SENEX: A CLOS/CLIM APPLICATION FOR MOLECULAR PATHOLOGY

Sheldon S. Ball and Vei H. Mah

University of Mississippi and Thomas Jefferson University
2500 North State Street and 130 South 9th Street, Suite 400
Jackson, MS 39216 and Philadelphia, PA 19107

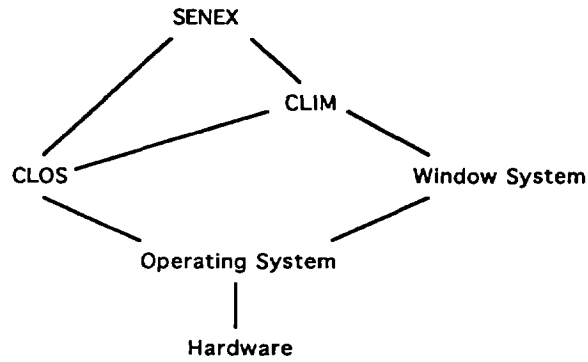
SENEX is a computer system under development to explore issues related to representation of molecular information, presentation of data, and reasoning with molecular information. It is written entirely in a portable programming environment supported by Common Lisp, the Common Lisp Object System (CLOS), and the Common Lisp Interface Manager (CLIM). SENEX contains information about molecules, molecular events and disease processes, and provides tools for reasoning with and displaying this information in useful ways.

Molecular pathology 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. However, the CLOS metaobject protocol allows a programmer to adjust the design and implementation of the language to fit an application domain. Thus the first objective of the SENEX project is to exploit this feature of the CLOS metaobject protocol in designing a language tailored to the domain of molecular pathology.

Graphical presentations can aid in a conceptual understanding of molecular biological phenomena. However, creation of graphical presentations can be quite time consuming and frequently provides little reusable code. It is possible to encode some of the reasoning processes used in producing graphics, for example, computing graphical presentations from symbolic representations, specializing, if desirable, particular features of the presentation. Thus the second objective of the SENEX project is to exploit features of CLOS and CLIM for computing graphical presentations from symbolic representations of molecular data.

Molecular pathology is a domain characterized by complex interactions of diverse structural elements and events. Much of our knowledge in this domain can be captured in terms of operations upon objects subject to specified constraints. For example, certain rules for ordering motifs within a protein can be employed, and details regarding the structure of motifs may be dependent upon the state of the molecule. Another example would be triggering a chain of molecular events in response to the presence of a hormone and alteration of that response in certain disease states. Perhaps most importantly, the ability to reason with information allows 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. Thus the third objective of the SENEX project is to exploit features of CLOS and the metaobject protocol for purposes of reasoning with molecular information.

CLOS and CLIM confer application portablility to SENEX.



A schematic diagram of how Senex is layered over the host system is shown above. The application uses tools provided by CLIM and CLOS. CLIM uses tools provided by CLOS and the host window system. CLOS and the host window system use tools provided by the operating system, and the operating system uses the available hardware. The same body of code constituting the entirety of SENEX currently runs on a Macintosh using MCL 2.0/Lucid CLIM 1.1 or on a Sun Sparc II Station using Lucid Common Lisp 4.1/CLIM 1.1. Compiler-specific access to the CLOS metaobject protocol (see below) constitutes the only nonportable fragment of Senex.

Molecules and molecular processes are represented in SENEX by means of Common Lisp objects.

THE SENEX CLASSIFICATION STRUCTURE

	# of Classes
Entities	5769
Organisms	636
Anatomic structures	262
Cells	143
Compartments	115
Molecules	4074
Proteins	2455
Genes	219
Motifs	284
Diseases	468
Events	110

A rough breakdown of the different classes of objects in the SENEX classification structure is shown above. An object in SENEX is an instance of an object class (Ball & Mah, 1992). For example, Homo sapiens is a class of organism and Sheldon S. Ball (first author) is an instance of the class Homo sapiens. Similarly, protein phosphorylation is a class of event and there are many specific instances of protein phosphorylation in SENEX.

Objects have slots which provide a means of describing an object in detail.

Slots are specialized descriptors of objects defined with the most generalized class to have that attribute or property. Slots may assume default values specified with class definitions and most slot values themselves are instantiated as objects. Slots and their default values are inherited through the classification structure. The basis of the SENEX classification structure is the MEDICAL SUBJECT HEADINGS (MeSH) tree structures. Thus the SENEX classification structure is a biological classification structure. There is a mapping of synonyms to the canonical forms used as class names.

Slot default values provide a means of programming biological knowledge into SENEX.

Molecules are classified in chemistry and biology largely on the basis of their properties. Thus classes of molecules share particular properties which may be represented as slot values. All members of a class of molecules may inherit properties as default slot values. Classes of molecules may have multiple supertypes, so that the properties of a class of molecule may be determined by inheritance of slot default values from multiple molecular supertypes.

Inheritance of slot values is specialized depending upon the slot.

Proteins contain structural elements which give rise to the function(s) of the molecule. These structural element are known as motifs. Most proteins consisting of a single polypeptide can be represented as an ordered set of motifs connected by peptide regions (see figure 1). Different classes of molecules contribute different motifs to their subclasses through slot default values. Thus when a class of molecules is defined with multiple supertypes, motifs are inherited from all supertypes. Inheritance of motifs is said to be 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 protein class.

