A Database for Globin Gene Expression Data

Webb Miller, Ahmed ElSherbini, Jamie Peck, Cathy Riemer,
Scott Schwartz, Nikola Stojanovic and Ross Hardison[1]
Departments of Computer Science and Engineering and
[1]Biochemistry and Molecular Biology
The Pennsylvania State University
University Park, PA 16802

Abstract

We describe a prototype database of sequence alignments and experimental results for the β-like globin gene cluster of mammals. This data repository is intended to help the international community of globin-gene biologists to plan experiments, to design models and, ultimately, to understand regulation of those genes. Moreover, the approaches and software developed for this project will be applicable to other sequence-analysis problems.

Our first steps were to develop a program that can simultaneously align a few very long sequences and to establish an e-mail server that formats any requested portion of the alignment, annotated to indicate highly-conserved regions and known sequence features. We are currently developing a World-Wide Web server that will provide access to the wealth of pertinent experimental data, in register with the annotated alignment, through a variety of query languages and graphical presentations.

1 Introduction

Molecular biology is facing an "informatics crisis". Data is being generated faster than it can be fully utilized with available tools. Dozens, and probably hundreds, of data collections have been established, ranging from narrowly-focused lists of a few hundred short sequences to comprehensive collections of all known DNA sequence data. Nucleic Acids Research [20] samples the available resources, and Letovsky and Berlyn [18] give an insightful discussion of the area.

We are building a biological database with several novel features. This database captures a variety of information related to the regulation of expression of genes in the β-like globin cluster of humans and other mammals, which we study for reasons described elsewhere [19]. Perhaps the most novel aspect of this database is the central role played by sequence alignments. Alignments have long served as an effective mechanism (indeed, the most effective tool) for exploring the function of DNA sequences. However, the alignments we use are noteworthy for their sheer size and for the way we are integrating them with data from other sources. When fully implemented, this database will allow users to view both the alignments and the experimental data in register, which is another novel feature of the database. This computer-aided, integrated view of alignments and experimental data, which we call electronic genetic analysis, allows users to more effectively plan experiments based on the full range of previous work.

Many of the underlying problems addressed by our project, in particular coordinated representation of heterogeneous data sets, are faced by other biological databases. We feel that our efforts are justified by both the importance of these particular data and by the potential of applying the ideas and software developed here to a variety of other tasks.

2 Biological Motivation

Almost any application of modern genetics, whether in medicine, agriculture, or industry, involves understanding regulation of a particular gene. Expression of a gene results in the synthesis within the cell of a product (protein or RNA) that carries out some physiological process. The pathway for gene expression includes activation of a chromatin domain that contains the gene, transcription of the gene by RNA polymerase, processing, splicing, transport of the mRNA into the cytoplasm, and translation of mRNA into protein. For many proteins, post-translational events such as cleavage or covalent modification are required to generate a functional product. Gene expression can be regulated at virtually any stage along the pathway, although for many genes the steps leading to initiation of transcription are the key control points. Thus the study of the regulation of gene expression (or gene regulation) has focused largely on the biochemistry of transcriptional initiation, which involves among other steps the binding of certain proteins (called transcription factors) to the DNA at particular sites and various protein-protein interactions.

In the case of β-like globin genes, the DNA sites involved in gene regulation are spread throughout a very large gene cluster. As illustrated in Figure 1, the human cluster has five active genes, ε, ${}^{G}\gamma$, ${}^{A}\gamma$, δ, and β, in a segment of about 75 kb (1 kb = 1000 nucleotides). The ε-globin gene is expressed in embryonic red cells, the γ-genes are expressed in fetal red cells, and the δ and β-globin genes are expressed in adult red cells, with β-globin being the major adult product. All these genes are in a domain of about 200 kb on the chromosome which is active in red cells (but not in other types of cells) throughout development, as shown by increased accessibility of the DNA in this region to DNaseI.

Within the active domain are many regulatory regions, some close to the genes (promoters and upstream regulatory sites) and some very far away. A locus control region (LCR) is located at the $5'$ end of the gene cluster, perhaps close to the start of the active domain. The LCR plays a key role in domain opening, insulation from effects of adjacent sequences, and enhancement of expression in a developmentally regulated manner. It is composed of at least five regions that are highly accessible to DNase I in chromatin (referred to as hypersensitive sites, or HSs); each of these HSs is a set of binding sites for specific proteins. The protein-DNA complexes at the LCR are thought to interact with the proteins bound to the proximal regulatory regions (e.g. promoters) to activate expression of individual genes [11]. Even more sequences besides the LCR, upstream regulatory sites and promoters, may be required for physiological regulation of the globin genes. Full regulation and "normal" developmental switching is observed when the entire gene cluster is placed in transgenic mice, whereas constructs containing just the LCR and a globin gene show some aberrations in developmental regulation[24, 21]. This indicates that regulatory sequences could be found throughout the gene cluster, so there are potentially 75 kb (LCR though β-globin gene) to 200 kb (approximate size of the chromatin domain) to contend with.

