

## Pattern Discovery in Gene Regulation; Designing an Analysis Environment.

Stella Veretnik and Bruce Schatz

Community Systems Laboratory  
Life Science South, 233  
University of Arizona  
Tucson, Arizona 85721 U.S.A.

phone: (602) 621-5079 e-mail: stella@csl.biosci.arizona.edu

### Abstract

Interactions that determine cellular fate are exceedingly complex, can take place at different levels of gene regulation and involve a large number of components (such as genes, proteins). 'Wet lab' biology has an inherent difficulty in considering multiple components within one experimental set-up. Thus, the individual experimental results may reflect the behavior of a sub-system and be missing some important information concerning interactions with other components. Computational tools can help in **simultaneously analyzing many different pieces of biological knowledge** from different data sources. Such tools will aid in comprehending the **whole system** (cell, organism) as a function of all of its components; this, in turn, will facilitate discovery of the global patterns in genetic regulation. The Worm Community System (WCS) which contains extensive knowledge from many different sources regarding model organism *C. elegans*, presents a suitable environment for development of the integrated analysis tools. Here we describe a working version of WCS and the strategies employed for the development of the global analysis tools within the System. The present paper deals with the construction of a highly interconnected information space (in the context of the problem) by introducing more sophisticated data objects representing knowledge about genetic regulation. We describe construction of the in-depth objects, development of the analysis tools and discuss the type of analysis feasible within such interconnected space. The analysis tools will serve as an ideal environment for 'dry biology' experimentation and provide a context for 'wet' experiments.

### The Worm Community System

The Worm Community System (WCS) is a software environment that enables members to manipulate all the knowledge of their community from the computer in their laboratory transparently across the national computer network (Schatz, 1991-92). It is part of larger effort to build integrated science information systems for biologists. WCS lets the user retrieve and add to "all" the knowledge about *C. elegans*, a non-parasitic nematode (Wood, et al., 1988; Roberts, 1990). The concentration thus far has been on literature, publishing and rudimentary analysis support. Features of WCS include search, display, entry, and linking to data.

Release 1 of WCS has been running in *C. elegans* labs for nearly two years and there are now about 25 sites, including some remotely across the Internet. WCS release 2, a complete revision will be released during the summer of 1993 as an X-windows package for UNIX workstations, Macintoshes, and PCs with a front-end for interaction and a back-end for search.

**Knowledge support.** The literature coverage in WCS includes the complete Worm Breeder's Gazette (WBG newsletter) as well as the abstracts from the majority of the articles in the community bibliography and the abstracts from the most recent *C. elegans* meetings. The search has a capability for selective search by type or by field and for matching of text phrases or boolean expressions. There is also a thesaurus tool which can suggest keywords commonly associated with a given term. References to other items within the text can be immediately followed (e.g. pointing to a gene name within a text will allow retrieving additional information about a gene). Information about genes, clones and maps is present within WCS and is connected to the literature.

**Analysis environment.** WCS supports the beginning of an analysis environment: an exploratory navigation across multiple sources to discover patterns within the data and literature. For example, you can search the literature for a topic of interest, follow links to find all mentioned clones, and display all genetic (or physical) map regions containing these clones. A combined genetic/physical map display allows easy cross-correlation of information between genetic and physical maps (Edgley and Riddle, 1990). There is also a traversable full lineage tree of *C. elegans* allowing follow up of the fates of individual cells during organism differentiation. Each cell within the lineage is associated with information regarding cell type and timing of the division.

**Electronic publishing.** A significant electronic publishing facility is supported within WCS. One of our major goals has been to collect and incorporate informal knowledge within the WCS community. The publishing facility provides a way to enter knowledge directly from the user's lab while using the system, have new entries be as fully linked to other information as the archival information, and provide selective



dissemination, if desired. (Existing technology such as the Internet Gopher package support items as flat text not as linked objects.) Any type of information can be entered as a real object within the system, then selectively published within the community. Publishing an entered object involves specifying privacy and editorial control: which members of the community are permitted to view or change any given entry, and which kind of quality checking should be performed on the entry (none, moderated like the newsletter, curated like the maps). For example, an item of biology data (e.g. a gene description) can be entered via a form with pre-defined fields. During entry, the user can check whether other referenced objects are already defined (e.g. the clone which contains the gene of interest). These references to the new object can be followed immediately within the system.