Senex uses reflective techniques in context of the CLOS metaobject protocol.

Senex uses two enabling technologies (Kiczales et al, 1991): reflective techniques which make it possible to expose the implementation of a language (in this case LISP), and object-oriented techniques which allow the implementation of the language to be locally and incrementally adjusted. The basic elements of CLOS - classes, methods, and generic functions - are accessible as metaobjects (objects that represent fragments of a program). A protocol operating on these metaobjects defines the behavior of CLOS. SENEX uses introspective protocols to access slot values of these metaobjects and intercessory protocols to change the behavior of CLOS in specializing the inheritance of motifs.

Ordinary functions, macros, and methods provide a means of programming biological knowledge into SENEX.

The inheritance of motifs by union necessitates processing of motif defaults to eliminate duplicates and identify specializations of generalized motifs. In addition, certain rules for ordering motifs within a protein can be employed, and details regarding the structure of motifs may be dependent upon the state of the molecule. This type of molecular information is programmed into SENEX using ordinary functions, macros, and methods defined on specific classes of molecules.

A CLIM presentation is a visual representation of an object linked directly to its semantics, thus facilitating the separation of the internal representation of objects from the presentation of data to users. For example, figure 1 shows a screen image obtained from browsing through the SENEX classification structure to find the amyloid precursor protein (APP). In the lower right hand corner of the screen, there is a window entitled Query that represents a menu of a sort. The Query window contains a set of commands for which

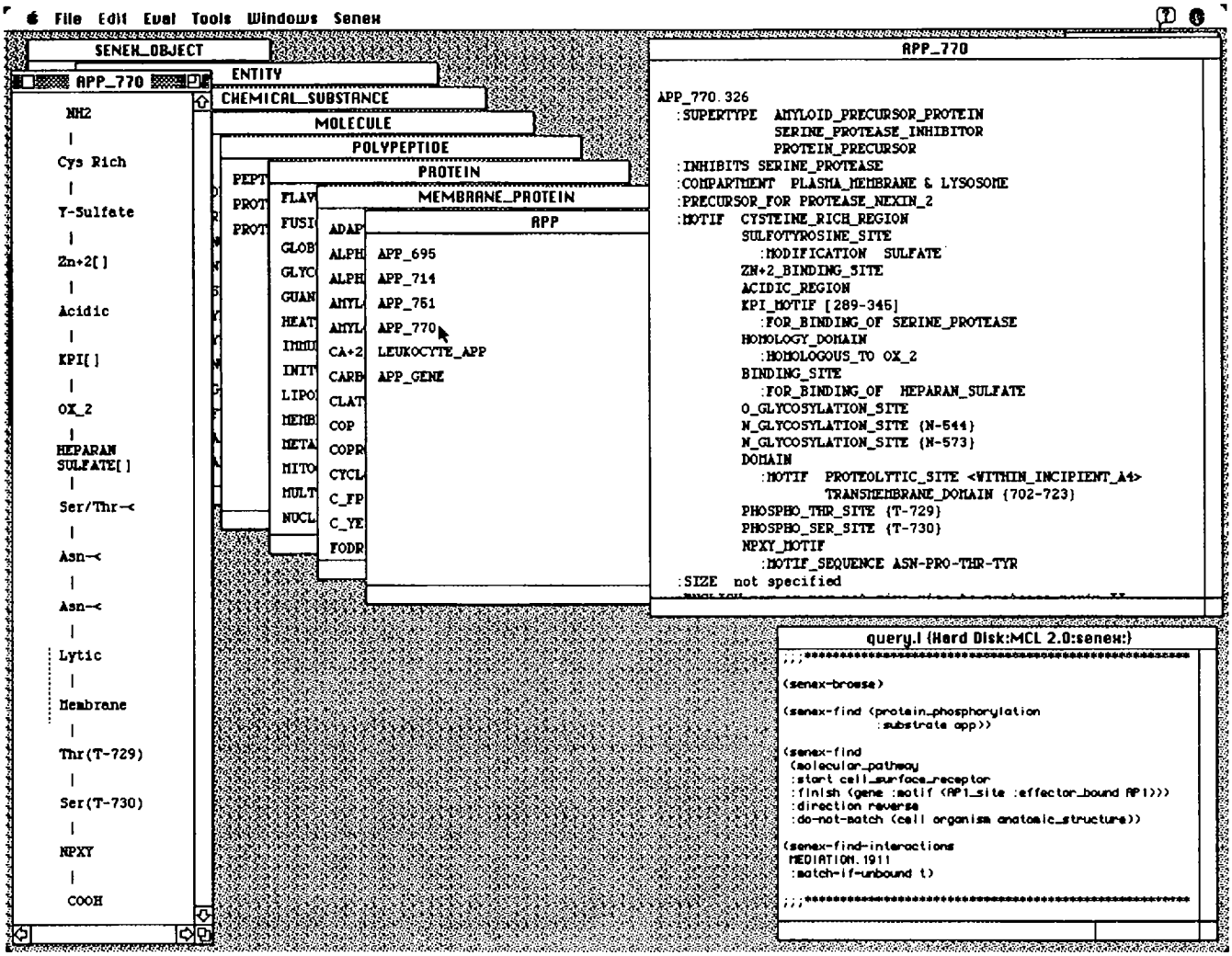


Figure 1

the user simply completes the details, places the cursor, and presses <enter>. All subsequent actions are through selection of mouse-sensitive objects in the windows that appear in response to the initially entered command.

