BIOINFORMATICS REVIEW

Vol. 16 no. 2 2000
Pages 79–95

The complexities of genome analysis, the Retroid agent perspective

Marcella A. McClure

Department of Microbiology and Center for Computational Biology, Montana State University, Bozeman, MT, USA

Received on February 23, 1999; revised on June 15, 1999; accepted on June 29, 1999

Abstract

Motivation: *The sequences of Retroid agents from a wide diversity of organisms constitute the largest set of complete genomes currently available for the study of genomic architecture and the transfer of information within and between organisms. These agents are ubiquitous in Eukaryotes, comprising 50–90% of the genomic information in some cases.*

Results: *Analyses conducted for over a decade illustrate that Retroid agents are engaged in a wide spectrum of molecular evolutionary events. A description of these complexities is presented as a three parameter conceptual framework that considers type, size, and mechanism of events that contribute to the evolution of genes, genomes, and organisms. The results of new data mining studies further illustrate the complexity of the network of relationships among and between Retroid agents and other organisms. A hidden Markov model construction strategy is presented that generates a multiple alignment more similar to those refined by human experts.*

Contact: *mars@parvati.msu.montana.edu*

Introduction

This article was written at the request of the editors of Bioinformatics. It is based on a keynote address, *The Complexities of Genome Analysis: Viruses, Transposable Elements and Introns*, given at the V International Meeting on Intelligent Systems for Molecular Biology (ISMB), Halkidiki, Greece in 1997. In an effort to update ideas presented at ISMB-97, aspects of my laboratory's work presented at the Genomic Informatics Workshop, 1998, and the Pacific Symposium on Biocomputing, 1999 are integrated into the paper. In converting the ISMB-97 address into a paper I hope to provide a historical perspective on the interplay between empirically derived biological data sources, bioinformatic tools and human decision making in the study of Retroid agent evolution. Regardless of the name given to this type of multidisciplinary *in silico* research (bioinformatics, genomics, computational biology, or biological informatics), the goal is the creation of new biological knowledge.

Thinking about genes and genomes as linear strings of characters that differ from one another due to point mutation is no longer adequate to describe the dynamic nature of genome evolution. The Retroid agents provide the largest model system of related genomes available to illustrate an evolutionary computational biology approach to the analysis of genes and genomes. The relationship between Retroid and host genomes is complex. In some cases, Retroid agents are involved in pathogenicity, while in others they have co-evolved to become integral to their hosts' survival. Given the ubiquity of these agents in Eukaryotes, studies in comparative genomics will necessarily include the analysis of the relationships among and between Retroid and host genomes (Table 1).

For more than 13 years I have tested and used the most advanced bioinformatic tools available to track and trace the evolutionary relationships among all genes common to Retroid genomes and their hosts. These studies have allowed me to suggest evolutionary models that not only account for sequence divergence, but also for genome construction. I have developed a conceptual framework, referred to as complex genome analysis (CGA), that allows one to think about the genome as a dynamic system evolving by a variety of mechanisms at different molecular levels. I will provide a brief historical background for the use of various terms to describe observed evolutionary events in genes and genomes. Data from studies on the relationships among the Retroid genomes will illustrate the usefulness of CGA approach. I will also provide a chronology of the various types of software methods used during the last 13 years and comment on the current state-of-the-art of analyzing extremely divergent genomic sequence information.

What is complex genome analysis?

The multidisciplinary nature of computational biology necessitates the use of a consistent language to describe the variety of evolutionary events and processes that give rise to genes and genomes. To the non-expert the complexity of evolution can appear even more bewildering

Table 1. Distribution of Retroid agents among Eukaryotes and Eubacteria. The +a denotes that *gypsy* can become an infectious virus. The +b denotes that other *gypsy*- and *copia*-Retrotransposons encode putative ENVs, although infectivity has not been demonstrated to date. Adapted from McClure (1999)

Retroid agents	Eukaryotes											Eubacteria	
	Human	Vertebrates	Invertebrates	Plants	Fungi	Protists				Plastids	Baculovirus	Genome	Conjugative transposons
						Slime Mold	Alga	Protozoa	Oomycetes				
Retroviruses	+	+	+a										
Pararetroviruses													
Caulimoviruses				+									
Badnaviruses				+									
Hepadnaviruses	+	+											
Transposons:													
Retrotransposons													
Gypsy-		+	+b	+	+	+		+	+	+	+		
DIRS1-			+		+	+							
Copia-		+	+b	+	+		+		+	+			
Retroprosons	+	+	+	+	+	+	+	+		+			
Retrointrons										+		+	+
Retroplasmids					+					+			
Retrons												+	
Retrophages												+	

due to the inconsistent use of terminology to describe events. The CGA framework is an attempt to present the dynamic nature of the genomic evolutionary process in a consistent terminology that can aid in development of new bioinformatic tools to analyze data.

In the CGA framework three parameters are defined to describe the evolution of genes and genomes. *Types of change* in sequence information can be due to point mutation, duplication, and insertion/deletion (indel) events. *Units of informational acquisition* can be: individual motifs; an ordered-series-of-motifs (OSM), or subset thereof, (modular evolution); an entire gene or sets of genes (segmental evolution); and entire genomes (horizontal evolution). *Modes of acquiring information* can be due to splicing, transposition, translocation, inversion, various types of recombination, and any other means of acquiring genetic information.

The concept of gene evolution by acquisition of 'gene pieces' or modules is well-documented (Blake, 1978; Gilbert, 1978). The term 'modular evolution' is applied to the genome evolution of bacteriophages by acquisition of entire genes or groups of genes (Botstein, 1980). This term is also used to describe distant sequence similarity among groups of genes, single genes, and subsets of genes found in different arrangements within positive-strand RNA viruses (Strauss and Strauss, 1988; Zimmern, 1988). Modular evolution is also used to refer to the rearrangement of exons (Dorit *et al.*, 1990; Baron *et al.*, 1991; Yanagawa *et al.*, 1993), and of catalytic and substrate specific regions of variable size not bounded by introns (Diaz *et al.*, 1990).

To bring some order to the plethora of these observed evolutionary events I use the term *modular evolution* to denote the acquisition of the OSM, or a subset thereof, conferring protein function or structural integrity (Figure 1). An OSM, that may span hundreds of residues, is defined as a set of conserved or semi-conserved motifs found in the same arrangement relative to one another in all the sequences of a protein family (Figure 1). A single motif is comprised of a contiguous string of 1–9 amino acid residues, but occasionally indels do occur. A single amino acid can only be recognized as a motif in the context of the OSM. Modular evolution, i.e., the acquisition of an OSM, may or may not correspond to individual structural domains (e.g., more than one motif may occupy a structurally defined domain). The spacing between the motifs can be highly variable, reflecting the regions of a protein that are less restricted by functional/structural constraints. These regions may evolve more rapidly, and be more subject to indel and duplication events, than the OSM (see McClure, 1992 and McClure *et al.*, 1994 for further discussion and examples). Nonetheless, these motif-intervening-regions (MIRs) can also provide evolutionary information regarding sub-class specific relationships and should not be ignored in multiple alignments used for phylogenetic reconstruction. The term *segmental evolution* denotes the acquisition by genomes of protein-unit-length or larger regions (Figure 1).

Within the CGA construct, even more specific events and issues must be considered. Deletions can be of various types. The loss of a few nucleotides may keep a reading frame intact, or cause a frame switch that may or may not be compensated for by the translational machinery. A new frame switch could lead to gain of function(s) while the original function becomes lost. Another example is the identification of an individual motif within a sequence that

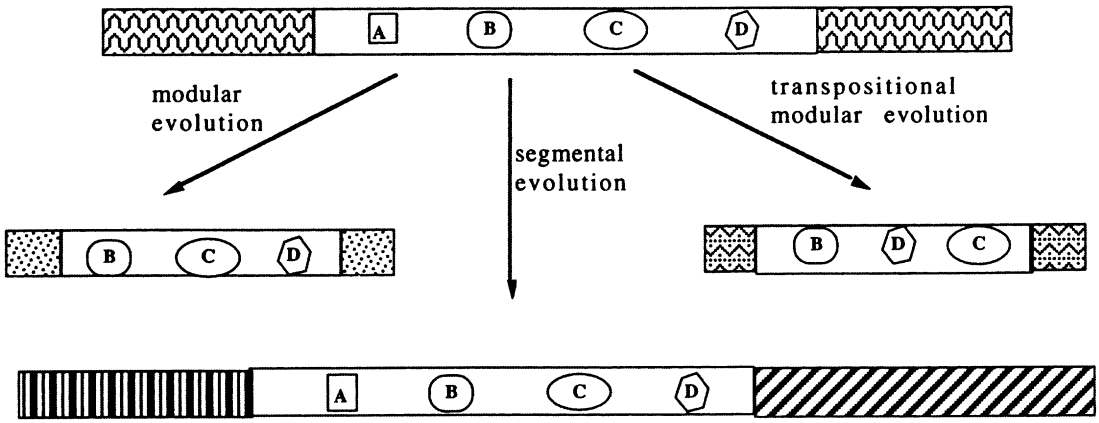

Fig. 1. Complex Genome Analysis. Schematic diagram of examples of various types of evolutionary events described in terms of the complex genome analysis parameters. Modular evolution is defined as the acquisition of a subset of the OSM. Segmental evolution is defined as the acquisition of an entire gene or set of genes. Transpositional modular evolution is the occurrence of a subset of the OSM in an arrangement that differs from the original order. The gene (white box) encodes individual amino acid motifs (A–D). The motifs comprise the OSM. The MIRs are the amino acids between the motifs. Patterned segments indicate different DNA backgrounds.

does not have any similarity to other motifs in the OSM. Convergence may account for such an observation rather than ancestry. Furthermore, various forms of active and passive acquisition of new information can occur within and between organisms. For example, a viral infection among cells or between organisms is an active acquisition, while engulfment of degradative cell products, including nucleic acids, is passive.

To date no single software system or group of programs can interpret sequences within the context of the variety of observed evolutionary events that give rise to gene and genome diversity. Throughout this paper various combinations of CGA parameters will be used to describe observations regarding the evolutionary process of RNA viruses, Retroid agents, and host genomes. I hope that examples of the new information generated by the integration of empirical observations and sequence analysis provided here will encourage development of new software methods to accelerate the pace of *in silico* research.

RNA genomes: model systems for evolution in the fastlane

The error prone polymerase, the lack of an RNA correction mechanism, and the high rate of replication give rise to RNA genomes that can mutate $1-10 \times 10^6$ times faster than DNA genomes. This high rate of accumulated mutation leads to the existence of virus populations as dynamic distributions of non-identical genomes, or quasispecies. A most fit sequence or master genome arises given a specific environment. This master genome may or may not be the average sequence of the population (for a recent review on viral population dynamics see Domingo and Holland, 1997). While it is well known that human immunodeficiency virus (HIV) populations exist as quasispecies, HIV is no different than Ebola, measles, or rabies viruses in this regard. Whether or not other Retroid agents rapidly mutate and create quasispecies is an active area of research. Although the heterogeneous nature of some agents is well documented there is little direct measurement of mutation rates. The rate of mutation has been directly measured for the *copia*-retrotransposon, Ty1, demonstrating that this agent can mutate as rapidly as a retrovirus (Gabriel *et al.*, 1996). One of the rare, active *copia*-retrotransposons of plants, Tnt1 of tobacco, is demonstrated to exist as a quasispecies-like population (Casacuberta *et al.*, 1997). Furthermore, RNA quasispecies can also undergo recombination events with other viruses leading to mosaic genomes and acquire unique functions from a variety of genetic sources that their relatives may not share.