### Underlying structure of WCS

A community system uses a particular kind of federated heterogeneous distributed object-oriented database, called an *information space*. An **information space** is a set of information units and their connections (Schatz and Caplinger, 1989; Schatz, 1991-92). Logically, it is a single uniform graph structure, although physically it may be composed of many different sources of data of many different types stored on many different machines in many different locations spread across a network.

Information spaces support uniform manipulation of heterogeneous data items by transforming them into homogeneous information units. The generation of an information space begins with data already existing in some external source. The format of this data is administratively transformed into a canonical internal representation called an **information unit** or **IU**. An information unit is an encapsulated object, in the sense of an object-oriented programming language, which has an associated set of operations to provide manipulation capability for its particular data type. Every "database" thus has a set of transformation routines and every "data type" has a set of data operations. Once the data items have become information units, there are a set of generic operations available for performing on them. These generic operations support uniform commands at the user level for such functions as search, display and grouping. Each information unit may be connected to other units to represent a semantic relationship and collections of information units may be grouped into new composite units. There are several levels of representation in an information space. Data exists in the external sources and is transformed into information within the space. Knowledge, in the sense of **community knowledge**, is represented by the different components of information units. Any IU can be annotated; a typical *annotation* is a note stating some additional feature of the encapsulated data (e.g. this gene may encode this function). Any two IUs can have a relationship specified between them; a

typical *connection* is a link to another IU supplying additional information (e.g. this article discusses this gene). Any collection of IUs can be grouped into a single composite IU which forms a region in the information space; a typical *region* is a set of IUs on the same topic (e.g. all genes coding for mechanosensory deficiencies). Since every IU has a unique identification within the entire space, it is possible to implement a uniform mechanism for forging and maintaining these groupings, even across sources. Every IU also has specification to provide publication control over the sharing of these groupings.

Anything accessible may potentially be incorporated into the space. That is, all data reachable via the underlying network for which appropriate transformation routines exist can reside logically within the information space. The major generic operations built into the system, as part of the IU class definition, are the support for grouping. These include connection links and region sets. Other operations, which provide support for the uniform user commands, are implemented at the individual subclass level (e.g. those for search and display). The object structure of information units enables an electronic community system to be extendable, with a base set of classes that can be augmented by specific classes for a specific community.

### Sample Session

Figure 1 shows a sample session with the Worm Community System (release 1). The session shows both data and the literature and some of the relationship links. The session begins with the user entering a search for the keyword "sensory". By browsing through the summaries of items the user selects one gene "mec-3" and zooms into its description, which is displayed in Figure 1. This description shows that mutations in the gene indeed make the worm insensitive to touch. From this gene, the physical map was selected and the interactive display of the DNA clones appeared centered around the gene of interest. Further information that can be retrieved includes DNA sequence. Note, that each of the items shown (literature, gene, map, sequence) comes originally from its own database, but the Worm Community System enables navigation across all this sources with single set of uniform commands.

### Developing analysis tools within WCS

We intend to build analysis tools to search for patterns in genetic regulation of the cell: strategies employed in the construction of the regulatory, uncovering hidden interdependencies between cellular processes, and associating the above properties with specific motifs within DNA/RNA/protein domains. Search for such patterns relies on a wealth of information regarding an object (such as gene) within WCS. While multifaceted information does exist within WCS regarding any object (such as gene) - it is dispersed among different bits of data and we currently have neither

the tools to extract this information nor the suitable format to store it (Lander et al., 1991). Our first priority is to create a **highly interconnected information space** in which different types of information regarding an object itself and other related objects are extractable and present in a format suitable for further analysis (Schatz, 1991-92).

