



90955410

Relais Request No. REG-20604693

Customer Code  
**91-3540**

Delivery Method  
**SED**

Request Number  
**RZDPLK11358 TSED48 COPYR**

Scan

Date Printed: 25-Oct-2005 23:04

Date Submitted: 24-Oct-2005 21:43

6082.230000

TITLE: NEW BIOLOGIST.

YEAR: 1992

VOLUME/PART:

PAGES:

AUTHOR:

ARTICLE TITLE:

SHELFMARK: 6082.230000

**Your Ref :**

RZDPLK11358 TSED48 COPYRT|NEW BIOLOGIST.|1992; 4/3; PP 247-60.|BOGUSKI MS,  
HARDISON RC, SCHWARTZ S, MIL|6082.230000 1043-4674



**DELIVERING THE WORLD'S KNOWLEDGE**

**This document has been supplied by the British Library**

**[www.bl.uk](http://www.bl.uk)**

The contents of the attached document are copyright works. Unless you have the permission of the copyright owner, the Copyright Licensing Agency Ltd or another authorised licensing body, you may not copy, store in any electronic medium or otherwise reproduce or resell any of the content, even for internal purposes, except as may be allowed by law.

The document has been supplied under our Copyright Fee Paid service. You are therefore agreeing to the terms of supply for our Copyright Fee Paid service, available at :

**[www.bl.uk/services/document/edd.html](http://www.bl.uk/services/document/edd.html)**

# Analysis of Conserved Domains and Sequence Motifs in Cellular Regulatory Proteins and Locus Control Regions Using New Software Tools for Multiple Alignment and Visualization

Mark S. Boguski,<sup>1</sup> Ross C. Hardison,<sup>2,3</sup> Scott Schwartz,<sup>4</sup> Webb Miller<sup>3,4,\*</sup>

With the tremendous expansion of molecular sequence data in recent years, multiple alignment is arguably one of the two most important analytic techniques (the other being fast database searching). A number of useful approaches to this problem have previously been developed, but often they are limited to only a subset of multiple-alignment applications and cannot easily deal with the complex structural organization seen in an increasing number of sequences. For example, a single sequence may contain several domains of different evolutionary origins, and the multiplicities and relative ordering of these domains may be quite different among related sequences. Here we describe an integrated set of interactive Unix tools that combines several multiple-alignment techniques with traditional "dot-plot" visualization to provide a flexible environment for approaching complex sequence analysis problems. We apply these tools to the identification and characterization of "catalytic" domains in ras and rho/rac GTPase-activating proteins, to "Src homology" (SH2, SH3) domains in cytoplasmic signaling proteins, to repetitive sequence motifs in the  $\alpha$  and  $\beta$  subunits of protein prenyltransferases, and to regulatory DNA sequences in the locus control region of the  $\beta$ -globin gene cluster.

Received November 18, 1991; revised January 15, 1992

It is widely appreciated that the volume of molecular sequence data is growing at a phenomenal rate. However, it must also be stressed that the data are changing qualitatively as well. Gone are the days when globins or cytochromes could be regarded as "typical" proteins and when simple pairwise alignment programs were all that one needed to analyze sequence data. Computing multiple alignments, i.e., simultaneously aligning more than two sequences, has become the rule rather than the exception and has important applications in protein modeling and structure prediction, studies of molecular evolution, and the detection and

quantitation of conserved patterns or motifs. The multiple-alignment problem is compounded by the fact that many sequences are modular and/or mosaic in structure, making global alignments (i.e., alignments that are required to extend from one end of the sequences to the other) misleading or meaningless in many cases. Internal repeats, intrasequence transpositions, and variations in local compositional complexity are increasingly common features of many sequence families, which necessitates computation of a number of local alignments (i.e., potentially involving only portions of the sequences) to reveal the similarities in such a family. New software tools are needed to deal with the richness and complexity of sequence analysis problems facing us today.

Multiple alignments are expensive to compute; it is currently infeasible to apply a rigorous alignment method to even as many as 10 protein sequences of average lengths (approx 250 residues). Still, the problem is so important that a wide variety of approaches have been proposed (see Doolittle, 1990; States and Boguski, 1991). On the other hand, pairwise alignments are far more tractable. Rigorous computer methods exist that can determine all of the local alignments of significant score between two sequences, even with sequences of lengths up to 100,000 (Huang et

<sup>1</sup>National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894.

<sup>2</sup>Department of Molecular and Cell Biology, The Pennsylvania State University, University Park, PA 16802.

<sup>3</sup>Institute for Molecular Evolutionary Genetics, The Pennsylvania State University, University Park, PA 16802.

<sup>4</sup>Department of Computer Science, The Pennsylvania State University, University Park, PA 16802.

\*To whom correspondence should be addressed.

1043-4674/92/0403-0008\$5.00/0

**KEY WORDS:** computer software/GTPase-activating proteins/locus control regions/multiple sequence alignment/prenylation/farnesyltransferase/sequence motifs/Src homology domains

al., 1990). Moreover, using the ubiquitous dot matrix method, or dot-plot, all of the strongly matching regions of two sequences are easily perceived. However, pairwise alignments are sometimes inadequate to reveal a motif shared by several sequences; when only two of the sequences are compared, the match involving the motif is lost in the noise created by a number of extraneous matches.

As commonly used, the term "dot-plot" often encompasses two independent notions that we want to distinguish. First, most dot-plot programs compute what amounts to the set of all gap-free alignments that satisfy some rather arbitrary criterion (such as alignments of length 30 that contain at least 19 matches). Second, these programs produce a two-dimensional plot that depicts the extents of the conforming alignments. That method of computing alignments is of limited value and we prefer to use either a method that permits gaps in alignments (Huang et al., 1990) or a method with a sound statistical basis (Altschul et al., 1990). However, the pictorial representation of alignments remains an unsurpassed visual metaphor for studying large sequences of complex structure. This "visualization" of a set of local alignments between two sequences is applicable to alignments regardless of how they are computed (Pearson and Lipman, 1988; Schwartz et al., 1991).

Here we describe an approach that combines some of the most useful features of other methods. Given a family of sequences that may contain regions that share common motifs, we first compute a number of pairwise alignments by the *sim* program (Huang et al., 1990), which permits gaps in alignments. Each pairwise alignment produces a number of nonintersecting local alignments between the two sequences. These pairwise alignments are then combined by a new program, called *pab*, that produces a set of aligned blocks, i.e., gap-free alignments of equal-length segments from each sequence. Neighboring blocks of this multiple alignment may identify segments of the sequences that can be meaningfully aligned in their entireties (allowing gaps) by the *msa* program (Lipman et al., 1989). At each stage, the restriction of an alignment to any two of its constituent sequences can be viewed in a dot-plot-like representation (Schwartz et al., 1991). A similar method was developed by Vingron and Argos (1991), but our software has advantages in ease of use, generality, and speed.

Table 1 lists the components of this software tool kit that are most relevant to this paper. For extreme ease of use, the programs *pblocks* and *dblocks* permit the user to specify simply a list of sequences (protein or DNA sequences, respectively); the programs then apply *sim* and *pab* in predefined ways (see Materials and Methods).