In general, all RNA-based genomes can mutate and replicate at such accelerated rates that they encode related but highly divergent proteins. In contrast to many host genomic homologues, sequence relationships within viral families are often less than 25% identical for proteins with common functions. This low level of relationship among viral homologues, (and in some cases between viral and cellular homologues) allows us to predict amino acid residues involved in the function and structural integrity of proteins by identifying the common subsequences, the OSM among the sequences. The analysis of such protein sequence data continues to challenge computational biologists and algorithm developers. Such extremes in

evolutionary dynamics also make the development of effective vaccines and anti-viral agents difficult.

Some RNA viruses encode additional proteins by alternative reading frames, RNA editing, and splicing. Molecular population genetics analysis of multi-coding capacity sequences from the order Monegavirales (e.g., measles) reveals the evolutionary dynamics of maintaining 1–3 functions in a single gene (Jordan *et al.*, 2000). The ability to utilize a diversity of molecular mechanisms and undergo a variety of evolutionary events complicates all levels of genome analysis. RNA-based genetic agents not only provide us with rapidly evolving model systems for the development and testing of bioinformatic tools; the results of analyzing such data are also important in medicine and agriculture.

Retroid agents: a system for studying genome evolution

Viruses and transposable elements that mediate their replication and/or movement by an RNA intermediate are Retroid agents, a classification comprised of all genetic agents that encode the ability to convert RNA genomes into DNA genomes by use of the reverse transcriptase (RT) enzyme (Figure 2). Although the negative effects of Retroid agent integration into host genes are well known, the examples of beneficial relationships are just beginning to emerge (for a recent review and complete reference list see McClure, 1999). Some agents provide regulatory sequences for control of host genes and effect host phenotype by insertional mutagenesis. Others mediate recombination, repair dsDNA breaks, or participate in telomere maintenance in *Drosophila*. In mammals, including humans, endogenous retroviruses appear to play a role in suppression of the female immune system during fetal implantation. Although Retroid agents are ubiquitous in plants, and can comprise 50–90% of the genomic DNA, they are rarely expressed unless stressful conditions exist. Perhaps one of the most revealing findings regarding these agents is their mobilization in the undermethylated genome of progeny resulting from interspecific hybridization between *Macropus eugenii* and *Wallabia bicolor* (Australian wallabies). This mobilization occurred in one generation, illustrating that Retroid agents play a role in the restructuring of genomes and are no doubt involved in speciation (Kidwell and Lisch, 1998; Waugh O'Neill *et al.*, 1998).

Many Retroid agents have co-evolved with their host species over such long evolutionary time spans that they are integral to the organisms' survival. Given the ubiquity of these agents in Eukaryotic genomes large-scale genomic comparison projects will encounter a variety of relationships and functions leading to an evolutionary network among and between host and Retroid genomes. How did such a co-evolutionary state arise? It has been suggested that Retroid agents are the descendants of the entities that encoded the RT function necessary for the transfer of genetic information from RNA to DNA. Ancient ancestry may explain how some Retroid agents came to perform essential roles in their hosts (for a detailed review and discussion of the co-evolutionary nature of Retroids and hosts see McClure, 1999).

The Retroid core enzymes

Retroid genomes are relatively easy to separate and purify from the host. They were among the earliest complete genomes sequenced because many of them are of medical significance. By 1988 the Retroid data set approached 100 complete genomes from a variety of hosts. These data afforded the opportunity to study not only the evolution of homologous genes, but also genome construction. Two important points emerged from early studies; extreme divergence can occur in genes encoding common functions, and variability in gene complement and order is found in various Retroid genomes.

The *pol* gene of the retrovirus genome encodes four catalytic functions distributed over three proteins: 1) the aspartic acid protease (PR); 2) the RNA-dependent DNA-polymerase (RDDP) composed of two functional domains, the RT, and the ribonuclease H, (RH); and 3) the integrase (IN) with endonuclease activity. The independent proteins for the PR and IN are cleaved from a polyprotein that is encoded as a *gag/pol* read-through transcript. The remaining cleavage product of the *pol* portion of the polyprotein is the RDDP, with the RT and RH functions residing in the amino and carboxyl portions of the protein, respectively (Figure 3).

Early sequence comparisons suggested that the retroviral PR was a member of the aspartic acid protease family (e.g., pepsin) (Toh *et al.*, 1985). The pepsin family sequence underwent an ancient tandem gene duplication physically linking the two sets of the OSM required for catalysis. Retroid agent PR segments have not undergone this duplication event. Sequence comparisons suggested that the retrovirus PR would be a dimer (Pearl and Taylor, 1987; Doolittle *et al.*, 1989). This prediction was confirmed by X-ray crystal determination for the Rous Sarcoma virus (RSV) PR (Miller *et al.*, 1989).

Comparison of the retroviral RDDP and the *Escherichia coli* RH sequences predicted the correct positioning of the RT and RH functions within the protein (Johnson *et al.*, 1986). It should be understood that sequence similarities between the retroviral functional RH domains of the RDDP and the *E.coli* RH are so limited that database retrieval methods did not exist in 1985 that could indicate any potential relationship between these sequences. Although the retroviral and *E.coli* RH sequences were

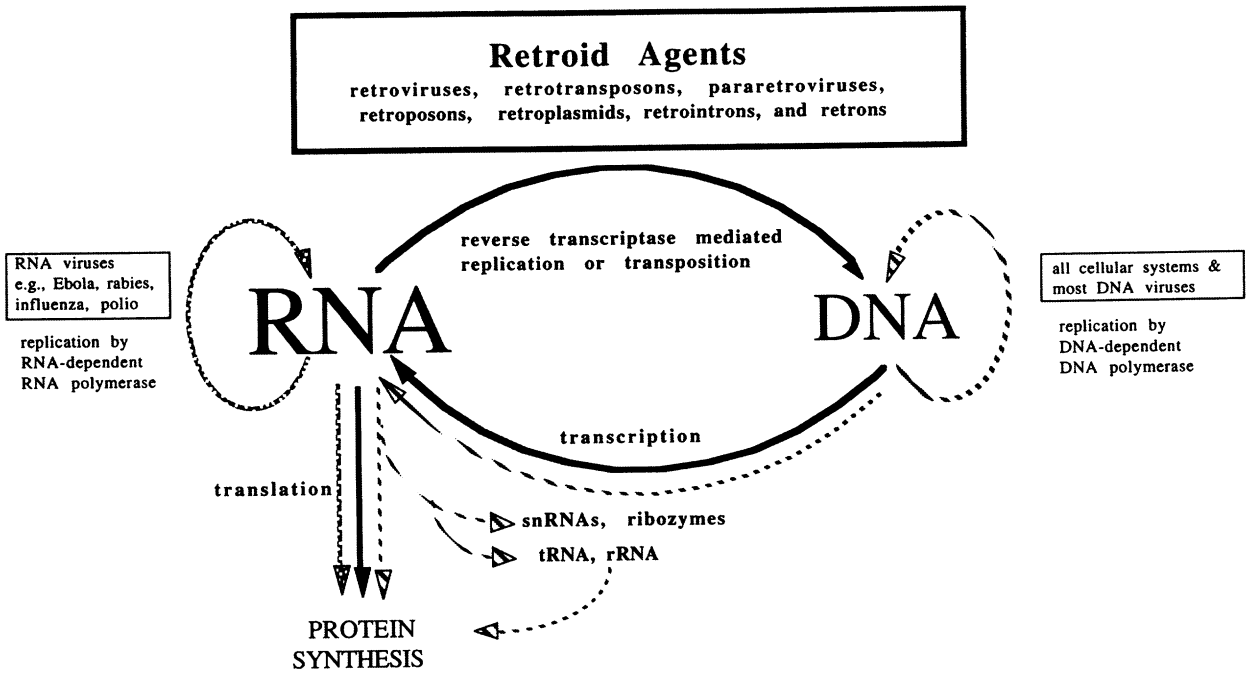

Fig. 2. RNA World Cycle. Schematic representation of the molecular life cycle for RNA- and DNA-based organisms. All Retroid agents use the reverse transcriptase in their life cycles (solid lines). True RNA viruses, i.e., no DNA stage, (strippled lines), and DNA viruses and host organisms (hatched lines) are indicated on the left and right side of the diagram, respectively. Adapted from McClure, 1992.

in the database they were retrieved by word search and visually inspected to identify any common subsequences. The RH sequences of several retroviruses and *E.coli* were then multiply-aligned in a text editor. The only readily available alignment approaches in 1985 were pairwise methods that proved to be inadequate for subsequence identification at this level of divergence. In addition, a Monte Carlo simulation of sequence similarities between any retroviral sequence and the *E.coli* RH failed the Dayhoff criteria (Schwartz and Dayhoff, 1978) for statistical confidence supporting homology (McClure, 1992). Nonetheless, our prediction regarding the position of the functional RT and RH domains by sequence analysis was confirmed by site-directed mutagenesis studies (Tanese and Goff, 1988).

It was obvious upon visual inspection that the RT functional domain possesses a hydrophobically embedded Asp-Asp (DD), the canonical motif of the RNA-dependent RNA-polymerases (RDRPs) of positive-strand RNA viruses (e.g., polio) (Kamer and Argos, 1984). These observations spawned a series of papers suggesting common ancestry by direct descent (no molecular or segmental evolution) for RDDPs and RDRPs (size range, approximately 200–2000 amino acids) (Argos, 1988; Poch *et al.*, 1989; Delarue *et al.*, 1990). While some of these cases maybe be statistically supportable, others are far from certain due to the lack of significant phylogenetic signals (Zanotto *et al.*, 1996). Suggested relationships based on structural similarities may be due to mechanistic analogy or convergence (Steitz, 1998), rather than common ancestry, when little or no statistical confidence can be generated in the sequence relationships. Even so, several analytical studies have tried to use the low similarity detected between some RDRP and RT sequences to root the RT tree (Xiong and Eickbush, 1990; Nakamura *et al.*, 1997; Nakamura and Cech, 1998). It has been suggested that all DNA polymerases share ancestry. Recent analysis of DNA polymerase structures indicates that there are four unrelated DNA polymerase families, one of which is the RT domain family (Brautigam and Steitz, 1998). Whether or not the RT domain shares common ancestry with the RNA-dependent polymerases is an open question.

