SENEX: CLOS IN MOLECULAR PATHOLOGY. UNCERTAINTY, GENERALIZATION AND COMPARISON OF OBJECTS.

Sheldon S. Ball
Dept. of Pathology
University of Mississippi
2500 North State Street
Jackson, MS 39216
ssb@fiona.umsmed.edu
(601) 984-1553 (phone), 984-2885 (fax)

Vei H. Mah
Dept. of Neurology
Thomas Jefferson University
130 South 9th Street, Suite 400
Philadelphia, PA 19107
(215) 955-8355 (phone), 923-6680 (fax)

Abstract

SENEX is an evolving set of computer tools for molecular pathology research, featuring display of molecular data through computed graphical presentations, and the actual customization of a computer language to fit the domain of molecular pathology. Flexibility in retrieval of molecular data and the capacity to reason with molecular information necessitates careful consideration of how an object is compared with object specifications. Domain-specific considerations include contextual information, the concepts of generalization and uncertainty, multiple layers of structure within molecules, and different logic for comparing lists of objects. Specialized methods of generic functions are used to compare or match object specifications with CLOS objects. Presentation of uncertainty in molecular data is a particularly difficult problem. The Common Lisp Object System (CLOS) in connection with the CLOS metaobject protocol provides a powerful and portable programming environment for representation and reasoning with information in molecular pathology.

Introduction

Molecular pathology deals with molecular aspects of disease. It is a discipline characterized by structures of variable complexity, events constrained by a variable number of factors, and incompletely understood phenomena. Representational issues inherent in the domain are complicated by the use of a language with a rigid/inflexible design. Nonetheless, much of our knowledge in this domain can be captured in terms of operations upon objects subject to specified constraints. The CLOS metaobject protocol (Kiczales et al, 1991; Bobrow et al, 1993) further enhances representational flexibility by allowing a programmer to adjust the design and implementation of the language to fit an application domain. Perhaps most importantly, the structuring of information sets the stage for reasoning with information, allowing a user to make suppositions or relax constraints in order to use the program for making novel predictions. These predictions may then serve as a basis for planning laboratory experiments.

SENEX allows a scientist/physician/student to ask a sequence of questions regarding the molecular bases of disease in a single interactive session. Feedback communication from SENEX is largely in the form of line drawings and images, with a minimum of text. SENEX communication is facilitated through use of Common Lisp Interface Manager (CLIM) presentations. Selection of a presentation with a mouse gesture reveals further details of the object represented by that presentation. This provides a means of allowing a user to explore many different aspects of a particular molecular structure or phenomenon. In some situations, SENEX uses the history of presentations selected by the user in the decision of what data to present.

SENEX also currently provides tools for assistance in data entry using terminology common to the domain. The new information is incorporated and processed in accord with what SENEX currently knows about the domain. For example, when specifying a motif for a reaction product (see below for description of motifs), the motif specifications are compared with defaults for the molecule, and the specified motif replaces the appropriate default or fits among the defaults accordingly. For polypeptides, polynucleotides and molecular complexes, SENEX computes graphical presentations from symbolic representations, providing a useful form of visual feedback during data entry.

Flexibility in retrieval of molecular data and the capacity to reason with molecular information necessitates careful consideration of how an object is compared with object specifications. Domain-specific considerations include contextual information, the concepts of generalization and uncertainty, multiple layers of structure within molecules, and different logic for comparing lists of objects. This paper presents an overview of SENEX and discusses issues related to representation of uncertainty, generalizations, and comparison of CLOS objects.

System Design

SENEX is written in the portable programming environment of Common Lisp, the Common Lisp Object System (CLOS), and the Common Lisp Interface Manager (CLIM) (Ball & Mah, 1992; 1993). SENEX has grown in size and complexity since its conception, and adapted to new developments in program design as well as in molecular pathology. These adaptations are not complete, nor will they ever be lest new developments in the disciplines of program design and molecular pathology cease.

Three major developments in programming design have shaped SENEX since its conception: 1) The ANSI standardization of CLOS; 2) Introduction of the CLOS metaobject protocol; 3) Development of CLIM. Developments in molecular pathology which have shaped the design of SENEX are too numerous to mention. However several domain associated concepts have provided significant influence on system design: 1) Exploration of molecular information in a graphical context; 2) Visualizing molecular information at increasing levels of detail; 3) Pursuit of tangential issues; 4) Molecular predictions for the purpose of designing laboratory experiments.

The SENEX Classification Structure

An object in SENEX is an instance of an object class (Keene 1989). For example, p60-src is a type or class of protein and there are many specific instances of p60-src in SENEX. To date, there are 7408 object classes and 15,438 unique objects in SENEX.