Regulatory elements are discovered by several complementary approaches. One is to search for evolutionarily conserved sequences. Sequences that are involved in regulation, e.g. by binding a particular protein, are not as subject to change over evolutionary time as are sequences with no precise role in encoding or expressing a gene, so they should be changing at a noticeably slower rate – i.e. they will be conserved. For example, two common components of promoters (TATA and CCAAT motifs) as well as the first enhancer of a cellular gene (immunoglobulin gene enhancer) were initially recognized by their conservation in comparisons of gene sequences [5, 7]. This general approach has been successful in examining promoter regions of globin genes. The conserved regions, or "phylogenetic footprints," have proven to be useful guides to sequences that specifically bind proteins and regulate expression [12, 13, 14, 25]. However, not every sequence that fits some objective criterion for "conserved" needs to be involved in regulation. How does one test for function of these conserved sequences?

Experimental tests of the roles played by these conserved sequences are critical to building an understanding of gene regulation. Analysis of naturally occurring mutations that cause thalassemia (inherited anemias due to an absence of one type of globin) has revealed many important sequences for expression and regulation, and has provided the clearest evidence that the LCR is required for opening a chromatin domain [8]. However, such thalassemic mutations are rare (fortunately) and of course not under the control of the investigator. Sites of DNaseI hypersensitivity are almost always involved in the structure or regulation of a gene or its chromatin domain [10], and mapping these sites provides a key initial step in examining a regulatory region [26]. Once a region has been identified as a candidate for a regulatory function, it can be mutated *in vitro* and the effects tested on expression. The mutated regulatory region *in cis* to an appropriate reporter gene can be placed in transgenic mice. This potentially allows the full range of developmental and tissue-specific control to be examined, as long as position effects do not obscure the view. Although a wealth of information can be obtained from a set of lines of transgenic mice for each mutant construct, the expense and labor of complete analysis precludes this as a general screening tool. The mutant construct can be placed in cultured cells to examine the effect on expression prior to integration into the host chromosome (transient expression) or after integration and selection for a co-transfecting marker

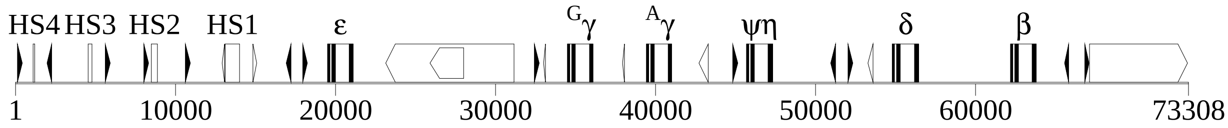

Figure 1: Schematic diagram of the human β-like globin gene cluster. The diagram encompasses the region covered by GenBank entry HUMHBB. Each of the five active genes and one inactive gene ($\psi\eta$) is indicated by a box containing three vertical bars, corresponding to the gene's three exons. Four DNaseI-hypersensitive sites are labeled HS1-HS4. Filled triangles denote Alu repeat sequences, and unfilled triangles and pointed rectangles depict L1 repeats.

(stable expression). However, the cell line(s) chosen cannot adequately mimic all aspects of globin gene regulation, and the problems with position effects and labor of screening are also found in this approach. Identification of specific binding proteins can be done more quickly using common techniques such as electrophoretic gel mobility shifts, DNaseI footprints, and methylation interference. These *in vitro* assays do not necessarily tell you which proteins are binding *in vivo*, and are highly skewed toward showing the most abundant proteins in an extract. These problems can be overcome by *in vivo* footprinting, although in these cases one only identifies the nucleotides that are interacting with some protein – a positive identification of the protein is not possible.

Although each of these approaches has its limitations, in combination they are a formidable arsenal of techniques. However, full utilization of the information requires that all data on a particular region be viewed in register. For example, is the region that binds an abundant nuclear factor well conserved, and is it included in a deletion or substitution that decreases the amount of expression? Such a compilation can be done readily for a short sequence, but to accomplish this for 75 kb is beyond the memory capacities of most investigators. We are developing the Globin Gene Server to provide computer-aided access to this information for the research community.

3 Critical Characteristics

The application of computers to assist investigations of the regulation of human globin genes differs in several important ways from other problems in computational molecular biology, and these factors have greatly influenced our database design. In some respects, our requirements are simpler than those in related areas, while in other respects they are more demanding.

A central aspect of our work is the need to compute high-quality simultaneous alignments of a few, very long sequences. The number of sequences involved is relatively small, because unlike sequence comparisons that span a wide range of organisms, detailed analysis of regulation of the human globin genes can probably be conducted using DNA data from mammals only. Evolution of these genes in non-mammalian animals occurred independently of that in humans, so it is reasonable to anticipate that sequence regions conserved across that evolutionary distance will not reveal subtle aspects of human globin switching. In any case, DNA sequences from, say, humans and chickens do not show measurable sequence conservation in the relevant regulatory regions [22]. On the other hand, each sequence needs to be longer than is typical for other projects; analysis of long sequences is necessary because of interactions between regions separated by more than 50,000 basepairs. Packaged alignment software of adequate sensitivity for our needs is incapable of handling sequences of anywhere near that length. Perhaps surprisingly, computer memory space, rather than computational time, is the constraining factor that makes these packages unsuitable for this project.

