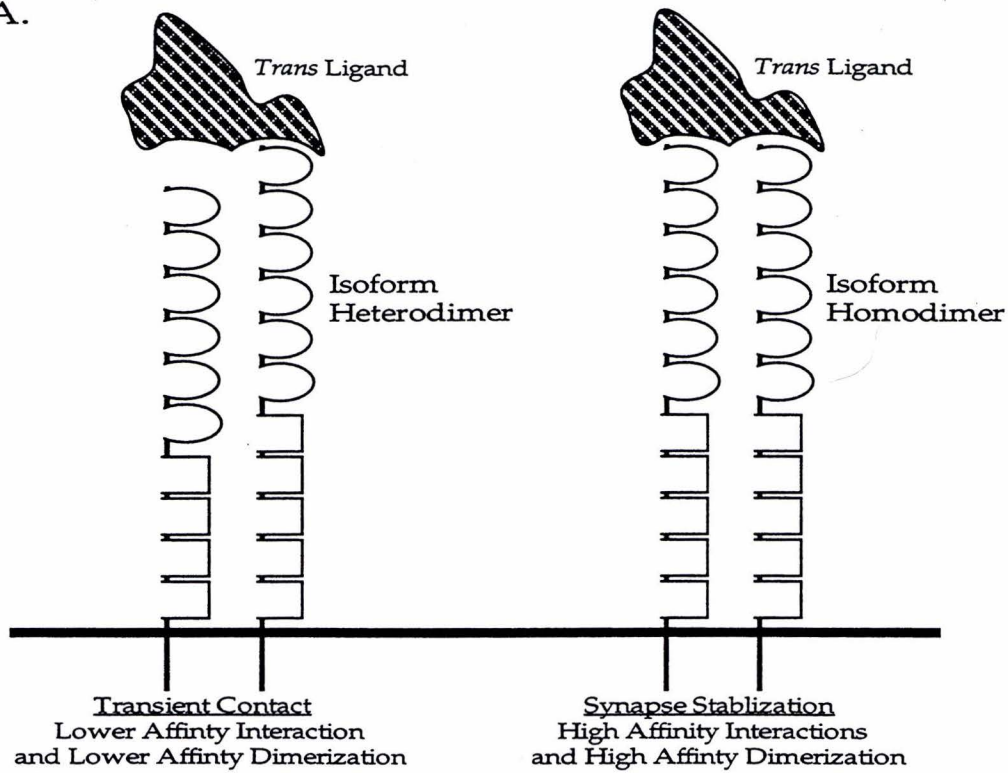
column volumes of running buffer was subsequently run through the column prior to specific competitive elution using one volume containing 1 mg free soluble Fn5. All fractions were analyzed by Western blots, staining with streptavidin alkaline phosphatase. A putative 70 kD ligand retained in the column is shown in competitively eluted fractions. This result is reproducible and is distinguished from negative controls (IgG affinity column, data not shown).



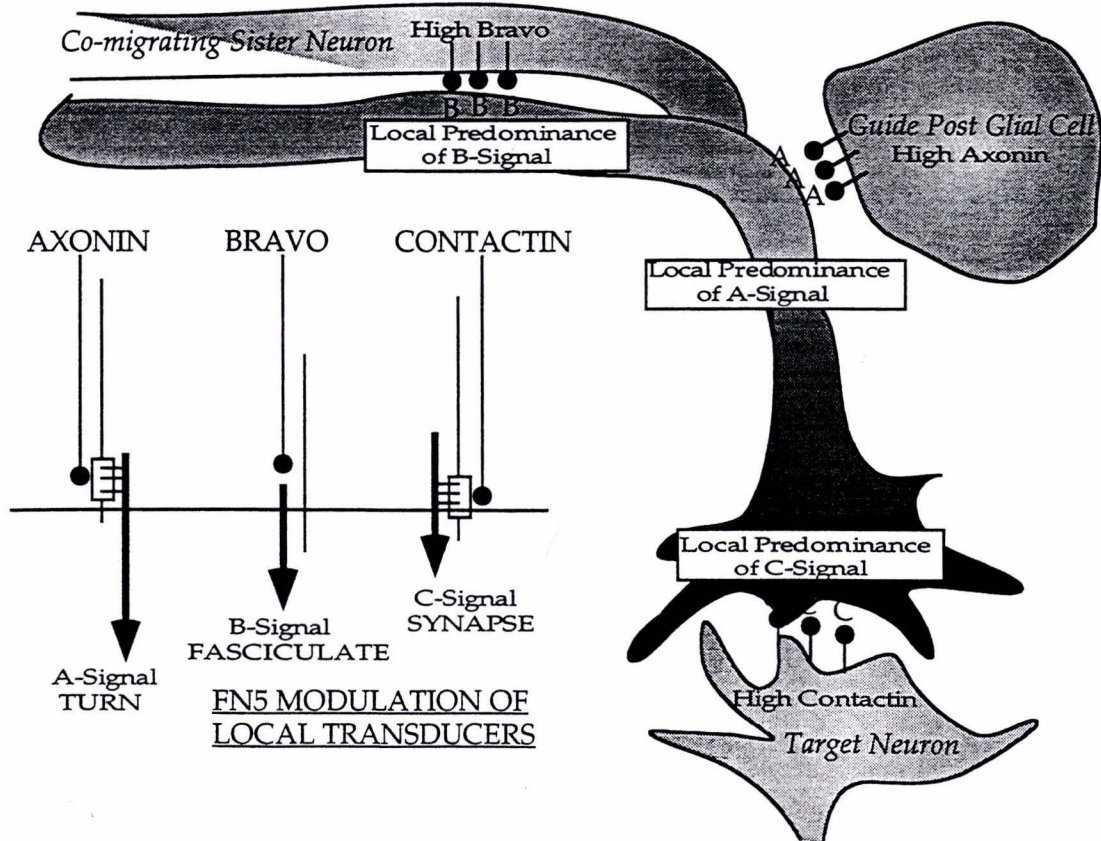
**Fig. 8. Functional Models.** Cartoons indicating two possible functional models of co-expressed Bravo/Nr-CAM isoforms. (A) The co-expressed Bravo isoforms may be regulated with respect to homodimer versus heterodimer *cis* interactions with functional consequence with respect to specific *trans* interactions. The model shown is one in which a transient ligand contact is reinforced by a shift from a low-affinity to high-affinity receptor complex, which might stabilize a synapse, for example. (B) The co-expressed Bravo isoforms may function in modulating different signal transduction consequences. In this model, the Fn5 domain shown to be capable of interacting with two different *cis* co-receptors that can generate two distinct signals, and these co-receptor interactions are dependent on specific *trans* ligand interactions. In the absence of the Fn5 domain, a third co-receptor interaction is possible, but this too is dependent on a specific *trans* interaction. In this scheme, the multiplicity of ligands (i.e., Bravo can interact with itself, Axonin-1 and Contactin) is integrated into discrete signals by virtue of specific isoforms. Ligands are described as being at high levels on various *trans* cells that contact the neurite, which causes local predominance of particular signal (and thus, in the example shown, local differences in neuronal behavior). However note that "high levels" might not be a concentration issue as much as an accessibility issue (i.e., by differential glycosylation or differential masking of functional epitopes by other cell surface proteins in the area, etc.).