Figure 1 shows that part of the SENEX classification structure containing the src family of proteins, of which p60-src is a subclass. All objects in SENEX are instances of classes embedded in this sort of finely granular classification structure. The basis of the SENEX classification structure is the Medical Subject Headings (MeSH) Tree Structures, a classification of terms used to index articles in the National Library of Medicine's MEDLINE database. Thus the SENEX classification structure is a biological classification structure. However, there are significant differences between the SENEX classification structure and MeSH, and a mapping between the two is implemented. There is a mapping of synonyms to the canonical forms used as class names, and a word completion facility for class names.

The SENEX classification structure serves several functions, including:

1. Structure for representation of biological information.
2. Inheritance of properties common to similarly classified objects.
3. Customized operations through methods defined on specific classes.
4. Representation of qualitative uncertainty & generalization.

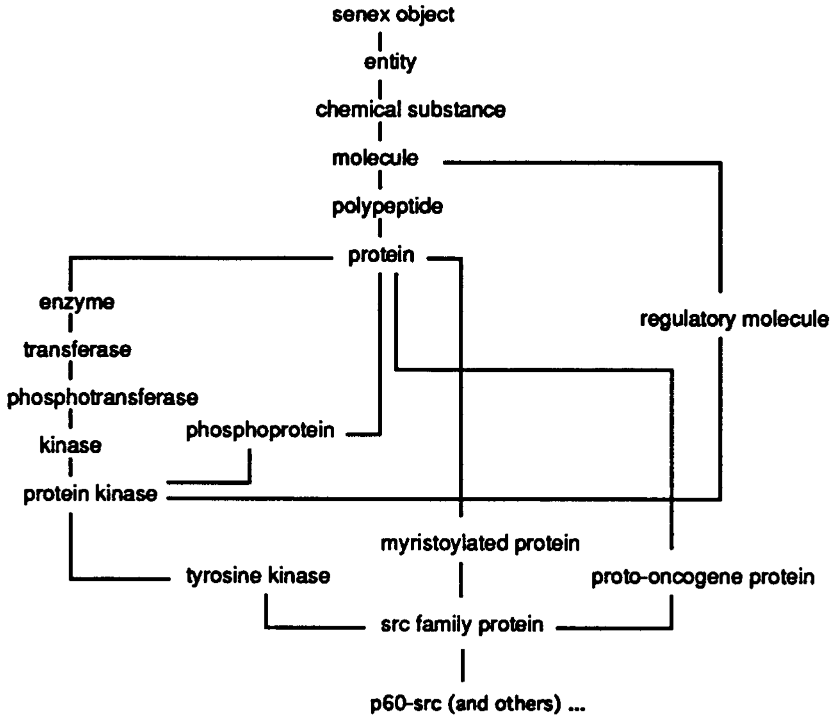


Figure 1. SENEX classification structure containing the class src family protein.

5. Entry points for specification of queries.

Queries

User interactions with SENEX are through facilities for exploratory browsing or specific queries. Queries are processed in essentially three steps: 1) translation of the query, 2) matching of objects with search specifications, and 3) presentation of retrieved objects for further inspection. Query translation provides flexibility in how a user might formulate questions. The queries are translated into object specifications which are used in SENEX searches.

The algorithm for matching objects is complicated by the expression of generalizations and uncertainty, the need for comparing different types of objects in different contexts, and optional search specifications. The matching or comparison of objects is facilitated by generic dispatch, i.e. a generic function match-objects with specialized methods for comparing different types of objects. Very briefly, the algorithm includes:

1) Decomposing target and test objects and recursively matching component objects.
2) Distinguishing logical "OR" from logical "AND" and "ORDERED SET" slots
3) Distinguishing generalization from qualitative uncertainty.
4) Identifying and matching appropriate substructures within objects that contain multiple structural layers.
5) Facilitating optional search specifications.

Comparison of Slot Values with Different Logic

Slots are specialized descriptors of objects defined with the most generalized class to have that attribute or property. For example, molecule has slots compartment, and size; molecular complex has an additional slot subunits. Slots may assume default values specified with class definitions and most slot values are themselves objects or lists of objects. Slots and their default values are inherited through the classification structure. The inherited slot default values may be accepted as is or further specialized.

Slot values provide a means of programming biological knowledge into SENEX. Slots are defined in SENEX as logical "AND" or logical "OR" slots. For example, the slot "compartment" is a logical "OR" slot, that is a molecule may be located in the cytoplasm or the nucleus, but not both at the same time. In contrast, the slot "motif" is a logical "AND" slot. Proteins contain structural elements known as motifs which give rise to the function(s) of the molecule. Most proteins consisting of a single polypeptide can be represented as an ordered set of motifs connected by peptide regions (figure 2). Comparison of motif slot values can be specified as a comparison of "ORDERED SETS", depending upon whether or not the ordering of motifs is to be considered. Slot logic directs differential comparison of slot-values consisting of lists of objects.