**Table 1. Some main components of our multiple-alignment tool kit.**

| Name           | Function  |
|----------------|---|
| <i>sim</i>     | Construct nonintersecting local alignments of two sequences         |
| <i>pab</i>     | Construct aligned blocks from a family of pairwise local alignments |
| <i>pblocks</i> | Construct aligned blocks for a list of protein sequences            |
| <i>dblocks</i> | Construct aligned blocks for a list of DNA sequences                |
| <i>msa</i>     | Compute a global (end-to-end) alignment of a list of sequences      |
| <i>laps</i>    | Display alignments on a PostScript printer                          |
| <i>lax</i>     | Display alignments on an X-Windows terminal                         |

We apply these integrated software tools to a number of complex sequence analysis problems. First, we examine the "catalytic" domains of two families of GTPase-activating proteins. Here significant sequence similarities are confined to rather small domains that occur at different locations within large, variable-length parent sequences. Next we analyze a group of proteins that contain two types of "Src homology" domains in different multiplicities and orders. Then we look at protein prenyltransferases, the  $\alpha$  and  $\beta$  subunits of which contain two types of variably conserved internal repeats. Last, we show that these methods can also be profitably applied to noncoding DNA sequences. We describe conserved sequence motifs in the locus control region (LCR) of  $\beta$ -globin gene clusters from three species.

## RESULTS AND DISCUSSION

### *GTPase-Activating Proteins*

p21<sup>ras</sup> and related proteins are membrane-associated GTPases that function as key control points in mitogenic signal transduction (Bourne et al., 1991). The intrinsic GTPase activity of ras proteins is low and accessory factors known as GTPase-activating proteins (or simply GAP) increase GTPase activity approximately 100-fold (McCormick, 1989). The interplay between ras and GAP constitutes a binary switch in the signal transduction control circuitry (Bishop, 1991). GAP proteins may also act as "downstream effectors" of ras action (Hall, 1990; Wigler, 1990).

Recently, the product of the human neurofibromatosis type 1 (NF1) gene (hereafter NF1GRP for NF1 GAP-related protein) has been shown to have GTPase-activating activity and joins an expanding family of GAP proteins that have been identified in yeast (Ira1, Ira2, sar1) and mammals (Wang et al., 1991). These proteins range in size from 765 to 3079 residues, and previous multiple-alignment studies have shown that

significant homology is confined to a comparatively small (~250-residue) catalytic domain (Wang et al., 1991). Previous analyses (Ballester et al., 1990; Wang et al., 1991) were quite tedious because they required the use of several alignment programs on different computers along with a good deal of manual editing. Here we show that the combined use of integrated multiple-alignment and visualization tools can greatly accelerate the analysis of such large, variable-length sequences that share only limited areas of significant similarity.

Figure 1 shows two superimposed multiple alignments displayed by *laps*. The thin lines correspond to three-way aligned blocks of human NF1GRP and *Saccharomyces cerevisiae* Ira1 and Ira2. Note that these three proteins share an extensive area of homology

spanning the central and carboxy-terminal regions of the sequences but that Ira1 and Ira2 are dissimilar to NF1GRP in the amino-terminal third of the sequence (see also Ballester et al., 1990; Marchuk et al., 1991). The thick lines located near the center of the plot represent five-way aligned blocks among NF1GRP, Ira2, Ira1, bovine GAP, and *Schizosaccharomyces pombe* sar1. Thus the significant sequence similarity common to all five proteins is confined to a domain that represents only about 10% of the average length of these sequences. This domain (called the GAP "catalytic domain"), when removed from any of its parent proteins, has been shown to genetically complement *ira1<sup>-</sup>/ira2<sup>-</sup>* yeast mutants and has also been shown to have GTPase-activating activity in vitro (Ballester et al., 1990; Martin et al., 1990; Xu et al., 1990).

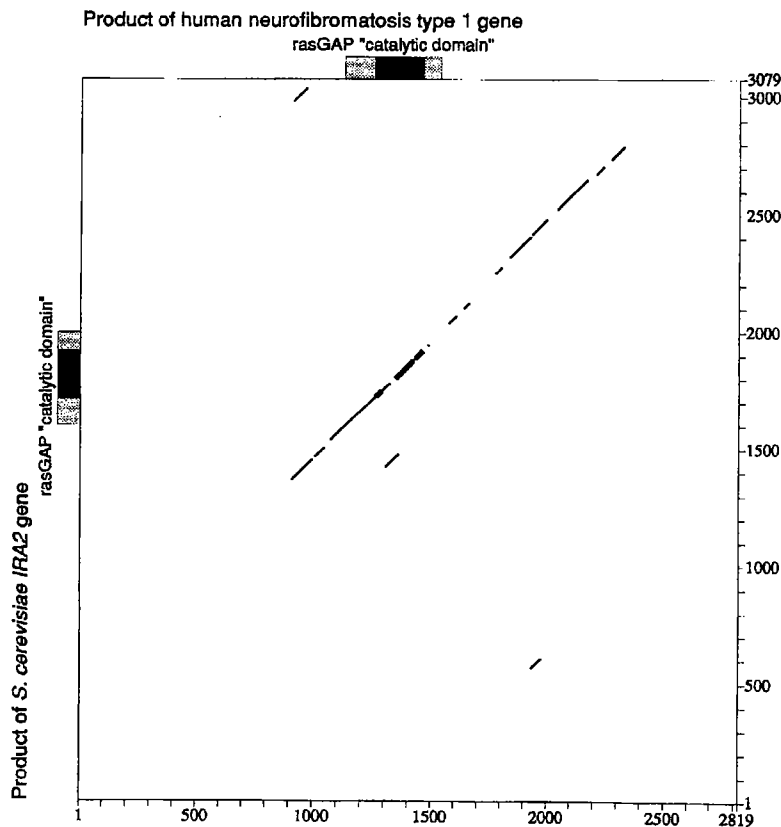


Figure 1. Graphical alignment of rasGAP-related proteins.

*Laps* plot of superimposed three-way and five-way *pblock* alignments among rasGAP-related proteins. The thin lines show the three-way aligned blocks for NF1GRP, Ira2, and Ira1. The thick lines show the five-way aligned blocks for NF1GRP, Ira2, Ira1, rasGAP, and sar1. The gray box labeled "rasGAP catalytic domain" corresponds to a fragment of NF1GRP that is sufficient to complement *ira1<sup>-</sup>/ira2<sup>-</sup>* yeast mutants in vivo (Ballester et al., 1990). The black box contained within corresponds to the most highly conserved region derived from a multiple-alignment analysis of the five proteins (Wang et al., 1991). Sources of the sequence data were as follows: human NF1GRP (GenBank/GenPept Accession No. M82814); *S. cerevisiae* Ira1 (PIR/NBRF Accession No. A30135); *S. cerevisiae* Ira2 (PIR/NBRF Accession No. A35656); bovine rasGAP (PIR/NBRF Accession No. S01966); *Sc. pombe* sar1 (GenInfo Backbone Accession No. B34137).

For purposes of molecular modeling and the identification of sequence motifs, Wang et al. (1991) manually extracted subsequences corresponding to the GAP catalytic domain from each of five proteins for gapped alignment using the program *msa* (Lipman et al., 1989). We have integrated *msa* with *lax* in such a way that one can use a computer mouse to select subsequences and run *msa* alignments in the background using an "external command" function. We illustrate the use of this interactive procedure in detail in a subsequent section.