A.



B.



Appendix I

Perspectives on the Evolution, Structure and Function of  
Fibronectin Domains  
(Written 1993)

...Fibronectin Domains Are People Too...



## Introduction

It is estimated that as much as 2% of all protein may contain a particular domain that was first identified on the extracellular matrix fibronectin molecule (Bork and Doolittle, 1992). This fibronectin domain module, approximately 100 amino acids in length, is found as a major component of many cell surface molecules so far characterized, including tenascin, perlecan, collagen, cytotactin, titan, and twitchin. In addition, fibronectin domains have been characterized on a wide scoping group of neural cell adhesion molecules, including NCAM, NgCAM/L1, Bravo/Nr-CAM, Neurofascin, Neuroglian, Contactin/F11, Tag1/Axonin1, Big1/Tag-like and Big2 proteins. And yet, as predominant as these domains are in the protein world, they have been rather treated like the "ugly sister" (or brother) to another domain-type found on many of these same proteins: the immunoglobulin (Ig) domain. At the risk of flirting with psychological issues, it seems timely for fibronectin domains to "come of age" which, like their more fashionable Ig domain partners, are beginning to emerge as important players in the cell recognition process.

### *General Structural Considerations*

Now that I have begun with the bold declaration that fibronectin (Fn) domains and Ig domains ought to have equal status in the extracellular fraternity, I should begin by saying that at the level of tertiary protein structure, these two domains are indeed essentially the same. Many more structures are known for Ig domains, including those found on IgG fragments (F<sub>ab</sub> and F<sub>c</sub>), T-cell co-receptors (CD8 and CD4), MHC class I molecules (HLA-A2), and the human growth hormone receptor (hGHR). Recently, structures have been reported for three fibronectin repeats: Leahy et al. (1992) solved the third fibronectin domain of tenascin, and Huber et al. (1993) solved the first and second fibronectin domain-pair of neuroglian. In each of these

structures, the basic topology is identical: seven antiparallel beta strands form two beta sheets (beta-barrel), with highly conserved hydrophobic residues maintaining this conserved core structure.

It is evident that although Fn and Ig-like structures share nearly identical folds, this similarity is almost certainly an example of convergent evolution: the highly conserved core residues for both are entirely distinct. Ig-like domains have conserved cysteines (disulfide bonds between cysteines are critical for proper Ig folds) and tryptophans; Fn-like domains have conserved tryptophans and tyrosines. The spacing of these core residues within the structure is also different between the two domains, further supporting the notion that they probably do not share a common evolutionary ancestor. It is perhaps interesting to note, however, that these core residues reside on the B, C and F strands in both domain types. Figure 1 shows the general Ig and Fn domain topology, the positioning of conserved core residues, as well as illustrating the two known variations of the fundamental beta barrel topology, with the tenascin Fn and hGHR Ig domain examples shown differing only in a C' strand switch.

#### *General Functional Considerations*

If it is true that Ig and Fn domains have converged upon a similar structure, then it must also be true that this general structural archetype is functionally advantageous. There are numerous functional studies of various kinds relating to molecules that contain these domains that reflect a unifying theme: Ig/Fn folds are utilized by molecules that are involved in intermolecular interactions and high-affinity ligand binding. Much of the functional information pertaining to Fn domains is from the study of the fibronectin molecule itself. Fibronectin is a cell adhesion protein which has soluble plasma isoforms made by hepatocytes and insoluble membrane forms on fibroblasts and other cells. Fibronectin has 14 fibronectin type III domains/repeats which can fold independently (Novokhatny et al.,



1992) and therefore may be thought of as modules on the full-length protein. In this regard, specific modules on the protein have been examined in various biochemical experiments, and while details of these studies are beyond the scope of this discussion, in general the results of these experiments demonstrate that specific ligand interactions map to specific modular regions of the protein. The C-terminal heparin-binding globule, for example, has an arginine-rich motif in Fn#13 whose positive charges are necessary in order to interact with the sulfate groups of heparin (Barkalow and Schwarzbauer, 1991). The central gelatin-binding region have fibronectin repeats that contain RGD sequences required to specifically interact with beta-1 integrin ligands in order to support cell adhesion to various substrates, including laminin (Piersbacher and Ruoslahti, 1984). Another 89 amino acid fibronectin type III domain (the alternatively spliced IIICS exon) contains minimal LDVPS/IDAPS/REDV motifs sufficient to support interactions with alpha4-beta1 integrins (Komoriya et al., 1991; Mould et al., 1991; Mould and Humphries, 1991). In each of these cases, the specific ligand interactions between fibronectin and other extracellular matrix proteins is supported by sequences within specific domain modules. It is perhaps interesting to note that in each of the examples given, these binding motifs contain numerous highly charged amino acids that may form strong electrostatic interactions *in vivo*.

While fewer biochemical details have been elucidated regarding other fibronectin-domain-containing proteins, it is clear that a diverse set of cellular functions have been mapped to specific Fn modules. Specific Fn domains have been shown to be critical for functions such as cell adhesion and attachment (Hayashi et al., 1992; Prieto et al., 1992; Lochter et al., 1991), mitogenic behaviors (Nagata et al., 1992), and muscle contraction (Labeit et al., 1992). From protein kinases to phosphatases, from receptors to matrix proteins, from prokaryotes to humans, fibronectin domains appear to be ubiquitous and functionally

diverse. One of the most wide-scoping use of fibronectin domains is among molecules involved in the assembly and post-developmental function of the nervous system, including NCAM-related Ig superfamily proteins. Many of these proteins have been implicated in various components of neuronal development, including cell adhesion, spreading, outgrowth, fasciculation and axon guidance (reviewed in Goodman and Schatz, 1993). The L1 protein, for example, has five fibronectin domains, and specific neurite outgrowth and cell adhesion functions have likewise been mapped to discrete domains (Appel et al., 1993). The NCAM protein has two fibronectin domains which appear to be required for the opposite effects of substrate adhesion and cell spreading (Frei et al., 1992); these contradictory functions may be reconciled *in vivo* by one domain supporting one function, the other domain supporting the other function, with the regulation between these two options possibly accomplished by selective masking of functional epitopes via glycosylation. It is also possible that these two domains may function synergistically, because it is also true that both domains together promote axon extension and outgrowth better than if only a single domain is present (Frei et al., 1992). And so on the one hand, individual domains demonstrate significant independence in terms of their folding and functions, on the other hand they are arrayed on the proteins that contain them and some have functions in which multiple modules are acting in concert.