Uncertainty and Generalization

Uncertainty and generalization are important representational concepts. Uncertainty will always be a part of scientific data, and generalization is a form of scientific organization that reflects an understanding of the domain. Uncertainty and generalization frequently share the same terminology and are distinguished by context.

Representation of qualitative uncertainty is facilitated by CLOS. For example, uncertainty in the identification of a particular protein can be represented by using a sufficiently general class to describe the protein. In addition, relationships among biologic entities are often understood at varying levels of detail. For example, sometimes it is known the binding of a receptor with its ligand is coupled to cellular events, but the details of that coupling are poorly understood. Nonetheless, this is useful information that carries with it a degree of uncertainty. SENEX uses specialized classes to represent this type of uncertainty (figure 3).

A general class is interpreted by SENEX to reflect uncertainty, unless otherwise specified. This becomes an issue any time an object is specified as an instance of a class for which there are more specific subclasses. Accordingly, assertions may be generalized to all subtypes of a class, allowing for exceptions to the generalization. Exceptions are specified in the same manner as other SENEX objects. Generality is distinguished from uncertainty during comparison of objects.

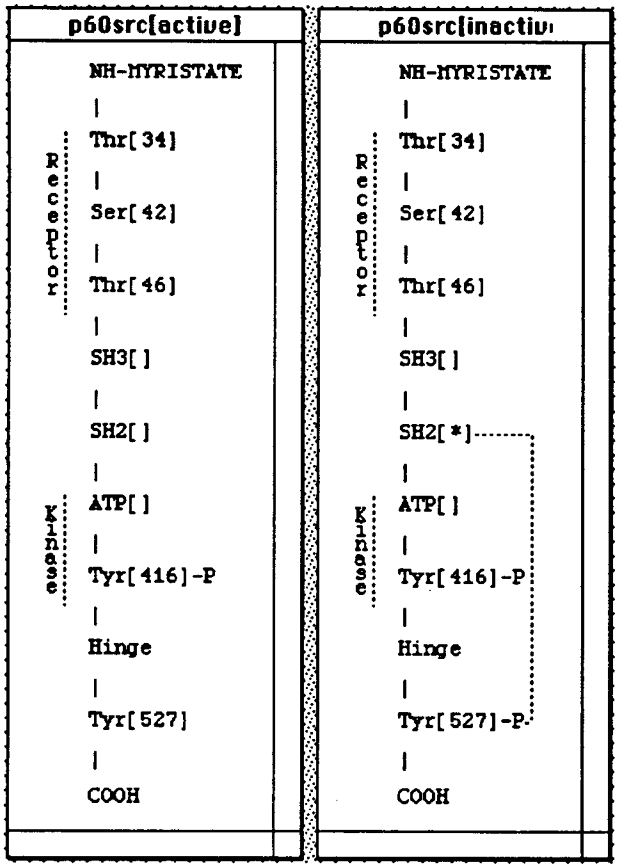


Figure 2. Line drawings of active and inactive forms of p60-src. The line drawings of src indicate that the proteins consist of an ordered set of motifs connected by peptide regions. Dashed lines to the left of the motifs represent domains containing the motifs. The more N-terminal (upper) domain is a binding site for binding of cell surface receptor containing three phosphorylation sites, Thr 34, Ser 42 and Thr 46. The more C-terminal domain is the kinase domain containing an ATP binding site and a tyrosine autophosphorylation site. In the inactive form of the enzyme, an intramolecular bond is seen involving phosphorylated tyrosine-527 in the C-terminal region with an SH2 domain N-terminal to the kinase domain. This intramolecular bond serves to mask the active site of the enzyme. Src activation occurs via phosphatase activity at tyrosine-527 or through displacement of the SH2-tyrosine phosphate bond by another tyrosine phosphate present in the cell surface receptor that interacts with src through the N-terminal region. These line drawings are CLIM presentations computed from the symbolic representations of the molecules.