### Interconnecting information space

Construction of the highly interconnected information space will be accomplished by creating conceptual links between different types of information existing within WCS regarding a chosen object. We will implement it by constructing in-depth objects with multiple fields to accommodate various types of information regarding the object. There are two major issues addressed within this section: 1.) the types of relevant information and its organization within the in-depth object; and 2.) generation of tools for extracting relevant information from WCS. We are attempting to make the process of interconnecting the objects within the database general enough to accommodate different types of questions to be addressed within existing interconnected space. However, different types of in-depth objects may be required to address different set of problems, for the **type of questions** one wants to analyze will define the **types of interconnections** among the objects. Different ways of connecting objects within an information space is essential; custom-constructed interconnections among objects in order to address specific problem will be supported.

#### Types of relevant information; format of in-depth object

Since our primary interest is in analysis of different aspects of interactions among genes and the effect of these interactions on the cellular fate, the basic **in-depth objects** in our implementation are **genes**. There are multitudes of gene attributes that can be extracted from the database and put within the field of the gene's object: information regarding the gene itself, information regarding other genes involved in interaction with the gene of interest, and information regarding the nature of the gene-gene interactions.

Two basic types of information will be available regarding an in-depth object: unprocessed information (which can be used in the currently existing WCS format; this include sequences of DNA, RNA or protein, alleles, strains, map positions) and processed information about the gene (this includes expression, localization, regulation, interactions with other genes, physical properties of the product, participation in pathways, etc.). The term '*processed*' refers to the fact that the information should be synthesized from the data existing within WCS. Processed information about the gene is viewed at two levels: general and molecular. The molecular level deals with the properties as they are determined by DNA/RNA/protein sequences (Fig. 2).

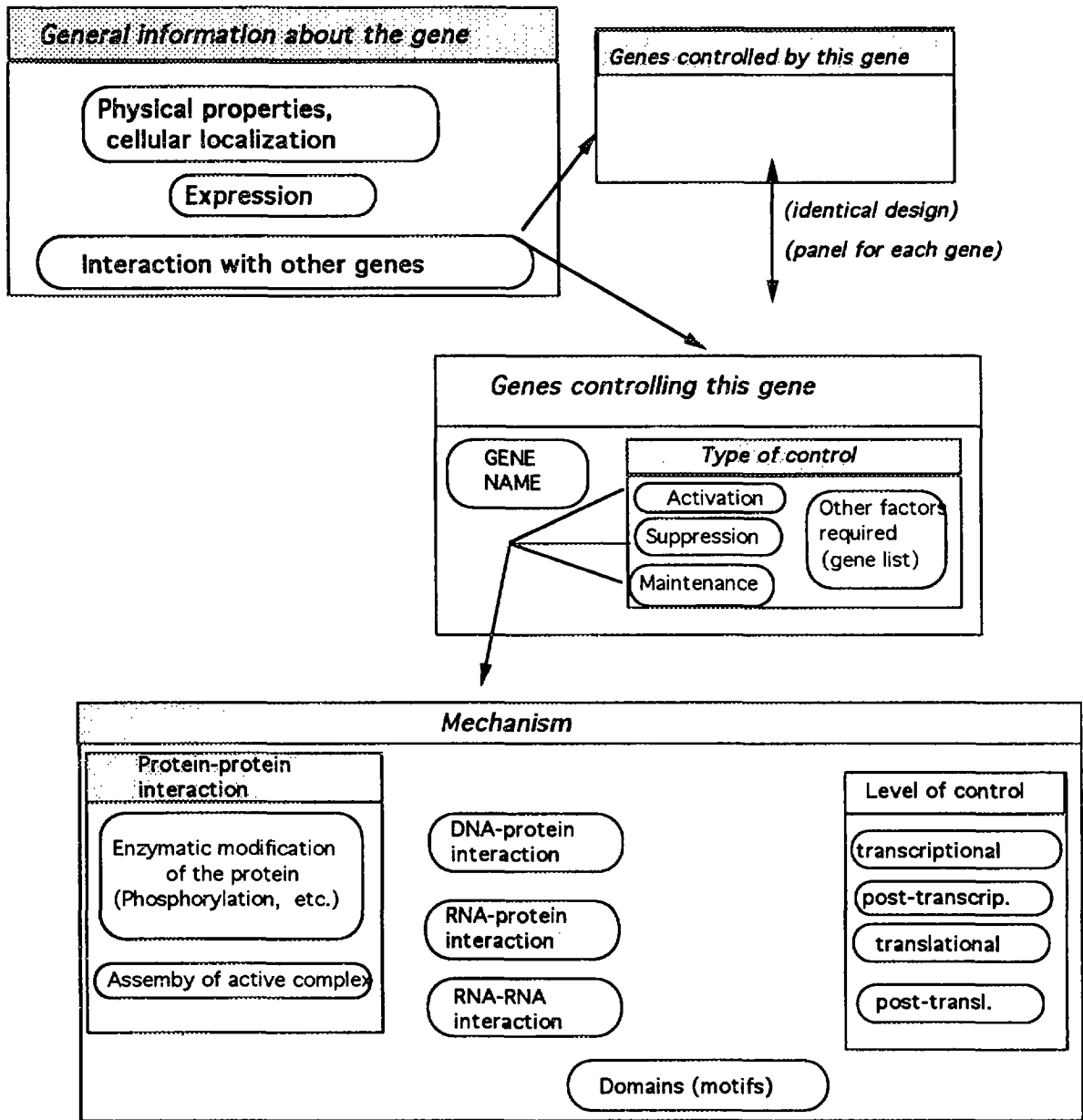
We have attempted to construct hierarchical ordering