A CLIM presentation is a visual representation of an object linked directly to its semantics, thus facilitating the separation of the internal representation of objects from the presentation of data to users. For example, figure 1 shows a screen image obtained from browsing through the SENEX classification structure to find the amyloid precursor protein (APP). In the lower right hand corner of the screen, there is a window entitled Query that represents a menu of a sort. The Query window contains a set of commands for which the user simply completes the details, places the cursor, and presses <enter>. All subsequent actions are through selection of mouse-sensitive objects in the windows that appear in response to the initially entered command.

The command used in this instance was the command (senex-browse). A window in the upper left hand corner of the screen entitled SENEX_OBJECT appeared (now partially hidden by a set of offset descending windows). A selection through a mouse gesture in this window produced the adjacent window entitled ENTITY (offset to the right & down). All mouse-sensitive objects in a window highlight when the mouse pointer is passed over the object. Choices from these windows are members of the SENEX classification structure, with each selection producing a new window showing subclasses of the selected class. If no subclass exists, no subclass window is produced and a general description of the class appears in a window in the upper right hand corner of the screen. Even if a class has subclasses, a description of the class may appear if specified conditions for the class are met. The description of the class is the symbolic representation for that class with associated slot

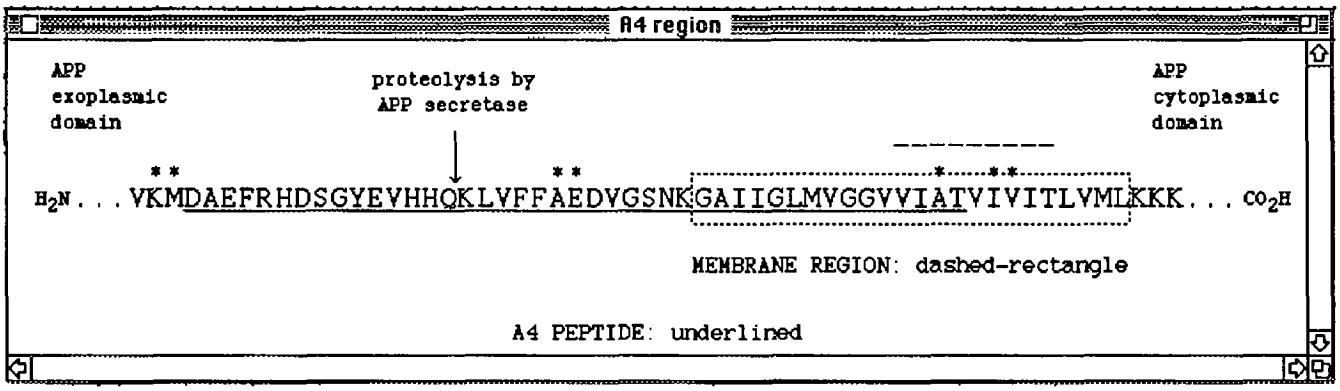


Figure 1a

default values. An English description and reference may accompany the symbolic representation.

Senex computes cartoon presentations from symbolic representations.

From this symbolic representation, SENEX computes a cartoon presentation of the object representing the class and displays this presentation in the lower left hand corner of the screen. The cartoon of APP indicates that the protein consists of an ordered set of motifs connected by peptide regions. These motifs, like the protein containing the motifs, are represented in SENEX as objects. Classes of motifs constitute an important branch of the SENEX classification structure. The motifs shown in the cartoon of APP (lower left) are mouse-sensitive, that is, details of the motif will appear in a new window if the user selects the motif with a mouse gesture.

Generic functions select methods appropriate for specialized classes.

The cartoon of APP shows an additional element of complexity. The dashed line to the left of the membrane region and proteolytic site represents a domain containing the two motifs. Selecting this domain with the mouse brings up a new window (figure 1a) showing the peptide sequence of this protein region. This is accomplished through use of a specialized method defined on a generic function *draw*. The cartoon of this domain shows the sequence which gives rise to the amyloidogenic A4 peptide, the normal cleavage site of the APP secretase, the region of the amyloid precursor protein traversing the plasma membrane, and several mutations in APP found in various hereditary disorders.

Figure 2 shows that part of the SENEX classification structure containing APP.

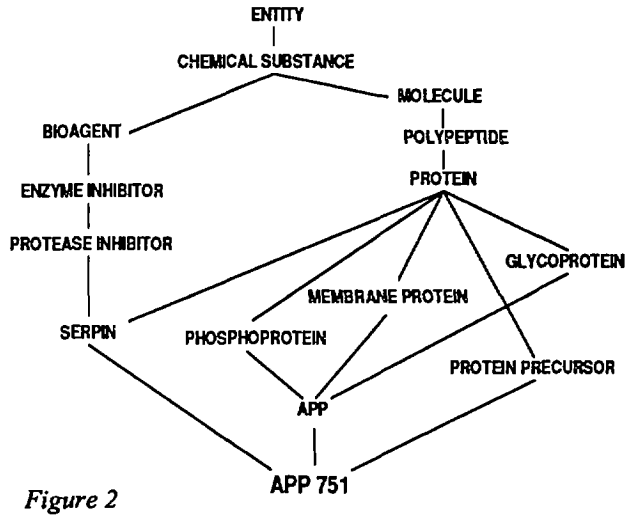


Figure 2

Methods defined on multiple classes provide a means of drawing objects in context of other objects.

