Manuscript based on a presentation at The First International Workshop on Integrative Approaches to Molecular Biology, Cuernavaca, Mexico, 20-24 February, 1994. Published as:

Robbins, R. J. 1996. Comparative genomics: A new integrative biology. In Collado-Vides, J., Magasanik, B., and Smith, T. F. (Eds) I*ntegrative Approaches to Molecular Biology*. Cambridge, Massachusetts: MIT Press. pp. 63-90.

COMPARATIVE GENOMICS:

A NEW INTEGRATIVE BIOLOGY

Robert J. Robbins

US Department of Energy
robbins@er.doe.gov

Johns Hopkins University
rrobbins@gdb.org

TABLE OF CONTENTS

COMPARATIVE GENOMICS: A NEW INTEGRATIVE BIOLOGY

Robert J. Robbins

ABSTRACT

Although reductionism has been wildly successful in molecular biology, it is now time to emphasize more integrative approaches. This will require the participation of those with special interests in integrative methods, since the intuition of bench researchers will not likely be fully adequate. Electronic information resources will play a crucial role in the integrative methods, provided that the data in them can be used as raw input for future analyses. If integration is to succeed, we must collectively take responsibility for maintaining databases with the same precision and rigor with which we maintain laboratory preparations and raw materials.

A new discipline of comparative genomics is emerging, facilitated by the technical advances accompanying the human genome project. This field will require analytical methods in which whole-genome data sets can be manipulated and analyzed. This, in turn, will require the availability of such data sets in formats and of a quality appropriate for computational analysis. Several impediments, both technical and conceptual, currently block the development of appropriate information resources.

The data-acquisition crisis of the 1980s has been largely overcome, but now we face a data-integration crisis of the 1990s. Proposals to build a federated information infrastructure for biology are promising, but the technical methods for implementing such a system are yet to be devised. Some of our fundamental biological concepts are in need of rethinking, for example "gene" and "genomic map."

Genetic nomenclature is unsystematic, especially across widely divergent taxa where the results are chaotic and unsatisfactory. A new and comprehensive comparative approach to genomic nomenclature is needed, with special emphasis on the provisional naming of genes. Geneticists should recognize that present nomenclature is mere scaffolding, sure to be supplanted when known genes number in the hundreds of thousands and full sequences are available for many species.

Manuscript based on a presentation at The First International Workshop on Integrative Approaches to Molecular Biology, Cuernavaca, Mexico, 20-24 February, 1994. Published as:

Robbins, R. J. 1996. Comparative genomics: A new integrative biology. In Collado-Vides, J., Magasanik, B., and Smith, T. F. (Eds) I*ntegrative Approaches to Molecular Biology*. Cambridge, Massachusetts: MIT Press. pp. 63-90.

COMPARATIVE GENOMICS: A NEW INTEGRATIVE BIOLOGY

Robert J. Robbins

THE CHALLENGE

Although reductionism in molecular biology has led to tremendous insights, we must now emphasize integrative activities. This will require special skills, since the intuition of the experimentalist alone is not likely to be adequate. Formalization is needed in database design, although not perhaps for bench research.

Considerable reductionist data, such as metabolic pathways, protein structures, gene sequences, and genomic maps are now available. Conceptual integration in molecular biology will depend upon access to properly managed and well integrated data resources, with databases providing the raw material for future analyses. For integrative efforts, databases do not merely provide summaries of previous findings and indices to the literature. Instead, they drive new scientific investigation, both helping to shape new bench research and also permitting a new kind of *in silico* studies (Danchin, this volume).

In private (and sometimes in public) biologists have sometimes claimed that database managers seem overly concerned with the niceties of data representation. Expert in separating signal from noise, many biologists suspect that computer scientists' concern with data modeling and data structures is just fussiness. However, if databases are to provide input for further analysis, then errors in data models and formats are not equivalent to simple untidiness — they are equivalent to the sloppy preparation of laboratory material.

A bench researcher would be appalled should a technician say, "There was a bit of mold on the plates, but we were in a hurry so I did the DNA extraction anyway." A

database designer reacts similarly to, "We knew the data model wasn't quite right, but we were in a hurry so we built the system anyway." We must collectively take responsibility for maintaining databases with the same precision and rigor with which we maintain laboratory preparations and raw materials.

Conceptual integration in molecular biology will be facilitated by a technical integration of data and of analytical resources. At present, however, there are impediments that interfere with both technical and conceptual integration. We will consider these impediments later in this paper.

The Human Genome Project

The international Human Genome Project (HGP) illustrates the need for integrative approaches to molecular biology. The original goals of the project were: (1) construction of a high-resolution genetic map of the human genome; (2) production of a variety of physical maps of all human chromosomes and of the DNA of selected model organisms; (3) determination of the complete sequence of human and selected model-organism DNA; (4) development of capabilities for collecting, storing, distributing, and analyzing the data produced; and (5) creation of appropriate technologies necessary to achieve these objectives (USDOE, 1990).

The ultimate goal is the integration of these diverse findings into a coherent understanding of the human genome and that of several model organisms. Physical map data must be integrated with genetic map data; analyses of the mouse genome must be merged with those of the human genome; sequencing information must be connected to map information. Success of the HGP will depend, in large part, upon integrative approaches to molecular biology.

In April of 1993, a group of informatics experts met in Baltimore to consider the role of community databases in the support of genome research. A report from this meeting (Robbins, 1994b) noted:

> The success of the genome project will increasingly depend on the ease with which accurate and timely answers to interesting questions about genomic data can be obtained.
>
> All extant community databases have serious deficiencies and fall short of meeting community needs.
>
> An embarrassment to the Human Genome Project is our inability to answer simple questions such as, "How many genes on the long arm of chromosome 21 have been sequenced?"

Although relating genes and sequences is central to the HGP, and although much consideration has been given to sophisticated integrative approaches to molecular information, the simple fact remains that we cannot now ask an integrative question as simple as that relating map and sequence data.[1]

---

[1] Since this meeting, the Genome Sequence Data Base (GSDB) has been created with the specific mission of managing DNA sequence data in the context of genomic research. Links between it and the Genome Data Base (GDB) are strong and

This report also provided examples of integrative queries that will become crucial for continued success in genome research.

> Return a list of the distinct human genes that have been sequenced.
>
> Return all sequences that map 'close' to marker M on human chromosome 19, are putative members of the olfactory receptor family, and have been mapped on a contig map of the region; return also the contig descriptions.
>
> Return all genomic sequences for which alu elements are located internal to a gene domain.
>
> Return the map location, where known, of all alu elements having homology greater than "h" with the alu sequence "S".
>
> Return all human gene sequences, with annotation information, for which a putative functional homologue has been identified in a non-vertebrate organism; return also the GenBank accession number of the homologue sequence where available.
>
> Return any annotation added to my sequence number #### since I last updated it.
>
> Return the genes for zinc-finger proteins on chromosome 19 that have been sequenced. (Note: answering this requires either query by sequence similarity or uniformity of nomenclature.)
>
> Return all sequences, for which at least two sequence variants are known, from regions of the genome within plus or minus one chromosome band of DS14#.
>
> Return all G1/S serine/threonine kinase genes (and their translated proteins) that are known (experimentally) or are thought (by similarity) also to exhibit tyrosine phosphorylation activity. Keep clear the distinction in the output.

As these examples show, there is a need to integrate data resources just within the HGP. The comparative studies that will arise as detailed genomic information becomes available for many species will depend upon easy information integration.

Molecular biological data are accumulating exponentially. Sequence database growth (Figure 1) is now such that 40% of the data in the database were provided in the last five months. Similar expansion is happening in other genome databases as well, such as the Genome Data Base, the Protein Data Bank (PDB), and others. These information resources must be structured so that data can be extracted and used for further analytic techniques without the need for manual adjustments or interpretation. If it isn't done right soon, it may never be done — the data volume will be too great. Although some may argue that *new* data is not accumulating that rapidly (many reported sequences are of proteins or genome regions previously

---

improving. Now, finally, it is possible to answer some of these questions. GSDB can be reached electronically at URL http://www.ncgr.org and GDB at http://www.gdb.org.

sequenced), *all* of the reported sequences must somehow be managed. Extracting and managing a non-redundant subset is yet another challenge.

Whatever one thinks of the HGP or of its goal of bulk sequencing human and model-organism DNA, the technological advances that it spawns will have profound effects upon biology. If the HGP even approximates meeting its goals, two things will happen: first, the amount of sequence data in the databases will increase more than 100 fold over present levels (even if no one besides genome researchers ever sequences another nucleotide), and second, sequencing will become so inexpensive that the production of very large sequences from non-genome laboratories will become commonplace.

When sequencing is reduced to pennies a base, or less, it will be possible to begin investigating new organisms by collecting large amounts of sequence from them. As this happens, we will need analytical tools that allow integrative studies of organisms for which little else is known, besides large amounts of sequence and information deduced from sequence.