### Structure Solutions for Fibronectin Type III Repeats

The first X-ray crystallography coordinates for a fibronectin domain were reported by Leahy et al. (1992). The solved structure was for the third fibronectin domain of the extracellular tenascin protein. Crystals with space group defined as  $P4_32_12$  were obtained. 91% of the reflections were measured out to  $1.8 \text{ \AA}$ , and 99% of the reflections were measured out to  $3.0 \text{ \AA}$ . Phasing was determined by multi-wavelength anomalous diffraction (MAD) and the figure of merit and



phasing power suggested that the data was unreliable beyond 3.8 Å. Electron density maps remained poor even after solvent flattening, however, maps produced from various resolution shells (4.0-3.0 Å) were used to trace the main chain. A final model was made and refined to an R-factor=0.196 using reflections out to 1.8 Å, which is very solid especially since only 74 waters were used. Geometry was also reasonable, with 0.012 Å rms deviation for bond lengths, 2.4 degree rms deviations for peptide bond angles, and allowable phi-psi angles throughout. It should be pointed out that an nmr structure was reported for another fibronectin domain just prior to the Leahy et al structure, which was in good agreement with this tenascin domain.

Recently, here at Caltech, our friend Andy Huber & Company (Huber et al., 1993) obtained high diffracting crystals for a tandem repeat of neuroglial fibronectin domain modules. Crystals with space group defined  $a=b=c=241.79$  Å (F432) were obtained with a single molecule per unit that diffracted up to 1.8 Å resolution in the native crystal and 2.4 Å in the heavy atom derivative. In the native crystal, 99% of reflections were measured up to 2.0 Å, and the  $R_{\text{sym}}=0.05$  shows that the intensity measurements were probably precise. Over 45000 unique reflections were measured and each unique measurement was 8-fold redundant suggesting that the  $R_{\text{sym}}$  figure is likely accurate. SIR data was collected with ethylmercuric phosphate, and 86% of the reflection data was utilized up to 2.6 Å; the  $R_{\text{sym}}=0.079$  likewise suggests reliability in the measurements. An electron density map (solvent flattened) was calculated with phase data and the mean figure of merit was low ( $<1$ ) as was the phasing power ( $<<< 2$  at most resolutions). The SIR phase data is therefore, somewhat unreliable making chain-tracing difficult, nevertheless, all 205 residues were traced with at least no advertised problems. The resulting model was refined with a final R-factor of 0.203 down to 2.1 Å with good geometry (0.0017 Å bond length rms deviation; 1.9 degree angle rms deviation), indicating a

solid correlation with the diffraction data and probably representing a very accurate set of coordinates.

### *Evolutionary Considerations*

From the structures discussed above, and from structures for very similar Ig-like domains, it is apparent that conserved residues are maintained for core structure and not for some common ligand-binding motif. Each of the conserved residues is hydrophobic and in the structure, extends its side chain between the strands and within the core between the two faces of the beta sheets. Figure 2 gives the structural sequences alignments, indicating the tightly conserved core, which includes a proline just before the A strand, and IxW motif in the middle of the B strand, a Y/F/W in the C strand, a M/Q/L in the EF loop, and a YxxxVxA motif in the F strand. When the three fibronectin domain structures are superimposed in TOM software, all of the best structural conservation is around these core residues. In the case of the hGHR:ligand co-structure, it is clear that the contact residues are distinct from this hydrophobic core (de Vos et al., 1992).

In contrast, there is significant divergence in both structure and sequence character outside this core, especially within the various loop structures. Sequence gaps are generally found in loops, while 1-2 amino acid beta bulges are commonly found especially towards the ends of strands. Charged amino acids (D/E/Y/K/R) make up almost 40% of the loop residues, and are especially concentrated in particular loops, such as the BC loop of neuroglian Fn#1. The now-famous integrin-binding RGD loop of tenascin Fn#3 is in the FG loop. Although this represents the only known convergence of biochemical/functional data with structural data for Fn domains and it is therefore tempting to assume that most if not all ligand interactions may map to these loop regions, it is also possible that the outside faces of the beta sheets may themselves present important functional epitopes. The outside facing residues of the F strands, for example,



indicate similar divergent characteristics among the three solved domains:

Ng Fn#1	xNxTxRxIxFx
Ng Fn#2	xKxLxKxVxIx
Tn Fn#3	xExExSxIxRx

These alignments illustrate the diverse character of structurally "homologous" residues, including examples of charge differences (K vs E), polarity differences (F vs R), and side chain length differences (R vs S). Receptor-ligand interactions among Ig domains are generally supported by beta-sheet face-to-face contact residues, and therefore it is possible that the outfacing diversity observed in fibronectin strand structures may be hinting that similar kinds of interactions are mediated here as well.

#### Structural Modeling of Bravo/Nr-CAM Fn#5

The fifth fibronectin domain (most membrane-near domain) of Bravo/Nr-CAM is alternatively spliced and may impart to those isoforms that contain the domain a particular modulated function. In order to glean further functional insights, the structure of this domain was modeled using TOM package software. Clearly, the most critical step in modeling is the determination of structurally "homologous" residues of solved structures (in this case, the tenascin and neuroglian coordinates). Overall, the Fn5 domain has only 24% identity with Ng-Fn#1, 19% identity with Ng-Fn#2, and 26% identity with Tn-Fn#3, making this determination difficult to say the least. Typically, alignments were accomplished by locating conserved core residues and identifying strands by alternating hydrophobic residues. One to two amino acid beta bulges were accepted in strands, but generally, differences in sequence length were assigned primarily to loop regions. The Blossum62 similarity matrix was used to otherwise make best estimates of the "homologous" residues in corresponding structures.



The Bravo/Nr-CAM domain aligned better with the Ng-Fn#1 domain than the Ng-Fn#2 domain. However, the long poly-proline II helix and abridged G-strand of Ng-Fn#1 aligned poorly to Bravo/Nr-CAM. Also considering that the 9-residue/3-turn helix of Ng-Fn#1 is rare (Adzhubei and Sternberg, 1993), it seems that at least this C-terminal region is better modeled against the Ng-Fn#2 domain. For this reason, the C-terminal FG loop and G strand structures were not modeled against the Ng-Fn#1 domain, rather the second domain was used as a blueprint for these specific structural elements only. The resulting model was refined several times with the ".FIT" algorithm in order to energy minimize side chains. The final model superimposed over the Ng-Fn#1 structure with an overall rms=0.1 Å. This model is shown in Figure 3.