The corresponding experimental data that we need to handle is currently relatively modest in amount (compared to e.g. all the sequences in GenBank). Determination of this data is extremely labor-intensive; several months of effort can frequently be summarized by just a few hundred bytes of information. Thus the volume of accumulated data will probably remain comparatively small, by database standards.

However, the structure of the data is quite complex, since a variety of experimental protocols are used and results are often expressed in units that make it difficult to compare two sets of experiments. One of the goals of our project is to tie together several types of data into a cohesive, integrated view of these regions. Our database will link data from sequences, alignments and conserved regions within those sequences, descriptive annotations of the regions' functions compiled from a variety of sources, specific experimental results with citations from the literature, and even pictorial views of the sequences and alignments.

Other factors driving our software design include the requirements and distribution of our potential user community. These biologists are geographically dispersed and typically might need only sporadic access to the data that we are collecting. Moreover, most individuals in the community do not have extensive expertise with computer technology. Thus, although many other projects have rightly focused on development of sophisticated graphical user interfaces for molecular biologists, e.g. [4], and the technology exists to make such tools portable across a wide range of hardware [29], it seemed ill-advised for us to expect that our work could be effectively delivered by a system that required installing and operating special-purpose software.

4 Initial Steps

We have devoted considerable effort to developing sequence-alignment software that suits our needs. As mentioned above, the main hurdle to be overcome concerned the space requirements of existing alignment software. Ultimately, we produced programs capable of aligning a 73,000-character sequence containing critical parts of the β-like globin gene cluster of humans with matching sequences from a few other mammals. To do this, we further developed an algorithmic idea proposed by Hirschberg [17]. Details of our methods are given in [6] and by references therein; we do not describe them here.

The computation of such an alignment, including preliminary calculations to determine which regions of sequence from other mammals should enter into the final multi-species alignment, involves hours of CPU time on currently available workstations. We have stored the resulting alignment in a compressed form, from which any selected region can be quickly formatted in a variety of styles. Portions of the alignment are annotated in various ways that are incompatible with typewriter-style output, so we have written a program that draws the alignment using the PostScript page-layout language, according to a flexible set of formatting commands. Figure 2 gives an example. Alternatively, the user can simply click the mouse on a picture like Figure 1.

PostScript output is well suited to the needs of our potential user community because it is both powerful and widely supported. Sending a simple electronic-mail message of the form "`send postscript 8640 8790`" (where in general the desired starting and ending addresses are given) to our server at `globin@globin.cse.psu.edu` gets an image like Figure 2. For those without access to a PostScript printer (or wishing to process the alignment with other programs), a plain-text version is obtained by replacing `postscript` by `ascii` in the above command, although this will sacrifice some of the annotation. Alternatively, we will supply hard-copy alignments to interested parties. A more detailed description of various aspects of this e-mail server is given in [16], though automatic annotation from the Transcription Factor Database [9] was added subsequently.

5 The User's View of the Prototype Database

The World-Wide Web [2, 3], abbreviated "WWW", provides a mechanism for accessing our database that seems well-suited to our potential user community. Anyone with access to one of several popular computer types and an Internet connection can browse a wealth of information that includes our nascent database. Our WWW URL (i.e., address) is `http://globin.cse.psu.edu/`. Other WWW resources of particular interest to our users include the major biosequence databases (e.g., GenBank) and MedLine summaries of medicine-related publications. All these resources are tapped using a single, intuitively appealing graphical interface. Biologists are flocking to the WWW community, e.g. [28], and we expect the trend to continue.

A few words and pictures should convey the flavor of what it's like to interact with our WWW server at the time of this writing, but please keep in mind that because this project is still in the development stage, it is undergoing frequent improvements and enhancements, so it may look a bit different by the time you try it out. The precise visual presentation depends on the user's particular WWW "client" program; Figure 3 was made using the Mosaic program for X Windows on a Sun SparcStation. The user is first presented with a view like Figure 3. The computer's mouse can be used to select any of nine options, which appear in Figure 3 as underlined phrases. For example, clicking on "alignments of those sequences" summons up screens that allow the user to pick the

```
                         |            |  NFE2     |  NFE2     |  nfe2     |           |  CACBP    |  E box
 8640:  CTGGCTCAAGCACAGCAATGCTGAGTCATGATGAGTCATGCTGAGGC----------TTAGGGTGTGTGCCCAGATGTTC   human
 4966:  ....TC.C....G....G...........C.........T....T.ATGCTGTT--........G...G.........T   rab.lcr
 5623:  .CAC...C.A......C.......A.....C..............T.ATGCTGAGGC...........G..........   goat.lcr
12028:  .A....TG.........G............C................----------...........G.........T   mouse
                  match: 1        match: 2-4  match: 5-7                             match: 8

                         |  gata1     | GATA1     |           |           |    YY1    |
 8710:  TCAGCCTAGAGTGATGACTCCTATCTGGGTCCCCAGCAGGATGCTTACAGGGCAGATGGCAAAAAAAAGGAGAAGCTGAC   human
 5044:  ......CT..........AG.........G..........A........AA.........G.....-......T....   rab.lcr
 5703:  ......CT.CC..T....CT....................G.........A.....TA........--.....ATAC..   goat.lcr
12098:  .....TGT.......C.G.G...........T.T..G....A.TCC......A..G..AA..G...----.T...T.TG.   mouse
                             match: 9-12                         match: 13
```