There is growing evidence that each subfamily of ras homologs may have its own type of GTPase-activating factor(s). The NF1-related GTPase-activating proteins that we have been discussing are now referred to as rasGAPs to distinguish them from these other proteins. For example, Rubinfeld et al. (1991) have cloned an activating protein (rapGAP) that is specific for Krev-1 p21 and has absolutely no similarity to NFIGRP, Ira1, Ira2, bovine GAP, or sar1 (or any other known protein). Similarly, a protein encoded by the human breakpoint cluster region (or Bcr) on the Philadelphia chromosome associated with chronic myelogenous leukemia has been shown to serve as a GAP for the rho/rac family of p21 GTPases (Diekmann et al., 1991). The Bcr protein is not related to either rasGAP or rapGAP but does have significant local similarities to both chimaerin (Hall et al., 1990) and the human and bovine phosphatidylinositol 3-kinase-associated p85 protein (also called GRB-1 or p85alpha) (Otsu et al., 1991; Skolnik et al., 1991). Furthermore, the carboxy-terminal 401 residues of Bcr are able to stimulate the GTPase activity of p21<sup>ras</sup> in vitro (Diekmann et al., 1991) and might thus be considered the catalytic domain by analogy with the rasGAP family of sequences.

We have used *pblocks*, *lax*, and *msa* to characterize highly significant ( $P < 10^{-5}$ , Karlin and Altschul, 1990) sequence motifs in rhoGAP-related sequences as Wang et al. (1991) have done for the rasGAP family. Significant sequence similarity among Bcr, chimaerin, and GRB-1 is confined to a rather small, 140-residue homology block; the region is even more restricted when the bovine p85beta sequence (Otsu et al., 1991) is included (Fig. 2A). A textual alignment of this region is shown in Fig. 2B. We suggest that this region may contain some of the essential determinants of rhoGAP activity. Similar studies on the rasGAP family of sequences have directed in vitro mutagenesis studies to a small number of essential residues (Wang et al., 1991). Incidentally, in the course of our studies, we also identified a putative CaLB or "C-2" motif in the Bcr sequence (Fig. 2). CaLB is believed to be a phospholipid-binding motif involved in the Ca<sup>2+</sup>-dependent translocation of certain regulatory proteins (including rasGAP) to membranes (Clark et al., 1991). (See Maru

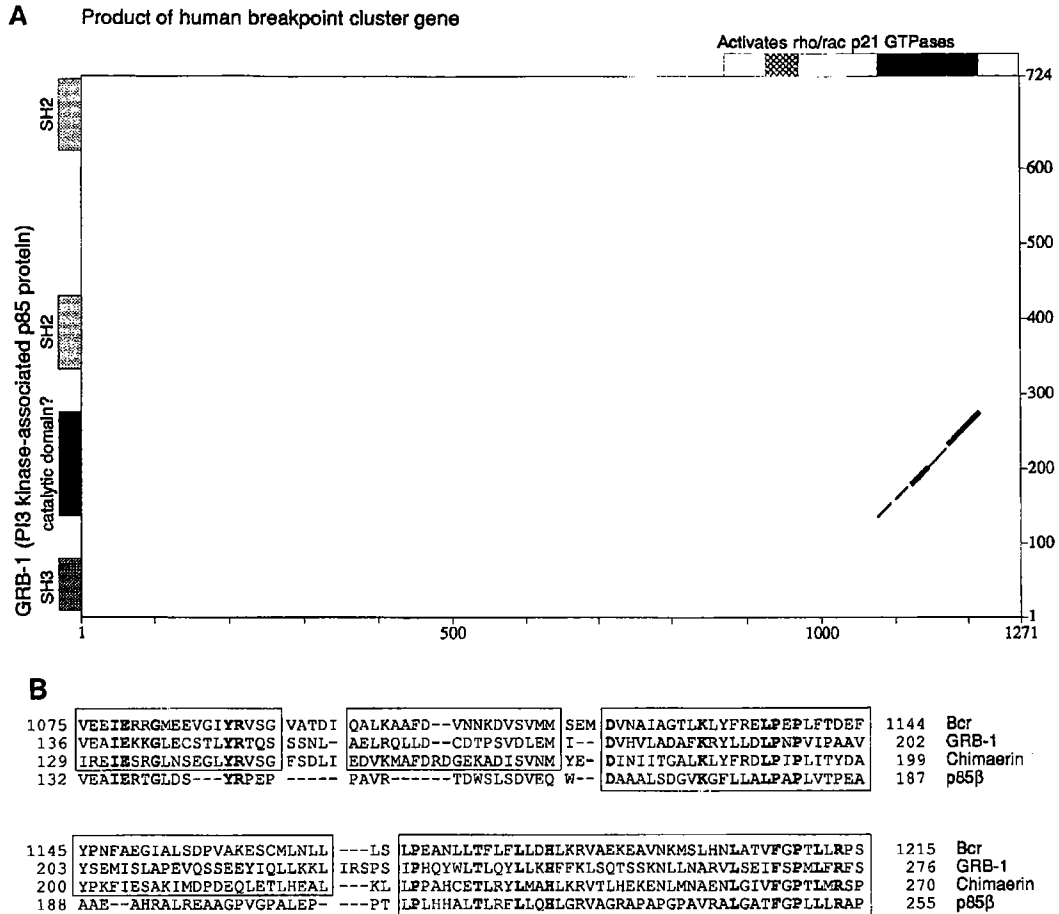
and Witte, 1991, for additional motifs and activities associated with Bcr.)

### Src Homology Domains

Named for the oncogene product in which they were first observed, "Src homology" domains are protein sequences that appear to mediate protein-protein interactions in cytoplasmic signaling pathways that respond to growth factor stimulation (Koch et al., 1991). There are two distinct varieties of Src-related sequences, SH2 and SH3. SH2 domains are approximately 100 residues in length and bind tyrosine-phosphorylated proteins. SH3 domains are approximately 45 residues in length and may be involved in interactions between the cytoskeleton and plasma membrane. From a sequence analysis point of view, there are two challenging aspects of SH2/3 domains. First, they are difficult to detect and align because the conserved sequence patterns are short and interrupted by multiple, variable-length gaps. Second, SH2 and SH3 domains may exist in multiple copies within a single sequence, and the relative number and ordering of these domains are variable such that global alignments of the parent sequences have no meaning. The combined use of *sim*, *pab*, *lax*, and *msa* aids greatly in the identification and characterization of SH2/3 domains. Indeed these problems are approached only with great difficulty using other methods and have forced some investigators to rely on visual inspection for identification and alignment (Koch et al., 1991).

GRB-1 (see previous section) is described as containing three widely spaced Src homology domains in the following order: SH3-SH2-SH2 (see Fig. 4 in Skolnik et al., 1991). They also compare the domain organization of GRB-1 with four other SH2/3-containing proteins and present a multiple alignment of intra- and interscience repeats. We reproduce their analysis here using *sim*, *pab*, *lax*, and *msa* and show how quickly and efficiently such complex analyses are performed using these tools.