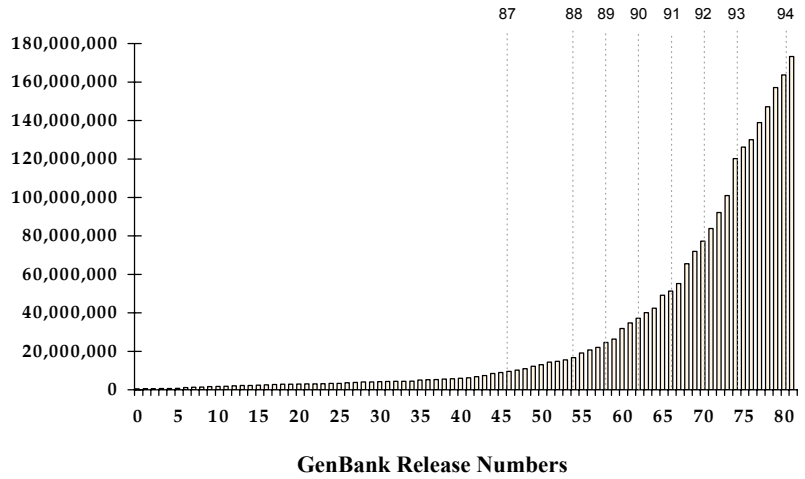


Figure 1. The growth of base pairs of sequence data in GenBank,[1] from Release 1 through Release 82. The year boundaries beginning with 1987 are given at the top of the figure. The

[1] Looking at this figure, it is hard to believe that in the mid 1980's there was a crisis in the sequence databases, with the rate of data production apparently far outstripping the ability of the databases to keep up (Lewin, 1986). Then, data took more than a year to go from publication to entry into the databases. Now, data appear in the databases within a few days of submission. Technical and sociological changes were required to solve the problem. The development of software to allow scientists to prepare and submit data directly to the databases provided the technical fix. Sociologically, many journals began to require that scientists take personal responsibility for submitting data prior to publishing.

database is currently doubling in size every two years. (Data provided by Los Alamos National Laboratory and by the National Center for Biotechnology Information.)

Comparative Genomics: An Integrative Biology

Comparative genomics will emerge as a new scientific discipline as a result of the success of the genome project. To be sure, some journals already publish genomic papers with comparative content, but these usually involve relatively small genomic fragments. Here, "comparative genomics" refers to studies that involve data sets with entire genomes as their scope.

For example, imagine that complete human and mouse sequences were available. Whole-genome comparisons to look for chromosomal rearrangements, conserved linkage groups, etc., could provide remarkable insights. In principle, one might do such comparisons with a giant dot plot that compared one three-billion-base-pair sequence against the other.

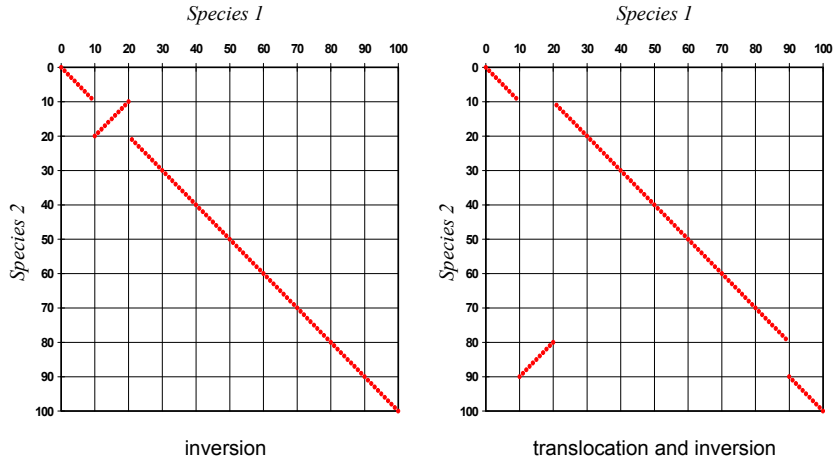


Figure 2. Hypothetical whole genome map dot plots. On the left is the pattern expected from two maps that differ only by a simple inversion. On the right is the plot for two maps that differ by a translocated inversion.

Although this might not be computationally possible, similar results could be obtained through more practical means. For example, whole-genome *map* dot plots can be constructed now. The axes represent linear versions of the genetic maps from two species, and homologous genes are represented as points, with the x coordinate corresponding to the map position in species 1 and the y coordinate that for species two. Perfectly congruent maps produce a single diagonal set of dots. Maps that differ by simple chromosomal rearrangements show recognizable patterns corresponding to those rearrangements (Figure 2).

To construct such a plot we need only have maps and homology documentation for the markers on the maps. Sufficient data are available for some bacteria. Figure 3 shows whole genome map dot plots comparing *Escherichia coli*, *Bacillus subtilis*, and *Salmonella typhimurium*.

These plots were prepared to demonstrate that, in principle, interesting biological comparisons can be done using whole-genome-sized data sets. To do such comparisons, however, we must have access to whole-genome data sets upon which we can begin to compute straight-away, without extensive manual adjustments. Even now when we have at most a few thousand markers per organism, the requirement of significant manual adjustments is a strong deterrent to comparative genomic analysis. Soon, with tens or hundreds of thousands of known markers per genome, the requirement of manual data adjustments would render the process essentially impossible. Present nomenclatural inconsistencies now makes the automatic production of such data sets difficult. This is discussed below in the section entitled "Conceptual Impediments".

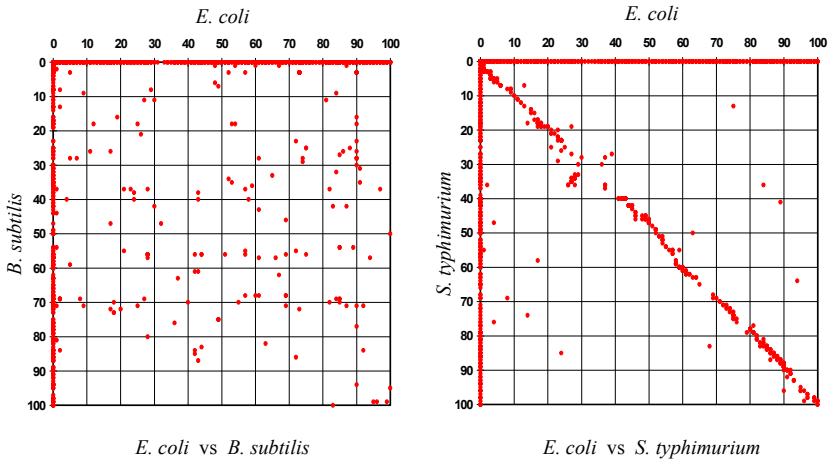


Figure 3. Whole genome map dot plots show essentially random distributions of homologous genes when the maps of *E. coli* and *B. subtilis* are compared. However, a similar plot of *E. coli* and *S. typhimurium* shows an obvious diagonal and gives some evidence of the well known inversion in the 30-40 minute region (Sanderson and Hall, 1970; Riley and Krawiec, 1987). Individual outliers may indicate small rearrangements or simply errors in the data. (Homology data from Abel and Cedergren, 1990.)

TECHNICAL IMPEDIMENTS

Developing the necessary information infrastructure for integrative molecular biology will be hard work, with many technical difficulties to be overcome. First, basic design and implementation of information systems is always challenging.

Second, achieving some semantic consistency across multiple, independently operated information resources will be a continuing challenge that has both technical and conceptual components. Finally, generating interoperability, getting the different components of an information infrastructure for molecular biology to work together smoothly, will be the biggest technical challenge.

Basic Design Challenges

Building a database is surprisingly difficult. Without adequate design and appropriate plans for integration of the components, efforts can fail spectacularly. In the 1980's the General Services Administration tried to develop a database of the properties owned and occupied by the federal government. Although this may seem trivial (how difficult can it be to implement a database about buildings?), the project ultimately collapsed with no deliverables, despite having consumed more than $100 million dollars (Levine, 1988):

> The General Services Administration said last week it will discontinue development of Stride, a complex computer system to automate the Public Buildings Service. GSA has spent $100 million since 1983 trying to make the system work — a figure that does not include $ 78 million spent on the system Stride was intended to replace.
>
> GSA officials said Stride fell apart largely because of a failure to create a workable systems-integration plan for it. The most glaring problem with Stride was that it lacked an integration design. Stride went directly from the functional design to work packages without the intermediate step of a detailed design to show how all of the packages would fit together.

To those unfamiliar with the risks of software development, the idea of spending this much money with no results seems inconceivable. To appreciate how this can happen one must realize that some early decisions in building databases are so fundamental that if they prove to be incorrect, the entire effort is wasted. An equivalent problem in architecture might be the discovery, upon completion of a new research facility, that another 12 inches of real clearance is required on each floor, if the building is to house the equipment planned for it. How does one add 12 inches of real clearance to each floor of a completed building? The answer is, one does not. The only options are to abandon the original equipment or to abandon the building. Unfortunately, such disasters occur all too frequently in software projects.

Such abject failure rarely occurs in architecture, because much of architectural design and construction involves the assembly of previously designed and well understood components into new combinations, with the entire enterprise bounded by the reality-check limitations of time and space. Software, like poetry, is constructed of pure thought. Individual programs and even complete software systems are frequently constructed entirely *de novo*. Architectural construction would be more like software development if architects and contractors themselves had to devise most of the components used during construction.