Figure 2: A portion of an alignment covering 73,000 nucleotides (letters) of DNA sequence from the β-like globin gene cluster of humans and a few other mammals. The particular region is called "hypersensitive site 2" (HS2). Each dot denotes the letter in the corresponding position of the human sequence. Boxes are automatically drawn around regions meeting a user-defined criterion for "conserved", in this case blocks of length at least six with at most one mismatch per column. Annotations above the alignment and associated single underlines indicate sequence features entered manually. Automatically-determined regions that match consensus binding sites for entries of the Transcription Factor Database are indicated by e.g. "match:2-4" below the alignment and by double underlines; more information, including literature citations, is given on additional pages of output (not shown).

region to be drawn and to specify formatting options.

Another option on the screen of Figure 3 is to access "a database of experimental results" about the β-like globin genes. Two query languages are currently provided. One is a typical "relational calculus" language (illustrated by Figure 7), and the other is based on (and implemented in) Prolog. Prolog supports very general kinds of database queries [27]. As an alternative to queries in these languages, which, frankly, we expect most biologists will find daunting, the user can elect to fill out a simple form to select desired information.

Frequently, users will want a graphical summary of results extracted from the database; we will soon be implementing the ability to generate a PostScript file like Figure 4 in response to user instructions. The display in Figure 4 shows in register the conserved sequences (top line), known or hypothesized binding proteins and other markers (bottom line) and effects of mutations on the ability of reporter genes to be expressed (lines in the center of the figure). As an example of electronic genetic analysis, the viewer can determine which conserved regions or binding sites fall within constructs that have been mutated but retain some of the functions of the LCR.

Among the other options available in Figure 3 is access to a collection of documents describing available services. This currently includes descriptions of the available query languages, directions for preparing and submitting data, instructions for obtaining a program that users can run locally to facilitate data entry, and several pictures summarizing various aspects of the globin gene cluster.

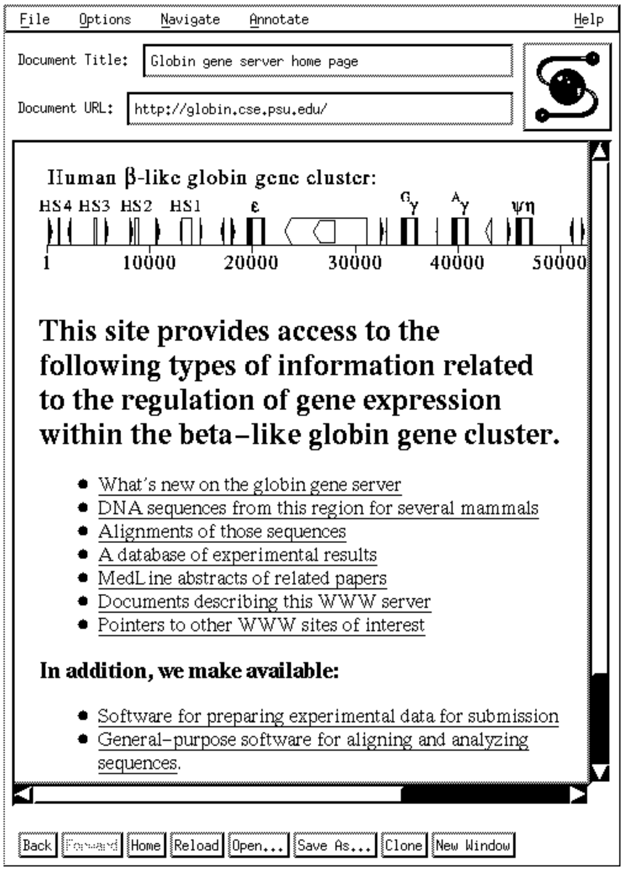

Figure 3: Initial view provided by our World-Wide Web database server.