Events and subtypes of events are represented in SENEX as objects. Finding events that satisfy particular specifications, for example, phosphorylations of the general class of protein APP (includes any subtype of APP) is facilitated with the command (senex-find (protein_phosphorylation :substrate APP)). SENEX collects all instances of the class PROTEIN_PHOSPHORYLATION testing for additional user defined target specifications (in this case :substrate APP). All instances in SENEX have associated unique instance id's. A concise symbolic representation of the phosphorylation of APP catalyzed by Ca+2/Calmodulin-

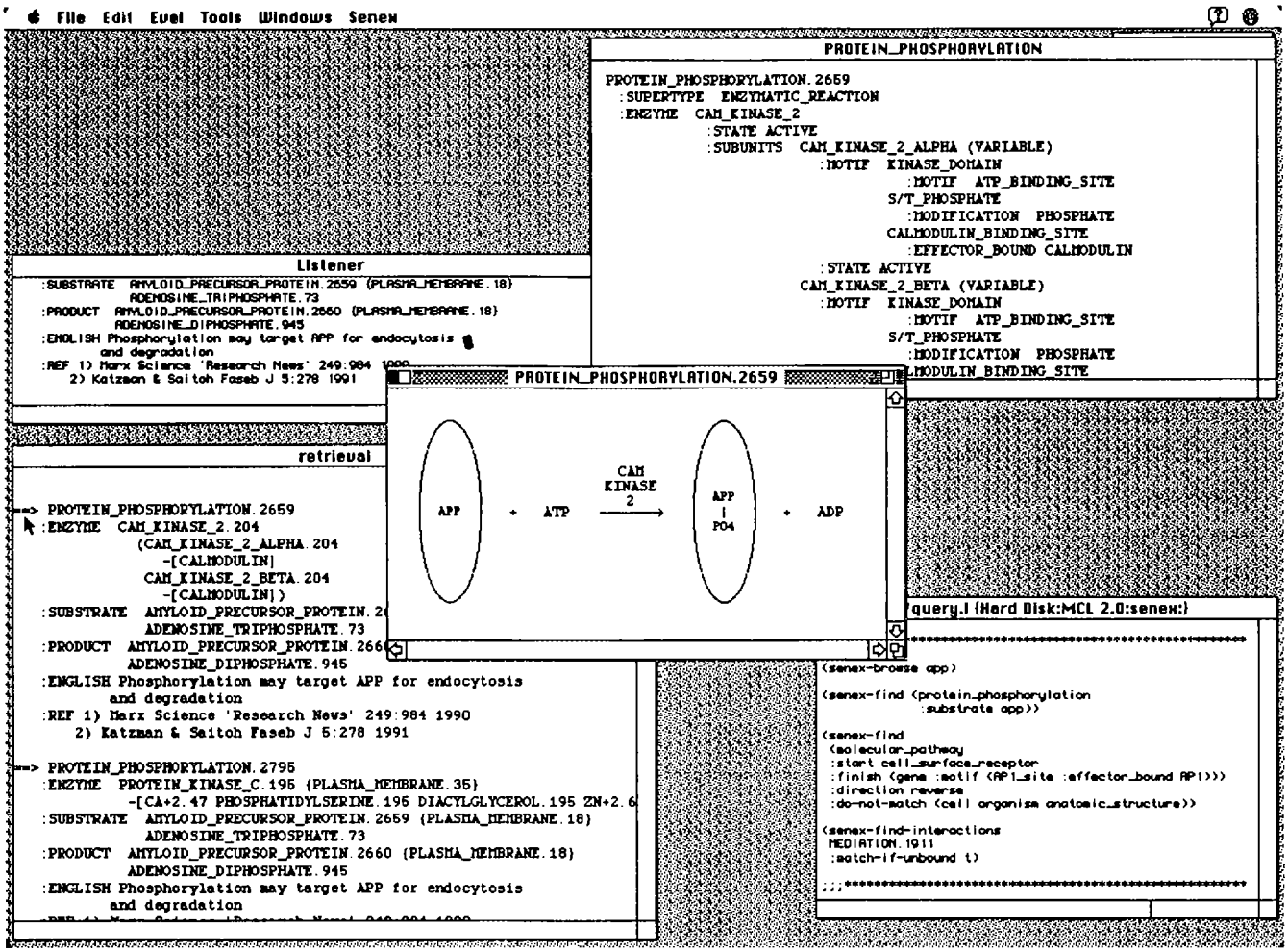


Figure 3

dependent protein kinase-2 along with a cartoon of the reaction is shown in figure 3. The ellipses in the cartoon represent proteins, with substrates on the left and products on the right. The enzyme for the reaction is shown above the arrow. APP in the context of a reaction now appears as a simple ellipse rather than as a stick figure illustrating all of its motifs as in figure 1. Thus, in figure 1 APP is drawn in context of itself, and in figure 3, it is drawn in context of a reaction it undergoes. We have also seen examples of a domain drawn in context of a protein and in context of itself (figures 1 and 1a). This feature serves to provide the user with an appropriate level of detail.