Early sequence analysis of both RT and IN sequence relationships indicated the conservation of residues among retrovirus proteins (Johnson *et al.*, 1986). These studies were extended by the identification of subsequences common between proteins of Retroid lineage I and II, (Figure 4) (Doolittle *et al.*, 1989). Full alignments, refined by visual inspection, revealed the OSM (McClure, 1991). In these latter studies the minimal number of possible residues involved in RT/RH and IN functions was inferred

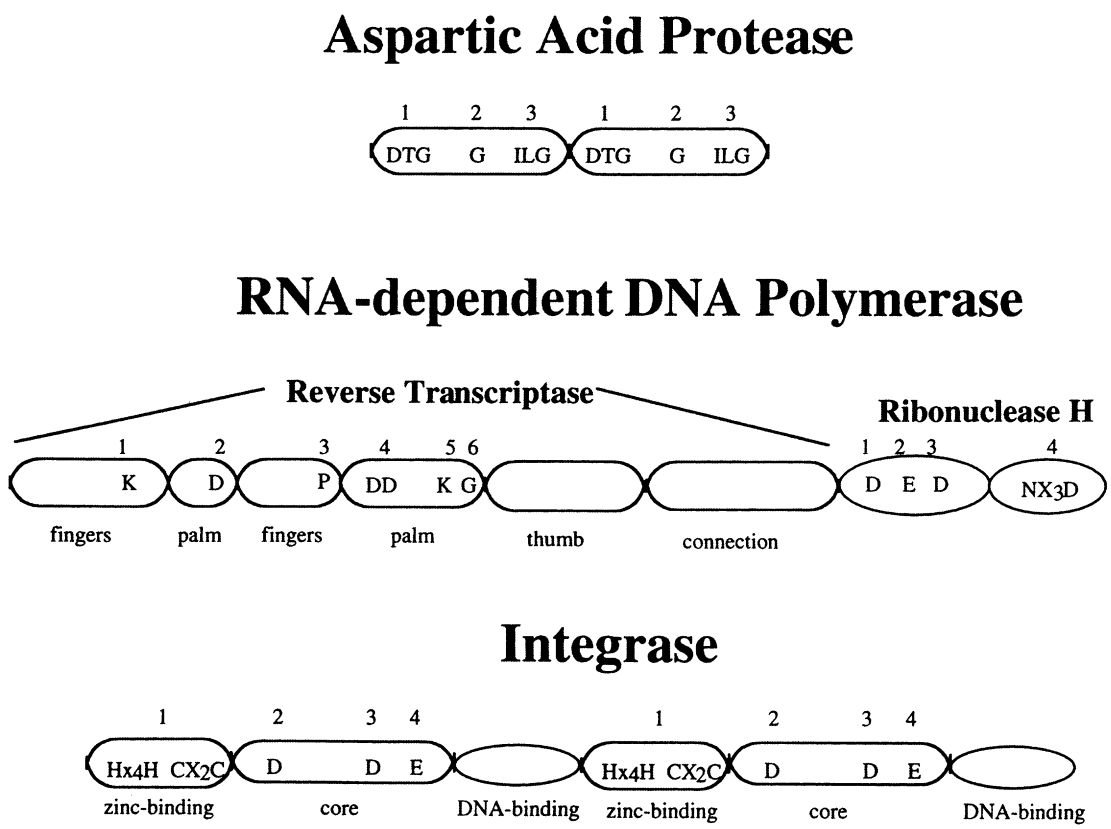

Fig. 3. Core Enzymes of the Retroid genome. The most highly conserved residues of the OSM of the Retroid aspartic acid protease (PR) are positioned within the homodimer. The RNA-dependent DNA polymerase (RDDP) is a monomer comprised of two functional domains, reverse transcriptase (RT) and ribonuclease H (RH). The most highly conserved residues of the OSM of the RT functional domain are placed within the structural domains (fingers, palm, fingers, palm, thumb, and connection) defined by the HIV-1 RT crystal. The most highly conserved residues of OSM of the RH domain are placed within the structural RH domains based upon comparison of the HIV-I and *E.coli* RH crystal structures. The integrase (IN) is divided into three domains: a zinc-binding region, H/H-C/C; the catalytic core that contains the D-D-E residues, also found in the polynucleotidyl transferases, and a DNA-binding domain. The most conserved residues of the IN OSM are placed in the homodimer. Amino acids are in the single letter code. × indicates approximate number of variable residues within motifs. Each motif of the OSM in each protein is labeled numerically. Adapted from McClure, 1999.

by comparing the sequences of lineage I and II (Figure 4). Based on the crystal structure, the RT functional domain of the RDDP is divided into six structural domains (fingers, palm, fingers, palm, thumb, and connection) (Kohlstaedt *et al.*, 1992). The second palm domain contains three of the six motifs of the ordered series found in the RT functional domain (Figure 3, motifs 4–6). The three aspartic acid (D) residues of the two palms form part of the catalytic site. Within the RH functional domain the two D and the glutamic acid (E) containing motifs are in the same proposed structural domain (Figure 3, motifs 1–3) (Davies *et al.*, 1991). These three residues are involved in metal binding (Kanaya *et al.*, 1990; Katayanagi *et al.*, 1990; Davies *et al.*, 1991). The first domain of the IN is a zinc-finger metal binding region (Burke *et al.*, 1992) as predicted by sequence analysis (Figure 3, motif 1). The second domain of the IN protein is the core region that contains three motifs that conserve the canonical DDE residues of the polynucleotidyl transferases (Figure 3, motifs 2–4) (Dyda *et al.*, 1994). Experimental evidence suggests that residues involving DNA-binding are located in the carboxyl portion of the IN, although some studies also implicate the core in this function (Brown, 1997). Both crystal structure and NMR studies show that IN functions as a homodimer.

How were sequence patterns of these early studies determined? The development of software to identify subsequences or motifs in highly divergent sequences was in the early stages (Waterman, 1984, 1986; Waterman *et al.*, 1984). One of the early multiple alignment methods was developed in response to the difficulties of analyzing Retroid protein sequences (Feng and Doolittle, 1987).

This method uses the Needleman-Wunch algorithm in the progressive multiple alignment approach (Waterman and Perlwitz, 1984). Only with the use of unpublished parameters, however, can subsequences be identified within and between Retroid sequences and cellular homologues.

Many sequences were visually inspected for common subsequences. From this work emerged the idea of an OSM defining residues important in function and structural integrity. This was not a new idea. In 1983, Dayhoff posed the question 'Are identical residues (4–9) in related proteins clustered together?' (Dayhoff *et al.*, 1983). The concept of an OSM is a natural extension of this idea.

Although the Retroid enzymes conserve the OSM that is integral to protein function and structural integrity, the sequence relationships between various lineages fail homology confidence tests as mentioned above. Attempts at generation of robust phylogenetic trees using several different published RT sequence alignments and various reconstruction approaches do not support homology either (Zanotto *et al.*, 1996). Given these facts I refer to the sequences that fail homology criteria but retain the OSM as *f*unctionally *e*quivalent *p*roteins (FEPs). The use of both homologous and FEP sequences in 'phylogenetic reconstruction' naturally leads to the idea of expressing this relationship as a mixture of supported phylogenetic lineages (within major lineages) and a *f*unctionally *e*quivalent *n*etwork, FEN (between major branches and lineages). Such a network could clearly delineate which paths relate statistically supported ancestors in contrast to those relationships based on conservation of the OSM and observed function. I use the term FEN to denote the deep relationships among the genes common among Retroid agents, and between these agents and their hosts (Figure 4).

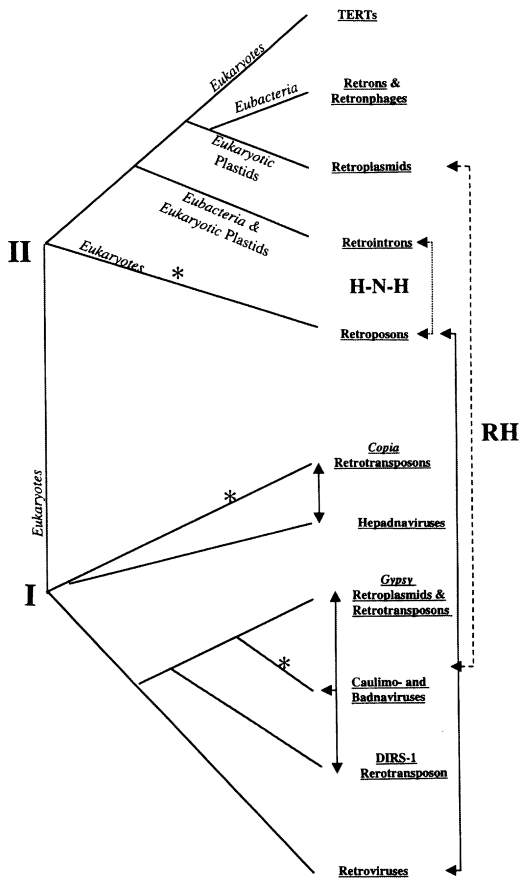

Fig. 4. The functionally equivalent network (FEN) of the Retroid agents. The topology of the FEN is based on the pairwise distance values of RT sequences representative of the Retroid agents and is unrooted. Two major branches are indicated I and II. Along these branches various genomes acquired or lost subsets of the retroviruses genomic complement (McClure, 1991, 1993, 1999). The RDDP sequence similarity among *gypsy*-retrotransposons and retroplasmids, the plant pararetroviruses (Caulimo/Badnaviruses) and DIRS1 retrotransposons, suggests that either these genomes acquired various subsets of genes from one another or lost genes from a hypothetical ancestor (three-way arrow). The PR and RDDP sequence similarity between Hepadnaviruses and *copia*-retrotransposons, suggests that one lineage acquired this part of the genome from the other or that the Hepadnaviruses lost some of the gene complement of the hypothetical ancestor (two-way arrow). Branch II bifurcates into the Eukaryotic-specific retroposons, and the Eubacterial lineage that also includes the Eukaryotic TERTs. Independent FEN topologies for RT and RH sequences indicate that the RH sequences of the Eubacterial retroplasmids are more similar to that of both *gypsy*-retrotransposons and plant pararetroviruses (dashed line, double-arrow bracket). The RH sequences of the Ingi3 and I factor, a subset of retroposons, are more similar to that of the retroviruses (solid line, double-arrow bracket). The presence of *in* genes is variable in some lineages. Asterisks (*) indicate genomes where the *in* genes are located upstream of the RT encoding region. In Ingi3-like retroposon genomes the first motif of IN functionally equivalent proteins (FEPs) also contain, the H-N-H motif common to the retrointrons and other intron and bacterial endonucleases (dotted line, double-arrow bracket). The names of Retroid agents and other abbreviations are defined in the text.

Genome complexities

Genome analysis can be done at various levels. Here we are concerned with patterns of conserved amino acids critical for protein function and structure that are revealed by comparing the highly divergent FEPs and homologous protein sequences (pairwise identity >30%). At the other extreme, the molecular population genetics analysis of closely related genomes (pairwise identity of nucleic acid sequences >70%) can provide information on the type of evolutionary selection operating on rapidly evolving genomes. Retroid agents provide us with a rich database exemplifying the complex evolutionary dynamics of genes and genomes that have co-evolved with all Eukaryotic genomes. The frequency of genomic acquisition and rearrangement, and the impact these changes have not only on pathogen diversity and host range, but also on organismal evolution and speciation is becoming apparent.

Modular and segmental evolution

While all Retroid agents encode the RT, variability in both complement and order of other genes is observed within different Retroid lineages. There is only one RT function, the Eukaryotic Telomerase Reverse Transcriptase (TERT), encoded by a cellular gene (Lingner *et al.*, 1997; Nakamura and Cech, 1998). The RT encoded by a *pol* gene is only found in the Eubacterial retrons and Eukaryotic TERTs as a solo function. The more common *pol* gene encodes both domains, RT/RH, of the RDDP. Cellular *rh* genes are ubiquitous in both Eukaryotes and Eubacteria. There is no evidence, currently, to support the notion that the initial RT/RH agent was derived by recombination of a retron or TERT *rt* gene, and an Eukaryotic or Eubacterial host *rh* gene, although such a relationship may no longer be detectable. Regardless of how the original RT/RH linkage arose, Retroid genome construction models have been hypothesized that are consistent with the available data (McClure, 1991, 1993, 1999). Phylogenetic reconstruction and relative rate analyses for all homologous proteins within the retrovirus lineage, as well as between the two major lineages of Retroid genomes, provides evidence for several cases of modular and segmental evolution. In addition to homologous recombination, xenologous recombination (replacement of a homologous resident gene with a homologous foreign gene), and acquisition of genes unique to a specific branch of a virus family also contribute to genomic evolution (Figure 4).