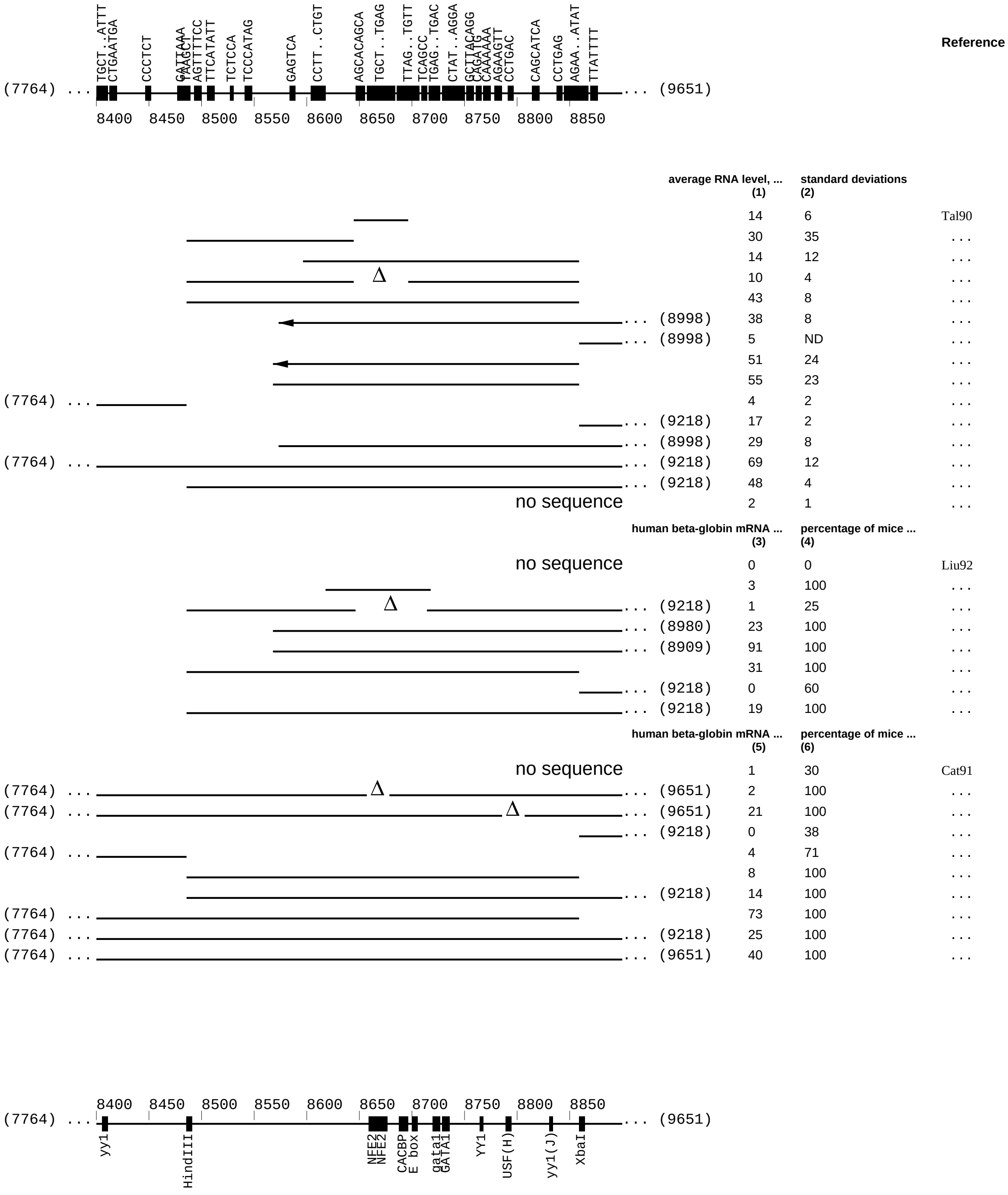

Figure 4: Graphical summary of some expression data. The top line shows positions and consensus sequences for the conserved regions that are enclosed in boxes in Figure 2. The bottom line indicates the manually-entered sequence features that are singly-underlined in Figure 2. Between them are found a number of graphical summaries of DNA transfer experiments; a horizontal line indicates the location of a segment of DNA (perhaps with an internal deletion, indicated by "Δ", and perhaps in reverse orientation, indicated by a left-pointing arrow-head), followed by numerical results of experiments and a literature citation. Additional pages of output (not shown here) give more complete details of the experiments and full literature references. The picture was generated automatically by a prototype program for communicating responses to database queries.

6 Implementation of the Prototype Database

The rapidly-changing nature of data in typical biological databases (and many other databases, as well) mandates an approach that permits the storage format (which is technically referred to as the "database schema") to be readily altered as the character of the stored data evolves. It must be possible to change this schema (at least, within prescribed limits) without changing the database software. Figure 5 gives a snapshot of part of our database schema at an early stage of development. New data elements, or even entire tables, can be added without changing the database software, since the software reads the schema and adapts its behavior accordingly.