of information; the top level presents most general information, while subsequent levels have more refined information regarding one of the aspects of the higher level. For example, the regulation of the gene is subdivided into two types: all genes that regulate the gene of interest and all the genes regulated by the gene of interest (Fig. 2). Each of those, in turn, is subdivided into types of regulation (activation, suppression, maintenance, Fig. 2), each type of regulation is further subdivided into molecular mechanisms (DNA-protein interaction, RNA-RNA interaction, etc.) and the level of the control (transcriptional, post-transcriptional, etc.). The level of control can be refined further (post-transcriptional regulation is subdivided into splicing, polyadenilation, etc., while splicing the field itself is refined into trans-splicing, self-splicing, alternative splicing, etc.). It is also possible to access the specific DNA or protein domains (on the level of sequence) that are involved in the regulation.

Such an in-depth approach to the information about the gene conceptually orders the existing data. It allows a somewhat logical access to the data by the researcher who manually searches this for in-depth objects or by an expert who connects the space by filling in various fields of the gene data (since the intention is to fill in as much data as possible by experts in the field as an alternative to a search of the literature, see below). Our implementation of this hierarchical refinement of information should allow easy addition of new fields of data at any of the levels as well as addition of the new levels - should those prove necessary. For example, little attention in the current model is given to the properties of the structural proteins; this information can be added to the field of *Structural* (Fig. 3).

#### Extracting relevant information from the WCS

Multiple bits of information existing within WCS regarding a given gene or its product **are not necessary conceptually associated together** in the current version of the WCS - they are often reported in the different types of journal articles due to the different type of problems addressed and different aspects of research performed. For example, one paper may discuss assembly of gene X product within basal membrane, the second paper may report on the formation of heteromer AB and its transcription activation of the battery of genes, among them gene X; and yet the third document may discuss differential expression of gene X as a function of alternative splicing. Often some subtle details regarding the gene of interest will be mentioned in the research centered around other gene(s) or phenomena; sometimes the gene will be mentioned in the discussion, when analogies between different systems are drawn. In order to be able to correlate any of gene's parameters, for example, effect of the transcriptional regulation on alternative splicing, it is essential to have an easy access to all types of the data regarding given gene. How can we access all the wealth of information already existing



**Figure 2. Conceptual subdivision of the information pertaining the gene object. Refinement of *Interaction with other genes* concept.**

in the literature, but is not filtered according to our interests?