The objects in the window showing a cartoon of the selected reaction are mouse-sensitive. Selecting any of the objects in the cartoon brings up further detail of the object. In the case of the ellipse representing APP, selection brings up a stick figure cartoon along with a symbolic description of the molecule as shown in figure 1.

CLIM separates the internal representation of objects from the presentation data to users.

When information about a gene for a protein is available, the gene appears as a choice when that protein is selected during browsing (see figure 1). The symbolic representation of the APP gene shown in figure 4 indicates that at least 6 different proteins are derived from this gene by alternative splicing of a single transcript. A chromosomal location for the gene (human if not otherwise specified) is shown as a value of the slot LOCUS. Internally, SENEX knows that genes are located in the cell nucleus but does not display this to the user since the fact is self-evident.

A cartoon presentation is computed from the symbolic representation shown in figure 4 when APP_GENE is selected during browsing and displayed in a separate window (see figure 4). The cartoon shows several gene regions including 5' enhancer and promoter regions, a coding region,

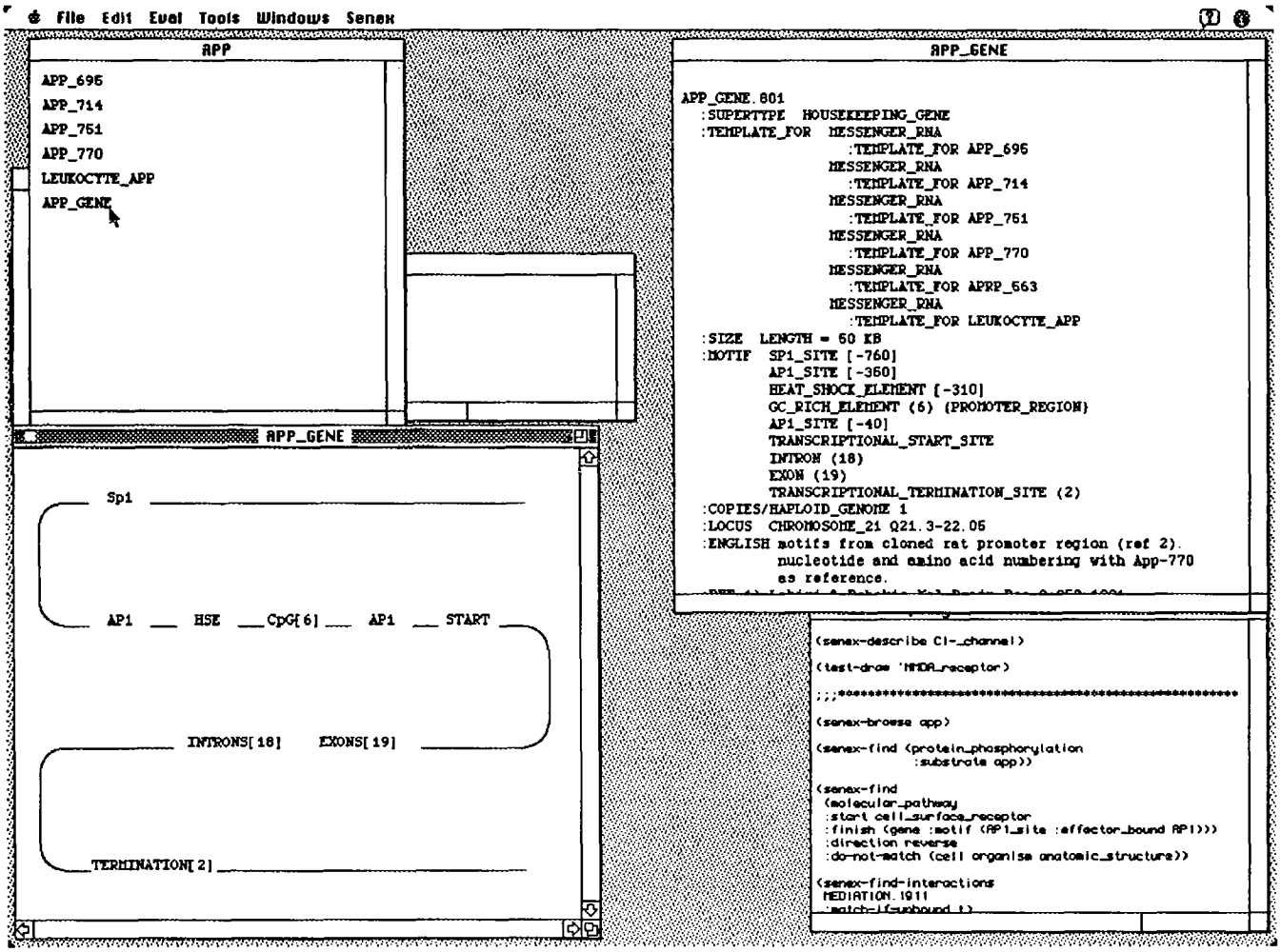


Figure 4

and a 3' enhancer region. DNA regulatory elements or motifs which control expression of the gene are shown in the enhancer and promoter regions. Regulatory elements embedded within an exon or intron are revealed upon selecting the coding region with a mouse gesture.

Senex provides a means of exploring tangential issues and increasing levels of molecular detail.