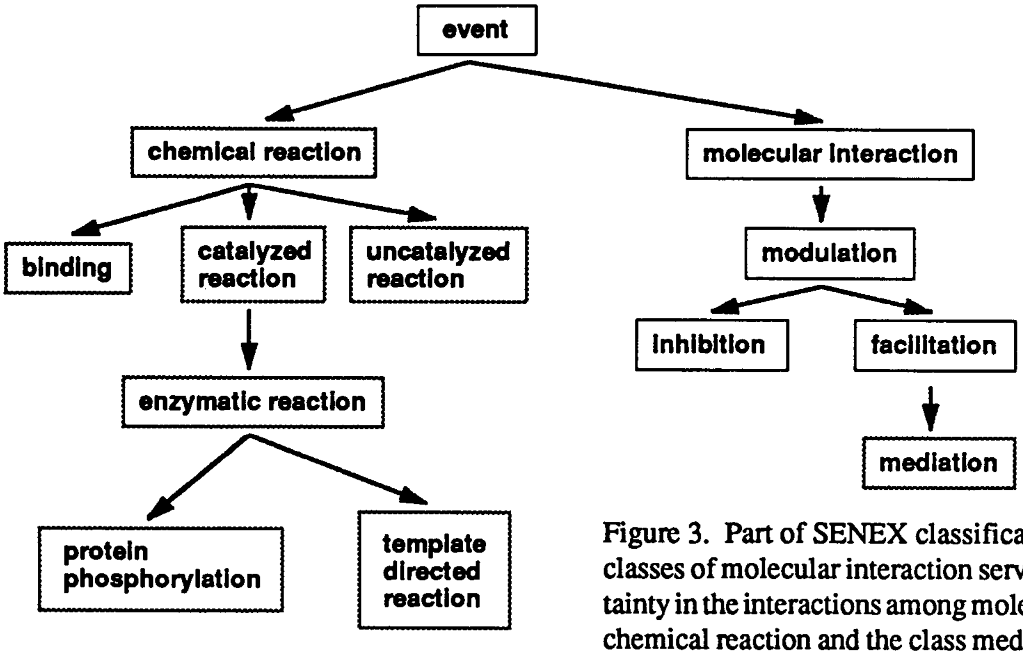


Figure 3. Part of SENEX classification of events. Subclasses of molecular interaction serve to represent uncertainty in the interactions among molecules. Subclasses of chemical reaction and the class mediation serve to represent events used in molecular pathway searches.

Quantitative uncertainty is simplistically represented through use of mean/standard deviation or range. Greater than or less than expressions of uncertainty are specializations of range. Thus objects of uncertainty types mean/standard deviation or range can be compared with a number. A quantitative value may be assigned to the certainty that a specific event may or may not occur, allowing all degrees of probability. Users can specify the degree of desired certainty in SENEX queries.

Other forms of generalization include lists of objects as values of logical "OR" slots and generalization through inheritance. Inheritance of slot values is specialized depending upon the slot. For example, motifs are inherited from all supertypes, i.e. inheritance by UNION as distinguished from inheritance by SHADOWING, the default method of CLOS inheritance. Motifs in addition to those motifs inherited from class supertypes may be specified as slot default values with the definition for the polypeptide or polynucleotide class. The inheritance of motifs by union necessitates processing of motif defaults to eliminate duplicates and identify specializations of more generalized motifs. In addition, certain rules for ordering motifs within a molecule can be employed, and details regarding the structure of motifs may be dependent upon the state of a molecule. Access to CLOS meta-objects (Kiczales et al, 1991; Bobrow et al, 1993) in connection with methods defined on specific classes of molecules provides a means of programming biological knowledge into SENEX through customized inheritance of motifs.

Substructure

Elements of substructure are shown in the line drawings of p60-src, a protein frequently abnormal in some kinds of cancer cells (figure 2). Dashed lines to the left of the motifs represent domains containing the motifs. The more N-terminal (upper) domain is a binding site for binding of cell surface receptor containing three phosphorylation sites, Thr 34, Ser 42 and Thr 46. The more C-terminal domain is the kinase domain containing an ATP binding site and a tyrosine autophosphorylation site.

Consider the case of comparing object specifications of an enzyme containing a tyrosine phosphorylation site and an N-terminal myristate with p60-src. Identifying the N-terminal myristate in src is no problem. It is a "top level" motif. However the tyrosine phosphorylation site of src is embedded in the kinase domain of the enzyme. Thus the matching algorithm needs to look for substructure within motifs for which this is allowed. Matching of substructures is facilitated by specific methods of a generic function "match-substructures".

Optional Specifications & Reasoning

Different search options facilitate prediction of novel molecular pathways. For example, we can tell SENEX to ignore cell type, anatomic considerations, and organism type in a query, so that we might piece together reactions known to occur, but to occur in different cell types, or in different organisms, in the same reaction pathway. It is through queries of this type that SENEX may be used to predict novel molecular pathways, in essence *generating hypotheses which may be tested in the laboratory.*

In the search for molecular pathways, SENEX first searches through known pathways to see if any match the query, then looks to see if a new pathway might be identified. Searches may be run in forward or reverse direction. This feature is useful for examining partial pathways which failed to reach their targets.