Figure 3 shows a snapshot of the computer screen during a session in which *lax* was used to analyze the five-way alignment generated by *pblocks* from c-src, v-crk, GAP, phospholipase C-γ (PLC-γ), and GRB-1. Note that the domain organizations of GRB-1 and GAP are readily interpretable after one gains some familiarity with *lax* (compare with Fig. 4C in Skolnik et al., 1991). Starting at the bottom left corner and proceeding up the vertical axis, the diagonals correspond to three Src homology domains in GRB-1 in the following order: SH3-SH2-SH2. Likewise, the horizontal dimension shows SH2-SH3-SH2 domains in bovine rasGAP. Any of the possible pairwise projections can be selected for viewing and thus one can readily obtain the locations of SH2/3 domains in each protein.



**Figure 2. Graphical and textual alignment of rhoGAP-related proteins.**

(A) *Laps* plot of three-way and four-way *pblock* alignments among rhoGAP-related proteins. The three-way blocks for the product of the breakpoint cluster (Bcr) gene, GRB-1 (phosphatidylinositol 3-kinase-associated protein), and chimaerin correspond to the thin lines on the plot. A paralogous sequence from bovine brain, p85beta, was added and the generated four-way blocks are represented by the superimposed thick lines. The gray boxes on the left vertical axis signify "Src homology domains" in GRB-1 as described in the following section. The long white box at the right end of the top axis represents the carboxy-terminal 401 residues of Bcr which alone are capable of activating the GTPase activity of rho/rac p21 proteins (Diekmann et al., 1991). The black box contained within (and the intersecting black box on the lower left vertical axis) signifies the extent of mutually consistent local alignments within the parent sequences. These subsequences were subject to gapped alignment (B). The crosshatched box corresponds to the putative CaLB domain (compare Bcr residues 927-969 with Fig. 5 in Clark et al., 1991). (B) *Msa* alignment of putative rhoGAP "catalytic" domain. Both three-way and four-way alignments are shown in boxes corresponding to the diagonals in (A). Invariant residues are shown in boldface type. Sources of the sequence data were as follows: human GRB-1 (GenInfo Backbone Accession No. B30308); human chimaerin (PIR/NBRF Accession No. S08242); human Bcr (PIR/NBRF Accession Nos. A26172/A29387); bovine p85beta (GenInfo Backbone Accession No. B30276).

The pointer icon marking a diagonal line in the upper left corner selects the blocks of the five-way alignment corresponding to this diagonal for inspection in the higher "schematic view" window (Fig. 3). In this case there are two five-way aligned blocks that project onto the chosen line segment, and they reveal that the SH2 domain is also duplicated in PLC-γ. The pointing-hand icon selects another block, and the schematic view and "alignment text view" of that block

are shown in windows at the middle and lower right of the screen. Comparison of the two schematic views shows that block 2 (rather than block 3) might be part of a longer alignment also containing block 26. This identifies subranges from the parent sequences for gapped alignment using *msa*—one simply supplies these block numbers as arguments to the "spanblocks" utility in the external command window (Fig. 3). *Msa* runs in the background and one can then view the

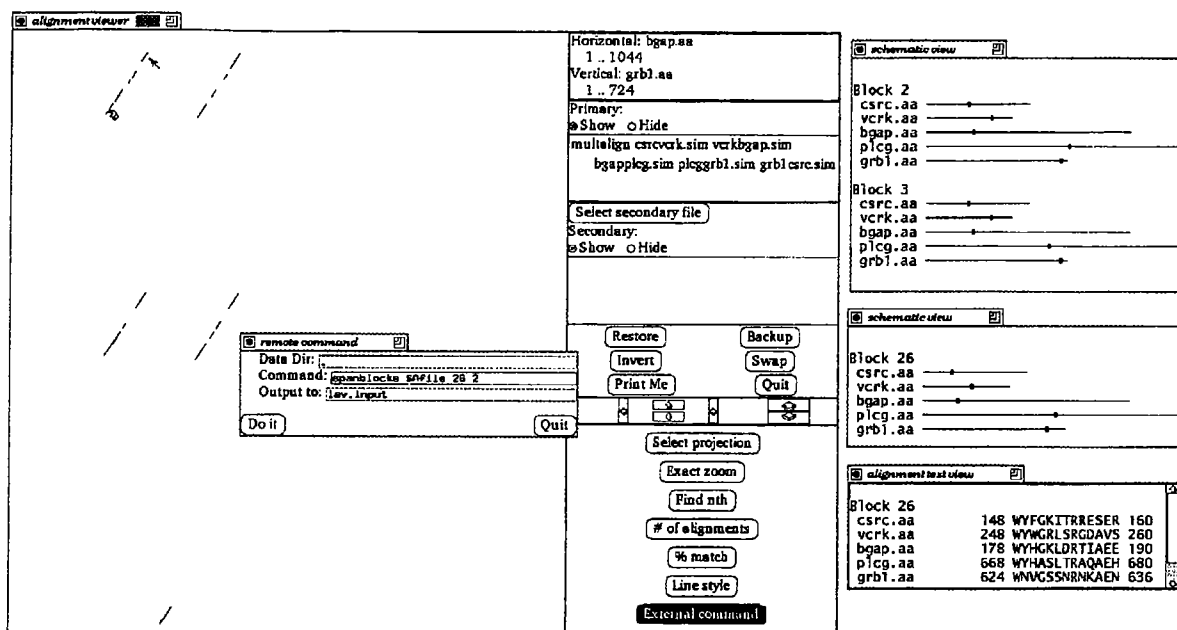


Figure 3. Interactive analysis of Src homology domains.

Contents of the computer screen when *lax* is used to examine the aligned blocks among five sequences containing SH2 and SH3 domains. *Pblocks* was applied to chicken *c-src* (PIR/NBRF Accession No. A00630), avian sarcoma virus *v-crck* (PIR/NBRF Accession No. S00872), bovine *rasGAP* (PIR/NBRF Accession No. S01966), human phospholipase C- $\gamma$  (PIR/NBRF Accession No. A36466), and GRB-1 (GenInfo Backbone Accession No. B30308). The vertical axis corresponds to GRB-1; there is an SH3 domain at the bottom, an SH2 domain in the middle, and an SH2 domain at the top (compare with the vertical axis labels in Fig. 2A). In bovine *rasGAP*, which corresponds to the horizontal axis, the domains appear close together in the order SH2-SH3-SH2. A pointing-hand icon and an arrow mark the ends of a chain of diagonal lines in the upper right corner. The arrow points to a diagonal line corresponding to two blocks, which differ only in their *plcg* segment (as indicated in the top "schematic view" window). The lower diagonal line corresponds to a single block (number 26) whose schematic view reveals that the block involves all segments that lie just before the corresponding segments of block 2. This suggests that blocks 26 and 2 might be meaningfully connected into a single gapped alignment.

*msa*-generated alignment in *lax* or inspect the textual form of the alignment.

We can further refine the analysis by adding self-comparisons to the arguments for *pab* for those sequences that contain more than one SH2 domain (i.e., *rasGAP*, PLC- $\gamma$ , and GRB-1). When this is done, all of the local alignments collapse into two diagonals (not shown) representing a multiple alignment of the eight SH2 domains found in the five proteins. As before, by using *lax*'s alignment text view facility to determine the block numbers corresponding to this eight-way alignment, we can select subranges from all five sequences and submit them for an *msa* alignment using the external command function. Such an alignment is shown in Fig. 4 (compare with Fig. 4A in Skolnik et al., 1991). Thus using the interactive and integrated tools embodied in *lax* and associated programs, complex analyses (that usually require extensive manual editing of sequence data) can easily be accomplished.