For example, Figure 5 specifies that some facts ("tuples" in database jargon) are of type "DNA_transfer_experiment". These facts have elements (known as "attributes") that include "location" (`REGION` is a special data type that permits specification of any segment of any of the aligned sequences), a "reference" pointing to an element of another table (the table is also named "reference"), etc. The general meaning of such table specifications should be immediately clear to anyone familiar with the basics of "relational" databases, though the use of attributes that are a `SET_OF` simpler items (e.g., "mutation" and "results" in Figure 5) departs somewhat from the classical relational approach.

Data (as opposed to the schema, which specifies the "shape" of the data) can be entered in a simple format (Figure 6). We expect that after reading a brief tutorial and a few encounters with our software (which will reject improperly specified data) biologists will find it easy to prepare data in this manner.

In addition to data access via "Query-by-Forms" and Prolog, a more-or-less standard relational language is being provided. Figure 7 shows a sample query in this language. We prefer a language patterned after the Quel language of the Ingress database system, but it would be a minor job to switch to the more traditional SQL syntax. Our implementation of the query processor, written in C, makes good use of the *byacc* parser generator, which is a freely-available version of the UNIX *yacc* program; formal grammars are specified for schema, for data, and for queries.

We have benefited from code-sharing among our tools. In particular, several of our tools use the query processor's code to access the schema and data; this includes the program that generates query forms directly from the database schema, the program that

```
DNA_transfer_experiment:  SET_OF {
   location:  REGION
   description:  OPTIONAL STRING
   reference:  OPTIONAL FOREIGN_KEY reference
   corrections:  OPTIONAL STRING
   contributor:  STRING
   orientation:  OPTIONAL (normal, reversed)
   assay:  (stable, transient, transgenic_mouse)
   cell:  OPTIONAL (K562, MEL, HeLa, 3T6)
   reporter:  STRING
   mutation:  SET_OF {
      mutation_location:  REGION
      mutation_type:  (replacement, substitution,
         insertion, deletion)
      mutation_description:  OPTIONAL STRING
   }
   result:  SET_OF {
      result_value:  STRING
      result_units:  STRING
   }
}
```

Figure 5: One table specification from an early schema for our database.

prepared Figure 4, and a program that converts our database into facts for the Prolog programming language.

7 Discussion

Several aspects of our database design deserve further discussion. Our decision against using commercial database software is a case in point, given the current popularity of Sybase for small biology databases. While we envision no use for "back-end" database functionality (such as simultaneous access by several users making updates and access-restriction mechanisms beyond those provided by the file system), Sybase (or an equivalent brand of database software) would have given a query mechanism and remote access capability. However, the loss from not adopting Sybase is relatively minor. Currently we have no proof that we or other users will regularly want to perform complex queries that span several databases. (For local use, we can implement complex modes of data access using Prolog or by writing special-purpose C programs.) Furthermore, writing our own query handlers for a useful subset of Quel (or equivalent SQL) is straightforward.

Our custom-built system has a number of advan-

```
BEGIN DNA_transfer_experiment
   location = (humhbb_orig, 8568, 8909)
   description = 'BalI to SnaI fragment of 5
      prime HS2 of human LCR was linked
      to the human beta-globin gene and
      microinjected into mice; the level
      of human beta-globin mRNA and mouse
      alpha-globin mRNA expression was
      measured in 16 day fetal liver'
   reference = Liu92
   contributor = 'Ahmed ElSherbini'
   assay = transgenic_mouse
   cell = NULL
   reporter = 'human beta-globin gene'
   BEGIN result
      result_value = 91
      result_units = 'human beta-globin
         mRNA in 16 day fetal liver,
         mean expression per copy of
         the transgene percentage
         relative to endogenous mouse
         alpha-globin mRNA'
   END result
   BEGIN result
      result_value = 100
      result_units = 'fraction of mice that
         expressed human beta-globin mRNA'
   END result
END DNA_transfer_experiment
```

Figure 6: A database entry describing one experiment.

```
range of x is binding_assay
range of r is reference

retrieve (r.title)
where
   r.year >= 1990 and
   r.id = x.reference and
   x.assay = methylation_interference and
   matches(consensus(x.location), "WGATAR")
```

Figure 7: A query in our Quel-like language. The query requests the titles of all papers published since 1990 that discuss a methylation interference experiment involving a conserved region that matches the pattern "WGATAR,", which is recognized by the DNA-binding protein GATA1. Built-in functions called `matches` and `consensus` are invoked; the latter accesses data on conserved regions that is generated by the program that prepared Figure 2.

tages. We are convinced that users would much prefer access to our database via World-Wide Web over, say, access via Sybase client-server functionality, for the reasons discussed by Schatz and Hardin [23]. Advantages for us include ease of handling *ad hoc* features of the query language which are specific to genetic sequence data (such as conversion of sequence addresses between various "coordinate systems") and code reuse (as mentioned above). Also, we are not constrained to use a standard data model or query language, which is a definite advantage here since the nature of our data is not yet completely understood. (We are currently using what is called the "nested relational data model" [1].) Moreover, keeping data as plain-text allows us to use UNIX tools/languages to manipulate it (e.g. when a schema change forces wholesale data changes) and obviates potential problems with making a complete version of the current database available to any interested party.