The algorithm for linking events in molecular pathways is summarized as follows:

```
Match products of chemical reaction with:
    substrate of other chemical reaction
        (except template-directed reaction)
    enzyme of enzymatic reaction
    template of template-directed reaction
    substrate, enzyme, or template of cause
    of mediation[1]

If match, put onto queue:
    product of chemical reaction
    product of effect of mediation

Check to see if molecule is already on queue.

Check to see if queue extension is reasonable.

If target of search is reached, save queue &
    continue.
If queue extension fails, remove molecule
    from queue.
Print queue.
```

Checking to see if the queue extension is reasonable is a step necessitated by generalizations. For example, consider the following partial pathway: A -> B* -> C. If A is a cell or compartment-specific molecule and B* is a molecule less specific in its cellular or compartmental location (i.e. the value of the slot compartment or cell is generalized or consists of a list of objects), the cell and compartment-specific constraints should be the same for A -> B* -> C as would be for A -> C. Thus it is necessary to bind slot values which assume generalized values with queue extension for evaluating whether or not further queue extensions are reasonable (i.e. do not violate cellular or compartmental constraints).

Miscellaneous Considerations

A decision was made in building SENEX to terminate further extensions of the classification structure beyond one class of protein for one gene. Thus polymorphic proteins, isoforms resulting from alternative splicing of a single gene transcript, and proteins with protein precursors are represented as instances of a class of protein with multiple forms rather than as separate classes. This adds still another dimension for representation of uncertainty and generalization and necessitates specializing the matching of objects for certain classes of proteins where different isoforms may have different functionality.

Presentation of Uncertainty in Molecular Data / Application to Alzheimer's Disease

Alzheimer's disease (AD) is a neurodegenerative disorder characterized by a progressive decline in memory, judgment, ability to reason, and intellectual function. The cause of AD is unknown and there is no cure available. More than 100,000 victims die annually of complications of AD making it the fourth leading cause of death in adults after heart disease, cancer, and stroke [Alzheimer's disease Fact Sheet, Alzheimer's Association, 1990]. Most victims are over 65; however, in rare cases and in Down's syndrome, the disease can strike those in their 40s and 50s. About 10% of people over the age of 65 are

[1] Mediation is a class of molecular interactions representing the observation that one molecular event causes another. Slots cause and effect have values which are molecular events.

afflicted with AD. This rises to 50% in those 85 and older.

Alzheimer's disease was described in 1907 by the German physician Alois Alzheimer, who found, at autopsy, abnormalities in the brain of a 51-year old patient with dementia. The abnormalities Alzheimer observed, senile plaques and neurofibrillary tangles, are the pathologic hallmarks of the disease. The senile plaques consist of extracellular deposits of amyloid surrounded by infiltrating glia and degenerating neurites. Neurofibrillary tangles are insoluble aggregates of paired helical filaments derived largely from tau, a microtubule-associated protein.

Degenerative diseases of the central nervous system generally affect specific populations of cells. Such selective neuronal vulnerability is a product of cell-specific expression of genes and interactions of gene products in the maintenance of cellular homeostasis and differentiation, all in the context of response to external stimuli.

Cells communicate with their environment in part through interaction of extracellular molecules with receptors on the cell surface. Interaction of cell surface receptors with their ligands in turn induces intracellular events (referred to as signal transduction) which lead to changes in intracellular enzyme activities and changes in gene expression. Figure 4 depicts signal transduction through the protein products of the src family and ras subfamily of proto-