The natural tendency would be to parse all the literature (formal and informal) and interpret all types of information required within fields of the in-depth object. While we do intend to use basic parsing, we also want to attempt another general strategy: to promote manual data entry by the researchers working with given phenomenon/gene/protein. This could be done if a meaningful format for presenting the in-depth data and user-friendly interfaces were developed. The majority of the information regarding the genes is **still to be discovered and reported**, so it is important to convince researchers now to enter it in the electronic form (just as DNA and protein sequences are currently submitted), so in the future we might avoid the bulk of the text parsing and data interpretation altogether. We do intend to develop simple tools for parsing text with various levels of sophistication. This will be important for early stages of the project (when we will need to input information, before the experts get convinced about the usefulness of the in-depth objects). Parsing of the full body journal articles will be essential for discovering **hidden interdependencies** between genes (or conserve motifs within the genes) existing within the current body of literature. It is especially important for **high level analogy** searches using the Thesaurus tool (see Analysis section). Relevant publications will be retrieved using a set of key-words; a larger set of the key-words will be able to narrow topic of search significantly. The retrieval will be done using an entire body of the article (which is not commonly done within current databases). Fields of in-depth gene objects can be associated with the predefined set of search key-words, this will allow automated search and retrieval of the relevant information. Once the relevant publication objects are retrieved, the interpretation of the information within them will be done manually. We hope to define all the fields of in-depth objects stringently, so that a person with general training in biology (such as undergraduate biology students) will be able to interpret and enter relevant data into the in-depth object fields.

### Analysis of the data

Search for patterns within existing data will involve applying a combination of specialized analysis tools (such as sequence comparison, protein folding) to the highly interconnected information space. Different analysis tools will be constructed from the set of available basic tools, the specific combination of basic tools applied will be defined by the problem under investigation. Ultimately, the integrated analysis package will be sufficiently modular to allow the user to construct a personal combination of the basic programs to approach a specific problem. We intend to develop analysis tools gradually, starting from simple sequence comparison programs called within WCS and eventually constructing multi-step analysis tools for more

sophisticated pattern searching. We will use existing analysis tools (GCG package, Genetic Computer Group, 1991) for individual steps of the analysis; the scaffold incorporating different application tools will be custom-written. Some of the fields within the in-depth gene object themselves require an analysis performed on the raw data (sequences); for example, finding all the regulatory domains in DNA and the proteins will require search program like BlastX or FastA (Pearson and Lipman, 1988) to compare the sequence of interest against all available regulatory consensus. We will use existing tools to perform recognition of acidic, hydrophilic, hydrophobic regions within proteins as well as for the formation of the secondary structure in DNA, RNA and proteins. PROSITE dictionary (Bairoch, 1992) along with databases such as EPD (Bucher, 1990), pkinases (Hanks et al. 1992), TFD (Ghosh, 1990) will be incorporated to aid in recognition of protein/DNA/RNA short diagnostic motifs. The stringency of the comparisons within the analysis tools will be 'tunable' and determined by the user. This is an essential point for recognizing less than perfect matches to the known motifs (a feature often lacking in the existing homology search tools) since the relevance of the identified sequence will be cross-correlated with other parameters. Weak sequence similarity are often avoided for they can be misleading by implying similar functions based on weak sequence similarity. Our system, however, uses sequence similarity as **one of many components** of the available types of information, since there are additional bits of data that can go into decision about function similarity (such as timing and localization of expression, co-expression with other known genes, etc.), weak sequence homologies can potentially be used. The process of considering multiple different components simultaneously (weak sequence homologies, expression, physical properties of the product), in turn, may identify a **subset of parameters** that are important for diagnosis of gene/protein functionality. Below we describe the type of analysis we intend to provide using a combination of basic analysis tools and in-depth objects within WCS.