There is a downside to use of WWW that needs to be considered. One issue is that the system is evolving rapidly. At the time of this writing, WWW clients for the MacIntosh and PC are not as mature, stable, or up to date as the UNIX ones. A second issue is that the protocol is stateless; each query is an entirely independent transaction with the server. While this simplifies the system, it makes it difficult to support a session involving a number of queries with shared state between them. For menu-driven queries to access the database or to customize display of alignments, a nontrivial amount of work was required to arrange for the previous entries to appear as defaults in anticipation of the next query. A final concern is security. Any powerful network service should be treated with suspicion, and WWW is no exception. Because a WWW server reads files and runs programs on the host system in response to uncontrolled input it must be very careful to validate incoming data before acting on it. A common source of vulnerability with systems like WWW and gopher involves quoting input in such a way that if it is passed as a command line argument to a shell command, the argument will be unexpectedly evaluated by the shell. We seek to avoid that problem by stripping troublesome characters from the input stream and also taking care that the shell is never used to execute programs external to the server. Similarly, one must be careful of those programs which are executed by the server. An early version of our relational query engine had an operator that would write results to a file. Clearly such a thing cannot be permitted in a publicly accessible anonymous service.

The Globin Gene Server (GGS), as proposed, will require two workstations for normal operation. One of these will be used solely as a computation host for the globin server, while the other will be used for data entry and other administrative functions. These two machines will be run in production mode, and dedicated to the GGS project. Our use of two machines rather than one is motivated by security concerns. Publicly announced repositories are natural targets for vandalism. To minimize our risk of exposure, we will run only globin related network services (ftp, WWW, smtp) on the server machine, and none of the normal user oriented UNIX services (like rlogin, NFS, X11). In the past each of these has been the source of a number of security problems; a cautious approach is to assume that the publicly accessible server machine will be compromised at some point, and to arrange that it is not trusted by either ourselves or our other machines. That means administrative functions like processing updates and installing new software will require a second, trustworthy, machine to be available.

8 Future Enhancements

As discussed above, our prototype database is evolving rapidly, and new features and capabilities are being added continuously. The direction we ultimately take with this project will depend on the needs of our user community.

A variety of additional types of data related to globin gene regulation are being considered for possible addition to the Globin Gene Server. For example, data concerning naturally occurring mutations, such as those causing thalassemia and hereditary persistence of fetal hemoglobin, are natural choices. Though less similar to the currently encompassed experimental results, data on concentrations of transcription factors as functions of tissue and/or developmental stage are also candidates for inclusion. It might even prove desirable to add structural data, alignments, and graphical representations of transcription factors to our WWW node. A more mundane, though perhaps more immediately useful type of information reports the availability of cloned DNA sequences from the globin gene clusters.

Another planned enhancement concerns the alignment of DNA sequences that underlies the Globin Gene Server. Although we have made a concerted effort to produce a meaningful alignment, there is no doubt room for improvement. It is possible, and perhaps desirable, to let users either edit the alignment manually or recompute portions (or all) of the alignment with different alignment parameters, additional sequence data, or alternative alignment algorithms.

9 Acknowledgement

This work was supported by PHS Grants R01 LM05110 to W.M. and R01 DK27635 and R01 HL44491 to R. H.

References

[1] Abiteboul, S., P. Fischer and H.-J. Schek (eds.), *Nested Relations and Complex Objects in Databases,* Lecture Notes in Computer Science 361, Springer-Verlag, 1987.

[2] Berners-Lee, T., R. Cailliau, J.-F. Groff and B. Pollerman, "World-Wide Web: the information universe," *Electronic Networking* 2:52-58, 1992.

[3] Berners-Lee, T., R. Cailliau, A. Luotonen, H. Nielsen and A. Secret, "The World-Wide Web," *Comm. ACM* 37:76-82. 1994.

[4] Boguski, M., R. Hardison, S. Schwartz and W. Miller, "Analysis of conserved domains and sequence motifs in cellular regulatory proteins and locus control regions using new software tools for multiple alignment and visualization," *The New Biologist* 4:247-260, 1992.

[5] Breathnach, R., and P. Chambon, "Organization and expression of eukaryotic split genes encoding for protein." *Annual Review of Biochemistry* 50:349-383, 1981.

[6] Chao, K.-M., R. Hardison and W. Miller, "Recent developments in linear-space alignment methods: a survey." To appear in *J. Computational Biology*, 1994.

[7] Emorine, L., M. Kuehl, L. Weir, P. Leder and E. E. Max, "A conserved sequence in the immunoglobulin Jk-Ck intron: possible enhancer element," *Nature* 304:447-449, 1983.

[8] Forrester, W. C., E. Epner, M. C. Driscoll, T. Enver, M. Brice, T. Papayannopoulou and M. Groudine, "A deletion of the human β-globin locus activation region causes a major alteration in chromatin structure and replication across the entire β-globin locus," *Genes Develop.* 4:1637-1649, 1990.