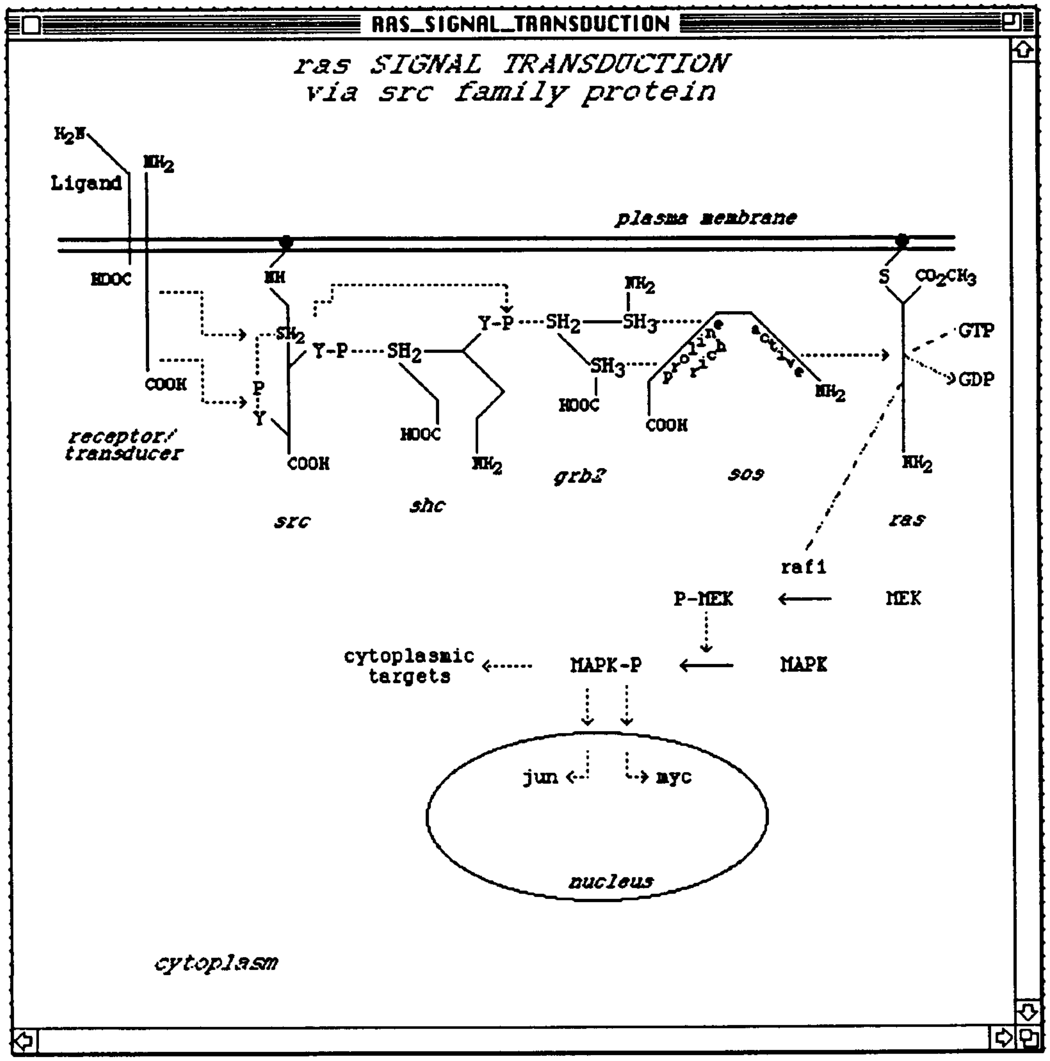


Figure 4. Signal transduction through src family and ras subfamily of proto-oncogene proteins.

oncogenes. Mutations in the src family and ras subfamily proteins are associated with a diversity of human malignancies. Figure 4 is a CLIM presentation of molecules presented in context of the signaling pathway. The features of the molecules presented serve to illustrate the interactions among these molecules in the pathway. More detail about specific molecules (i.e. src) is revealed with a mouse selection (figure 2).

Figure 4 presents both generalizations and uncertainty. For example, src family members are thought to transduce signals from cell surface receptor/transducer complexes to ras subfamily members. However, the receptor/transducers remain largely unidentified as are many of the events downstream of ras. For example, the Mitogen-Activated Protein kinases (MAP kinases or MAPK) are a subfamily of proline-directed kinases, members of the family of serine/threonine kinases. Several have been identified, one of which is p42MAP kinase. Among the proteins phosphorylated by p42MAP kinase is tau (Goedert, 1993). Phosphorylated tau is the major component of paired helical filaments (PHF) and neurofibrillary tangles, abnormal morphologic structures found in brains of patients with Alzheimer's disease. *In vitro*, p42MAP kinase phosphorylates tau on 12-16 serine/threonine residues, resulting in a form of tau identical to that found in PHF.