Based on sequences "aligned by eye," Koch et al.

(1991) have divided SH2 domains into a series of conserved motifs separated by intervening regions of greater sequence variability. The boxed residues in Fig. 4 correspond to these conserved motifs and it is instructive to compare the subjective alignments of Koch et al. (1991) and Skolnik et al. (1991) to a rigorous and objective analysis using the program *msa*. Particularly with respect to the placement of gaps, there are quite a few discrepancies (Fig. 4). Also, at least in this data set, some sequences in the "variable" regions appear to be nearly as well-conserved as the subjectively defined motifs (e.g., compare motif V with sequences flanking motifs II and III in Fig. 4).

### Protein Prenyltransferases

The covalent attachment of isoprenoid (or prenyl) groups to various proteins is important for the targeting and anchoring of these proteins to cellular membranes that are the sites of their function (Rine and Kim, 1990; Der and Cox, 1991). Both the heterotri-



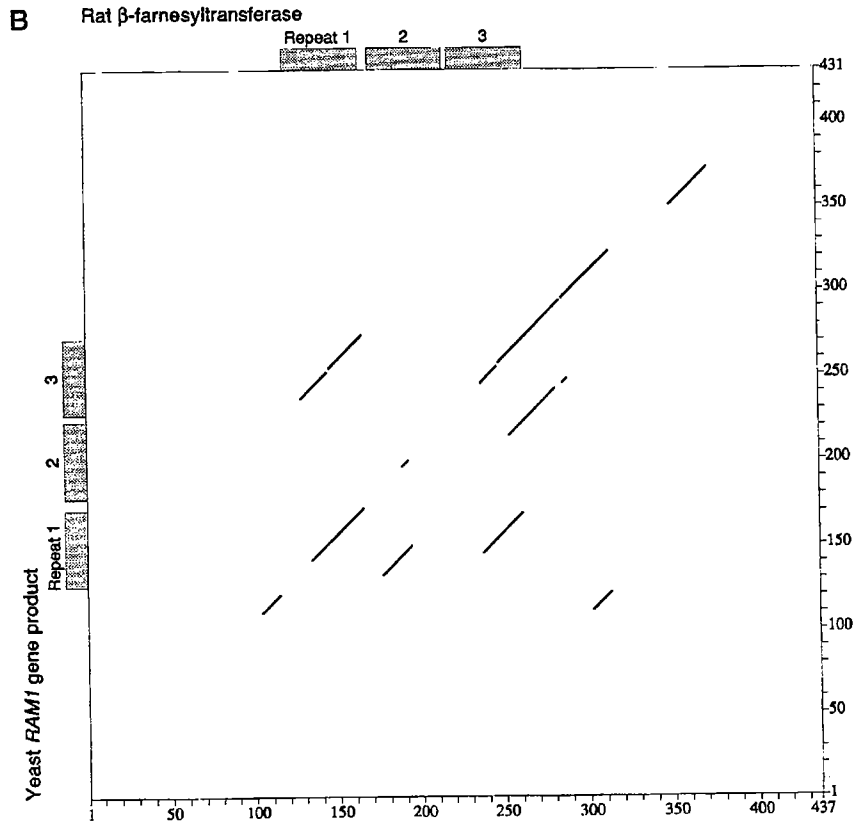
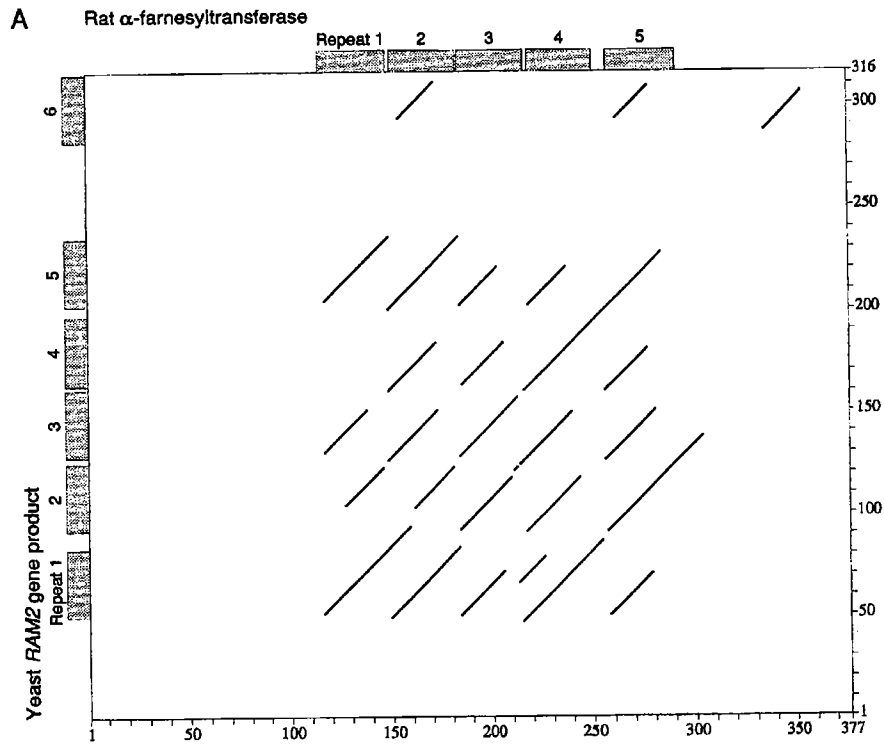




Figure 6. Textual alignments of protein prenyltransferases.

(A) *Msa* alignment of central repetitive sequence domain in  $\alpha$ -farnesyltransferase homologs. Residue numbers on the left correspond to subsequences of rat FT- $\alpha$  and yeast Ram2 and Mad2. The boxed areas labeled "Repeat 1," etc., correspond to the gray rectangles on the axes in Fig. 5A. (B) *Msa* alignment of central repetitive sequence domain in  $\beta$ -farnesyltransferase homologs. Residue numbers for rat FT- $\beta$  and yeast Ram1, Bet2, and Cdc43 are on the left. The boxed areas labeled "Repeat 1," etc., correspond to the gray rectangles on the axes in Fig. 5B.

### Locus Control Region of the $\beta$ -like Globin Gene Cluster

The power of the multiple-alignment software for analysis of DNA sequences is illustrated by its application to the locus control region of mammalian globin gene clusters. A dominant, *cis*-acting control region is

located at the 5' end of mammalian  $\beta$ -like globin gene clusters (reviewed in Orkin, 1990; Townes and Behringer, 1990). When attached to globin genes and stably transferred into cell lines or transgenic mice, this DNA sequence will confer high-level expression in erythroid tissues, regardless of the chromosomal position of the integrating genes. Thus the LCR has the properties of a powerful tissue-specific erythroid enhancer, and it appears to confer some degree of independence from position effects of the surrounding chromosomal DNA. However, the extent of such insulation from position effects is controversial; in particular, it is not clear whether the amount of expression of the linked globin gene is linearly related to its copy number in the integrated site (e.g., compare Grosfeld et al., 1987, and Reitman et al., 1990, with Ryan et al., 1989). Additionally, the LCR plays an important role in developmental regulation, although current data concerning the interaction of the LCR with promoters at progressive stages of development are in conflict (Enver et al., 1990; Dillon and Grosfeld, 1991). Thus, the LCR is involved in several different types of regulation, in combination with other regula-

Figure 5. Graphical alignments of protein prenyltransferases.