While the frequency of segmental and modular evolution among and between the genomes of hosts (which include transposons and endogenous viruses) and their intracellular obligate parasites (viruses) is just beginning to be observed; the high rate of homologous recombination within a specific retrovirus branch is well documented. HIV, however, was not thought to undergo homologous recombination because neither dual infection of cells (more than one type of HIV), nor superinfection (the ability to infect a cell a second time with a closely related virus), had been demonstrated *in vivo*. Early molecular genetics analyses demonstrated, however, that recombination occurs within HIV quasispecies (Delassus *et al.*, 1991) and between HIV subtypes (Li *et al.*, 1988). Since then a number of HIV 1 and 2 sequence analysis studies have demonstrated that both modular and segmental evolution are observed within human primate lentiviruses (Robertson *et al.*, 1995) creating mosaics of both genes and genomes (for review and references see McClure, 1996). Although recombination between HIV-1 and HIV-2 has not been observed, dual infections of these viruses are occurring in India (Grez *et al.*, 1994). Evidence for recombination between distantly related retroviral genomes is rare, but it does occur. Recombination is also observed between Ty 1 and 2 agents of *copia*-retrotransposon lineage (Jordan and McDonald, 1998, 1999). Given the observed acquisitions between HTLV-I and II (McClure *et al.*, 1988), it will not be surprising if both modular and segmental evolution occurs between the even more closely related HIV-1 and 2 viruses.

Compensatory and modular evolution: the 5' end of the pol *gene and the TAR element*

As more data accumulate on the various types of mosaicism observed in HIV genes, thinking of such events as modular may aid in understanding why specific regions of the genome can tolerate recombination while others cannot, and what selective advantages may be conferred by such recombinations. As mentioned above, much more complex evolutionary scenarios are observed than existing software can track. For example the evolutionary mapping of a functional association has been described (Gao *et al.*, 1996). The 5' end of the HIV-1 genome contains a regulatory region with known secondary structure, called TAR, to which one of the small virally encoded regulatory proteins binds. In HIV I, intersubtype A/E viruses, there is an association with the bulge size in the TAR element and the origin of the 5' region of the *pol* gene. Most HIV-1 genomes have a three base bulge in the TAR element. Those genomes that acquired the 5' region of the *pol* gene from HIV-1, subtype A, have only two bases in the TAR element. Such an association is not confined to HIV-1; it is also observed in SIV_{AGM}/SIV_{SM} hybrids. There is experimental evidence that supports the idea of a functional association between the RDDP and the TAR element. These data fit within the parameter description of CGA. In this case there appears to be compensatory evolution between a single point mutation in the TAR bulge and the modular evolution of the *pol* gene by homologous recombination, clearly illustrating the complexities of genome evolution.

Analogy and rearrangement: the endonuclease function

There are two cases of analogous function, endonuclease and ribonucleoprotein, found within various Retroid genomes (McClure, 1999). Only the endonuclease example will be described here. This function is necessary for the integration of Retroid DNA into a host genome. In retroviruses the IN function is encoded in the *pol* gene and it is produced by cleavage of the viral polyprotein (GAG/PR/RT-RH/IN). While the region encoding IN is in a homologous position in *gypsy*-retrotransposons it is found upstream of the RT-RH in *copia*-retrotransposons, petunia vein-clearing virus, a Caulimovirus (Richert-Poggeler and Shepherd, 1997) and the *s*pliced *l*eader *a*ssociated *c*onserved *s*equences, SLACS/CRE-like retroposons (Villanueva *et al.*, 1991). Although the position of the IN segment in the *pol* gene is variable, the RH domain

when present, is always downstream of the RT portion of the RDDP (Figure 5).

Among the retroposons not only is the position of IN variable, but so is the source of the endonuclease activity itself. Early sequence analysis of SLACS/CRE-like retroposons indicated that these agents encode the zinc-binding domain found in other Retroid IN sequences (Villanueva *et al.*, 1991). Motif identification analysis confirms the presence of this domain in SLACS/CRE-like retroposons and identifies the remaining motifs of the IN OSM just upstream from the RDDP (Figure 5). LINEs-like retroposons encode a protein with significant similarity to the cellular apurinic/apyrimidinc endonucleases (APEs) (Martin *et al.*, 1995) rather than the IN found in retroviruses. Although endonucleolytic activity has been demonstrated for both R2Bm and R1Bm, only the latter encodes an APE-like protein (Feng *et al.*, 1998). The R2Bm agent encodes an endonuclease function within the protein with RT activity (Xiong and Eickbush, 1988). The residues responsible for the endonuclease function in R2Bm, however, have not been experimentally identified.

Two types of putative endonucleases are encoded in the ingi-like TRS 1.6 genome in the same open reading frame (ORF) as the RDDP (McClure, 1999). In contrast to Ingi3, the TRS 1.6 genome has a longer ORF that encodes a putative APE, the RDDP, and IN-like sequences. In the Ingi3 genome the IN sequence is in the same reading frame as the RDDP and the APE is in an upstream reading frame (Figure 5). Although the coding sequences for each of these endonucleases are in different reading frames they are both interrupted by stop codons (Figure 5). It is well known that retroviral mRNA stop codons can be read through and frame shifts can occur during translation to produce polyproteins. Whether or not this occurs in the Ingi3 lifecycle to produce IN and/or APE proteins remains to be experimentally determined, but there is nothing to impede the production of both these putative endonucleases in the TRS 1.6 agent. Are the Ingi3-like agents the descendants of a genome that once encoded both APE and IN functions? All other retroposon genomes analyzed to date potentially encode either the IN, or the APE, however only the zinc-binding motif of IN is present in retrointrons (McClure, 1991) (Figure 5).

The IN zinc-binding domain is conserved in Ingi3-like retroposons and most mitochondrial, and bacterial retrointrons (Mohr *et al.*, 1993). Furthermore, a third endonuclease similarity has been noted between the H-N-H motif (H is histidine, N is asparagine) found in various bacterial endonucleases and the IN zinc-binding motif (Figures 3 and 6) present in many of the retrointrons, a subset of group II introns. Although there is no statistical support for common ancestry, these data provide the first indication of similarity between group I and II introns (Gorbalenya, 1994; Shub *et al.*, 1994). This similarity is observed as an overlap between the two motifs in which the second H of the IN zinc-binding motif is the first H of the bacterial endonuclease H-N-H motif. Mutagenesis studies on the zinc-binding domain of the retrointrons confirm that this region contains the endonuclease activity (Zimmerly *et al.*, 1995). Although the retrointrons do not encode the other two domains of the IN, the Ingi3-like agents not only encode these potential domains, they also have the H-N-H domain found in bacterial endonucleases (Corro and McClure, unpublished). It is remarkable to find similarity to three different types of endonucleases; Eukaryotic (APE), Eubacterial (H-N-H) and Retroid (IN) in a single genome. Are these features derived or acquired? These observations provide us with a glimpse of the evolutionary possibilities of endonuclease function. Only when empirical data are available will we know which of these region(s) of the Ingi3-like retroposon genomes encode an active endonuclease function(s).

Pathogen/Host gene networks: dut *gene evolution*

Early sequence analysis revealed significant similarity between two regions of the *pol* gene in all members of the mouse mammary tumor virus (MMTV)-like viruses, (including human endogenous retroviruses and the intercisternal particles of rodents) and all non-primate Lentiviruses but not any other retrovirus (McClure *et al.*, 1987). What is most surprising is the location of these two segments within their respective genomes. In the MMTV-like branch this segment is found immediately upstream of the PR portion of the *gag* gene (Power *et al.*, 1986). In the non-primate Lentivirus branch this segment is found between the RH domain of the RDDP and the IN (Figure 6).

Significant sequence similarity was detected between these two segments in polyphelytic lineages of retroviruses and regions of the DNA viruses Herpes and Vaccina (Mercer *et al.*, 1989; Slabaugh and Roseman, 1989). By 1990 the expanding sequence database and better search algorithms provided evidence that this segment in Herpesviruses was similar to the deoxyuridine triphosphatase (dUTPase) (McGeoch, 1990). In 1992 it was experimentally demonstrated that these novel retroviral gene segments encode a dUTPase (Elder *et al.*, 1992).

Our current dUTPase data set is comprised of 51 viral and 24 host sequences representative of all available viral and host Eukaryote, Archaea, and Eubacteria genomes (Baldo and McClure, 1999). A new hidden Markov model (HMM) construction strategy was employed to generate a multiple alignment of this divergent data set (see below). The results of our analyses indicate that multiple transfers of the *dut* gene have occurred between hosts and viruses, and among bacterial hosts. The evidence for horizontal transfer is particularly interesting