The occasional big failure in architectural construction (such as the 1981 walkway collapse in the Hyatt Regency Hotel in Kansas City that left 114 dead and over 200 injured) are often associated with first attempts at new designs or with the need for last-minute, on-the-spot design changes that characterize most software development (Levy and Salvadori, 1992; Petrosky, 1992).

Solving the design challenge for molecular biology databases requires a thorough understanding of the scientific subject matter and an appreciation of the subtleties of database development and information modeling. Some conceptual challenges that affect design will be discussed below in the "Conceptual Impediments" section. For a more extensive discussion of the requirements and design issues facing molecular biology databases, see Robbins, 1993.

Semantic Consistency

Reasonable interoperability among molecular biology information resources cannot be achieved unless some minimum level of semantic consistency exists in the participating systems. No amount of syntactic connectivity can compensate for semantic mismatches.

Developing an information infrastructure to support integrative approaches to molecular biology will require increased effort to ensure semantic consistency. Controlled vocabularies and common-denominator semantics are important. The same unique identifiers must be used for the same biological objects in all interoperating databases. Participating databases must provide stable, arbitrary external identifiers (accession numbers) for data objects under their curation and references to these objects in other databases should always be made via accession numbers, not via biological nomenclature. Linking data objects between databases requires that the other objects be unambiguously identifiable (accomplished via accession numbers) and relevant (accomplished via semantic consistency). Although perfect semantic consistency is probably unattainable, efforts to improve consistency are essential. In particular, community databases must document the *semantics* of their systems.

Interoperability Challenges

Databases have evolved from simple indices into a new kind of primary publication (Figure 4). Electronic data publishing (Cinkosky, et al., 1991; Robbins, 1994a) has transformed some areas of science so that data submission has become a requisite part of sharing one's findings with the community.

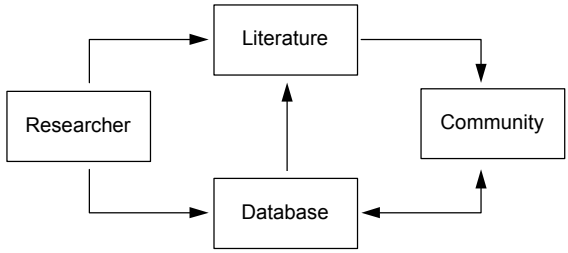


Figure 4. Mature electronic data publishing is an alternative form of publishing. Databases parallel, or even precede, the print literature. Authors are responsible for submitting their findings, authorship of submissions is recognized, and volunteer editors and reviewers help ensure the quality and trustworthiness of the resulting system.

Now we face a data-integration crisis of the 1990s. Even if the various separate databases each keep up with the flow of data, there will still be a tremendous backlog in the integration of information in them. The implication is similar to that of the 1980s: either a solution will soon emerge or biological databases collectively will experience a massive failure.

One possible solution for the data integration problem is the concept of "federation" (Robbins, 1994b, 1994d). In a federated information infrastructure, individual scientists would interact simultaneously with multiple resources, both for submitting and for accessing information (Figure 5).

Although this is being widely discussed in biology, the means for implementing such a federation is not immediately at hand. True federation is still a research topic in computer science (Sheth and Larson, 1990; Bright, et al., 1992; Hurson et al., 1994). However, great success has attended the spread of loosely coupled federations of text-server systems such as gopher, WAIS, and World Wide Web. Whether such a loosely coupled federated approach can be extended to include structured data from complex databases and, more importantly, to support true semantic joins across different databases remains to be seen (cf. Robbins, 1995).

A crisis occurred in the databases in the mid 1980s, when the data flow began to outstrip the ability of the database staff to keep up (Kabat, 1989; Lewin, 1986). A conceptual change in the relationship of databases to the scientific community, coupled with technical advances, solved the problem.

CONCEPTUAL IMPEDIMENTS

"Every database is a model of some real world system" (Hammer and McLeod, 1981). If the model is inadequate, the database will fail, sometimes totally. In scientific databases, the challenge is to represent the real world in a way that accommodates the subtlety of our present knowledge and that also can evolve gracefully with changing concepts.

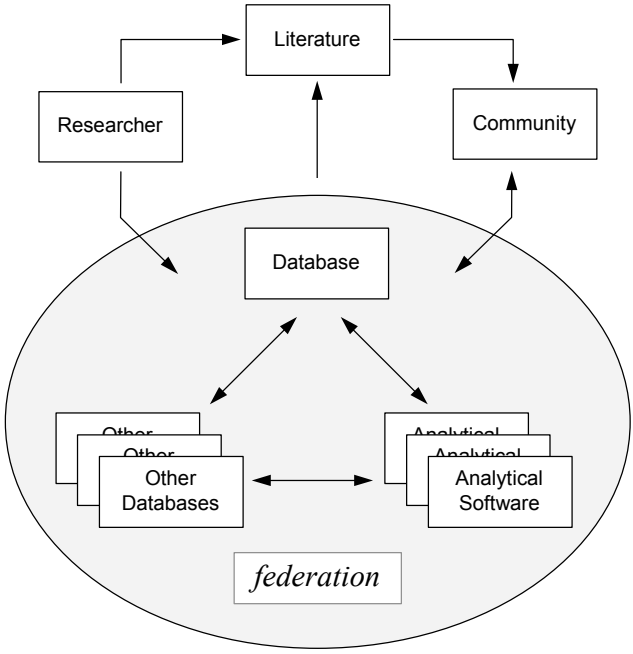


Figure 5. In a federated information infrastructure, individual databases interoperate with other databases and with online analytical tools. Individual researchers can interact with federated resources collectively, either in submitting or accessing information.

Scientific knowledge does change over time, sometimes dramatically. Examples of assertions that seemed unassailably true when made, yet which seem wildly wrong now are:

> If the genes are conceived as chemical substances, only one class of compounds need be given to which they can be reckoned as belonging, and that is the proteins in the wider sense, on account of the inexhaustible possibilities for variation which they offer. ... Such being the case, the most likely role for the nucleic acids seems to be that of the structure-determining supporting substance. (Caspersson, 1936)

> Fifty years from now it seems very likely that the most significant development of genetics in the current decade (1945-1955) will stand out as being the discovery of pseudoallelism. (Glass, 1955)

Undoubtedly, similarly misguided beliefs are now current. Identifying and avoiding them must be a key goal for a designer of scientific databases. Several such beliefs to be avoided exist for genomic databases, among them:

> **Gene** A hereditary unit that, in the classical sense, occupies a specific position (locus) within the genome or chromosome; a unit that has one or more specific

> effects upon the phenotype of the organism; a unit that can mutate to various allelic forms; a unit that codes for a single protein or functional RNA molecule.
>
> The ultimate ... map [will be] the complete DNA sequence of the human genome.

A database built according to these concepts (excerpted from the National Academy study that helped launch the HGP) would fail.

Inadequate Data Models

A database that contains information about a particular class of objects must contain some sort of "definition" of those objects. For example, a database of people must "define" persons in terms of their attributes, or possible attributes. Should a person be defined as having one, two, or an unlimited number of different telephone numbers? Does a person have one unchanging address, or can a person have more than one address? Definitions of database objects limit the capabilities of the database, with bad definitions yielding bad databases.

In genomics, definitional problems exist with our most fundamental concepts. "What is a gene?" and "What is a map?" are still questions without clear answers.

What is a Gene?

According to the classical definition, the gene was the fundamental unit of heredity, mutation, and recombination. Classical genes were envisioned as discrete, indivisible objects, each with its own unique location in the genome. Unchanging locations in the genome were believed to exist independently of the genes that occupied those locations: "The genes are arranged in a manner similar to beads strung on a loose string" (Sturtevant and Beadle, 1939). Mapping (Figure 6) involved identifying the correct order of the beads and the proper address (i.e., position on the string) for each bead. Addresses, i.e., loci, were considered points, since the beads were believed to be small and indivisible.

Although no biologist still employs this model when designing bench research, all early map databases (and some present ones still) indirectly use it by employing data structures that assign simple, single-valued addresses as attributes to genes. This is clearly beads-on-a-string revisited, since it assumes that a coordinate space (the string) exists independently of the genes (the beads) and that the genes are discrete, non-overlapping entities.

Later, physiological analyses led to the definition of genes through their products, first as "one gene, one enzyme," then "one gene, one polypeptide." Although we now know that very complex many-to-many relationships can exist between genes and their products, and between primary products and subsequent products (Riley, 1994), the notion that single genes can be unambiguously associated with single products still infects many genomic databases.

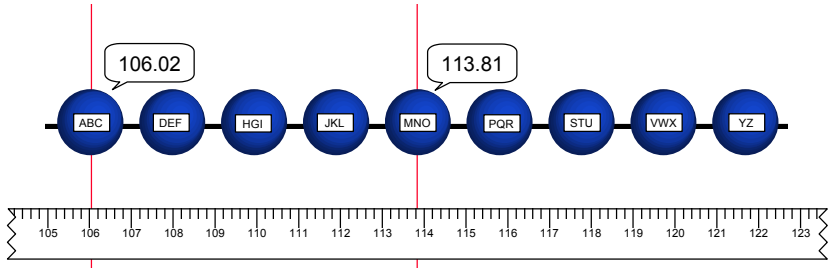


Figure 6. According to the "beads on a string" model of genes, mapping just involved determining the correct order of the beads and the correct address, or locus, for each bead.