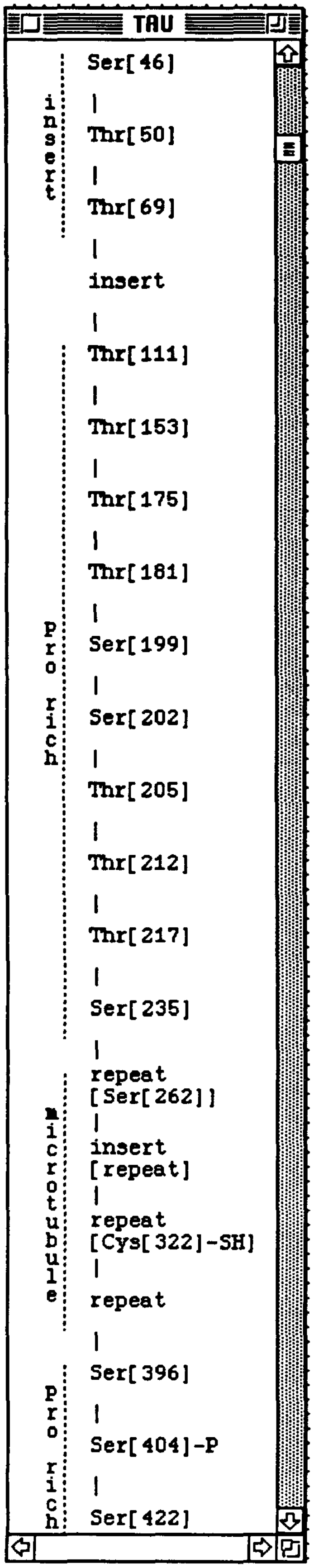
Tau confers polarity to microtubules and plays a major role in the differentiation of neurites into axons (Goedert, 1993). Human tau exists in one of 6 isoforms derived from alternative splicing of a single gene transcript (figure 5). Cys[322] and Cys[291] (embedded in the repeat of insert 3) within the microtubule binding domain of tau may play a significant role in the pathology of Alzheimer's disease (Trojanowski & Lee, 1994). It has been suggested that tau of paired helical filaments (PHF tau) consists of phosphorylated tau homodimers, oriented in an antiparallel fashion and covalently linked through disulfide bond(s) involving cysteine residues in the microtubule-binding domains of the tau polypeptides. Figure 6 shows the structure that SENEX computes for PHF tau from the symbolic description of the molecular complex.

Uncertainty and generalization characterize tau phosphorylation by MAP kinase. There is uncertainty associated with which MAP kinases in addition to p42MAP kinase phosphorylate tau and whether p42MAP kinase actually phosphorylates tau *in vivo*. There is generality in the observation that p42MAP kinase phosphorylates all of the tau isoforms. There is uncertainty in the number and the identity of the serines and threonines phosphorylated in PHF tau. In fact, phosphorylation in PHF tau may vary from molecule to molecule.

A great deal is known about the molecular pathology of Alzheimer's disease, much of which can be represented by interactions among CLOS objects. However, that which is unknown inevitably vastly exceeds that which is known, and uncertainty characterizes much of the data currently in hand. One of the representational challenges of SENEX is defining what information is desired for a complete, accurate and detailed description of the phenomena involved in disease processes, and how collection and organization of incomplete or partial information can function to place pieces of the puzzle in their proper perspective. Thus the design of SENEX has been oriented towards integrating data obtained from disparate sources by representation of partial information with associated uncertainty and remodeling that information as new data becomes available.