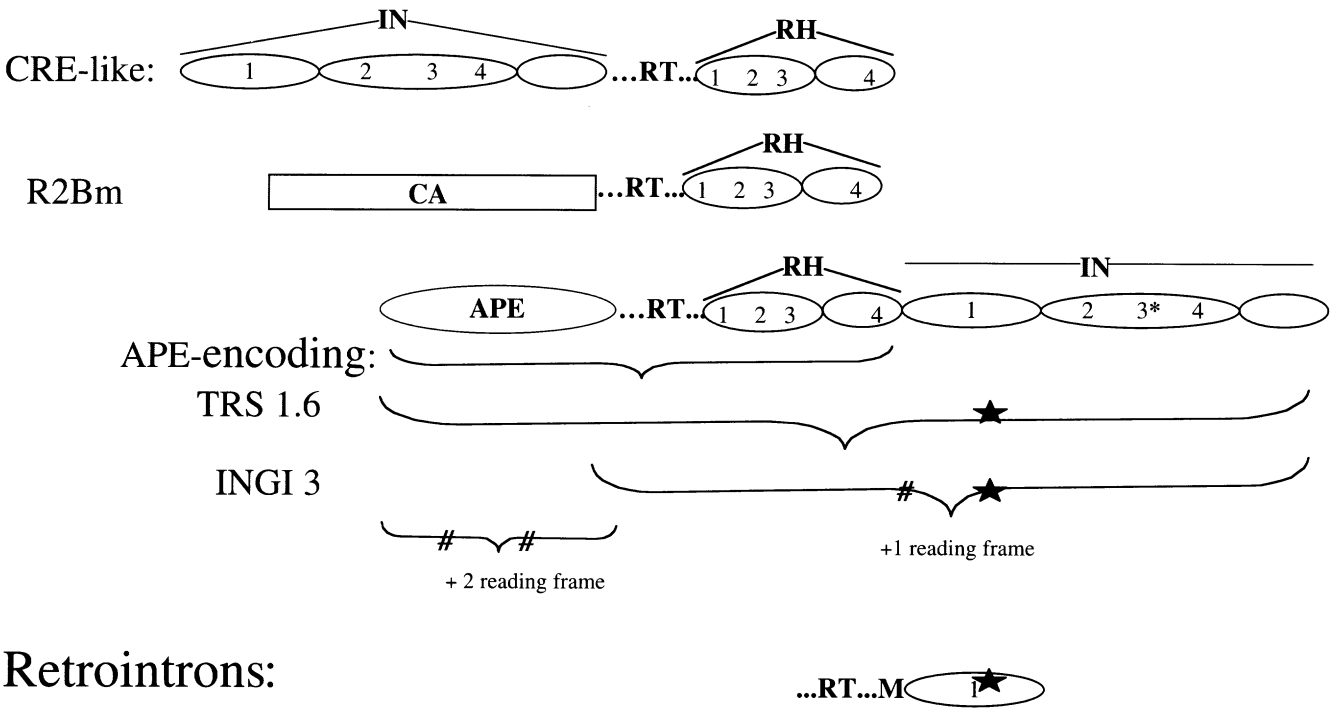

Fig. 5. Analogy and rearrangement: the endonuclease function. Schematic representation of variable coding positions for the different endonuclease activities found in retroposons. The CRE-related retroposons encode an IN protein upstream of the RT-RH. The R2Bm genome does not encode an APE; a region upstream of the RT domain, is similar to the capsid protein (CA) sequence of retroviruses while the region downstream conserves the OSM of the RH domain. Many retroposons encode an APE upstream of the RT and the RH downstream, although some lack the RH. In TRS 1.6, however, the APE is found upstream and adjacent to the RT-RH domains while a segment conserving the IN protein OSM is downstream. In Ingi3 the APE is in the +2 reading frame with two stop codons while the OSM of the IN protein is immediately downstream from the RH domain with one stop codon. In contrast the retrointrons encode the RT, a maturase (M) and the first domain of the IN protein. The stars indicate the H-N-H motif of the bacterial endonucleases located within the zinc-binding motif of the IN found in the TRS1.6/Ingi3 agents and retrointrons. Each motif of the OSM in each protein is labeled numerically (see Figure 3 for the highly conserved amino acids of each motit). Asterisks (*) indicate that although the most conserved residue is missing surrounding residues are conserved in specific motifs. The pound sign (#) indicates the position of stop codons. Adapted from McClure, 1999.

as Eukaryotic *dut* genes have introns, while DNA and RNA virus versions do not. This implies that a Retroid agent can mediate the horizontal transfer process between host mRNA and viruses in Eukaryotes. In addition most *dut* genes exist as a single copy, although tandemly duplicated and tandemly triplicated versions are found in Herpesviruses and *Caenorhabditis Elegans*, respectively. We also mapped secondary structures from single-copy dUTPase crystals onto the aligned duplicate and triplicate sequences, and hypothesize assembly structures. The results of these studies exemplify the type of new biological knowledge that is created by the development and application of recent bioinformatic tools to biological knowledge and databases. As comparative genomic analysis continues various types of evolutionary networks among and between obligate parasites and hosts will further reveal the co-evolutionary nature of genomes.

Application of new bioinformatic tools to genome analysis

The ability to analyze large amounts of sequence data *in silico* allows us to track and hypothesize plausible evolutionary networks of gene and genome construction. Currently available software, however, cannot take into account all possible evolutionary scenarios in the evaluation of sequence relationships in an automated fashion. The results described above, regarding the complexity of Retroid genome evolution, are the result of expert decision making and visual refinement of data when software limits were determined.

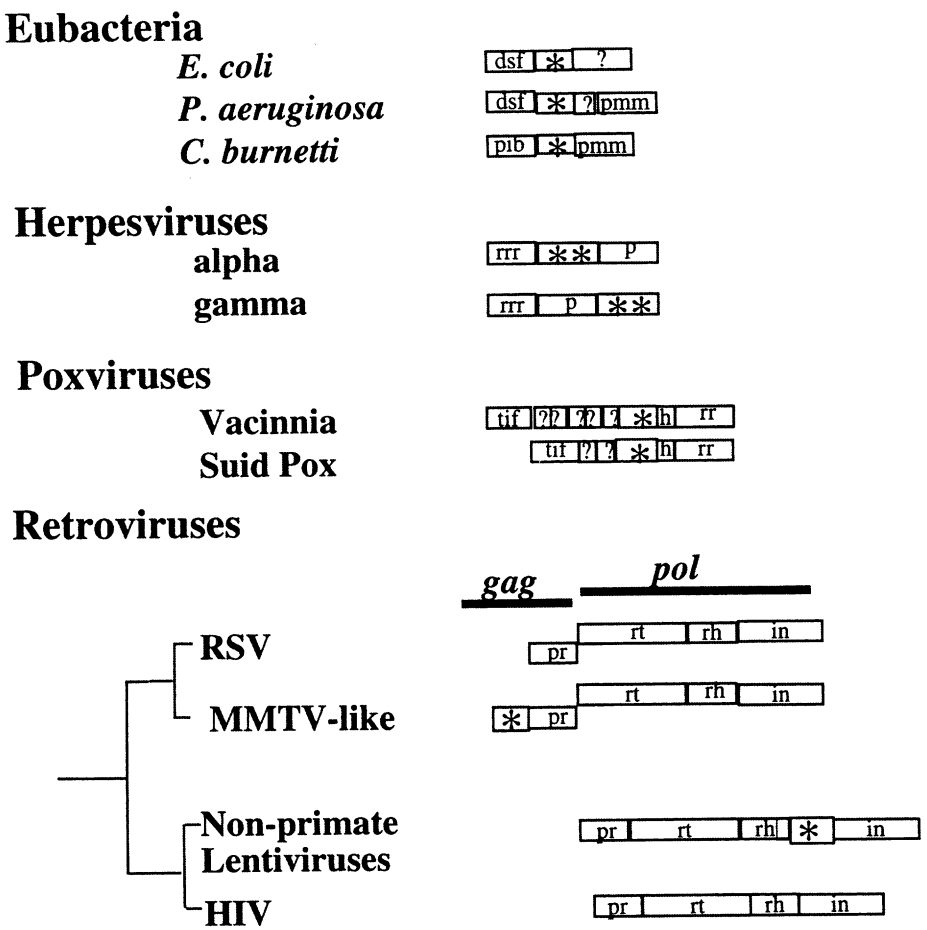

Fig. 6. The location of the *dut* gene is variable in the genomes of Eubacteria, DNA viruses, and retroviruses. Closely related gram negative Eubacteria *E.coli*, *Pseudomonas aeruginosa*, and *Coxiella burnetti*, all encode a dUTPase as do their more distant relatives. Eubacterial gene map comparison indicates that the type and number of genes surrounding *dut* varies. Alpha and gamma herpesviruses both encode a tandomly duplicated dUTPase, but the *dut* gene is not in the same position relative to the primase in these two lineages. Poxviruses have substituted genes upstream of *dut*. Between *dut* and the ribonucleotide reductase there is a single homologous but unidentified, ORF in both Vaccinia and Suid Pox. Between the transcription initiation factor gene and *dut*, however, neither of the two ORFS in Suid Pox are homologous to any of the five ORFs in Vaccinia. Two lineages of retroviruses (MMTV-like and non-primate lentiviruses) transcribe *dut* as part of the *gag* and *pol* polycistrons, respectively. The *dut* gene is not encoded by either neighboring taxon of MMTV-like and non-primate lentiviruses, RSV (Rous sarcoma virus) nor HIV, respectively. Asterisks designate single or tandemly duplicated *dut* genes. Gene abbreviations are as follows. Eubacteria: dsf is a DNA-synthesis flavoprotein; ? denotes an ORF of unknown function; pmm is phosphomannomutase, and pib is an outer membrane protein. Herpes virus genes: rrr is a ribonucleotide reductase related protein; and p, is a primase. Poxviruses: tif is a transcription initiation factor; h denotes ORF homologues of unknown function; and rr, is the ribonucleotide reductase. Abbreviations for retroviral genes are defined in the text. Adapted from Baldo and McClure (1999).

Although the software for analysis of biological sequence information has vastly improved over the last 10 years there are still problems to be resolved at all levels of gene and genomic *in silico* research. When all that remains of the similarity between two sequences is the OSM, current database search algorithms cannot find all common function sequences with a single query. Our RT data set (approximately 600 sequences) was initially collected by word searches and single sequence queries. As our Retroid sequence collection increased we developed a representative set of query probes, based on cluster analysis, for each of the enzymatic functions. These probes (including host FEPs; TERT, RH and PR) are used to routinely update the Retroid database.

No matter how one retrieves the data, the most common first stage of protein sequence analysis is to generate a multiple alignment. Given the variability of gene position in Retroid genomes it is necessary to identify each homologue/FEP in each polyprotein sequence, and estimate the functional-unit-length protein before any type of

analysis can be conducted. This is essentially a conceptual polyprotein cleavage guided by empirical knowledge of Retroid PR amino acid target specificities. The degree of sequence diversity observed among the sequences of the multiple alignment determines the type of hypotheses that can be addressed or generated. Many homologous protein relationships (all pairwise relationships >30% identity) are revealed by current multiple alignment programs. The ability to correctly align FEP sequences to reveal the OSM, however, is method dependent. Recall that an optimal alignment is not necessarily the biologically informative one. Most multiple alignment methods developed to date can align protein sequences greater than 50% identical well enough for most analytical needs. Some methods can align the OSM of some distantly related sequences with less than 30% identity, although visual inspection is still necessary to find some of the motifs when sequence identity (8–20%) and similarity are very low (McClure *et al.*, 1994).

New software development allows the incorporation of existing information from various sources into data representations that can be used to address a variety of biological questions. An example of a novel approach to the multiple alignment problem is the use of a HMM to represent the data (Baldi *et al.*, 1994; Krogh *et al.*, 1994; Eddy, 1995; Hughey and Krogh, 1996). A HMM is a stochastic production model consisting of a linear series of nodes. After training the model with sequence data each node contains the observation probabilities for match and insert states, and the transition probabilities between match, insert and delete states. In the SAM/HMM implementation the Baum-Welch algorithm guarantees the likelihood of the model will increase with each training iteration (Krogh *et al.*, 1994). Sequences are then aligned to the model using the Viterbi algorithm (Rabiner, 1989).

A *de novo* HMM is one that only incorporates observed amino acid frequencies as *a priori* information, while more refined models can include a variety of additional information. Nodes of a *de novo* model representing the MIRs are not modeled because available HMM implementations only model motifs common to the entire data set, i.e., the OSM. *De novo* HMMs are easy to generate and they are usually sufficient for identification of potential function. Such a model detected the significant H-N-H motif similarity between intron and bacterial endonucleases described above. It also detected this pattern in the retrointron IN motif, although not at a statistically supported level (Shub *et al.*, 1994).

There is little testing to validate the automated alignment quality generated by HMMs, e.g., are all residues of a verifiable OSM correctly identified in all sequences used to train a specific model, as well as the validation set? It is expected that *de novo* HMMs may not be robust enough to retrieve all similar sequence patterns from the database. Although HMM approaches outperform other multiple alignment methods in identifying the OSM in benchmark tests they are not 100% accurate (McClure and Raman, 1995; McClure *et al.*, 1996).