With the recognition of DNA as the hereditary substance, some began to define genes in terms of sequence, for example: "the smallest segment of the gene-string consistently associated with the occurrence of a specific genetic effect." Benzer's (1956) fine-structure analysis resulted in an operational definition, the *cis-trans* test, and in the concept of the *cistron*. This idea mapped nicely to DNA and, for many, provided the final definition of a gene: a transcribed region of DNA, flanked by upstream start regulatory sequences and downstream stop regulatory sequences (Figure 7).

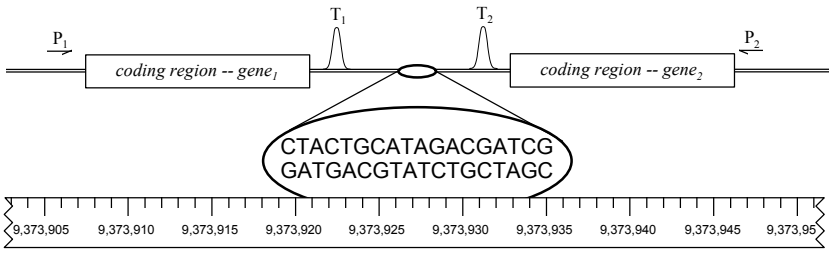


Figure 7. A simplistic view of the discrete coding-sequence model of the gene. Here, the fact that genes may be oriented in either direction in DNA is shown by reversing the placement of the promoter and terminator signals for the second gene. By this model, the locus of a gene corresponds to the address in base pairs of its start and stop signals.

This discrete-coding-sequence model of a gene was extended to include mapping by asserting that a genome can be represented as a continuous linear string of nucleotides, with landmarks identified by the chromosome number followed by the offset number of the nucleotide at the beginning and end of the region of interest.

Although this simplistic concept ignores the fact that human chromosomes may vary in length by tens of millions of nucleotides and that some regions of the genome exhibit very complex patterns of expression, it continues to be widely

espoused today. A major molecular-genetics textbook has carried the following definitions unchanged, through five editions spanning 1983-1994:

> **Gene** (cistron) is the segment of DNA involved in producing a polypeptide chain; it includes regions preceding and following the coding region (leader and trailer) as well as intervening sequences (introns) between individual coding segments (exons).
>
> **Allele** is one of several alternative forms of a gene occupying a given locus on a chromosome.
>
> **Locus** is the position on a chromosome at which the gene for a particular trait resides; locus may be occupied by any one of the alleles for the gene.

This particular book has been a leader in emphasizing the molecular approach to genetics, yet these definitions carry much conceptual baggage from the classical era. By implying a single coding region for a single product, they derive from the "one gene, one polypeptide" model. By logically distinguishing the gene from its location, they even connect to the beads-on-a-string approach.

Whether or not the "gene as sequence" should include just the coding region or also the up- and down-stream regions such as promoters and terminators is also unresolved. Geneticists studying prokaryotes routinely restrict the concept of gene to the coding region and eukaryotic geneticists just as routinely extend it to include the promoter and terminator and everything in between.

These differing views are so strongly held within their respective communities that they are considered to be virtually self evident. Prokaryotic geneticists are usually astounded to hear that some would include flanking regions in the concept of gene, and eukaryotic geneticists are equally astounded at the notion they should be excluded. Despite these differences across different biological communities, databases like GenBank are expected to include consistent concepts of the "gene" in their annotation of sequences from all possible taxa.

Many regions in prokaryotic and eukaryotic genomes are now known to possess a level of functional complexity that renders simplistic definitions wholly inadequate. Figure 8 illustrates the MMS operon in *E. coli*, a complex region of overlapping control and coding regions. If control regions were considered part of the gene, there are no simple genes as nonoverlapping sequences here.

Eukaryotic systems have provided us with examples of fragmented genes (exon and introns), alternative splicing (different protein products from the same transcript), and nested genes (complete genes carried on the introns of other genes). The concept of gene as *discrete* sequence of DNA is no longer viable.

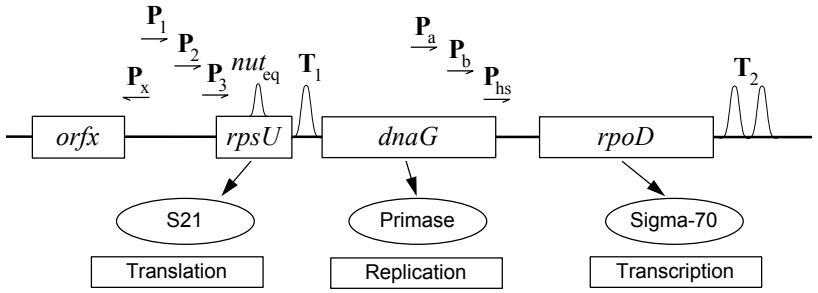


Figure 8. The MMS operon in *E. coli* contains a complex set of control and coding regions. Transcription of the *rpoD* coding region can begin from any of a half dozen promoters, some of which are embedded in the coding region of the *dnaG* gene. (Lupski, et al., 1989)

In humans, the *UGT1* locus is actually a nested gene family (Figure 9). The first exon and promoter of an ancestral five-exon gene have apparently been replicated five times, with functional divergence occurring in the multiple first exons. Alternative splicing does not seem to be involved, rather each promoter initiates the transcription of a transcript that is processed to yield a single mRNA. If the cis-trans test were applied to mutations of phenol UDP-glucuronosyl-transferase and bilirubin UDP-glucuronosyltransferase, the determination of whether or not they were encoded by the same cistron would depend upon which exon carried the mutations.

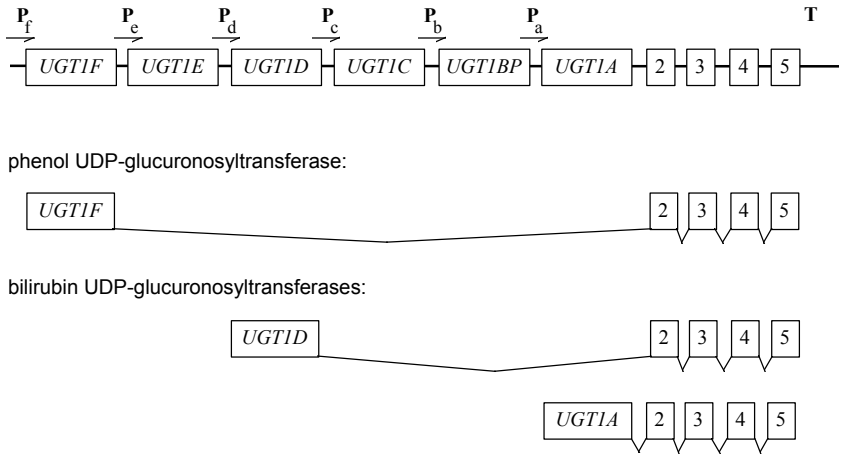


Figure 9. The *UGT1* locus in humans is actually a nested gene family that yields multiple transcripts through alternative promotion. Each promoter produces a transcript that is spliced so that the exon immediately adjacent to the promoter is joined with the four terminal exons shared by all of the transcripts. (Ritter, et al., 1992)

Must cistron now be redefined as equivalent to exon, not to gene? Given its original intent as the shortest stretch of contiguous DNA comprising a functional genetic unit, such a definition for "cistron" would seem appropriate. However, that interpretation is not common.

Such complexities have led some authors to declare that a single definition for gene is no longer possible. For example, Singer and Berg (1991) say:

> The unexpected features of eukaryotic genes have stimulated discussion about how a gene, a single unit of hereditary information, should be defined. Several different possible definitions are plausible, but no single one is entirely satisfactory or appropriate for every gene.

Most biologists readily accommodate new complexities into their mental models of genome function and proceed to interpret their experimental results accordingly, albeit without ever developing a formal modification of their linguistic definition of a gene. This is fine for guiding bench research, but a database cannot be built to represent genes if genes cannot be defined. What is needed is a new, more abstract definition that can accommodate complexity.

Some authors have started to rethink the essence of "gene" and some fairly radical concepts can be found casually presented in textbooks. For example, Watson, et al. (1992) offered these thoughts:

> DNA molecules (chromosomes) should thus be functionally regarded as linear collections of discrete transcriptional units, each designed for the synthesis of a specific RNA molecule. Whether such "transcriptional units" should now be redefined as genes, or whether the term gene should be restricted to the smaller segments that directly code for individual mature rRNA or tRNA molecules or for individual peptide chains is now an open question.

Although this holds to the established notion of *discrete* transcriptional units, it also suggests a radical redefinition of "gene" to mean *unit of transcription*. Restricting the concept of gene to functions that occur closest to the level of DNA is appealing, but not widespread.