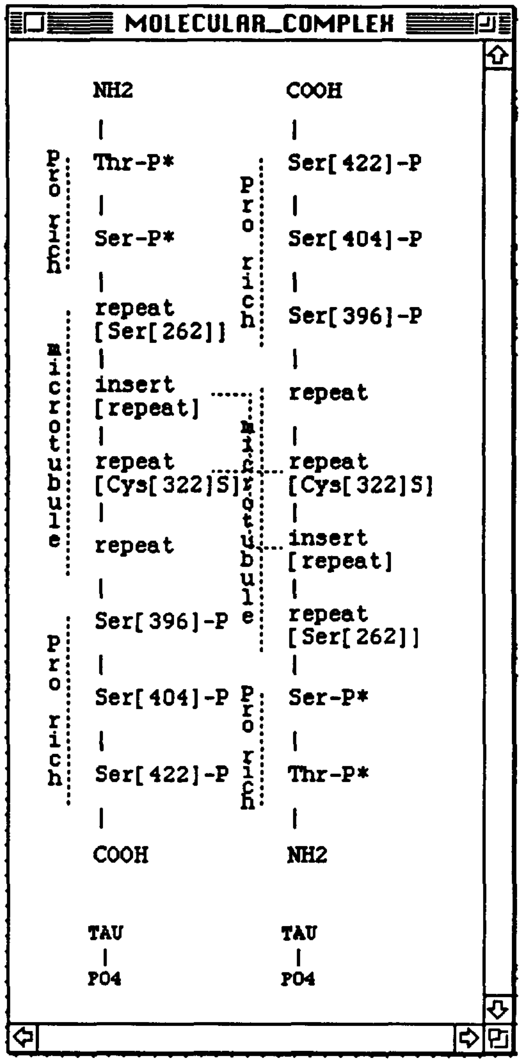
Summary

Flexibility in retrieval of molecular data and the capacity to reason with molecular information necessitates careful consideration of how an object is compared with object specifications. Domain-specific considerations include contextual information, the concepts of generalization and uncertainty, multiple layers of structure within molecules, and different logic for comparing slot-values consisting of lists. Specialized methods of generic functions are used to compare or match object specifications with CLOS objects. Generalizations necessitate special considerations, particularly in sequential operations simulating reasoning, such as identifying molecular pathways. Presentation of uncertainty in molecular data is a particularly difficult problem.



```
Ser[46]
|
Thr[50]
|
Thr[69]
|
insert
|
Thr[111]
|
Thr[153]
|
Thr[175]
|
Thr[181]
|
Ser[199]
|
Ser[202]
|
Thr[205]
|
Thr[212]
|
Thr[217]
|
Ser[235]
|
repeat
[Ser[262]]
|
insert
[repeat]
|
repeat
[Cys[322]-SH]
|
repeat
|
Ser[396]
|
Ser[404]-P
|
Ser[422]
```

Figure 5. The longest isoform of tau contains 3 inserts, 2 in the N-terminal region, and one in the microtubule-binding domain. The NH2 and COOH termini have scrolled off to the top & bottom of the screen, respectively. When a presentation in SENEX is too large to fit on the screen, the window size for the presentation is maximized, and the presentation coerced to scroll in the window. Tau has either 3 or 4 repeats in the microtubule-binding domain depending on the presence or absence of the 3rd insert (Goedert, 1993). C-terminal to the microtubule-binding domain is a proline-rich region containing 3 serine phosphorylation sites. N-terminal to the microtubule-binding domain is a proline-rich region containing several serine and threonine phosphorylation sites. Ser[404] appears to be constitutively phosphorylated in normal adult tau. Phosphorylation of Ser[396] by MAP kinase is suggested to diminish the affinity of tau for microtubules (Trojanowski & Lee, 1994). Cys[322] & Cys[291] (embedded in the repeat of insert 3) may play a significant role in the pathology of Alzheimer's disease.



```
NH2            COOH
|              |
Thr-P*         Ser[422]-P
|              |
Ser-P*         Ser[404]-P
|              |
repeat         Ser[396]-P
[Ser[262]]
|              |
insert         repeat
[repeat]
|              |
repeat         repeat
[Cys[322]S]    [Cys[322]S]
|              |
repeat         insert
               [repeat]
|              |
Ser[396]-P     repeat
               [Ser[262]]
|              |
Ser[404]-P     Ser-P*
|              |
Ser[422]-P     Thr-P*
|              |
COOH           NH2

TAU            TAU
|              |
PO4            PO4
```

Figure 6. PHF tau consists of a homodimer of highly phosphorylated tau polypetides linked by disulfide bond(s) in the microtubule binding domain. The tau polypeptides are oriented antiparallel in PHF tau. The * indicates several phosphorylation sites. Exactly which sites are phosphorylated is uncertain. PHF-tau with insert 3 within the microtubule binding domain provides 2 cysteines on each polypeptide for intermolecular disulfides. Two different structures with 2 intermolecular disulfides are possible. The dashed lines connecting motifs on the individual subunits represent intermolecular bonds. Cys[291] is embedded within the repeat of insert 3 and is not shown at this level of detail.

References

Ball SS. Mah VH. CLOS/CLIM in Biomedical Education. In: Proceedings of the 3rd Annual Lisp Users and Vendors Conference, August 9-13, Cambridge, MA, 1993

Ball SS. Mah VH. Symbolic Representation in Molecular Pathology. In: Proceedings of the 2nd Annual Lisp Users and Vendors Conference, August 10-14, San Diego. CA, 1992

Bobrow DG. Gabriel RP. White JL. CLOS in perspective. In: Object-Oriented Programming: The CLOS Perspective. A Paepcke (ed) MIT Press, Cambridge MA, 1993

Goedert M. Tau protein and the neurofibrillary pathology of Alzheimer's disease. Trends in the Neurosciences 16:460-65, 1993

Hyman BT. Elvage TE. Reiter J. Extracellular signal regulated kinases. Localization of protein and mRNA in the human hippocampal formation in Alzheimer's disease. Am J Pathol 144:565-572 1994

Keene SE. Object Oriented Programming in COMMON LISP, Addison Wesley, Reading, MA, 1989

Kiczales G. des Riveres J. Bobrow DG. The Art of the Metaobject Protocol, MIT Press, Cambridge, MA, 1991

Trojanowski JQ. Lee VM-Y. Paired helical fliament tau in Alzheimer's disease. The kinase connection. Am J Pathol 144:449-53, 1994