(A) *Laps* plot of three-way aligned blocks among farnesyltransferase  $\alpha$ -subunit homologs. *Pblocks* was applied to rat FT- $\alpha$ , yeast Ram2, and yeast Mad2. Gray rectangles along the axes signify 34-residue internal repeats as further detailed by Boguski et al. (in press). (B) *Laps* plot of four-way aligned blocks among farnesyltransferase  $\beta$ -subunit homologs. *Pblocks* was applied to rat FT- $\beta$ , yeast Ram1, yeast Bet2, and yeast Cdc43. Gray rectangles along the axes signify 45-residue internal repeats (Boguski et al., in press). Sources of the sequence data were as follows: rat farnesyltransferase  $\alpha$  subunit (Chen et al., 1991a); *S. cerevisiae* RAM2 gene product (He et al., 1991); *S. cerevisiae* MAD2 gene product (GenBank/GenPept Accession No. M73821); rat farnesyltransferase  $\beta$  subunit (GenBank/GenPept Accession No. M69056); *S. cerevisiae* RAM1 (DPRI) gene product (PIR/NBRF Accession No. S07864); *S. cerevisiae* BET2 gene product (GenInfo Backbone Accession No. B32756); *S. cerevisiae* CDC43 (CAL1) gene product (GenBank/GenPept Accession No. M74109/M29471).

tory sequences, but the mechanisms by which that regulation is exerted are not understood.

The mammalian  $\beta$ -globin gene LCR is a rather long segment of genomic DNA (about 15 kb) that is marked by at least four DNase I hypersensitive sites in nuclei of erythroid cells (Tuan et al., 1985; Forrester et al., 1987). The DNase I cleavage sites in HS2 and HS3 have been mapped at high resolution, and binding sites for transcription factors have been localized. HS2 has binding sites for the erythroid transcriptional activators NFE-2, which binds to a sequence similar to that of AP-1, and GATA-1 (Talbot et al., 1990). The NFE-2 (AP-1) sites will act as strong enhancers of globin gene expression (Ney et al., 1990), and a 1.9-kb fragment containing only the HS2 region will confer position-independent expression of globin genes in transgenic mice (Caterina et al., 1991). HS3 has binding sites for GATA-1 and a protein binding to a CACC motif (Philipsen et al., 1990).

The mapped binding sites for both HS2 and HS3 are confined to a region of about 100 to 200 bp.

However, when the sequence of the  $\beta$ -globin gene LCR in humans is compared with that in goats, sequence matches are found to extend a few thousand base pairs around the known binding sites (Li et al., 1990). The pattern of matching sequences in the HS3 region was examined more closely by comparing the sequences from three different species (human: Li et al., 1985; goat: Li et al., 1991; and rabbit: R. Hardison, J. Xu, J. Mansberger, and O. Selifonova, unpublished). Figure 7 shows the positions of matching sequences between goat and human identified by *sim* (thin lines); the segments that also align between rabbit and human and between rabbit and goat are shown as thick lines. The consistently aligned segments in all three species extend over about 3000 bp, interrupted primarily by an *Alu* repeat in humans and *Nla* repeats in goat. These consistently aligned regions were then searched for recognition sites for a selected battery of transcription factors; these recognition sites are marked by the short, thin lines perpendicular to the *sim* alignment lines. The previously mapped factor binding sites (Philipsen et

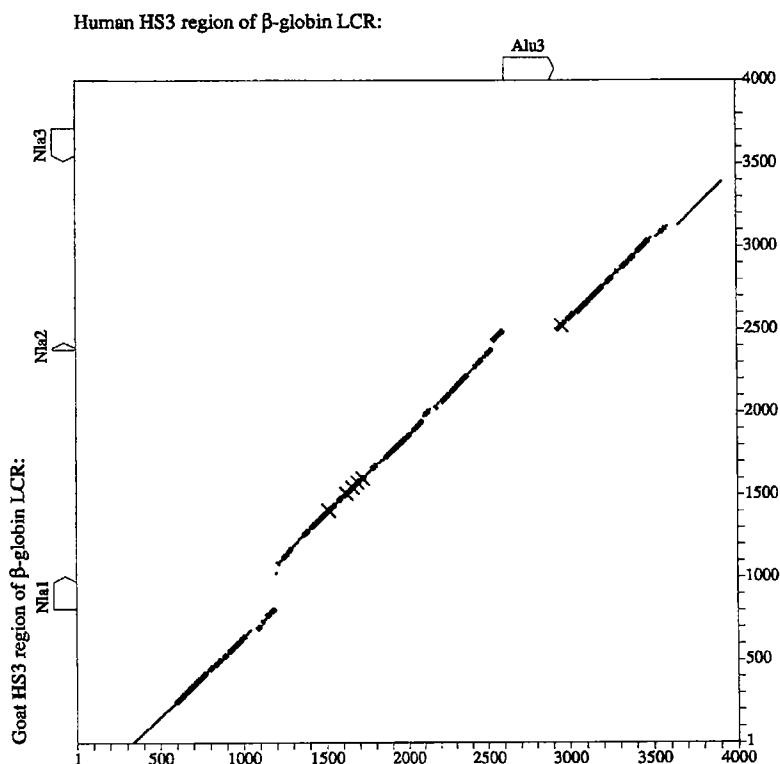


Figure 7. Graphical alignment of locus control regions.

*Laps* plot of the HS3 region of the  $\beta$ -globin gene locus control region. The thin lines (positive slope) show the positions of *sim* alignments between human (Li et al., 1985) and goat (Li et al., 1991; GenBank Accession Number M37648). The thicker lines show the portions of those alignments that are consistent with *sim* alignments between human and rabbit and between rabbit and goat, as computed by *pab*. The locations of six transcription factor binding sites common to the three sequences are indicated by short perpendicular lines; these sites were computed automatically by a program called *DNAsites* that reads *lax*-style alignments.

al., 1990) are located between 1618 and 1718 in the human sequence; this corresponds to the central four sites in Fig. 7. The search for consistently aligning sites in the three species reveals an additional site (for NFE-2/AP-1 at 1511) just before the previously mapped sites, and another site (for a CCAAT-binding protein) located 3' to the human *Alu* repeat (Fig. 7).

Thus the pattern of extensive matches in the LCR is a common feature in three different mammals; the fact that the multiple sequence alignments extend almost as far as the pairwise alignments is a strong indication of conservation. It is not clear why such extensive sequences are conserved in the LCR; for example, enhancer elements are usually only 50 to 100 bp long (multiple sites for binding transcription factors). We have considered the possibility that some protein could be encoded here, but this is not supported by inspection of the alignments. These conserved sequences are excellent candidates for elements involved in various aspects of regulation, such as insulation from position effects. Experimental analysis of this hypothesis is in progress.