Despite denying that "gene" can be defined, Singer and Berg adopted a working definition:

> For the purposes of this book, we have adopted a molecular definition. A eukaryotic gene is a combination of DNA segments that together constitute an expressible unit, expression leading to the formation of one or more specific functional gene products that may be either RNA molecules or polypeptides.

This approach has potential for database design, in that it abandons the concept of gene as discrete DNA sequence and explicitly embraces the potentially many-to-many relationship between genes and their products. In fact, an extension of this to "A map object consists of a set of not necessarily discrete and not necessarily contiguous regions of the genome" may well prove to be the definition upon which database designs converge.

Such a definition would include, essentially, any subset of the genome to which one might wish to attach attributes, with "gene" being just one subclass of such map objects. Dislodging the gene concept from a central position in a database of genes may be distasteful to geneticists, but it is proving essential to achieving working database designs.

That biologists assimilate complex new findings without necessarily reducing them to precise definitions has been known for some time. After attempting to formalize Mendelian genetics, Woodger (1952) noted that the *language* of geneticists is usually not as complex as their *thoughts*:

> Geneticists, like all good scientists, proceed in the first instance intuitively and ... their intuition has vastly outstripped the possibilities of expression in the ordinary usages of natural languages. They know what they mean, but the current linguistic apparatus makes it very difficult for them to say what they mean. This apparatus conceals the complexity of the intuitions. It is part of the business of [formalizing] genetical methodology first to discover what geneticists mean and then to devise the simplest method of saying what they mean. If the result proves to be more complex than one would expect from the current expositions, that is because these devices are succeeding in making apparent a real complexity in the subject matter which the natural language conceals.

In short, what biologists say about biology does not match what they understand about it. Anyone attempting to design biological databases must recognize that essential truth.

What is a Map?

A genetic map is some ordered relationship among different genetic or physical markers in the genome. Devising appropriate data models for maps depends upon the attributes of the objects being mapped (Figure 10). Complete maps of indivisible beads can be simple ordered lists. Partial maps can be represented with directed graphs. Partial maps of complex overlapping objects might involve directed graphs of end points of atomic components of complex objects.

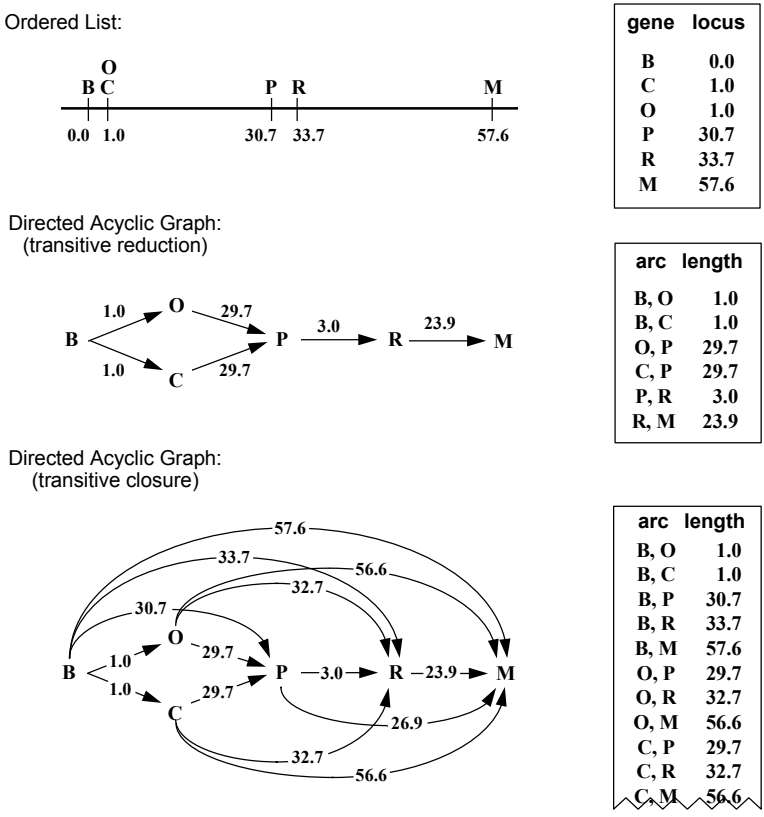
Even the notion of "map" itself is problematic, since generic descriptions of genome structure have much more in common with anatomies than with true maps. Maps describe the actual properties of individual objects, whereas anatomies describe the average properties of populations of objects (Robbins, 1994c).

Inadequate Semantics

Databases often exhibit semantic differences in their treatment of related material. For example, information about human beta-hemoglobin can be found in several databases, such as PIR-International, SwissProt, GenBank, GDB, OMIM, and others. Although it would seem a simple matter to provide links that allow the easy traversal of these entries, these databases may have fundamental semantic differences that interfere with the development of sensible links. In the past, PIR-

International data objects were proteins in the chemical sense so that any two proteins with the same structure were the *same* protein.

Thus, the PIR-International entry for human beta-hemoglobin was also the entry for human, chimpanzee, and pygmy chimpanzee beta-hemoglobin. Although this policy has been discontinued by PIR-International, it is still evident in SwissProt release 28.0, where entry P02023 is for beta-hemoglobin for all three species, with cross references to the three different entries in PIR-International.



gene	locus
B	0.0
C	1.0
O	1.0
P	30.7
R	33.7
M	57.6

arc	length
B, O	1.0
B, C	1.0
O, P	29.7
C, P	29.7
P, R	3.0
R, M	23.9

arc	length
B, O	1.0
B, C	1.0
B, P	30.7
B, R	33.7
B, M	57.6
O, P	29.7
O, R	32.7
O, M	56.6
C, P	29.7
C, R	32.7
C, M	56.6

Figure 10. Data structures for genetic maps reflect the underlying conceptual models for the maps. Many geneticists still think of maps as ordered lists, and ordered list representations are used in many genome databases. Directed acyclic graph (DAG) data structures can be represented pictorially (left) or tabularly (right). Depending on their use, DAGs may be represented as transitive reductions or transitive closures.

In GenBank, objects are reported sequences, which may or may not correspond precisely with a gene or particular protein. GenBank may have hundreds or thousands of entries of genomic RNA, cDNA, DNA, or even individual exon sequences that relate in some way to human beta-hemoglobin.

In GDB, objects include genes, probes, and polymorphisms. There will be one GDB entry for the beta-hemoglobin *gene*, but multiple entries for associated polymorphisms and probes.

In OMIM, objects are essays on inherited human traits, some of which are associated with one locus, some with multiple loci, and some whose genetic component (if any) is unknown.

Different community databases vary in the richness of their semantic concepts. GDB has more subtleties in its concept of a gene than does GenBank. GenBank's concept of nucleotide sequence is richer than that of other databases. To facilitate integration, participating databases should attempt to accommodate the semantics of other databases, especially when the other semantics are richer or more subtle.

Inadequate Nomenclature

A future goal of the genome project is comparative analyses of results in human and selected model organisms. To test the feasibility of an integrative approach in which gene homologies were deduced from data other than sequence similarities, I extracted information about 16,500 genes from nine different organismal information resources (Table 2).

Table 2. The taxon distribution of the 16,500 gene records from different species used in the test. (The data on *E. coli* were provided directly by Mary Berlyn, those on *S. typhimurium* by Ken Sanderson, those on yeast by Mike Cherry, those on corn from Stan Letovsky. The others were obtained from networked servers.)

Taxon	Genes
E. coli	1.391
S. typhimurium	754
S. cerevisiae	1,427
C. elegans	229
D. melanogaster	5,684
Z. mays	1,035
A. thaliana	236
M. musculus	1,361
H. sapiens	4,383

The data were reduced into a single comparable format that included taxon, gene symbol, gene name, map position, and gene-product function. Enzyme name and Enzyme Commission number were also included for genes with known enzymes as products. Then, the data were examined to see how readily one might detect potentially homologous genes. In some cases, homologies were already indicated in the databases but these were ignored since the purpose of the test was to see if there were enough information present in the basic data to allow even the

identification of candidate homologies. The examples in Table 3 were chosen to illustrate a number of patterns found in the data.

Table 3. Examples of similar gene symbols as used in different organisms. (At = *A. thaliana*, Ce = *C. elegans*, Dm = *D. melanogaster*, Ec = *E. coli*, Hs = *H. sapiens*, Mm = *M. musculus*, Sc = *S. cerevisiae*, St = *S. typhimurium*, Zm = *Z. mays*.)