The OSM defines a pattern among the sequences that allows the possibility of common function and ancestry. These patterns populate motif databases. As mentioned earlier MIRs contain information regarding nearest-neighbor relationships important to the reconstruction of the phylogenetic history of the sequences. These regions can also define sub-class functional specificities and motifs. To access the maximum information contained in primary structure data both the OSM and MIRs should be aligned as precisely as possible. All positions in the alignment can provide data to test and generate a variety of evolutionary hypotheses regarding gene and genome construction. Automated generation of a multiple alignment of large numbers of highly divergent homologous and functionally equivalent protein sequences remains a challenge in the field of bioinformatics.

I am interested in the construction of HMMs that adequately reflect the evolutionary relationship for the entire length of all sequences of a given protein class. The correct identification of the OSM that defines membership in a specific protein class is the first criterion for constructing a robust HMM. There has been significant development in methods for motif identification in recent years. Six new methods have been analyzed for the ability to identify the OSM of benchmark protein families. The analysis of RT sequences suggests that the Probe (Neuwald *et al.*, 1997) and Meme (Bailey and Elkan, 1994) methods outperform others, including the SAM/HMM, in OSM identification (McClure *et al.*, 1998; Hudak and McClure, 1999).

The idea of distinguishing between the motifs common to a set of sequences (the OSM) and the intervening regions (the MIRs) in multiple alignment strategies is not new (Martinez, 1988). A prototype HMM construction strategy has been tested for alignment of both the OSM and MIRs using a low identity, low similarity data set representative of the entire RT sequence set. The 'robustness' of a model is accessed by comparing alignments by a scoring function designed to reflect the types of changes made by a human expert in refining a multiple alignment (McClure *et al.*, 1998). Constraining the OSM in the same position within a set of HMMs representing sub-classes of the data creates a series of models that generate better multiple alignments representing highly divergent sequences than a single model with or without OSM constraint (McClure and Kowlask, 1999). Further development of this strategy demonstrates that independent, sub-class modeling of MIRs (Figure 7) increases the retrieval of information within the MIR because similar sequences influence the model more than distant ones (McClure *et al.*, 1998). These studies indicate that both

sub-class clustering and OSM/MIR partitioning of the data increases the multiple alignment score.

Our current prototype HMM strategy combines the results of the studies described above. The HMM implementation, SAM (Hughey and Krogh, 1996) is used to generate a multiple alignment as described in Figure 7. OSM identification can be determined by any means. Currently we use Probe, a combined Gibbs sampling/genetic algorithm method, and Meme, a HMM approach, for cross-validation of *de novo* OSM identification. Sequences are sub-classed by an in-house clustering method. OSM information is used to partition each sub-class into MIRs. At this stage a second cluster analysis is performed on each MIR sequence set to detect any potential modular evolution, i.e., clustering for entire sequence and MIRs must be congruent. Each MIR model is independently generated by modeling the sequences within and between sub-classes by differential weighting (McClure *et al.*, 1998). This strategy essentially allows for independent modeling of the OSM versus the MIRs and a check for modular evolution. It is clear from this analysis that an automated HMM approach that distinguishes between OSMs and MIRs will provide a better approximation of alignments that have been refined by human experts. Once robust HMMs are constructed for homologous and FEP sequences they will provide data representations that can be used not only for functional determination, but also to test a variety of evolutionary ideas. While the HMM approach to data representation provides new ways of developing and testing hypotheses, the development of algorithms for tracing the network of genome construction is in its infancy.

Due to the growing body of literature that documents the deviation from the assumption of tree-like behavior in the evolution of genomes there is interest in developing bioinformatic tools to detect and reconstruct evolutionary scenarios that account for modular, segmental and horizontal evolution. While methods to detect various levels of identity discontinuities (i.e., potential recombinants) between sequences exist they are currently limited to analysis of closely related genes (e.g., HIV strains) (Stephens, 1985; Churchill, 1989; Sawyer, 1989; Hein, 1990; Smith, 1992; Hein, 1993). Other methods have been developed to detect recombination, convergence, and horizontal transfer (von Haeseler and Churchill, 1993); recombination and transversions (Holloway and Cull, 1994) and recombination (Komatsoulis and Waterman, 1997). These methods have not been analyzed to determine the limits of detectable recombination as a function of sequence divergence.

Once various types of sequence rearrangements have been identified, several combinatorial algorithms exist to reconstruct a most likely network of relationships among the gene segments (Bafna and Pevzner, 1993; Kececioglu

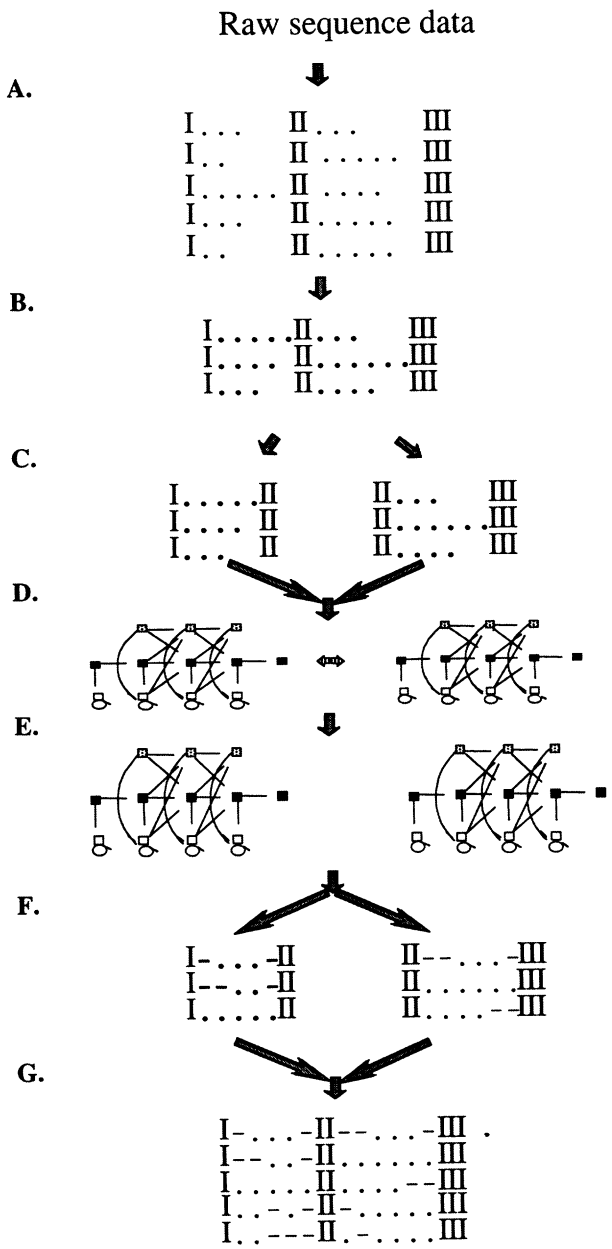

Fig. 7. Schematic representation of the strategy used to generate robust alignments from HMMs. A) Probe and Meme identify the OSM in raw sequence data. B) Sequences are clustered into similarity sub-classes. C) Division of sub-class sequences at OSMs to create sub-classed MIR data. Consistency check for congruent clustering of complete and MIR sequences. D) Each MIR dataset is independently modeled within and between sub-classes. E) Generation of a set of MIR sub-class models with amino acid probabilities at each node representing the constraining OSM and the MIR. F) Alignment of each MIR sequence to its respective MIR model. G) Concatenation of MIR alignment segments and stacking of sub-class alignments to create a single comprehensive alignment of the OSM and MIRs. A final HMM can then be generated from this alignment. The HMM is the basic three-state, left-right architecture. Roman numerals represent the OSM, and periods represent amino acid residues in the MIRs. Dashes represent indels.

and Sankoff, 1993, 1994). A method of this type has been successful in inferring the network of gene rearrangements observed in Herpesviruses (Hannenhalli *et al.*, 1995).

Conclusions and perspective

Retroid agents will continue to provide a unique probe to explore both empirically and *in silico* the relationship within and between the genomes of organisms. Whether or not these agents are the descendants of a RNA lifeform that mediated the transfer of information from RNA to DNA may be unknowable. Regardless of their origin, they are intimately involved in a great variety of Eukaryotic genomic events and processes. Given that some Retroid agents can move horizontally within and between species, comparative genomics includes the network of their relationships. The distinction between a 'host' or 'Retroid agent' origin of a specific sequence may blur as genome networks develop. Furthermore, the existence of RNA-based lifeforms that accumulate mutations millions of times faster than DNA-based systems allows us to address even more difficult molecular evolutionary questions. If the only similarity common to homologues and FEPs is the OSM, the most functionally constrained region or deep structure of the protein, can we discriminate between divergence and convergence given the extraordinary ability of RNA genomes to explore sequence space? In the evolution of a single motif how often does the same 'solution' to the motif's role in function arise? Are there discrete variations in the composition of motifs that suggest sequence convergence occurs at least at the local level? Is actual sequence convergence responsible for the statistically unsupported similarity observed between the RDDP and the RDRPs, or the observed overlap in the Retroid IN motif I and the bacterial endonuclease H-N-H motif? Of course, MIR analysis addresses some of these issues by providing information about near neighbor similarities, but these are also the more divergent regions that provide the opportunity for modular evolution.

Given that modular and segmental evolution is taking place at measurable frequencies in genomes, the assumption of tree-like behavior in the reconstruction of phylogeny may lead to erroneous conclusions regarding common descent. Phylogenetic reconstruction should be considered at two levels; 1) the network of relationships that gave rise to the genes or genomes of interest, and 2) the accumulated change that has occurred since these genes or genomes came into their current configuration. The development of new approaches for identification of deviation from tree-like behavior and the reconstruction of relationships among mobile gene and genome segments is an area of computational biology that needs rapid development given the pace of genome sequencing.

Recent approaches to the development of algorithms collectively referred to as 'intelligent systems' are just beginning to provide tools. Expert systems based on human decision-making processes are under explored for their potential uses in sequence interpretation and biological inference. Artificial intelligence systems are being developed for uses in crystal structure determination (Glasgow *et al.*, 1993; Conklin *et al.*, 1996; Leherte *et al.*, 1997) and molecular data mining of protein structural databases (Glasgow *et al.*, 1999). It will be quite a challenge to develop automated approaches to the complex decision making processes necessary to provide appropriate biological interpretation at various levels of gene and genomic evolution.

The establishment of databases, development of algorithms and analysis of sequence information was initiated over 30 years ago. The second wave of the multidisciplinary field of computational biology allows for the rapid development of *in silico* research approaches to address a variety of biological problems (Boguski, 1998). In the last few years experimental and computational biologists, systems scientists, and mathematicians have begun collaborations necessary for this second wave to reach maturity. New methods are being developed. Adequate testing of approaches with benchmark data is becoming standard, although biologically meaningful benchmarks are needed for all stages in development of tools for analysis. Sequence data, for example, is becoming available for empirically detected recombination events. These data sets need to be assembled so that methods can be developed and tested for detecting recombination as a function of sequence divergence. While it appears that systems scientists are aware of the limitations in existing software, laboratory-based biologists are often surprised to learn of the difficulties in developing and implementing algorithms that can accurately detect the various levels of evolutionary complexity readily observed through experimentation. As new tools that have survived benchmarking tests for genomic analysis in a broad evolutionary context become available the biological community at large will gain a full appreciation of the *in silico* approach to questions regarding the evolution of genes and genomes.

Acknowledgements

This work is supported by NIH grant AI 28309 and a Research Career Development Award. I thank Drs. Angela Baldo and King Jordan, and Seanna Corro, Julianna Hudak, and Ben Sutter for discussion and critical reading of the manuscript.