Because Bravo/Nr-CAM's Fn5 domain is only 93 amino acids, it aligns with fewer gaps to the Tn-Fn#3 structure, and a second model was constructed using these coordinates, which superimposes with an overall rms=0.9 Å. In this model, no polyproline II helix was included, and the sequence alignment in the entire FG loop and G-strand is strong (Fig 2). It is interesting to speculate about this C-terminal region between the F-strand and the end of the domain. The F-strand is the most conserved of all strands/loops in the structure because it contains most of the structurally required core residues. The FG loop to follow appears to be highly diverged structurally, which may contain the unusual polyproline II helix of various lengths or an important ligand-binding RGD motif for example. And the final G-strand is dramatically different in length for the 3 solved structures, and like the loop that precedes it, represents a structurally diverse region of the domain. It is perhaps noteworthy that all four domains (including Bravo/Nr-CAM Fn#5) have a serine residue conserved either at the beginning of a polyproline II helix or G-strand. So, which of these two options is the Bravo/Nr-CAM serine preceding? The overall sequence similarity is closer to the Tn-Fn#3 domain in this

region, implying the absence of the helix and rather an extended G-strand. Bravo/Nr-CAM also lacks prolines in this region, further suggesting the absence of the helix. However, like neuroglian, Bravo/Nr-CAM does not have a hairpin FG loop, and contains sequence that has striking similarity to the helix-spanning region of Ng-Fn#2 (VAAEE versus RSSED). In this regard, I think the question remains open as to whether this region of Bravo/Nr-CAM actually looks more like tenascin or neuroglian.

*How useful are these sequence-derived models?*

Clearly, models derived from sequences are only as good as the alignments themselves, and even at that, they are mere fantasy until the structure is actually solved. When the two separately-determined models (versus Ng-Fn#1 and Tn-Fn#3) were superimposed, they do so with an rms=3.9 Å. This comparably poor alignment suggests significant disagreement in the placement of the carbon-alphas, which may indicate that one or both efforts contain error beyond rescue. Having said this, both models look very similar with respect to general characteristics. In both models, the side chains for the core residues are appropriately buried in the core structure, and even in loops, hydrophilic side chains are energy-minimized to face outward from the core structure. The ABE and CC'FG outward-facing sheets, for example, have numerous long-side chain, charged amino acids (D/E/K, etc.). The bottom AB/CC'/EF loops in the model have energy-minimized side chains that include polycharged KED and RSSED motifs extending outward towards the solvent environment. The top BC/C'E loops in the model form an electron-dense region, including SRS and DEY side chains that likewise extend outwards. And so while these models may not represent anything close to reality with regard to specific coordinates, it is likely that they do illustrate the character of structural landscapes with their striking side-chain features that may indeed, provide strong hints to functionally important residues.



### Closing Perspectives

Bravo/Nr-CAM and other molecules that contain fibronectin domains are generally complex proteins, often a mosaic of multiple domain modules. The twitchin molecule of *C. elegans* for example, has 31 fibronectin domains in tandem. The neuroglial tandem structures described here, provide hints as to how these multiple domains may be oriented to each other. These fibronectin domains are separated by a stretch of amino acids that align the domains in a 175 degree orientation that, if this relationship were maintained, would generate a corkscrew-like zig-zagging of adjacent domains resembling a long, twisting rod (Huber et al., 1993). If this is true, these multiple domains may rigidly align orthogonal to the cell membrane and therefore each domain might remain accessible to its own unique set of ligands. On the other hand, if Ig domain structural conclusions are transferable, one might expect that flexible hinges might intervene between pairs or groups of fibronectin domains so that the twisting rod might kink or bend. If this is true, this bending might permit groups of domains to interact with common ligands over a much larger interface. A specific tenascin-contactin interaction, for example, has been shown to be sensitive to the spacing of particular fibronectin domains (Zisch et al., 1992), which may indicate the importance of multiple domains making multiple ligand contacts over a discrete and well-defined space. Finally, alas, there is another scenerio that at this point can not be dismissed: multiple fibronectin domains may rarely bind ligands at all, and are more structural in nature, merely permitting the more N-terminal Ig domains to extend far from the cell membrane. And if so, fibronectin domains will remain the "ugly sister" (or brother) of Ig domains, once and for all stuck in obscurity, forever maligned in the lackluster realm of liver and spinach, wrinkled movie stars, and the insufferable abyss of the housekeeping genes.

Appendix III

References

Adzhubei, A.A. and Sternberg, M.J.E. (1993). Left-handed polyproline-II helices commonly occur in globular proteins. *J. Mol. Biol.* 229, 472-493.

Appel, F., Holm, J., Conscience, J.F., and Schachner, M. (1993). Several extracellular domains of the neural cell adhesion molecule L1 are involved in neurite outgrowth and cell body adhesion. *J. Neurosci.* 13, 4764-4775.

Barkalow, F.J.B., and Schwarzbauer, J.E. (1991). Localization of the major heparin-binding site in fibronectin. *J. Biol. Chem.* 266 (12), 7812-7818.

Bork, P., and Doolittle, R.F. (1992). Proposed acquisition of an animal protein domain by bacteria. *Proc. Natl. Acad. Sci. USA* 89, 8990-8994.

de Vos, A.M., Ultsch, M., and Kossiakoff, A.A. (1992). Human growth-hormone and extracellular domain of its receptor - crystal structure of the complex. *Science* 255, 306.

Frei, T., Halbach, F.V., Wille, W., and Schachner, M. (1992). Different extracellular domains of the neural cell adhesion molecule (N-CAM) are involved in different functions. *J. Cell. Biol.* 118 (1), 177-194.

Goodman, C.S., and Schatz, C.J. (1993). Developmental mechanisms that generate precise patterns of neuronal connectivity. *Cell* 72 (Suppl), 77-98.

Hayashi, K., Madri, J.A., and Yurchenco, P.D. (1992). Endothelial cells interact with the core protein of basement membrane perlecan through

beta-1 and beta-3 integrins - an adhesion modulated by glycosaminoglycan. *J. Cell. Biol.* 119 (4), 945-959.

Huber, A.H., Wong, Y.M.E., Bieber, A.J., and Bjorkman, P.J. (1993). Crystal structure of tandem type III fibronectin domains from *Drosophila* Neuroglian at 2.0 Angstrom. *Neuron* 12 (4), 717-731.

Komoriya, A., Green, L.J., Mervic, M., Yamada, S.S., Yamada, K.M., and Humphries, M.J. (1991). The minimal essential sequence for a major cell type specific adhesion site (CS1) within the alternatively spliced type III connecting segment domain of fibronectin is leucine-aspartic acid-valine. *J. Biol. Chem.* 266 (23), 15075-15079.

Labeit, S., Gautel, M., Lakey, A. and Trinick, J. (1992). Towards a molecular understanding of titin. *EMBO J.* 11 (5), 1711-1716.

Leahy, D.J., Hendrickson, W.A., Aukhil, I., and Erikson, H.P. (1992). Structure of a fibronectin type III domain from tenascin phased by MAD analysis of the selenomethionyl protein. *Science* 258 (2084), 987-991.

Lochter, A., Vaughan, L., Kaplony, A., Prochiantz, A., Schachner, M., and Faissner, A. (1991). J1/tenascin in substrate-bound and soluble form displays contrary effects on neurite outgrowth. *J. Cell. Biol.* 113,(5), 1159-1171.