It is noted that the gene for APP contains two AP1 sites, binding sites for the transcription factor AP1. We can use the browse facility to obtain information about AP1 (Figure 5). A cartoon presentation of AP1 is shown in figure 5. It is noted that AP1 is a heterodimer of p55fos and p39jun each of which contain motifs which hold the proteins together in a specific configuration. Thus, the cysteine-containing basic helix-loop-helix motif (CbHLH) on each protein is a half site which aligns with the analogous site on its companion protein to form a redox-sensitive DNA-binding motif. Similarly, the leucine zipper motifs on the two proteins are half sites which align to hold the proteins together by hydrophobic bonding of heptad repeats of leucine residues. The cartoon presentation of AP1 thus shows the alignment of these motifs.

The PEST region of p55fos is selected with a mouse gesture to produce a new window which contains a symbolic representation of this motif. An English description and references may be the most useful aspect of this window.

An obvious limitation to this cartoon presentation of AP1 is apparent. From the cartoon, it appears that p39jun is larger than p55fos. This occurs because the length of the polypeptide in the cartoon is a function of the number of motifs identified, not the size of the protein. An innovative solution to this problem is a goal of this project.

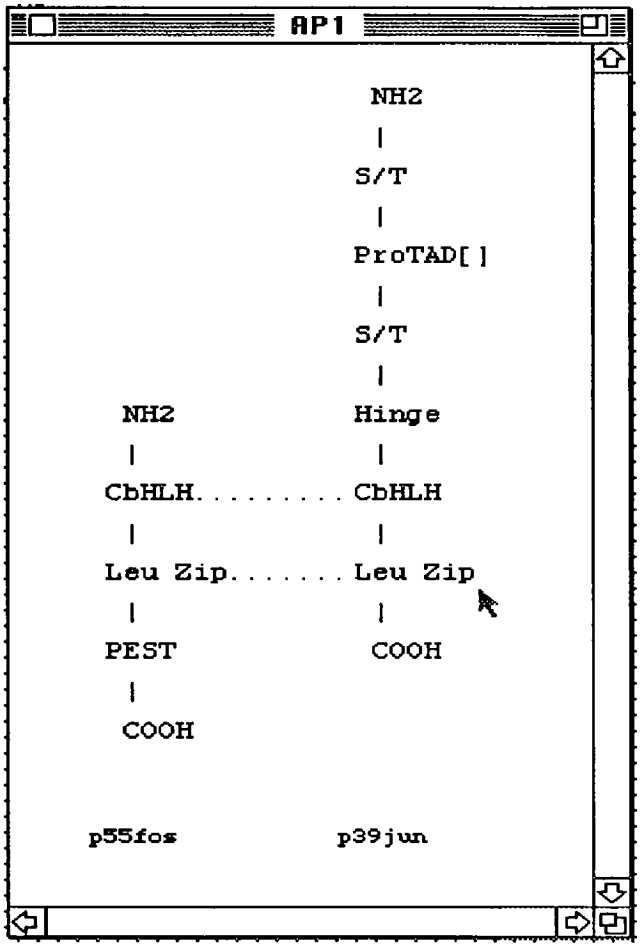


Figure 5

Representation of events as objects facilitates identification of molecular pathways.

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 can lead to changes in gene expression. Transcription of genes is regulated through binding of specific nuclear proteins to regulatory elements within promoter or enhancer regions of the gene. Thus, we may be interested in signal transduction pathways that stimulate transcription of the APP gene.

SENEX is a tool of discovery. Search options facilitate prediction of novel signal transduction pathways.

We can tell SENEX to ignore cell type, anatomic considerations, and organism type 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 signal transduction pathways, in essence generating hypotheses which may be tested in the laboratory.

Such a query seeking regulation of APP gene expression by extracellular signals, yields twenty-two such possible pathways, one of these originating from the beta-2 adrenergic receptor (see figure 7). The queue at each step of the search along with a summary of pathways found is printed to the listener window. Any one of the pathways as well as any of the events that comprise a particular pathway may be chosen for further examination. Figure 6 shows a screen shot showing translation of c-fos messenger RNA, obtained by selecting one of the events of a selected pathway.

Regulation of signal transduction involves complex networks of interactions.

It is also possible to identify events which interact with other events or with molecular pathways. These interactions may be further specified as inhibitory or stimulatory.
The algorithms employed in SENEX searches are discussed in Ball et al, 1991.

Specialization of molecular presentations illustrates compartmental relationships of molecules.

Cell surface receptors which convey the signal of their extracellular ligand to the interior of the cell via activation of G proteins (serpentine receptors), have a common structure of 7 transmembrane domains. The ligand binding regions and sites of interaction with G proteins lie within the membrane (see figure 7). Regulatory sites may be found within the cytoplasmic loops and cytoplasmic tail of the receptor. These features may be elucidated when the user clicks on one of the many mouse sensitive features in the cartoon.

Figure 7 shows a cartoon presentation of the beta adrenergic receptor computed from the symbolic representation of the molecule. This is accomplished through specialized methods defined on two generic functions, *draw* and *draw-cartoon*. The common structure of serpentine receptors facilitates drawing regions of the generalized structure and specialized drawing of particular regions for particular receptors. In the case of the beta-2 adrenergic receptor, the exoplasmic N-terminus, cytoplasmic tail and 3rd cytoplasmic loop were drawn using specialized methods of the generic function *draw*.