## CONCLUSIONS

We have described a new method for deriving multiple alignments from consistent pairwise local alignments among a group of sequences. Vingron and Argos (1991) presented a method of comparing all possible pairwise alignments (or dot-plots) between a set of sequences to produce an alignment between two sequences that is consistent with the other alignments. Their underlying notions of alignment and consistency are essentially the same as those used here. Our approach, which involves a "backtracking" algorithm, has the advantages over Vingron and Argos's that we do not require all pairwise comparisons and that our algorithm is much more efficient (e.g., the *pab* runs for Figs. 1-7 took less than a second each on a workstation). On the other hand, the approach of Vingron and Argos has the advantage of extending readily to handle (1) "agreement of at least  $k$  of the  $n$  alignments" and (2) alignments that provide a numerical "degree of match" with each aligned pair. Gotoh (1990) presented an algorithm for computing an alignment that is consistent with the three pairwise global alignments of three sequences; his notion of an alignment is much more restricted than ours.

Our method, implemented in the program *pab*, has been integrated with previously described software tools for sequence comparison and visualization in an X-Windows/Unix environment (Schwartz et al., 1991). Furthermore, we have added the ability to interactively select subranges of sequence data for gapped alignment with the program *msa* (Lipman et al., 1989). These integrated software tools have been applied to

several complex sequence analysis problems. We described conserved domains in two families of p21 GTPase-activating proteins. We also showed how these methods can readily reveal the presence and organization of Src homology domains—an application in which ungapped local alignment methods (e.g., Schuler et al., 1991) are inadequate. We characterized the centrally located repetitive sequence domains of prenyltransferase subunits. Finally we studied conserved DNA sequences in the  $\beta$ -globin locus control region.

## MATERIALS AND METHODS

Most of the programs listed in Table 1 have a number of options. Furthermore, *laps* utilizes a special-purpose language for specifying the sequence features on the top and left borders (see Figs. 1, 2A, 5, and 7) and *lax* has an extensive repertoire of mouse-activated commands to aid the user in examining alignments. Often one of these programs calls another program, which the user may want to occasionally invoke directly to solve some special problem.

The programs are easy to use for simple tasks. For example, the following command sequence generates a copy of Fig. 5A but without the labels on the left and top borders. First, the three sequences are placed in files named, say, *rfta*, *Ram2*, and *Mad2*. Each of these files consists of an optional line beginning with any character other than a letter (the line is considered to be a descriptive header and ignored) followed by lines containing letters that specify the sequence of residues. Then a command like

```
pblocks rfta Ram2 Mad2 > 3way.a
```

sends encoded versions of aligned blocks to the file *3way.a*. The command

```
laps 3way.a > 3way.ps
```

creates a PostScript version of the simplified Fig. 5A in file *3way.ps*. That file can be sent directly to a printer. Alternatively, or in addition, the command

```
lax 3way.a
```

initiates interactive viewing of the alignments. In what follows, we describe in more detail how these programs were used.

### *Sim*

The *sim* program (Huang et al., 1990; Huang and Miller, 1991) computes any specified number of local alignments between two sequences. Like other dynamic-programming methods, it optimizes a precisely defined score, but it improves upon other methods by requiring only space proportional to the longer of the two sequences. For amino acid sequences, we followed a suggestion of Altschul (1991) and scored substitutions with the PAM200 matrix; we somewhat arbitrarily set the gap-open penalty to 12 and the gap-

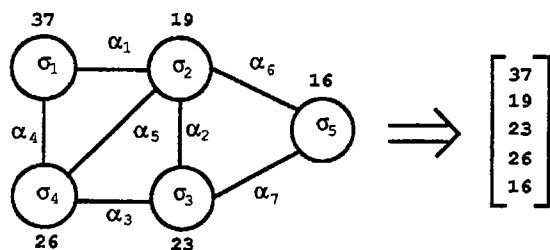


Figure 8. Graph for a connected family of pairwise alignments, with a consistent column.

$\alpha_1$  aligns sequences  $\sigma_1$  and  $\sigma_2$ , etc. Position 37 of sequence  $\sigma_1$  is aligned to position 19 of  $\sigma_2$  by alignment  $\alpha_1$ , etc.

extension penalty to 4. For nucleotide sequences, matches scored 1, mismatches scored -1, the gap-open penalty was 6, and the gap-extension penalty was 0.2.

### Pab

Here we describe the task performed by *pab* (an acronym for "pairwise alignments to aligned blocks"); the method that *pab* uses to carry out that task will be presented elsewhere. *Pab* reads a list of "alignment files," each of which contains a set of local alignments between two fixed sequences. The alignments must be given in a particular compressed format, but they can be generated by *sim*, by a traditional dot-matrix approach, or by some other alignment process. *Pab* treats each alignment file as simply a set of "points"  $(i, j)$ , where  $i$  specifies a position in one sequence and  $j$  specifies a position in the other sequence. A position of the first sequence, say residue 37, can be aligned to zero, one, or several positions in the second sequence.

A family of pairwise alignments consists of a set of sequences  $S = \{\sigma_1, \sigma_2, \dots, \sigma_n\}$  and a set of pairwise alignments  $A = \{\alpha_1, \alpha_2, \dots, \alpha_m\}$  such that each  $\alpha \in A$  is an alignment (in the sense of the preceding paragraph) between two distinct sequences in  $S$ . (The model can be extended to include alignments between a sequence and itself, but the details are not given here.) A family of pairwise alignments is naturally interpreted as an undirected graph; each node corresponds to a sequence and each edge to an alignment.

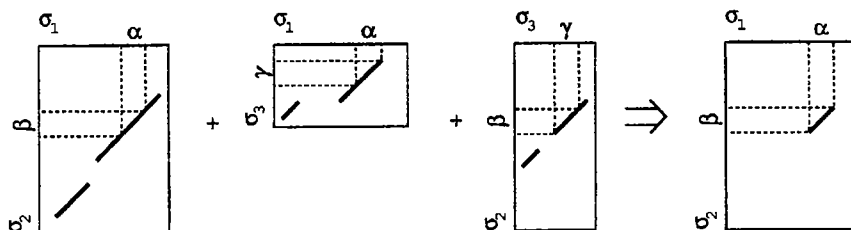


Figure 9. Three pairwise alignments that generate a single aligned block.

Region  $\alpha$  of sequence  $\sigma_1$  is matched with region  $\beta$  of  $\sigma_2$  by the first alignment. The second alignment matches  $\alpha$  with region  $\gamma$  of  $\sigma_3$ , and the third alignment pairs  $\gamma$  with  $\beta$ . This gives the aligned block  $\alpha$ - $\beta$ - $\gamma$ , which is shown projected on  $\sigma_1$  and  $\sigma_2$  in the rightmost plot. The remaining portions of the leftmost plot are absent from the rightmost plot because they lack consistency with the other two pairwise alignments.

The family is *connected* if the graph is connected, i.e., if any two nodes are connected by a path.

Consider a fixed family  $(S, A)$  of pairwise alignments and its associated graph. A *consistent column* is an assignment of integers (sequence positions) to nodes, one integer per node, so that the integers on any pair of nodes joined by an edge are related by the corresponding alignment. Figure 8 gives an example, and Fig. 9 presents a graphical interpretation of a trivial case. The *aligned blocks generated by  $(S, A)$*  consist of the set of all consistent columns arranged into runs of "consecutive" columns. The *pab* program computes these blocks efficiently for any connected family  $(S, A)$ .