Mould, A.P., Komariya, A., Yamada, K.M., and Humphries, M.J. (1991). The CS5 peptide is a 2nd site in the IIICS region of fibronectin recognized by the integrin alpha-4-beta-1 - inhibition of alpha-4-beta-1 function by RGD peptide homologs. *J. Biol. Chem.* 266 (6), 3579-3585.



Mould, A.P., and Humphries, M.J. (1991). Identification of a novel recognition sequence for the integrin alpha-4-beta-1 in the COOH-terminal heparin-binding domain of fibronectin. *EMBO J.* 10 (13), 4089-4095.

Nagata, S., et al. (1991). *Prog. Growth Factor Res.* 3 (2), 131-141.

Novokhatny, V., Schwarz, F., Atha, D., and Ingham, K. (1992). Domain-structure and domain domain interactions in the carboxy-terminal heparin-binding region of fibronectin. *J. Mol. Biol.* 227 (4), 1182-1191.

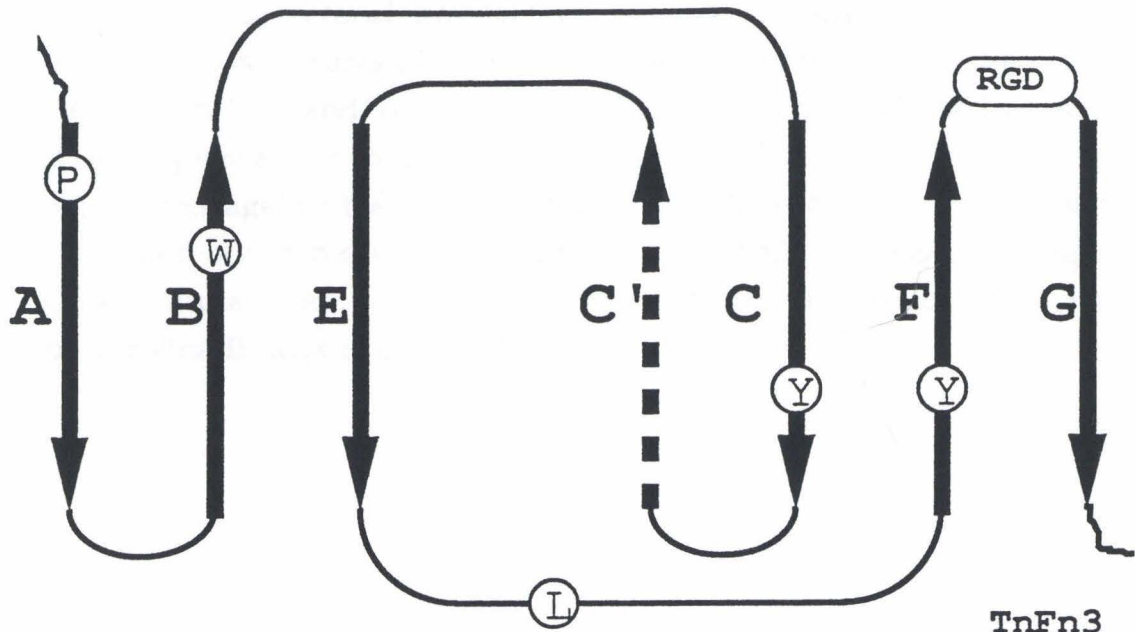
Pierschbacher, M.D., and Ruoslahti, E. (1984). Cell attachment activity of fibronectin can be duplicated by small synthetic fragments of the molecule. *Nature* 309, 30-33.

Prieto, A.L., Anderssonfiscone, C., and Crossin, K.L. (1992). Characterization of multiple adhesive and counteradhesive domains in the extracellular matrix protein cytactin. *J. Cell Biol.* 119 (31), 663-678.

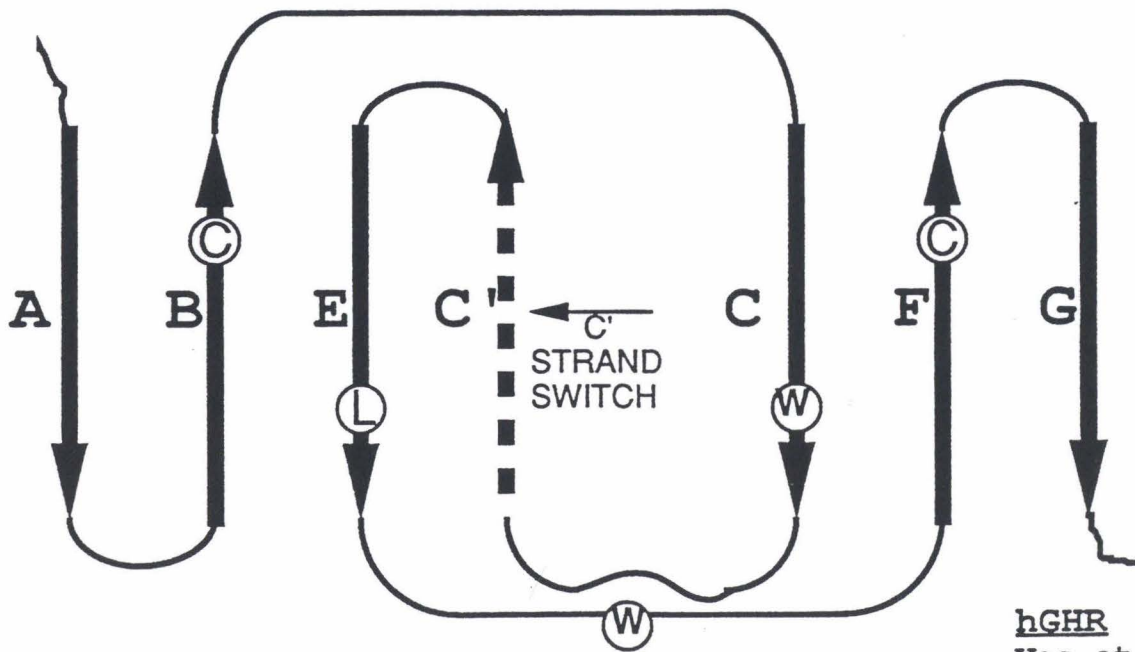
Zisch, A.H., Dalessandri, L., Ranscht, B., Falchetto, R., Winterhalter, K.H., and Vaughan, L. (1992). Neuronal cell adhesion molecule contactin/F11 binds to tenascin via its Immunoglobulin-like domains. *Journal of Cell Biology* 119, 203-213.

**Fig. 1.**      *Ribbon diagrams of two known variations of the fibronectin/ Ig-like beta-barrel folds.*    A C' strand-switch defines this topological distinction, and two representative examples are shown: tenascin Fn3 (Leahy et al., 1992) and hGHR (de Vos et al., 1992).





TnFn3  
Leahy et al



hGHR  
Vos et al

**Fig. 2.**        *Structural alignments.* Structural alignments for the three solved structures of fibronectin type III repeats: Neuroglial Fn1, Neuroglial Fn2, and Tenascin Fn3. The Bravo Fn5 is aligned against these sequences for modeling purposes. The Bravo domain seems to align better against the Neuroglial Fn1 domain, except for the FG loop - G strand which more resembles Neuroglial Fn2. Strands A-through-G are underlined; conserved core residues are bold; the rare polyproline II helix is italicized.