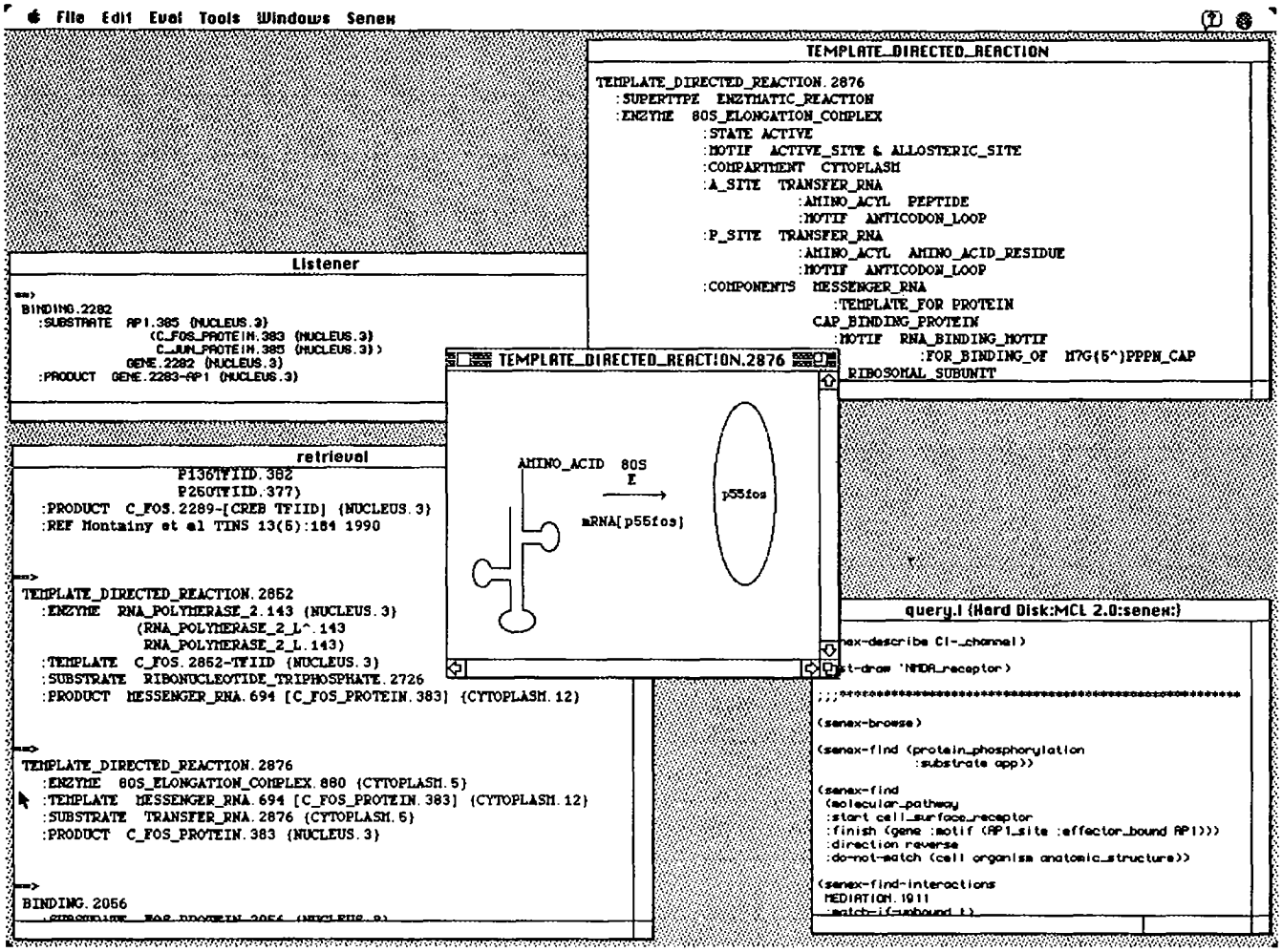


Figure 6

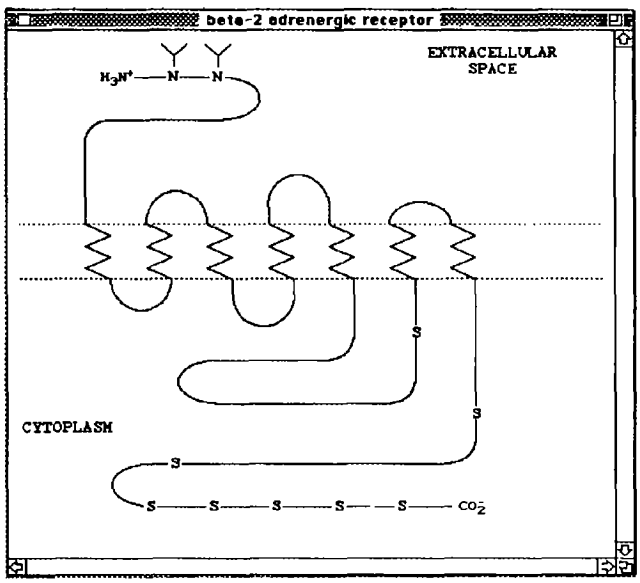


Figure 7

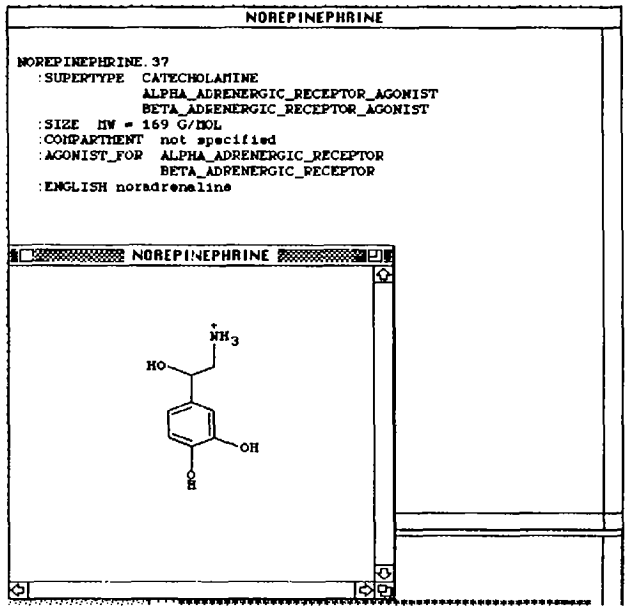


Figure 7a

Visualizing chemical structures of small molecules helps to understand their biological functions.

Ligands for receptors when known appear as choices when a receptor is selected during browsing. Selecting norepinephrine brings up a description of the principal CNS ligand for the adrenergic receptors (see figure 7a). Visually identifying the hydrophobic aromatic ring of norepinephrine with its polar substituents helps to understand how such a ligand might fit into the transmembrane pockets of the adrenergic receptors.

Presentation of molecular pathways and cascades is facilitated by generic functions which select methods appropriate for specialized objects and methods defined on multiple classes which provide a means of drawing objects in the context of other objects.

Figure 8 shows a presentation of the coagulation cascade with its associated amplification loops. This figure represents a general level of detail. All of the objects in the presentation are mouse-active. Thus visualizing further detail about particular aspects (molecules, events, or disease manifestations) of coagulation are visualized by selecting appropriate objects in the sequentially revealed windows.

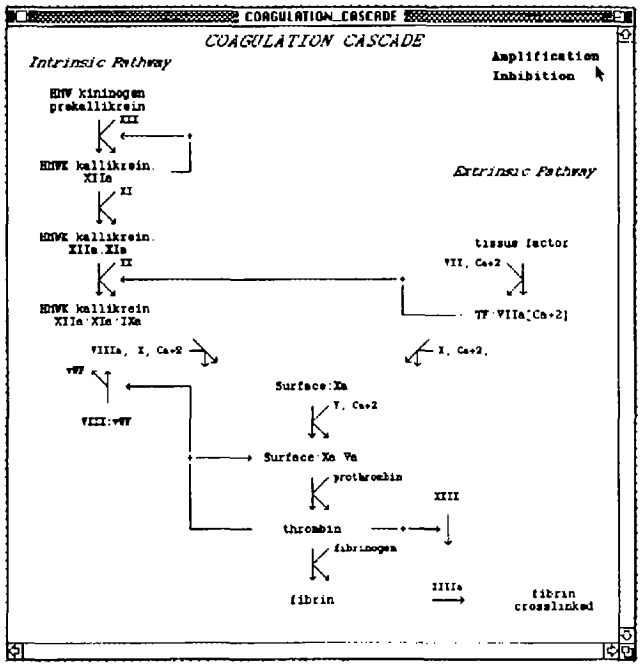


Figure 8

Summary/Discussion

SENEX is an expanding CLOS/CLIM application in the domain of molecular pathology. SENEX contains information about molecules, molecular events, and disease processes, and provides tools for reasoning with and displaying this information in a useful way.