- At the very basic level, the data stored within the in-depth gene objects can be analyzed manually. At the manual level, the researcher can access data from any of the fields of the gene object and correlate it with the similar data of the different gene(s). Comparison among genes can be performed by looking **simultaneously at several different fields** of a gene object. This is a laborious, but possible way to do manual cross-correlations between genes or gene parameters. For example, if the researcher wants to identify any DNA/protein sequences that may affect alternative splicing, he could search for all the genes in *C. elegans* that exhibit an alternative splicing (by checking the field *splicing*), then check for the presence of DNA/RNA sequences (by checking the field *cis-acting DNA, RNA sequences*) common to all or some of members of the

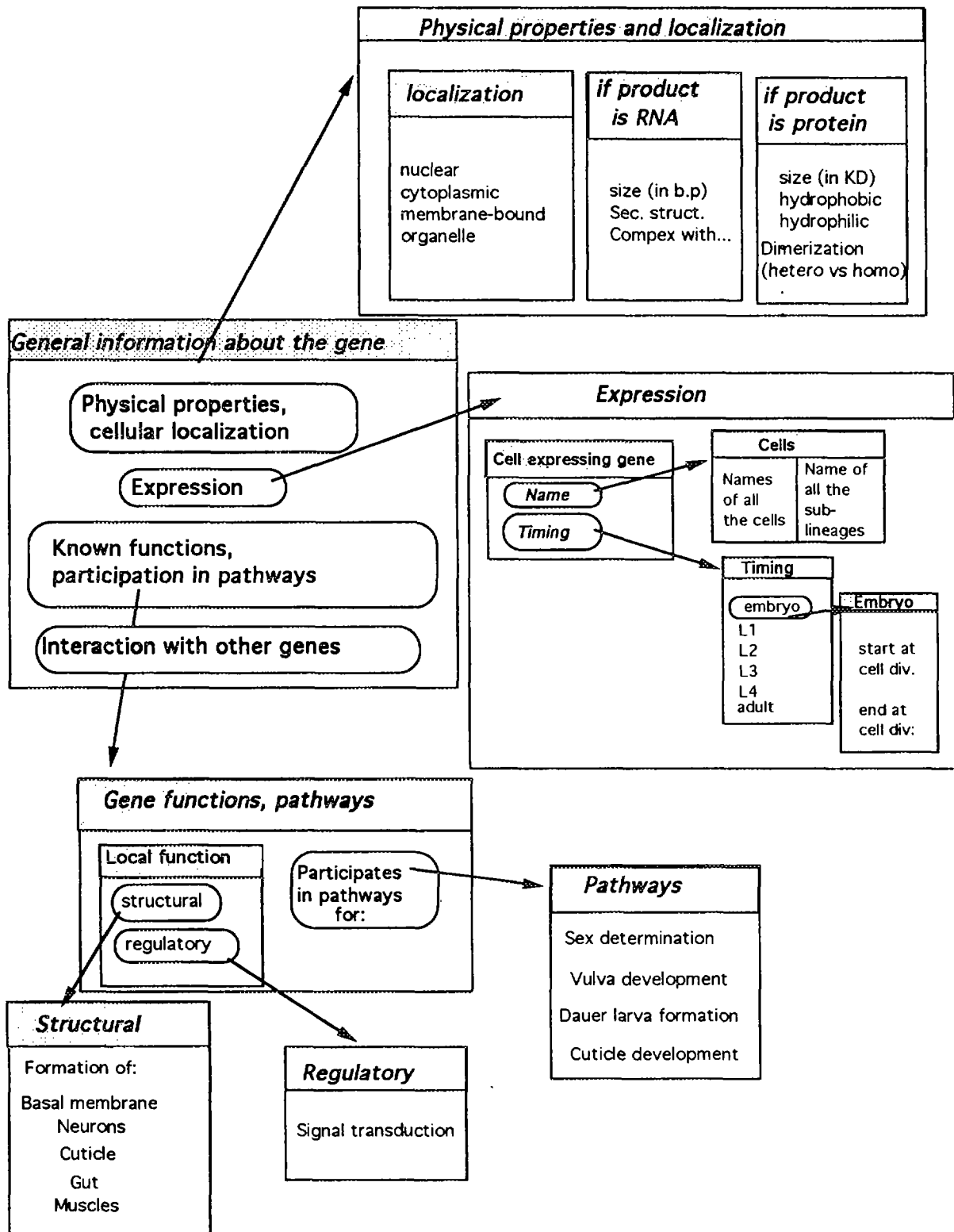


Figure 3. Conceptual subdivision of the information pertaining the gene object. Refinement of the *General information* concept.

alternative splicing subset. Finding **theoretical correlation** between phenomenon of alternative splicing and presence of a particular DNA sequence **does not necessarily prove functional correlations**, but presents a distinct possibility. Relevant experiments can be performed to test the hypothesis.