Symbol	Taxon	Gene Name
acd	Ec	Acetaldehyde CoA deHase
acd	At	accelerated cell death
acd	Mm	adrenocortical dysplasia
ag	Dm	agametic
ag	At	agamous
CHS1	Hs	Cohen Syndrome 1
CHS1	Sc	chitin synthetase
cal	Dm	coal
Cal	Dm	Calmodulin
cal	Ce	CALmodulin related
cal	At	cauliflower (defective inflorescence)
Cat	Dm	Catalase (EC: 1.11.1.6)
CAT	HS	catalase (EC: 1.11.1.6)
cat	Ce	CATecholamine abnormality
Cat	Mm	dominant cataract
cat1	Zm	catalase1 (EC: 1.11.1.6)
cat1	Sc	catabolite repression
dw	Dm	dwarf
dw	Mm	dwarf
ft	Dm	fat
ft	At	late flowering
ft	Mm	flaky tail
Gad1	Dm	Glutamic acid decarboxylase 1 (EC: 4.1.1.15)
GAD1	Hs	glutamate decarboxylase 1 (EC: 4.1.1.15)

In many cases (such as *acd*, *ag*, and *CHS1*) identical gene symbols were assigned to wholly unrelated genes. In others, symbols that differed only in case[1] (e.g., *cal/Cal* and *cat/Cat/CAT*) were assigned to unrelated, related, or similar genes, but with no consistency on the use of case. Some genes with phenomenological names had identical symbols and names (e.g., *dw*) but no homology at all, and others had identical symbols but complete different names

[1] However, case cannot be safely ignoredin all species. In *D. melanogaster* there are more than 100 pairs of genes whose symbols differ only in the use of case. Examples for loci with one-letter symbols: *a* = arc, *A* = abnormal abdomen; *b* = black body, *B* = bar eye; *d* = dachs, *D* = dichaete; *h* = hairy, *H* = hairless; *j* = jaunty, *J* = jammed; *p* = pink, *P* = pale; *r* = rudimentary, *R* = roughened; *s* = sable, *S* = star; *w* = white, *W* = wrinkled; *z* = zeste, *Z* = Zerknittert.

(e.g., *ft*). Sometimes the names were similar, the symbols differed only in case, and the EC numbers suggested homology (e.g., *Gad1/GAD1*). Although a biologist should easily recognize the equivalence of glutamic acid decarboxylase 1 and glutamate acid decarboxylase 1, writing software to do so for every possible pair of equivalent enzyme names would be nearly impossible.

EC numbers proved to offer the best identifier of potential homologies, but they are far from wholly adequate for such purposes. Because EC numbers are actually attributes of chemical reactions that relate to enzymes only transitively through the catalysis of a particular reaction by a particular protein, a single protein can have multiple EC numbers and the same EC number may be assigned to multiple polypeptides. The connection of EC numbers to genes is even more indirect.

EC numbers are unrelated to gene names and symbols. Of 3886 pairs of genes (from different organisms) with the same EC number, only 105 used the same symbol for both species (183, ignoring case). Also, numbers were poorly correlated with canonical enzyme names. Of 1824 records accompanied with an EC number, only 250 had an associated enzyme name that matched the canonical name according to PIR-International. If different use of hyphens was ignored, that rose to 314. If substring/superstring differences (i.e., if the canonical name was a substring of what appeared in the database) were also ignored, the number of matching names was still only 411 out of 1824.

Possible solutions for this dilemma are as technologically simple as they sociologically impossible. What is needed is the development of a common genetic nomenclature that is independent of the taxon in which the gene is identified. What is also needed is the recognition that a provision nomenclature should be used until most genes are known. Although some efforts have been made to standardize genetic nomenclature within some taxonomic groups (e.g., for enteric bacteria, for mammals), no efforts have even been seriously proposed to do so across all life.

TECHNICAL ADVANCES

Continuing technical improvements in data-handling capacities are needed, with scalable systems and processes especially important. A better technical approach to data resource integration is essential. A federated approach to information infrastructure may provide a solution.

Scalable Data Entry Systems

The data acquisition crisis of the 1980s was solved through direct data submission. However, as the data volume in molecular biology continues to grow, a new data acquisition crisis may occur, unless further changes are made in the procedures for getting data into the databases.

Direct data submission involved the research community in the process of loading the database by allowing them to prepare electronic files for submission.

This method is similar to old batch processing technology, where computer users prepared files of punched cards for submission to the computer. What is needed is an advance in data submission equivalent to the development of direct, on-line computing. Researchers must shift from direct data *submission* to direct data *entry*.

The availability of direct, on-line data entry systems will also allow improvements in the data entry tools to propagate immediately. With direct submission tools, each researcher must obtain his or her own copy of the software. If the software is upgraded, it takes some time before the upgrade propagates to all users. With direct data entry, the entry software resides on one or more centralized systems. Improvements to the software become available immediately to all users.

The next generation of data-entry software should be tightly coupled with data analysis tools, so that researchers develop a submission at the same time that they analyze the sequence. Requiring that researchers use one set of software to analyze a sequence in their laboratory, then use a separate set to copy the results of that analysis into a format appropriate for submission is wasteful.

The Federation Concept

Some have addressed the interoperability challenge by calling for a federated approach to information resources (DOE Informatics Summit, reported in Robbins, 1994b):

> We must think of the computational infrastructure of genome research as a federated information infrastructure of interlocking pieces.
>
> Each database should be designed as a component of a larger information infrastructure for computational biology.
>
> Adding a new database to the federation should be no more difficult than adding another computer to the Internet.
>
> Any biologist should be able to submit research results to multiple appropriate databases with a single electronic transaction.

This vision is attractive, especially when one recognizes that different biological communities have overlapping information infrastructure needs. Access to databases is required for genome research, for molecular biology, for ecosystems work, for a national biological survey. Many of these involve similar sorts of data, for example nucleotide sequences.

The trend in information-system development has been from the specific to the generic. Originally, each new information resource was developed as a stand-alone system. Anyone wanting access to it had to obtain a copy of the software and the data and install them on a local machine. The process was onerous and few were willing to invest considerable time just to test the system. Often, the system was developed for a particular hardware and software platform, so even the interested user might not be able to acquire the system without making special purchases.

Next came a move toward networked, client-server systems. Developers produced data resources that were available over the networks, so local users needed only to obtain copies of the client software. Still, the client software was dedicated for accessing a particular data resource and often required specific hardware on which to run. In addition, client systems frequently involved embedded commercial software so that users were obliged to purchase appropriate licenses before accessing the system.

Now generic client-server systems are appearing, in which each data-resource developer "publishes" his or her resource onto the networks via some generic server system. Users need only obtain one copy of the generic client software and then can use this to access multiple, independent data resources.

The first big successes for this approach came with simple text and file retrieval using gopher and WAIS. Now, with Mosaic as the generic client and World Wide Web as the generic server, the ability to produce inter-resource integration through the establishment of hypertext links is available. This has been used to develop data-resource accession systems with considerable integration. For example, the GenQuest server[1] now available through Johns Hopkins allows users to carry out sequence homology searches using the algorithm of choice (FASTA, BLAST, Smith-Waterman) and then receive in minutes the results of the analysis with live hot links to all other referenced databases. If the search identifies a potential protein homologue for which a structure is available, a simple mouse click retrieves the structure and launches an appropriate viewer so that the three-dimensional image may be viewed and manipulated.

Although many molecular biology databases are now publishing their data in WWW servers, complete integration is not yet possible, since WWW browsers do not process queries that involve real joins across multiple databases, nor do they process *ad hoc* queries to the underlying databases. However, many sites are at work developing middle-ware systems to meet these needs, and we can expect WWW clients to acquire increasing sophistication.

Almost certainly, a federated information infrastructure for molecular biology will not be achieved through a massive, centralized, single-site development project. Instead, we will see an increasing trend toward the integration of components developed at multiple sites. Although we are not yet at a point where such interoperability is truly plug and play, we are approaching a "plug, tap, tweak, and play" situation, where the required tapping and tweaking is declining.

CONCEPTUAL SOLUTIONS

As technical solutions become available from a variety of sources, the limiting factors for data integration in molecular biology are likely to be conceptual and sociological. Better data models are needed, especially in the

---

1 Available as a choice on the main WWW page at URL http://www.gdb.org.

conceptual sense. A more consistent approach to genetic and genomic nomenclature is essential. Although there is some movement towards a more generic approach to nomenclature within some taxonomic groups, such as mammals, the overall approach to gene nomenclature is still phenomenological at base and independent in application. Indeed, the independence in genetic nomenclature approaches anarchy. For example, imagine that a new organism is discovered. A biologist who proposed a new name for the species would be called a taxonomist, whereas one who completely proposed new names for its anatomical components, or for its enzymes, would be pronounced a lunatic. Yet, any biologist who offered new phenomenological names for all its genes would simply be deemed a classical geneticist.

Better Data Models

Most developers of biological databases now recognize the need to employ subtle and complex data representations to represent biological knowledge. To meet the various needs of many users, databases should employ internal data structures that most individual users consider too complex. Users of and advisors to biological databases must come to recognize this. User needs must guide the design of the system's external behavior, but user opinions should not be allowed to dictate internal data structures. These should be designed to accommodate diversity, to support local requirements for viewing data, and to facilitate complex integrative analyses.

The problems associated with the various definitions of "gene" and "genetic map" discussed above offer examples where better data models are needed. Although no biologists still use the beads-on-a-string model to drive their experimental work, many biologists still fall back on that model when thinking about data models for genetic maps. One prominent database researcher was dissuaded from addressing the problems of genomic databases when an equally prominent biologist informed him that, "There are no more than 100,000 genes in the human genome, each with relatively few attributes. The final genetic map will just be a list of these genes in the correct order."