There are many other computer applications in molecular biology including works of Brtulag & Galper (1990), Karp (1989), Kazic et al (1990), and Koile &Overton (1989). However, these works bear little resemblance to SENEX.

SENEX uses many features of CLOS and CLIM for representation of molecular information, presentation of data, and reasoning with molecular information. These features include CLIM presentations, classes defined on multiple supertypes, generic functions and methods defined on multiple classes, inheritance and specialization, and reflection on the CLOS metaobject protocol.

References

Ball, S.S. and Mah, V.H. (1992) Knowledge Representation in Molecular Pathology. Proceedings of the 16th Annual Symposium on Computer Applications in Medical Care. McGraw Hill, NY, 1992 pg 371-375.

Brutlag. D.L. and Galper. A.R. (1990) Simu;ating DNA metabolism: A knowledge-based approach. Bioinformatics, Integration of Organismic and Molecular Data Bases, and Use of Expert Systems in Biology, July 9-11, 1990, George Mason University

Karp, P. (1989) Hypothesis Formation and Qualitative Reasoning in Molecular Biology, Ph.D. thesis, Stanford University, Stanford CA, Tech Report 1263.

Kazic, T., Liebman, M.N. and Overbeek, R.A. (1990) Steps towards a computational model of Escherichia coli physiology. Bioinformatics, Integration of Organismic and Molecular Data Bases, and Use of Expert systems in Biology, July 9-11, 1990, George Mason University.

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

Koile, K. and Overton G.C. (1989) A qualitative model for gene expression. In: Proceedings of the 1989 Summer Simulation Conference, Society for Computer Simulation, July 1989.