BrFn5	VQP LYPRIDNVT T A	AAETYANISWEYEG	PDH	ANFYVEYGV
NgFn1	IVQDVP NAPKLTGIT C Q	AD KAEIHWEQOQ	DNRSPI	LHYTIOFNT
NgFn2	PDVPFKNPDDNVVGO GTE P N	NLVISWTPMPEIE	HNAPNEHYVYSWKR	
TnFn3	RL DAPSQIEVKDV T DT	TALITWEKPL A EI	DGIELTYGI	
BrFn5	VQP LYP RIRNVTT AAADT	YANISWEYEG P DH	ANFYVEYGV	

C

A

B

BrFn5	AGSKEDWKKELVNG	SRSEFVLKGLTPGTAYKVRVGAEG
NgFn1	SFTPASWDAAAYEKVP	NTDSSEVVO MSPWANYTFRVIAFN
NgFn2	DIPAAAWEEN N NIFDWRQNNIVIADQ	PTFVKYLIKVVAIN
TnFn3	KDVPGDRTT I DLT	EDENOYSIGNLKPDTEYEVSLISRR
BrFn5	AGSKEDWKKELVNG	SRS FFVLKGLTPGTAYKVRVGAEG

C'


E

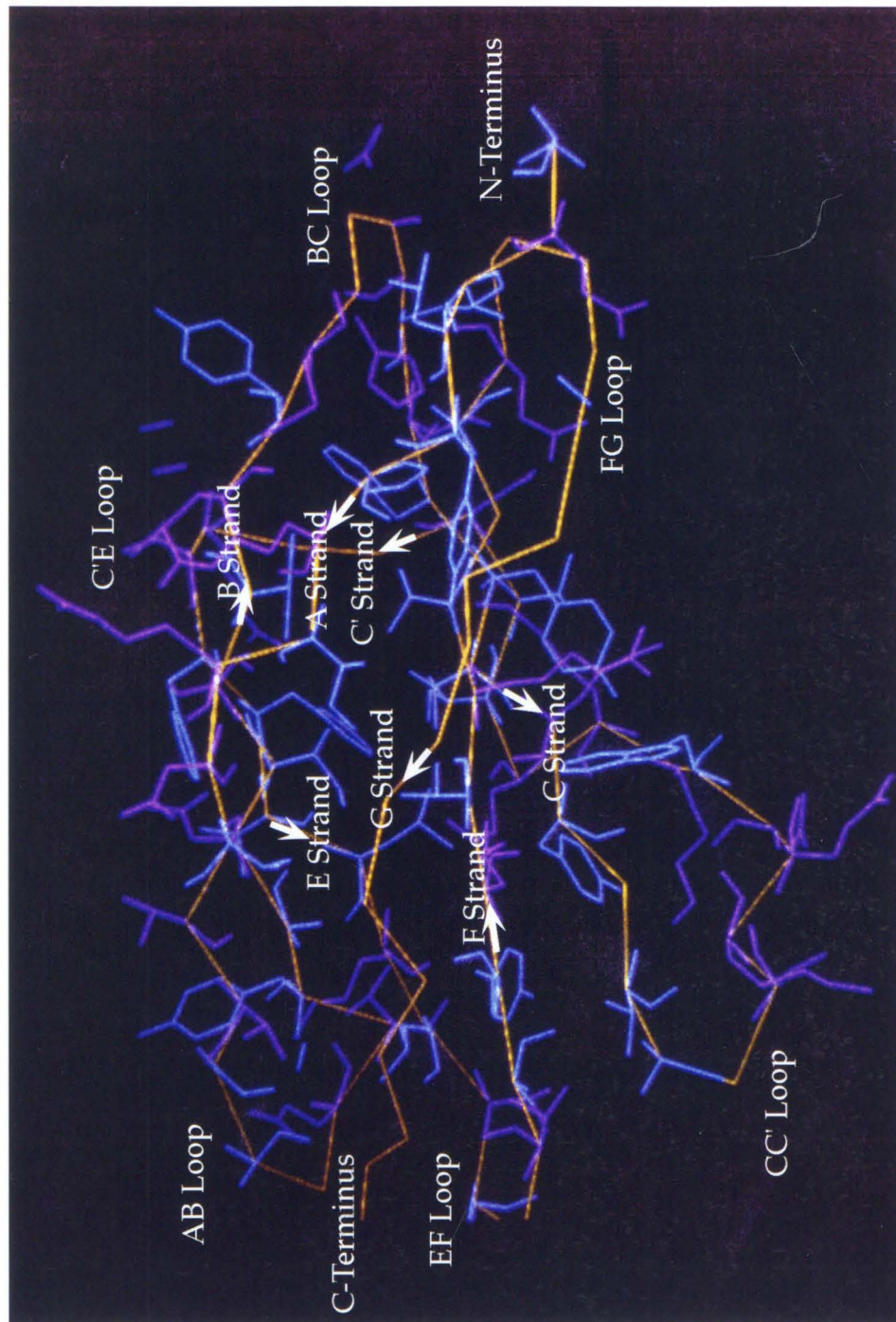
F

BrFn5	LSGLSFRSSE DLFETGP
<b>NgFn2</b>	DRGESNVAAE EVVGYSGEDR
NgFn1	KIGAS PPSA HSDSCTTQ
TnFn3	GDMSS NPA KETFTT
BrFn5	LSGDRSS EDL FETGP

G

**Fig. 3.** *TOM-generated structural model of the Bravo Fn5 domain (print of an SGI snapshot).* The domain is modeled against the Neuroglian Fn1, except for the C-terminal portion (FG-G strand), which is modeled against Neuroglian Fn2 (because it is unlikely that the Bravo domain contains a polyproline II helix). The model was refined several times and the ".FIT" algorithm was used to energy-minimize side chains. Loops and strands are indicated.







Appendix IIDo DNA Rearrangements Play a Role in Development?

DNA changes definitely happen. Genome content is not constant, as was initially postulated based on nuclear transplantation experiments of Spemann and others. The question is not whether or not these genomic rearrangements occur, rather it seems the more appropriate question is, to what extent? Moreover, a more interesting question to me is, do these rearrangements have any universal developmental significance? In the plant and animal kingdoms, where it has been studied these DNA changes seem widespread but sporadic: ciliates splice out DNA sequences in a developmentally regulated fashion (Prescott, 1992), maize transposable element jumping is developmentally programmed (Gierl *et al.*, 1989), *Ascaris* chromosome diminution is a significant developmental mechanism in asymmetric cell division (Muller *et al.*, 1991), yeast mating type switching involves regulated genomic interconversion events (Klar, 1990), cyanobacteria likewise exhibit programmed genomic changes when forming the nitrogen-fixing heterocyst (Haselkorn, 1992), and the most well studied example of all, V(D)J recombination results in cell-specific genome differences among antibody-producing cells (initially proposed in Dreyer and Bennett, 1965). These genetic switches, therefore, range from bacteria to ciliates, plants to yeast, nematodes to humans, but, do they represent random and unrelated outcroppings of a highly unusual phenomenon, or are they homologous and do they hint at universal importance?