- The above process of a manual search can be automated, which will dramatically increase the speed of the search and will allow simultaneous correlation between a larger number of components. This type of search can allow rapid correlations between **unusual combinations** of the gene attributes as well as correlation among **large numbers of the parameters** simultaneously. For example, we can look for all the genes that possess the following set of attributes: presence of the leucine zipper domain in the protein product (Struhl, 1989), trans-splicing of the RNA message (Smith et al. 1989), expression during larval stages L1-L2 and co-expression with any other genes that contain upstream regulatory sequence 'CTGGTAA'. Correlation between multiple parameters currently requires a prolonged manual search through the huge amount of literature. It is often not feasible within the realm of 'wet biology'. Once the in-depth objects and simple analysis tools are established within WCS, interactive search of this type will take seconds.

- An important part of research with the in-depth objects will be the ability to locate **all possible regulatory motifs** within the DNA/RNA/protein. Greater sensitivity might be reached if instead of matching an entire sequence present in the databank to the sequence of interest, the match will be performed against a set of known short regulatory motifs. Identifying **all potential motifs** will point to the **potential levels of regulation and interactions** that may exist for the given gene and that have not (yet) been observed experimentally. Existence of such proposed regulation can be tested in a 'wet lab' environment. Increases in the potential number of strategies by which the gene can be regulated will lead to more sophisticated models of regulation and may also explain some of the already observed, but obscure phenomena. We intend to compile a database with all known regulatory domains across organisms similar to the idea of PROSITE (Bairoch, 1992); each motif will be associated with all the available information regarding this motif (such as stringency of homology, potential functions, domains it interact with, etc.). Every sequence within WCS will be searched against this database of reported motifs, and all potential motifs will be recorded within gene object.

- In-depth gene objects within WCS will present an ideal environment for finding potential information about the unknown sequence based on its similarity to the sequence of any of the existing gene objects. Such information as expression, function, physical properties, etc. of the gene homologous to the sequence of interest can hypothetically be extended to the sequence of interest, providing some additional information about the

sequence of interest. Thus, the data from the project such as EST can be easily and richly interpreted once the homology to the gene(s) within WCS is found.

- Another line of research that is possible within this system is finding similarities between genes not necessarily by **sequence comparison**, but by matching of **any subset of the gene's attributes**. For example, the user can define similarity between two gene products based on similarity in their cellular localization, co-expression within same set of cells and possession of the hydrophobic domain in the carboxy-terminal region.

- Regulatory network among interacting genes can be constructed by recursively traversing links established between interacting genes (gene object is linked to all the genes involved in interactions with it). This can be done by accessing fields, *Interaction with other genes*, for the gene of interest, and sub-field, *Type of control* (Fig. 2). A tool that interprets this information and graphically displays the resulting regulatory network can be developed. Such regulatory networks will be detailed and complete for they will include all the genes involved in the interactions, will present genes removed by several functional steps and will indicate the type of regulation (induction, repression, maintenance). Regulatory networks can be further abstracted to the interactions between regulatory domains on the level on DNA/protein sequence (by accessing field, *Domains*, within *Mechanism* level, Fig. 2). Comparison between different regulatory networks can be performed when searching for similar strategies in regulation.