[9] Ghosh, D. "New developments in the transcription factor database," *Trends in Biochemistry* 16:445-447.

[10] Gross, D., and W. Garrard, "Nuclease hypersensitive sites in chromatin," *Ann. Rev. Bioch.* 57:159-197, 1988.

[11] Grosveld, F., M. Antoniou, M. Berry, E. de Boer, N. Dillon, J. Ellis, P. Fraser, O. Hanscombe, J. Hurst, A. Imam, M. Lindenbaum, S. Philipsen, S. Pruzina, J. Strouboulis, S. Raguz-Bolognesi and D. Talbot, "The regulation of human globin gene switching," *Phil Trans. R. Soc. Lond. B* 339:183-191, 1993.

[12] Gumucio, D. L., H. Heilstedt-Williamson, T. A. Gray, S. A. Tarle, D. A. Shelton, D. Tagle, J. Slightom, M. Goodman and F. S. Collins, "Phylogenetic footprinting reveals a nuclear protein which binds to silencer sequences in the human γ and ε globin genes," *Mol. Cell. Biol.* 12:4919-4929, 1992.

[13] Gumucio, D. L., D. A. Shelton, W. J. Bailey, J. L. Slightom and M. Goodman, "Phylogenetic footprinting reveals unexpected complexity in trans factor binding upstream from the ε-globin gene," *Proc. Natl. Acad. Sci. USA* 90:6018-6022, 1993.

[14] Hardison, R., K.-M. Chao, M. Adamkiewicz, D. Price, J. Jackson, T. Zeigler, N. Stojanovic and W. Miller, "Positive and negative regulatory elements of the rabbit ε-globin gene revealed by an improved multiple alignment program and functional analysis," *DNA Sequence* 4:163-176, 1993.

[15] Hardison, R., and W. Miller, "Use of long sequence alignments to study the evolution and regulation of mammalian globin gene clusters," *Mol. Biol. Evol.* 10:73-102, 1993.

[16] Hardison, R., K.-M. Chao, S. Schwartz, N. Stojanovic, M. Ganetsky and W. Miller, "Globin Gene Server: a prototype E-mail database server featuring extensive multiple alignments and data compilation for electronic genetic analysis," *Genomics* 21:344-353, 1994.

[17] Hirschberg, D. S., "A linear space algorithm for computing maximal common subsequences," *Comm. ACM* 18:341-343. 1975.

[18] Letovsky, S. I., and M. B. Berlyn, "Issues in the development of complex scientific databases," *Proc. 27th HICSS*, 5-14, 1994.

[19] Miller, W., S. Schwartz and R. Hardison, "A point of contact between computer science and molecular biology," *Computational Science and Engineering* 1:69-78, 1994.

[20] Special issue on databases. *Nucleic Acids Research* 21(3), 1993.

[21] Peterson, K., C. Clegg, C. Huxley, B. Josephson, H. Haugen, T. Furukawa and G. Stamatoyannopoulos, "Transgenic mice containing a 248-kb yeast artificial chromosome carrying the human β-globin locus display proper developmental control of human globin genes," *Proc. Natl. Acad. Sci. USA* 90:7593-7597, 1993.

[22] Reitman, M., J. Grasso, R. Blumenthal and P. Lewit, "Primary sequence, evolution, and repetitive elements of the *Gallus gallus* (chicken) β-globin cluster," *Genomics* 8:616-626, 1993.

[23] Schatz, B., and J. Hardin, "NCSA Mosaic and the World Wide Web: global hypermedia protocols for the Internet," *Science* 265:895-901, 1994.

[24] Strouboulis, J., N. Dillon and F. Grosveld, "Developmental regulation of a complete 70-kb human *beta*-globin locus in transgenic mice," *Genes Devel.* 6:1857-1864, 1992.

[25] Trepicchio, W., M. Dyer and M. Baron, "Developmental regulation of the human embryonic β-like globin gene is mediated by synergistic interactions among multiple tissue- and stage-specific elements," *Mol. Cell. Biol.* 13:7457-7468, 1993.

[26] Tuan, D., W. Solomon, Q. Li and I. M. London, "The β-like globin gene domain in human erythroid cells," *Proc. Natl. Acad. Sci. USA* 82:6384-6388, 1985.

[27] Yoshida, K., C. L. Smith and R. Overbeek, "A primer on rapid prototyping of genomic databases in Prolog," In *Biocomputing: Informatics and Genome Projects*, D. W. Smith, ed., Academic Press, 1994.

[28] Zehetner, G., and Hans Lehrach, "The Reference Library System – sharing biological material and experimental data," *Nature* 367:489-491, 1994.

[29] Zhang, J., J. Ostell and K. Rudd, "ChromoScope: a graphic interactive browser for *E. coli* data expressed in the NCBI data model," *Proc. 27th HICSS*, 58-67, 1994.