In the above examples, in some cases the answer seems to be that these DNA rearrangement phenomena are analogous as opposed to homologous: functionally similar but not evolutionarily-related genomic events. On the one hand, gene conversion and homologous recombination drive DNA change events (yeast mating type switches),

while at the opposite extreme, site-specific recombination drives a quite different splicing mechanism (bacteria, phage and the vertebrate immune system). Because the mechanisms that underlie these rearrangement events appear in these specific cases to be non-homologous, it would be grossly incorrect to extrapolate and assume that similar events are ubiquitous. Yet, the fact that these specific organisms have evolved analogous programmed (as opposed to incidental or random) DNA splicing mechanisms suggest that the result (i.e., genetic switching) provides a powerful and important advantage.

In general, it is not difficult to fathom the advantages of such rearrangement phenomena as independently arrived-at solutions for programmed cell differentiation or phenotypic events, and these advantages might include efficient use of genetic material, generation of sequence diversity, control over genotype repertoire, and a potent means for cell commitment and heritability (reviewed in Plasterk, 1992). Most argue, however, that despite these advantages, similar events are not likely to be a component of general development for a primarily two reasons: 1) it may not be necessary (integration of transcriptional, post-translational, and epigenetic information may be sufficient code for cell type determination), and 2) it is risky (incorrect switching may be fatal to the individual, and furthermore, devastating to the species if the genome is not absolutely protected).

Nevertheless, a distinction can be made which might change this perspective. In the immune system, DNA rearrangements are involved in generating diversity and permit the organism to produce novel gene products in order to better combat antigens that it has never before been exposed to in its evolutionary history. This kind of DNA change is indeed rather haphazard and in my estimation, unlikely to be utilized in other vertebrate developmental systems that require highly ordered cell specificity and recognition. The kind of DNA change that seems a more likely candidate is one in which the change is used to



"mark" the DNA: programmed and systematic excision, inversion or recombination of sequences within promoter regions might underlie differentiated cell commitment events. Furthermore, a mechanism of this type would also permit the observed robust nature of cell heritability (after all, DNA is the vehicle of inheritance). The distinction I am making is one based on outcome (i.e., combining coding gene elements versus editing non-coding control regions), not mechanism. It is entirely possible within this framework that immune system-like recombination mechanisms might have directly evolved from this proposed more general use of the machinery in promoter regulation and/or cell fate control.

In this regard, it is worth noting that the site-specific mechanism of V(D)J recombination is ancient and widespread (Dreyfus, 1992). In some cases, it is homologous (i.e., involves homologous proteins and recombination signal sequences) to the site-specific recombination events in lambda phage. Recently, it was also reported that ciliates share homologous components in some of its DNA editing machinery (Doak *et al.*, 1994). A by-product of both immune system and phage recombination is the splicing out of an intervening ring of DNA, and interestingly, changes in DNA ring populations have been correlated with differentiation events in specific cell types (see for example, Yamagish *et al.*, 1983). Transposable elements and other repeat sequences monopolize most genomes, and while unlike phage and retroviruses these elements do not encode the ability to escape the cell, they do encode integrases which permit excision and movement within the genome. Might specific transposable elements be involved with differentiation and development? Most again argue that these sequences are "junk" or "selfish", and if they serve any function to the host it is at the level of species and population and not the individual; namely, they are thought to provide a means to "shuffle" the genome at times of increased stress in the environment, perhaps by moving

promoter sequences to novel places in the genome in order to expand the regime of specific transcription factors.

Nevertheless, for "junk" there appears in some cases, to be astonishing levels of developmental specificity built into the system. Dali Ding, a former graduate student in the Lipshitz lab here at Caltech, recently published the expression of specific *Drosophila* transposable elements (Ding and Lipshitz, 1994). Each and every transposon that he tested was expressed in a developmental pattern that, if this were an enhancer trap experiment, would generate great excitement and enthusiasm (for examples, see Fig. 1). One of the questions I decided to ask experimentally was: What exactly is being expressed? Transposable elements encode an entire set of genes, including reverse transcriptase, integrase, and envelope genes. Might only the integrase be expressed in these striking patterns? Using *Drosophila* libraries of different tissues and developmental stages, and generating probes of only those transposable elements that exhibited intriguing expression profiles in more than a single fly strain (Gypsy and 412 to begin with), several cDNA clones were isolated and preliminarily sequenced. In every case examined, it appears that the entire full-length element is expressed and there is no specific gene product preferentially transcribed. While these results provide no further insight or clue, the question as to what exactly these transposable elements might be doing in these discrete embryonic territories remains open and provocative.

The second set of experiments along this line, were inspired by sequence data that was reported for the *Drosophila* "Suppressor-of-Hairless" gene (Furukawa et al., 1991). In development, the null mutation of this gene results in an intriguing patterned phenotype in bristle formation. Specifically, the developmental process seems to stall at a precursor fate as if the stereotyped cell divisions were decoupled from differentiation events resulting in the normal number of cells (4), all of which adopt precursor phenotypes; i.e., the differentiation component in these normally asymmetric cell divisions



failed to occur (Posakony, 1994). The cloning and sequencing of this gene showed a remarkable level of conservation to an integrase motif found in the mouse RBP-J<sub>k</sub> immune system protein. Here at last, was a developmental phenotype and an integrase protein united somewhere outside the immune system. I was curious as to how widespread this presumed integrase might be (as well as its phenotype in other genetic organisms), and so using both the mouse and fly sequences, a probe was designed in an attempt to clone the homolog in *C. elegans*. While a Southern blot indicates that a homologous sequence may exist in worms (Fig. 2), efforts to clone the gene by both PCR and library screening failed to identify the worm RBP gene (a single cDNA was isolated that, although strongly positive in the screen, shares no homology with any known Genbank sequences). About the time I was ready to resign anyway, the Posakony lab published its paper declaring that the integrase motif of the fly RBP gene did not, after all, function as a recombinase (Schweisguth *et al.*, 1994).