References

Argos,P. (1988) A sequence motif in many polymerases. *Nucl. Acids Res.*, **16**, 9909–9916.

Bafna,V. and Pevzner,P.A. (1993) Genome rearrangements and

sorting by reversals. *34th IEEE Symposium on the Foundations of Computer Science.* IEEE Computer Society Press, pp. 1–27.

Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In Altman,R., Brutlag,D., Karp,P., Lathrop,R. and Searls,D. (eds), *Proceedings of Second International Conference on Intelligent Systems for Molecular Biology.* AAAI Press, Stanford University, Stanford, CA, pp. 28–36.

Baldi,P., Chauvin,Y., Hunkapiller,T. and McClure,M.A. (1994) Hidden Markov models of biological primary sequence information. *Proc. Natl Acad. Sci. USA*, **91**, 1059–1063.

Baldo,A.M. and McClure,M.A. (1999) Evolution and horizontal transfer of dUTPase encoding genes in viruses and their hosts. *J. Virol.*, **73**, 7710–7721.

Baron,M., Norman,D.G. and Campbell,I.D. (1991) Protein modules. *Trends Biochem. Sci.*, **16**, 13–17.

Blake,C.C.F. (1978) Do genes-in-pieces imply proteins-in-pieces? *Nature*, **273**, 267.

Boguski,M.S. (1998) Bioinformatics-a new era. *Trends Guide to Bioinformatics*, 1–3.

Botstein,D. (1980) A theory of modular evolution for bacteriophages. *Ann. N.Y. Acad.*, **354**, 484–491.

Brautigam,C.A. and Steitz,T.A. (1998) Structural and functional insights provided by crystal structures of DNA polymerases and their substrate complexes. *Curr. Opin. Struct. Biol.*, **8**, 54–63.

Brown,P.O. (1997) Integration. In Coffin,J.M., Hughes,S.H. and Varmus,H.E. (eds), *Retroviruses.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, pp. 161–203.

Burke,C.J., Sanyal,G., Bruner,M.W., Ryan,J.A., La Femina,R.L., Robbins,H.L., Zeft,A.S., Middaugh,C.R. and Cordingley,M.G. (1992) Structural implications of spectroscopic characterization of a putative zinc-finger peptide from HIV-1 integrase. *J. Biol. Chem.*, **267**, 9639–9644.

Casacuberta,J.M., Vernhettes,S., Audeon,C. and Greandbastien, M.-A. (1997) Quasispecies in retrotransposons: a role for sequence variability in Tnt1 evolution. *Genetica*, **100**, 109–117.

Churchill,G.A. (1989) Stochastic models for heterogeneous DNA sequences. *Bull. Math. Biol.*, **51**, 79–94.

Conklin,D., Fortier,S., Glasgow,J.I. and Allen,F.H. (1996) Conformational analysis from crystallographic data using conceptual clustering. *Acta Crystallographica*, **B52**, 535–549.

Davies,J.F., Hostomska,Z., Hostomsky,Z., Jordan,S.R. and Matthews,D.A. (1991) Crystal structure of the ribonuclease H domain of HIV-1 reverse transcriptase. *Science*, **252**, 88–95.

Dayhoff,M.O., Barker,W.C. and Hunt,L.T. (1983) Establishing homologies in protein sequences. *Meth. Enzymol.*, **91**, 524–545.

Delarue,M., Poch,O., Tordo,N., Moras,D. and Argos,P. (1990) An attempt to unify the structure of polymerases. *Protein Eng.*, **3**, 461–467.

Delassus,S., Cheynier,R. and Wain-Hobson,S. (1991) Evolution of human immunodeficiency virus type 1 nef and long terminal repeat sequences over 4 years in vivo and in vitro. *J. Virol.*, **65**, 225–231.

Diaz,E., Lopez,R. and Garcia,J.L. (1990) Chimeric phage-bacterial enzymes: a clue to the modular evolution of genes. *Proc. Natl Acad. Sci. USA*, **87**, 8125–8129.

Domingo,E. and Holland,J.J. (1997) RNA virus mutations and fitness for survival. *Annu. Rev. Microbiol.*, **51**, 151–78.

Doolittle,R.F., Feng,D.-F., Johnson,M.S. and McClure,M.A. (1989) Origins and evolutionary relationships of retroviruses. *Quart. Rev. Biol.*, **64**, 1–30.

Dorit,R.L., Schoenbach,L. and Gilbert,W. (1990) How big is the universe of exons? *Science*, **250**, 1377–1382.

Dyda,F., Hickman,A.B., Jenkins,T.M., Engelman,A., Craigie,R. and Davies,D.R. (1994) Crystal structure of the catalytic domain of HIV-1 integrase: similarity to other polynucleotidyl transferases. *Science*, **266**, 1981–1985.

Eddy,S. (1995) Multiple alignment using hidden Markov models. In Rawlings,C., Clark,D., Altman,R., Hunter,L., Lengauer,T. and Wodak,S. (eds), *Third International Conference on Intelligent Systems for Molecular Biology.* AAAI Press, Cambridge, England, pp. 114–120.

Elder,J.H., Lerner,D.L., Hasselkus-Light,C.S., Fontenot,D.J., Hunter,E., Luciw,P.A., Montelaro,R.C. and Phillips,T.R. (1992) Distinct subsets of retroviruses encode dUTPase. *J. Virol.*, **66**, 1791–1794.

Feng,D.-F. and Doolittle,R.F. (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.*, **25**, 351–360.

Feng,Q., Schumann,G. and Boeke,J.D. (1998) Retrotransposon R1Bm endonuclease cleaves the target sequence. *Proc. Natl Acad. Sci. USA*, **95**, 20083–20088.

Gabriel,A., Willems,M., Mules,E.H. and Boeke,J.D. (1996) Replication infidelity during a single cycle of Ty1 retrotransposition. *Proc. Natl Acad. Sci. USA*, **93**, 7767–7771.

Gao,F., Robertson,D.L., Morrison,S.G., Hui,H., Craig,S., Decker,J., Fultz,P.N., Girard,M., Shaw,G.M., Hahn,B.H. and Sharp,P.M. (1996) The heterosexual HIV-1 epidemic on Thailand is caused by an intersubtype (A/E) recombination of African origin. *J. Virol.*, **70**, 7013–7029.

Gilbert,W. (1978) Why genes in pieces? *Nature*, **271**, 501.

Glasgow,J.I., Fortier,S. and Allen,F.H. (1993) Molecular scene analysis: crystal structure determination through imagery. In Hunter,L. (ed.), *Artificial Intelligence and Molecular Biology.* AAAI/MIT Press, pp. 433–458.

Glasgow,J.I., Steeg,E. and Fortier,S. (1999) Motif discovery in protein structure databases. In Wang, Shapiro, and Shasha (eds), *Pattern Discovery in Molecular Biology: Tools, Techniques and Applications.* Oxford Press, in press.

Gorbalenya,A.E. (1994) Self-splicing group I and II introns encode homologous (putative) DNA endonucleases of a new family. *Protein Sci.*, **3**, 1117–1120.

Grez,M., Dietrich,U., Balfe,P., von Briesen,H., Maniar,J.K., Mahambre,G., Delwart,E.L., Mullins,J.I. and Rubsamen-Waigmann,H. (1994) Genetic analysis of human immunodeficiency virus type 1 and 2 (HIV-1 and HIV-2) mixed infections in India reveals a recent spread of HIV-1 and HIV-2 from a single ancestor for each of these viruses. *J. Virol.*, **68**, 2161–2168.

Hannenhalli,S., Chappey,C., Koonin,E.K. and Pevzner,P.A. (1995) Scenarios for genome rearrangements: Herpesvirus evolution as a test case. *Genomics*, **30**, 299–311.

Hein,J. (1990) Reconstructing evolution of sequences subject to recombination using parsimony. *Math. Biosci.*, **98**, 185–200.

Hein,J. (1993) A heuristic method to reconstruct the history of sequences subject to recombination. *J. Mol. Evol.*, **36**, 369–405.

Holloway,J.L. and Cull,P. (1994) Aligning genomes with inversions

and swaps. In Altman,R., Brutlag,D., Karp,P., Lathrop,R. and Searls,D. (eds), *Second International Conference on Intelligent Systems for Molecular Biology.* AAAI Press, Stanford, CA, pp. 195–202.

Hudak,J. and McClure,M.A. (1999) A comparative analysis of computational motif-detection methods. In Altman,R.B., Dunker,A.K., Hunter,L., Klein,T.E. and Lauderdale,K. (eds), *Proceedings of the Pacific Symposium on Biocomputing, 99.* World Science, New Jersey, pp. 138–149.

Hughey,R. and Krogh,A. (1996) Hidden Markov models for sequence analysis: extension and analysis of the basic method. *CABIOS*, **12**, 95–107.

Johnson,M.S., McClure,M.A., Feng,D.-F., Gray,J. and Doolittle,R.F. (1986) Computer analysis of retroviral genes: assignment of enzymatic functions. *Proc. Natl Acad. Sci. USA*, **83**, 7648–7652.

Jordan,I.K. and McDonald,J.F. (1999) Phylogenetic perspective reveals abundant Ty1/Ty2 hybrid elements in *Saccharomyces cerevisiae* genome. *Mol. Biol. Evol.*, **16**, 419–422.

Jordan,I.K., Sutter,B. and McClure,M.A. (2000) Molecular evolution of Paramyxoviridae and Rhabdoviridae multiple protein encoding P gene. *Mol. Evol. Bio.*, **17**, 75–86.

Jordan,K.I. and McDonald,J.F. (1998) Evidence for the role of recombination in the regulatory evolution of *Saccharomyces cerevisiae* Ty elements. *J. Mol. Evol.*, **47**, 14–20.

Kamer,G. and Argos,P. (1984) Primary structural comparison of RNA-dependent polymerases from plant, animal and bacterial viruses. *Nucl. Acids Res.*, **12**, 7269–7282.

Kanaya,S., Kohara,A., Miura,Y., Sekiguchi,A., Iwai,S., Inoue,H., Ohtsuka,E. and Ikehara,M. (1990) Identification of the amino acid residues involved in an active site of Escherichia. *J. Biol. Chem.*, **265**, 4615–4621.

Katayanagi,K., Miyagawa,M., Matsushima,M., Ishikawa,M., Kanaya,S., Ikehara,M., Matsuzaki,T. and Morikawa,K. (1990) Three-dimensional structure of ribonuclease H from *E.coli*. *Nature*, **347**, 306–309.

Kececioglu,J. and Sankoff,D. (1993) Exact and approximation algorithms for the reversal distance between two permutations. In Apostolico,A., Crochemore,M., Galil,Z. and Manber,U. (eds), *4th Ann. Symposium on Combinatorial Pattern Matching.* Padova, Italy, pp. 87–105.

Kececioglu,J. and Sankoff,D. (1994) Exact and approximation algorithms for sorting by reversals, with application to genome rearrangements. *Algorithmica*, **13**, 180–210.

Kidwell,M.G. and Lisch,D.R. (1998) Transposons unbound. *Nature*, **393**, 22–23.

Kohlstaedt,L.A., Wang,J., Friedman,J.M., Rice,P.A. and Steitz,T.A. (1992) Crystal structure at 3.5 A resolution of HIV-1 reverse transcriptase complexed with an inhibitor. *Science*, **256**, 1783–1790.