The Enzyme Nomenclature Solution

In the 1950s, enzyme nomenclature was spiraling out of control because of its phenomenological basis and its independent application (Webb, 1993):

> [W]orkers were inventing names for new enzymes, and there was a widespread view that the results were chaotic and unsatisfactory. In many cases the enzymes became known by several different names, while conversely the same name was sometimes given to different enzymes. Many of the names conveyed little or no idea of the nature of the reactions catalyzed, and similar names were sometimes given to enzymes of quite different types.

Responding to requests from leaders in the community, the International Union of Biochemistry in 1956 established an International Commission on Enzymes. The Commission produced an interim report in 1958 and a final report

in 1961 that included a standardized approach to enzyme nomenclature and a hierarchical numbering system for identifying reactions catalyzed by enzymes. Continuing efforts have resulted in the publication of five editions of *Enzyme Nomenclature*, which provide summaries of the current classification and nomenclature of known enzymes .

The Commission was reasonably successful, but the effort not without some problems and controversy. Webb summarized the goals of the commission and its difficulties as:

> A major part of this assignment was to see how the nomenclature of enzymes could best be brought into a satisfactory state and whether a code of systematic rules could be devised that would serve as a guide for the consistent naming of new enzymes in the future. ... [T]he overriding consideration was to reduce the confusion and to prevent further confusion from arising. *This task could not have been accomplished without causing some inconvenience, for this was the inevitable result of not tackling the problem earlier.* [emphasis added]

Figure 11 plots the number of known enzymes, and of known human genes, over time. At the time that Webb claims was too late a start, there were only about 600 known enzymes. Now there are more than 4000 known human genes and more than 10,000 known genes in other organisms. If efforts to standardize enzyme nomenclature began too late when there were only a few hundred enzymes, how would one describe the present situation with regard to genetic nomenclature?

In addition to standardizing the names of existing enzymes, the Enzyme Commission and its successor have offered advice on the naming of new enzymes. In particular, the habit of applying phenomenological names has been strongly disapproved (Nomenclature Committee of the International Union of Biochemistry, 1984):

> In this context it is appropriate to express disapproval of a loose and misleading practice that is found in the biological literature. It consists in designation of a natural substance ... that cannot be described in terms of a definite chemical reaction, by the name of the phenomenon in conjugation with the suffix *-ase*... Some recent examples of such *phenomenase* nomenclature, which should be discouraged even if there are reasons to suppose that the particular agent may have enzymatic properties, are: *permease, translocase, replicase*, ... etc.

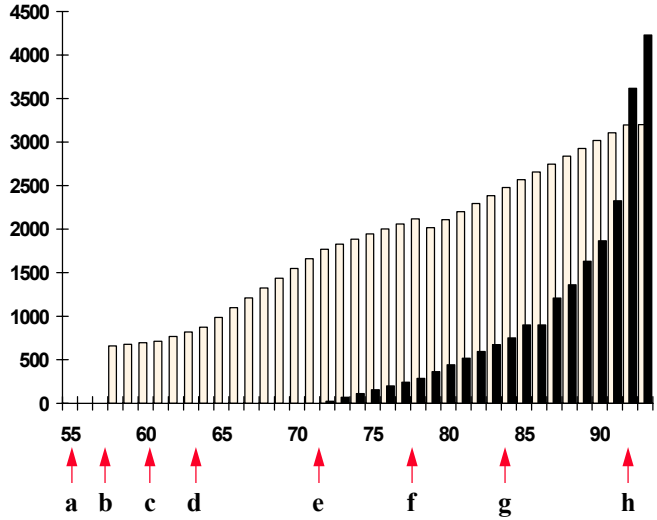


Figure 11. The increase in number of known enzymes (shaded bars) and known human genes (solid bars). The arrows indicate significant events in the history of enzyme nomenclature: a = Enzyme Commission formed, b = commission preliminary report, c = commission final report, d - h = appearance of five editions of *Enzyme Nomenclature*.

A review of titles and abstracts of papers in MedLine from 1986 - 1993 shows that this advice has been honored mainly in the breach. “Permease” appears in 520 citations, “translocase” in 160, and “replicase” in 362.

Even though all the injunctions regarding enzyme nomenclature are not being followed, it is the case that the efforts of the International Union of Biochemistry to standardize this nomenclature have been generally successful. Without the efforts of the Commission, enzyme nomenclature might have collapsed in chaos. And, even where the nomenclature has not yet been fully standardized, the development of a structured numbering scheme allows a relatively stable approach to characterizing enzymes that catalyze particular reactions.

A Modest Proposal

A new, integrated approach to genetic nomenclature is needed. It makes little sense to have systematic rules and standards for the naming of organisms, for the naming of parts of organisms, and for the naming of proteins in organisms, yet to have no equivalent standards for the naming of genes in organisms. A more rigorous approach to the molecular notion of homology is also needed, since clear concepts of homology will be essential for a consistent genetic nomenclature.

Currently, some molecular workers use “homology” in the strict sense of meaning “similar by descent,” whereas others use it only to mean “similar in sequence” (cf., Donoghue, 1992). Genomic databases should consider that providing sufficient data to allow the detection of probable homologies to be part

of their mission. One or more databases specifically dedicated to maintaining and publishing (electronically) asserted homologies and the underlying evidence would be useful.

Although full sequences are available for some viruses (c.f., Kutter, this volume), a complete sequence for a free-living organism has not yet been obtained. Significant efforts are being made to identify and characterize all of the coding regions for some microbes (e.g., Neidhardt, this volume), at present, only a small percentage of genes have been identified and characterized in any species. Until many more are described, and until considerable comparative data becomes available, it will be impossible to develop a coherent plan for a global genetic nomenclature. Now that genomic research is discovering new genes rapidly, it might be wise to adopt an explicitly provisional approach to nomenclature while awaiting the development of a better scheme

To be sure, some proposals have been made for new approaches to the provisional naming of genes. For example, in microbiology the following suggestions have been offered (Nierlich, 1992):

1. Where applicable, the new gene may be given the same name as a homologous gene already identified in another organism.

2. The gene may be given a provisional name based on its map location in the style *yaaA*, analogous to the style used for recording transposon insertion. [That is, *y* designates a provisional name, the next two letters represent the map position, and the final upper-case letter is a serial number indicating occurrence. For example, *ybbC* would be the third provisionally named reading frame in the 11-12 minute interval of the map.]

3. A unique, provisional name may be given in the Demerec style.

However, this plan only addresses the need for short-term provisional nomenclature to allow the temporary naming of an open reading frame until a full name can be assigned. And, all of these recommendations are flawed in one way or another, especially in long-term adequacy.

1. Adopting names from other taxa may sometimes work, but it would also allow, perhaps encourage, bad nomenclature to propagate from species to species.

2a. Embedding semantic content into arbitrary identifiers is never a good idea. If it turns out that a provisionally identified gene was incorrectly located, either its name must change (thereby invalidating the whole point of assigning a name) or the embedded semantics of map position will no longer be valid (thereby invalidating the whole point of embedding the semantics).

2b. This approach allows only 26 provisionally named genes per minute of map. Yura et al. (1992) found more than 26 open reading frames in one minute of the *E. coli* chromosome.

3. The recommendation that provisional names be syntactically correct provides no specific guidance for provisional naming at all. If we use up a reasonable name space for provisional names, later to be revised, there will be no names left, if debilitating synonymies and homonymies are to be avoided.

An Example from Taxonomy

More than a century ago, de Candolle (1867) recognized the conflict between provisional and final naming in botanical taxonomy:

> There will come a time when all the plant forms in existence will have been described; when herbaria will contain indubitable material of them; when botanists will have made, unmade, often remade, raised, or lowered, and above all modified several hundred thousand taxa ranging from classes to simple varieties, and when synonyms will have become much more numerous than accepted taxa. Then science will have need of some great renovation of its formulae. This nomenclature which we now strive to improve will then appear like an old scaffolding, laboriously patched together and surrounded and encumbered by the debris of rejected parts. The edifice of science will have been built, but the rubbish incident to its construction not cleared away. Then perhaps there will arise something wholly different from Linnaean nomenclature, something so designed as to give certain and definite names to certain and definite taxa.
>
> That is the secret of the future, a future still very far off.
>
> In the meantime, let us perfect the binomial system introduced by Linnaeus. Let us try to adapt it better to the continual, necessary changes in science, ... drive out small abuses, the little negligences and, if possible, come to agreement on controversial points. Thus we shall prepare the way for the better progress of taxonomy.

Geneticists first need to achieve a Linnaean-like comprehensive, formal approach to genetic nomenclature. Then they could do far worse than to heed the advice of de Candolle and recognize that present nomenclature is mere scaffolding, sure to be supplanted when known genes number in the hundreds of thousands and full sequences are available for many species.

SUMMARY

It is now time to emphasize more integrative approaches to molecular biology, and this will undoubtedly require the participation of those with special interests in integrative methods, since the intuition of bench researchers cannot be fully adequate across the range of materials to be brought together. Electronic

information resources will play a crucial role in the integrative methods, provided that the data in them are of sufficient quality and in proper format so that they can be used as raw input for future analyses.