As a stubborn warrior, where might one go from here? My instincts lead me to consideration of specific genomic regulatory regions, especially within gene clusters. Gene clustering and maintenance of gene order relationships over evolutionary time is observed in at least two families of genes: the Hox clusters critical for establishing developmental identities, and the olfactory receptor clusters which are regulated with exquisite resolution at the level of single cells. In the former case, the array of Hox genes also exhibits principles of colinearity: the temporal (and spatial) order of expression is correlated with 3'-to-5' position in the cluster. At least two intriguing models that relate gene grouping and order, to transcriptional regulation, are conceivable. The commonly proposed model for Hox clusters, for example, is described as a "latching relay" or autoregulatory model: genes share intervening *cis* regulatory sequences in order to generate specific patterns of overlapping expression. An alternative model is a mechanistic model whereby the

DNA is "marked" sequentially down the chromosome: specific mother cells may be born in a temporal order, each with a progressively more 5' "marking" of promoter regions, perhaps specified by site-specific changes to intervening sequences within the cluster. This non-stochastic mechanism of sequential activation might guarantee representation of the entire gene family, explicitly define a specific gene or gene set per cell (i.e., differential cell-specific phenotypes), and provide an autonomous mechanism for mother cell receptor "memory" and heritability. Not coincidentally, my immediate future scientific plans are to study the evolutionary and regulatory features of the olfactory clusters in collaboration with Lee Hood and Richard Axel, an adventure that can begin just as soon as you sign that form.

Appendix IV  
References

Ding, D. and Lipshitz, H.D. (1994). Spatially regulated expression of retrovirus-like transposons during *Drosophila* embryogenesis. *Genetical Research (Cambridge)* 64: 167-181.

Dreyer, W.J., and Bennett, J.C. (1965). The molecular basis of antibody formation: a paradox. *Proc. Natl. Acad. Sci. USA* 54 (3), 864-869.

Dreyfus, D.H. (1992). Evidence suggesting an evolutionary relationship between transposable elements and immune system recombination sequences. *Molec. Immunol.* 29(6): 807-810.

Doak, T.G., Doerder, F.P., Jahn, C.L., and Herrick, G. (1994). A proposed superfamily of transposase genes: transposon-like elements in ciliated protozoa and a common "D35E" motif. *Proc. Natl. Acad. Sci. USA* 91, 942-946.

Furukawa, T., Kawaichi, M., Matsunami, N., Ryo, H., Nishida, Y., and Honjo, T. (1991). The *Drosophila* RBP-J<sub>k</sub> gene encodes the binding protein for the immunoglobulin J<sub>k</sub> recombination signal sequence. *Journal of Biological Chemistry* 266 (34), 23334-23340.

Gierl, A., Saldler, H., Peterson, P.A. (1989). Maize transposable elements. *Annu. Rev. Genet.* 23,, 71-85.

Haselkorn, R. (1992). Developmentally regulated gene rearrangements in prokaryotes. *Annu. Rev. Genet.* 26, 113-130.

Klar, A.J.S. (1990). Regulation of fission yeast mating-type interconversion by chromosome imprinting. *Development (S)*, 3.

Muller, F., Wicky, C., Spicher, A., and Tobler, H. (1991). New telomere formation after developmentally regulated chromosomal breakage during the process of chromatin diminution in *Ascaris lumbricoides*. *Cell* 97, 815-822.

Posakony, J.W. (1994). Nature versus nurture: asymmetric cell division in *Drosophila* bristle development. *Cell* 76, 415-418.

Plaskterk, R.H.A. (1992). Genetic switches: mechanism and function. *Trends in Genetics* 8 (12), 403-406.

Prescott, D.M. (1992). Cutting, splicing, reordering and elimination of DNA sequences in hypotrichous ciliates. *BioEssays* 14 (5), 317-324.

Schweisguth, F., Nero, P., and Posakony, J.W. (1994). The sequence similarity of the *Drosophila* Suppressor of Hairless protein to the integrase domain has no functional significance *in vivo*. *Developmental Biology* 166, 812-814.

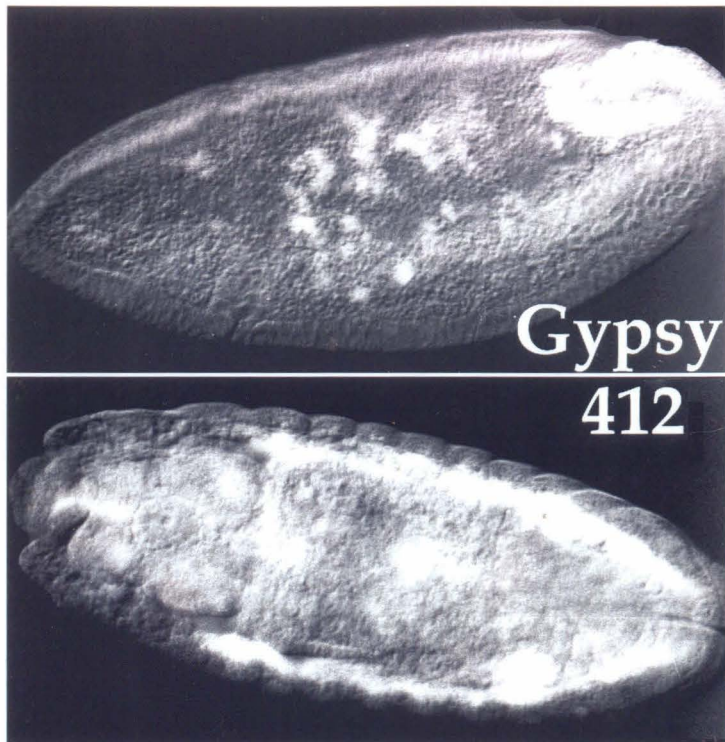


**Fig. 1. (Top) *Expression of transposable elements in Drosophila development.***

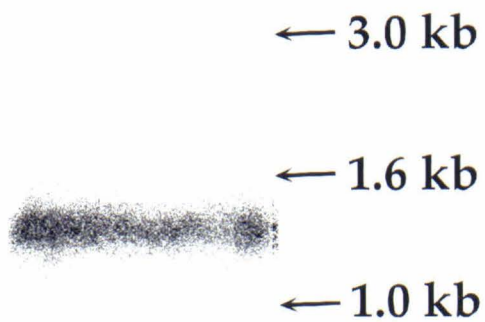
Spatial expression of Gypsy transposable element at germ band retraction. Transcripts first accumulate in a small set of outer-ventral cell in the stomodeal invagination as well as the yolk nuclei. The expression of the 412 transposable element is quite different, and includes segmentally-repeated clusters in the mesoderm on either side of the ventral midline. These patches gradually extend toward the ventral midline and eventually fuse. In addition, the 412 element is heavily expressed in the gonad. These figures are reproduced from Ding and Lipshitz, 1994.

**Fig. 2. (Bottom) *RBP Southern blot.***

Southern blot of *C. elegans* EcoRI-digested genomic DNA and probed with a mixed *Drosophila* and mouse RBP-J<sub>k</sub> PCR product that spans the highly conserved integrase motif. A single dominant band (approximately 1.2 kb) was visualized after medium stringency conditions (30°C, 35% formamide); with longer exposure times, several very faint higher molecular weight bands in the 4-6 kb range are evident on the original film. This same probe was used to screen a cDNA library and it hybridized strongly and reproducibly to a clone that shares no sequence homology with any other translated or untranslated sequences in the Genbank database.



## Southern Blot



**C. elegans Genomic DNA  
Probed with RBP Sequences**