Komatsoulis,G.A. and Waterman,M.S. (1997) Chimeric alignment by dynamic programming: algorithm and biological uses. *Proceedings of the First Annual International Conference on Computational Molecular Biology.* The Association for Computing Machinery, pp. 174–180.

Krogh,A., Brown,M., Mian,I.S., Sjolander,K. and Haussler,D. (1994) Hidden Markov models in computational biology: applications to protein modeling. *J. Mol. Biol.*, **235**, 1501–1531.

Leherte,L., Glasgow,J.I., Fortier,S., Baxter,K. and Steeg,E. (1997) Analysis of three-dimensional protein images. *JAIR*, **7**, 125–159.

Li,W., Tanimura,M. and Sharp,P. (1988) Rates and dates of divergence between AIDS virus nucleotide sequences. *Mol. Biol. Evol.*, **5**, 313–330.

Lingner,J., Hughes,T.R., Shevchenko,A., Mann,M., Lundblad,V. and Cech,T.R. (1997) Reverse transcriptase motifs in the catalytic subunit of telomerase. *Science*, **276**, 561–567.

Martin,F., Maranon,C., Olivares,M., Alonso,C. and Lopez,M.C. (1995) Characterization of a non-long terminal repeat retrotransposon cDNA (L1Tc) from trypanosoma cruzi: homology of the first ORF with the APE family of DNA repair enzymes. *J. Mol. Biol.*, **247**, 49–59.

Martinez,H.M. (1988) A flexible multiple sequence alignment program. *Nucl. Acids Res.*, **16**, 1683–1691.

McClure,M.A. (1993) Evolutionary history of reverse transcriptase. In Skalka,A.M. and Goff,S.P. (eds), *Reverse Transcriptase.* Cold Spring Harbor Laboratory Press, pp. 425–444.

McClure,M.A. (1991) Evolution of retroposons by acquisition or deletion of retrovirus-like genes. *Mol. Biol. Evol.*, **8**, 835–856.

McClure,M.A. (1992) Sequence analysis of eukaryotic retroid proteins. *Math. Comp. Modeling*, **16**, 121–136.

McClure,M.A. (1996) The complexities of viral genome analysis: the primate lentiviruses. *Curr. Opin. Genet. Devel.*, **6**, 749–756.

McClure,M.A. (1999) The Retroid Agents: Disease, Function, and Evolution. In Domingo,E., Webster,R. and Holland,J. (eds), *Origin and Evolution of Viruses.* Academic Press, London, pp. 163–195.

McClure,M.A., Hudak,J. and Kowalski,J. (1998) Low identity, low similarity protein sequences: independent modeling of the ordered-series-of-motifs and motif-intervening-regions. In Miyano,S. and Takagi,T. (eds), *Genome Informatics 1998.* Universal Academy Press, Inc, Tokyo, Japan, pp. 183–192.

McClure,M.A., Johnson,M.S. and Doolittle,R.F. (1987) Relocation of a protease-like gene segment between two retroviruses. *Proc. Natl Acad. Sci. USA*, **84**, 2693–2697.

McClure,M.A., Johnson,M.S., Feng,D.-F. and Doolittle,R.F. (1988) Sequence comparisons of retroviral proteins: relative rates of change and general phylogeny. *Proc. Natl Acad. Sci. USA*, **85**, 2469–2473.

McClure,M.A. and Kowlask,J. (1999) The effects of ordered-series-of-motifs anchoring and sub-class modeling on the generation of HMMs representing highly divergent protein sequences. In Altman,R.B., Dunker,A.K., Hunter,L., Klein,T.E. and Lauderdale,K. (eds), *Proceedings of the Pacific Symposium on Biocomputing, 99.* World Science, New Jersey, pp. 162–170.

McClure,M.A. and Raman,R. (1995) Parameterization studies of hidden Markov models representing highly divergent protein sequences. In Hunter,L. and Shriver,B. (eds), *28th Annual Hawaii International Conference on System Sciences.* IEEE Computer Society Press, Hawaii, pp. 184–193.

McClure,M.A., Smith,C. and Elton,P. (1996) Parameterization studies for the SAM and HMMER methods of hidden Markov model generation. In States,D., Agarwal,P., Gaasterland,T., Hunter,L. and Smith,R.F. (eds), *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology.* AAAI Press, Menlo Park, CA, pp. 155–164.

McClure,M.A., Vasi,T. and Fitch,W.M. (1994) Comparative analysis of multiple protein-sequence alignment methods. *Mol. Biol.*

Evol., **11**, 571–592.

McGeoch,D.J. (1990) Protein sequences, comparisons show that the 'pseudoproteases' encoded by poxviruses and certain retroviruses belong to the deoxyuridine triphosphatase family. *Nucl. Acids Res.*, **18**, 4105–4110.

Mercer,A.A., Fraser,K.M., Stockwell,P.A. and Robinson,A.J. (1989) A homologue of retroviral pseudoproteases in the Parapoxvirus, Orf virus. *Virology*, **172**, 665–668.

Miller,M., Jaskolski,M., J.K., M.R., Leis,J. and Wlodawer,A. (1989) Crystal structure of a retroviral protease proves relationship to aspartic protease family. *Nature*, **337**, 576–579.

Mohr,D.A., McKay,L.L. and Lambowitz,A.M. (1993) Evolutionary relationships among group II intron-encoded proteins and identification of a conserved domain that may be related to maturase function. *Nucl. Acid Res.*, **21**, 4991–4997.

Nakamura,T.M. and Cech,T.R. (1998) Reversing time: Origin of telomerase. *Cell*, **92**, 587–590.

Nakamura,T.M., Morin,G.B., Chapman,K.B., Weinrich,S.L., Andrews,W.H., Lingner,J., Harley,C.B. and Cech,T.R. (1997) Telomerase catalytic subunit homologs from fission yeast and human. *Science*, **277**, 955–959.

Neuwald,A.F., Liu,J.S., Lipman,D.J. and Lawrence,C.E. (1997) Extracting protein alignment models from the sequence database. *Nucl. Acids Res.*, **25**, 1665–1677.

Pearl,L.H. and Taylor,W.R. (1987) A structural model for the retroviral proteases. *Nature*, **329**, 351–354.

Poch,O., Sauvaget,I., Delarue,M. and Tordo,N. (1989) Identification of four conserved motifs among the RNA-dependent polymerase encoding elements. *EMBO J.*, **8**, 3867–3874.

Power,M., Marx,P., Bryant,M., Gardner,M., Barr,P. and Luciw,P. (1986) Nucleotide sequence of SRV-1, a type D simian acquired immune deficiency syndrome retrovirus. *Science*, **231**, 1567–1572.

Rabiner,L.R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, **77**, 257–285.

Richert-Poggeler,K.R. and Shepherd,R.J. (1997) Petunia vein-cleaning virus: a plant pararetrovirus with core sequences for an integrase function. *Virology*, **236**, 137–146.

Robertson,D., Hahn,B. and Sharp,P. (1995) Recombination in AIDS viruses. *J. Mol. Evol.*, **40**, 249–259.

Sawyer,S. (1989) Statistical tests for detecting gene conversion. *Mol. Biol. Evol.*, **6**, 526–538.

Schwartz,R.M. and Dayhoff,M.O. (1978) Matrices for detecting distant relationships. In Dayhoff,M.O. (ed.), *Atlas of Protein Sequence and Structure.* National Biomedical Research Foundation, Washington, DC, pp. 353–358.

Shub,D.A., Goodrich-Blair,H. and Eddy,S.R. (1994) Amino acid sequence motif of group I endonucleases is conserved in open reading frames of group II introns. *Trends Biochem. Sci.*, **19**, 402–404.

Slabaugh,M.B. and Roseman,N.A. (1989) Retroviral protease-like gene in the vaccinia virus genome. *Proc. Natl Acad. Sci. USA*, **86**, 4152–4155.

Smith,J.M. (1992) Analyzing the mosaic structure of genes. *J. Mol. Evol.*, **34**, 126–129.

Steitz,T.A. (1998) A mechanism for all polymerases. *Nature*, **391**, 231–232.

Stephens,J.C. (1985) Statistical methods of DNA sequence analysis: detection of intragenic recombination or gene conversion. *Mol. Biol. Evol.*, **2**, 539–556.

Strauss,J.H. and Strauss,E.G. (1988) Evolution of RNA viruses. *Ann. Rev. Microbiol.*, **42**, 657–683.

Tanese,N. and Goff,S.P. (1988) Domain structure of the Moloney murine leukemia virus reverse transcriptase: mutational analysis and separate expression of the DNA polymerase and RNAase H activities. *Proc. Natl Acad. Sci. USA*, **85**, 1777–1781.

Toh,H., Ono,M., Saigo,K. and Miyata,T. (1985) Retroviral protease-like sequence in the yeast transposon Ty1. *Nature*, **315**, 691.

Villanueva,M.S., Williams,S.P., Beard,C.B., Richards,F.F. and Aksoy,S. (1991) A new member of a family of site-specific retrotransposons is present in the spliced leader RNA genes of *Trypanosoma cruzi*. *Mol. Cell. Biol.*, **11**, 6139–6148.

von Haeseler,A. and Churchill,G.A. (1993) Network models for sequence evolution. *J. Mol. Evol.*, **37**, 77–85.

Waterman,M.S. (1984) General methods of sequence comparison. *Bull. Math. Biol.*, **46**, 473–500.

Waterman,M.S. (1986) Multiple sequence alignment by consensus. *Nucl. Acids Res.*, **14**, 9095–9102.

Waterman,M.S., Arratia,R. and Galas,D.J. (1984) Pattern recognition in several sequences: consensus and alignment. *Bull. Math. Biol.*, **46**, 515–527.

Waterman,M.S. and Perlwitz,M.D. (1984) Line geometries for sequence comparison. *Bull. Math. Biol.*, **46**, 567–577.

Waugh O'Neill,R.J., O'Neill,M.J. and Marshall Graves,J.A. (1998) Undermethylation associated with retroelement activation and chromosome remodelling in an interspecific mammalian hybrid. *Nature*, **393**, 68–72.

Xiong,Y. and Eickbush,T.H. (1988) Functional expression of a sequence-specific endonuclease encoded by the retrotransposon R2Bm. *Cell*, **55**, 235–246.

Xiong,Y. and Eickbush,T.H. (1990) Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J.*, **9**, 3353–3362.

Yanagawa,H., Yoshida,K., Torigoe,C., Park,J.-S., Sato,K., Shirai,T. and Go,M. (1993) Protein anatomy: functional roles of barnase module. *J. Biol. Chem.*, **268**, 5861–5865.

Zanotto,P., Gibbs,M.J., Gould,E.A. and Holmes,E.C. (1996) A reevaluation of the higher taxonomy of viruses based on RNA polymerases. *J. Virol.*, **70**, 6083–6093.

Zimmerly,S., Guo,H., Eskes,R., Yang,J., Perlman,P.S. and Lambowitz,A.M. (1995) A group II intron RNA is a catalytic component of a DNA endonuclease involved in intron mobility. *Cell*, **83**, 529–538.

Zimmern,D. (1988) Evolution of RNA viruses. In Domingo,E., Holland,J.J. and Ahlquist,P. (eds), *RNA Genetics, Retroviruses, Viroids, and RNA Recombination, Vol. II* CRC Press, Inc., Boca Raton, pp. 211–240.