- Finally, **analogies between different systems** regarding similarity in sequences, strategies of regulation, interactions among functional domains, etc. are frequently observed by the researcher and are brought up within the discussion section of the publication. Analogy search of the literature can be done on the level of the frequency of the terms' association in the text. We are intending to use the Worm Thesaurus tool which is being developed in collaboration with the researchers in the Management Information Systems (MIS) department, to search for the co-occurrence between a given term and all other terms within the literature. The Worm Thesaurus tool performs an analysis on an entire set of the literature (or any other data) and returns the concepts most commonly associated with the concept in the query. By tuning the frequency of the terms' association it is possible to use the Thesaurus tool as a very general search mechanism for **association between high level concepts** that may not be obvious if only the subset of available information is known (such as reading only a subset of publications regarding a gene/phenomenon). After frequency analysis has been performed, the literature objects mentioning the term of interest that is associated with query term, can be tagged and accessed by the user for further manual analysis. High level analogies can be tested then on lower levels by performing specialized search of the in-depth objects



in the database. Consider the following hypothetical example: the query is performed on the term 'heterochronic'. The term 'leucine zipper' comes up among others that are associated with term 'heterochronic'. The association is surprising to the researcher and can be interpreted as heterochronic gene(s) having a 'leucine zipper' domain. An alternative interpretation is that the gene(s) interacting with the heterochronic gene(s) possess a 'leucine zipper' domain. Accessing the tagged literature will clarify this question. The information provided within the literature can be now tested on a more general level by checking all the heterochronic genes or all the genes interacting with heterochronic genes for the presence of the 'leucine zipper' domain. If indeed it is the second possibility (the genes that interact with the heterochronic genes possess a 'leucine zipper' domain, then all the genes within the database that possess 'leucine zipper' can be tested for the possibility of interaction with heterochronic genes. This can be theoretically tested by checking pattern and timing of expression of all the 'leucine zipper' containing genes and the heterochronic genes. Also, all the genes containing 'leucine zipper' can be compared to the set of genes that are known to interact with heterochronic genes in search for other similar features between them.

The above examples rely on a highly interconnected information space; presence of knowledge and analysis tools from many different sources. Our current version of WCS supports rich and flexible connections among the literature sources. We are currently working on integrating sequence analysis tools into WCS.

## References

- Bairoch, A. 1992. PROSITE: a dictionary of sites and patterns in proteins. *Nucleic Acids Research* 20: 2013-2018.
- Bucher, P. 1990. The Eucaryotic Promoter Database EPD. *EMBL Nucleotide sequence Data Library Release* 33, Postfach 10.2209, DD6900 Heidelberg.
- Edgley, M.L. and Riddle, D.L. 1990. *Genetic Maps* 5: 3.
- Genetic Computer Group Inc. 1991. GCG Package, Version 7.
- Ghosh, D. 1990. A relational database of transcription factors. *Nucleic Acids Research* 18(7): 1749-1756.
- Hanks, S., Quinn, A.M., Hunter, T. 1992. Protein kinases catalytic domain database references. The Salk Institute for Biological Studies.
- Lander, E., Langridge, R., and Saccocio, D. 1991. Computing in Molecular Biology; Mapping and integrating biological information. *Computer* 11: 6-13.
- Pearson, W. and Lipman, D. 1988. Improved tools for biological sequence analysis. *Proceedings of the National Academy of Sciences USA* 85: 2444-2448.
- Roberts, L. 1990. The Worm Project. *Science* 248: 1310-1313.
- Schatz, B. and Caplinger, M. 1989. Searching a Hyperlibrary. Proc 5th IEEE Int. Conf. on Data Engineering, Los Angeles (Feb.), 188-197.
- Schatz, B. 1991-92. Building an Electronic Community System. *J. Management Information Systems* 8: 87-107.
- Smith, C., Patton, J., and Nada-Ginard, B. 1989. Alternative splicing in the control of gene expression. *Annual Review of Genetics* 23: 527-577.
- Struhl K. 1989. Helix-turn-helix, zinc-finger, and leucine-zipper motifs for eukaryotic transcriptional regulatory proteins. *Trends Biochem Sci* 14: 137-140
- Wood, W., and the community of *C. elegans* Researchers eds. 1988. *The nematode Caenorhabditis elegans*. Cold Spring Harbor Laboratory.