A new discipline of comparative genomics is emerging, facilitated by the technical advances accompanying the human genome project. This field will require new analytical methods that permit whole-genome data sets to be manipulated and analyzed. This, in turn, will require the availability of appropriate data. Several impediments, both technical and conceptual, currently block the development of appropriate information resources.

Having overcome the data-acquisition crisis of the 1980s, we now face a data-integration crisis of the 1990s. Proposals to build a federated information infrastructure for biology are promising, but the technical methods for implementing such a system are yet to be devised. Conceptual advances will be required before full semantic integration can be achieved. Some of our most basic biological concepts, for example “gene” and “genomic map,” are in need of rethinking.

Genetic nomenclature is unsystematic, especially across widely divergent taxa where the results are chaotic and wholly unsatisfactory. A new and comprehensive comparative approach to genomic nomenclature is needed, with special emphasis on the provisional naming of genes. Geneticists should recognize that present nomenclature is mere scaffolding, sure to be supplanted when known genes number in the hundreds of thousands and full sequences are available for many species.

BIBLIOGRAPHY

Abel, Y., and Cedergren, R. (1990). The Normalized Gene Designation Database. Available electronically via ftp from ncbi.nlm.nih.gov.

Benzer, S. (1956). The Elementary Units of Heredity. In: McElroy, W. D. and Glass, B. (eds). *A Symposium on the Chemical Basis of Heredity*. Johns Hopkins University Press, Baltimore, MD. pp. 70-93

Bright, M. W., Hurson, A. R., and Pakzad, S. (1994). *Multidatabase Systems: An Advanced Solution for Global Information Sharing*. Los Alamitos, California, IEEE Computer Society Press.

Caspersson, T. (1936). Über den chemischen Aufbau der Strukturen des Zellkernes. *Acta Med. Skand*., 73, Suppl. 8, 1-151.

Cinkosky, M. J., Fickett, J. W., Gilna, P., and Burks, C. (1991). Electronic data publishing and GenBank. *Science*, 252:1273-1277.

de Candolle, A. (1867). *Laws of Botanical Nomenclature*, quoted in Nicolson, D. H. (1991). A history of botanical nomenclature. *Annals of the Missouri Botanical Garden*, 78:33-56

Donoghue, M. J. (1992.). Homology. In: Keller, E. F., and Lloyd, E. A. (eds). *Keywords in Evolutionary Biology*. Cambridge, Massachusetts, Harvard University Press. pp. 170-179

Glass, B. (1955). Pseudoalleles. *Science*, 122:233.

Hammer, M., and McLeod, D. (1981). Database description with SDM: A semantic database model. *ACM Transactions on Database Systems* 6:351-386

Henikoff, S., Keene, M. A., Fechtel, K., and Fristrom, J. W. (1986). Gene within a gene: Nested *Drosophila* genes encode unrelated proteins on opposite strands, *Cell* 44:33.

Levine, A. S. (1988). GSA razes $100M PBS data project. *Federal Computer Week*, 2:1,53.

Lewin, R. (1986). DNA databases are swamped. *Science*, 232:1599.

Levy, M, and, Salvadori, M. (1992). *Why Buildings Fall Down*. New York, W. W. Norton & Company.

Lupski, J. R., Godson, G. N. (1989). DNA→DNA, and DNA→RNA→Protein: Orchestration by a single complex operon, *BioEssays*, 10:152-157.

Nierlich, D. (1992). Genetics nomenclature enters the computer age. *ASM News*, 58:645-646.

Nomenclature Committee of the International Union of Biochemistry. (1984). *Enzyme Nomenclature*. Academic Press, New York.

Petrosky, H. (1992). *To Engineer is Human*. New York, Vintage Books

Riley, M. (1993). Functions of the gene products of *Escherichia coli*. *Microbiological Reviews* 57:862-952

Riley, M., and Krawiec, S. (1987). Genome organization. In: Neidhardt, F. C., Ingraham, J. L., Low, K. B., Magasanik, B., Schaechter, M., and Umbarger, H. E. (eds). *Escherichia coli and Salmonella typhimurium*. Washington, DC, American Society for Microbiology. pp. 967-981.

Ritter, J. K., Chen, F., Sheen, Y. Y., Tran, H. M., Kimura, S., Yeatman, M. T., and Owens, I. S. (1992). A novel complex locus *UGT1* encodes human bilirubin, phenol, and other UDP-glucuronosyltransferase isozymes with identical carboxyl termini, *J. Biol. Chem.* 267:3257-3261.

Robbins, R. J. (1992). Database and Computational Challenges in the Human Genome Project. *IEEE Engineering in Medicine and Biology Magazine*., 11:25-34.

Robbins, R. J. (1993). Genome informatics: Requirements and challenges. In: Lim, H. A., Fickett, J. W., Cantor, C. R., and Robbins, R. J. (eds). *Bioinformatics, Supercomputing and Complex Genome Analysis*. Singapore: World Scientific Publishing Company.

Robbins, R. J. (1994a). Biological Databases: A New Scientific Literature. *Publishing Research Quarterly*, 10:1-27.

Robbins, R. J. (1994b). Genome informatics I: Community databases. *Journal of Computational Biology*, 1:173-190.

Robbins, R. J. (1994c). Representing genomic maps in a relational database. In: Suhai, S. (ed). *Computational Methods in Genome Research*. New York: Plenum Publishing Company. pp 85-96.

Robbins, R. J. (1994d). Genome Informatics: Toward a Federated Information Infrastructure (keynote address). The Third International Conference on Bioinformatics and Genome Research; Tallahassee, Florida; 1-4 June 1994.

Robbins, R. J. (1995). An information infrastructure for the human genome project. *IEEE Engineering in Medicine and Biology Magazine*., in press.

Sanderson, K. E., and Hall, C. A. (1970). *Genetics*, 64:215-228.

Sheth, A. P., and Larson, J. A. (1990). Federated database systems for managing distributed heterogeneous, and autonomous databases. *ACM Computing Surveys*, 22:183-236.

Singer, M., and Berg, P. (1991). *Genes & Genomes*. University Science Books, Mill Valley, California.

Sturtevant, A. H. (1913). The linear arrangement of six sex-linked factors in *Drosophila* as shown by their mode of association, *Journal of Experimental Zoology*, 14:43-59.

Sturtevant, A. H., and Beadle, G. W. (1939). *An Introduction to Genetics*. W. B. Saunders Company, Philadelphia,

United States Department of Energy. (1990). *Understanding Our Genetic Inheritance. The U. S. Human Genome Project: The First Five Years.*

United States National Academy of Sciences, National Research Council, Commission on Life Sciences, Board on Basic Biology, Committee on Mapping and Sequencing the Human Genome. (1988). *Mapping and Sequencing the Human Genome*. Washington, DC: National Academy Press.

Watson, J. D., Hopkins, N. H., Roberts, J. W., Steitz, J. A., and Weiner, A. M. (1992). *Molecular Biology of the Gene*. Benjamin/Cummins Publishing Company: Menlo Park, California.

Webb, E. C. (1993). Enzyme nomenclature: A personal retrospective. *The FASEB Journal*, 7:1192-1194.

Woodger, J. H. (1952). *Biology and Language*. Cambridge University Press, Cambridge.

Yura, T., Mori, H., Nagai, H., Nagata, T., Ishihama, A., Fujita, N., Isono, K., Mizobuchi, K., and Nakata, A. (1992). Systematic sequencing of the *Escherichia coli* genome: analysis of the 0-2.4 min region. *Nucleic Acids Research*, 20:3305-3308.

KEY SOURCES

For a compelling vision of how information infrastructure is affecting biology, read: Gilbert, W. 1991. Towards a paradigm shift in biology. *Nature,* 349:99.

Are databases becoming a new scientific literature? For thoughts on this, see: Cinkosky, M. J., Fickett, J. W., Gilna, P., and Burks, C. (1991). Electronic data publishing and GenBank. *Science*, 252:1273-1277, and Robbins, R. J. (1994). Biological Databases: A New Scientific Literature. *Publishing Research Quarterly*, 10:1-27.

Two special issues of journals provide a wealth of information: A special issue of *Trends in Biotechnology* (1992, volume 10, number 1) carried many articles on biological information management. A forthcoming special issue of *IEEE Engineering in Biology and Medicine* (1995, volume 14, number 6, Nov/Dec) is scheduled that will feature a number of papers on computational issues relating to genome informatics.

Useful general background on computational implications of genomics may be found in: Frenkel, K.A. 1991. The human genome project and informatics. *Communications of the ACM*, 34:41-51, Lander, E.S., Langridge, R. and Saccocia, D.M. 1991. Mapping and interpreting biological information. *Communications of the ACM*, 34:33-39, and Robbins, R. J. (1992). Database and Computational Challenges in the Human Genome Project. *IEEE Engineering in Medicine and Biology Magazine*., 11:25-34.

Systematists have been managing large amounts of complex information for centuries. Useful insights from that community may be found in: Hawksworth, D. L. (ed.) 1988. *Prospects in Systematics*. Clarendon Press, Oxford