### Pblocks and Dblocks

These programs apply *sim* and *pab* in a systematic way to, respectively, a family of protein sequences or a family of DNA sequences. In particular, they determine (1) which pairwise alignments are computed and (2) the number of local alignments computed for a given sequence pair, as described in detail below.

Given a set of  $n$  sequences, there are  $\frac{1}{2}n(n-1)$  potential pairwise alignments between different sequences. We have found it effective and efficient to have *pblocks* and *dblocks* compute simply the "circular" family of  $n$  pairwise alignments  $\sigma_1$  vs  $\sigma_2$ ,  $\sigma_2$  vs  $\sigma_3$ , ...,  $\sigma_n$  vs  $\sigma_1$ . There is no particularly compelling rationale for this choice—others may work as well or better in a given situation.

In addition to two sequences as input, the *sim* program requires the user to supply the desired number of nonintersecting local alignments to be computed. We find it desirable to have an automatic and objective criterion to decide how many alignments to save. Specifically, *pblocks* and *dblocks* retain only those *sim* alignments that score at least  $\tau$ , where  $\tau$  is chosen so that the probability is 0.99 that random sequences matching the given sequences in length and (amino acid or nucleotide) composition have a gap-free alignment scoring at least  $\tau$ . The threshold  $\tau$  is computed by the method of Karlin and Altschul (1990). This cutoff criterion is far too lenient if the pairwise alignment is an end in itself, and may not be sufficiently stringent if only three sequences are being compared (for example, see the extraneous blocks in Fig. 1). On the other hand, it is unlikely to exclude potentially interesting matches and worked acceptably for multiple

alignments involving more than three sequences. (By default, *pblocks* and *dblocks* compute at most 50 alignments, and they issue a warning message if that number is insufficient. For long DNA sequences, the user should change 50 to a larger number, with a command like "pblocks K=200 sequence1 sequence2...")

For the three DNA sequences from the locus control region we applied a similar threshold, but replaced 0.99 by 0.05. This much more restrictive criterion was chosen to permit a meaningful pairwise alignment of a highly conserved region (Fig. 7).

We emphasize that there is no rigorous statistical basis for using these cutoff criteria with alignments that include gaps. On the other hand, the criteria are entirely objective (although arbitrary) and have been more effective in practice than other automatic cutoff criteria that we have tried.

### Msa

Once regions of similar lengths and encompassing one or more *pab* blocks were identified in protein sequences, the regions were extracted from parent sequences by a *lax* command and aligned globally by the *msa* program (Lipman et al., 1989). An important advantage of *msa* is that it contains a sequence weighting procedure to control for data redundancy (Altschul et al., 1989).

### Laps and Lax

Alignments were viewed by modified versions of the software tools *lad* and *lav* (Schwartz et al., 1991), which were enhanced to handle multiple alignments. The new program for drawing pictorial representations of alignments is called *laps* (an acronym for "local alignments to PostScript") and the new program for interactive viewing of alignments on a computer screen is called *lax* ("local alignments to X-Windows").

### Availability

These programs, the programs they call, and certain related programs (e.g., a primitive tool, called *DNAsites*, that was used to locate the conserved transcription factor binding sites for Fig. 7) are freely available by anonymous ftp from groucho.cs.psu.edu, where they reside in the TNB subdirectory. Whereas *lax* currently runs on only Sun workstations, *laps* runs on computers with a C++ compiler. *Sim*, *pab*, and *msa* require only a C compiler. *Pblocks* and *dblocks* are Unix "shell scripts" that are easily rewritten for most other systems; they run *sim*, *pab*, and some other C programs that we provide.

### Acknowledgments

We thank Drs. Joseph Goldstein and Scott Powers for providing sequence data prior to publication. R.C.H. was supported by PHS Grant DK27635 and an RCDA DK01589. S.S. and W.M. were supported in part by Grant R01 LM05110 from the National Library of Medicine. The referees made a number of suggestions that improved the presentation of this paper.

## REFERENCES

- Altschul S (1991): Amino acid substitution matrices from an information theoretic perspective. *J Mol Biol* 219:555-565
- Altschul SF, Carroll RJ, Lipman DJ (1989): Weights for data related by a tree. *J Mol Biol* 207:647-653
- Altschul S, Gish W, Miller W, Myers E, Lipman D (1990): A basic local alignment search tool. *J Mol Biol* 215:403-410
- Ballester R, Marchuk D, Boguski MS, Saulino A, Letcher R, Wigler M, Collins F (1990): The NF1 locus encodes a protein functionally related to mammalian GAP and yeast IRA proteins. *Cell* 63:851-859
- Bishop JM (1991): Molecular themes in oncogenesis. *Cell* 64:235-248
- Boguski MS, Murray AW, Powers S (1992): Novel repetitive sequence motifs in the  $\alpha$  and  $\beta$  subunits of prenyl-protein transferases and homology of the  $\alpha$  subunit to the MAD2 gene product of yeast. *New Biol.* 4 (in press)
- Bourne HR, Sanders DA, McCormick F (1991): The GTPase superfamily: Conserved structure and molecular mechanism. *Nature* 349:117-127
- Caterina JJ, Ryan TM, Pawlik KM, Palmiter RD, Brinster RL, Behringer RR, Townes TM (1991): Human  $\beta$ -globin locus control region: Analysis of the 5' DNaseI hypersensitive site HS 2 in transgenic mice. *Proc Natl Acad Sci USA* 88:1626-1630
- Chen WJ, Andres DA, Goldstein JL, Brown MS (1991a): Cloning and expression of a cDNA encoding the  $\alpha$  subunit of rat p21<sup>ras</sup> protein farnesyltransferase. *Proc Natl Acad Sci USA* 88:11368-11372
- Chen WJ, Andres DA, Goldstein JL, Russell DW, Brown MS (1991b): cDNA cloning and expression of the peptide-binding  $\beta$  subunit of rat p21<sup>ras</sup> farnesyltransferase, the counterpart of yeast DPR1/RAM1. *Cell* 66:327-334
- Clark JD, Lin LL, Kriz RW, Ramesah CS, Sultzman LA, Lin AY, Milona N, Knopf JL (1991): A novel arachidonic acid-selective cytosolic PLA<sub>2</sub> contains a Ca<sup>2+</sup>-dependent translocation domain with homology to PKC and GAP. *Cell* 65:1043-1051
- Der CJ, Cox AD (1991): Isoprenoid modification and plasma membrane association: Critical factors for Ras oncogenicity. *Cancer Cells* 3:331-340
- Diekmann D, Brill S, Garrett MD, Totty N, Hsuan J, Monfries C, Hall C, Lim L, Hall A (1991): Bcr encodes a GTPase-activating protein for p21<sup>ras</sup>. *Nature* 351:400-402
- Dillon N, Grosveld F (1991): Human  $\gamma$ -globin genes silenced independently of other genes in the  $\beta$ -globin locus. *Nature* 350:252-254
- Doolittle RF (ed) (1990): Aligning protein and nucleic acid sequences. In *Methods in Enzymology*, San Diego, Academic Press, Vol. 183, pp 352-474
- Enver T, Raich N, Ebens AJ, Papayannopoulou T, Costantini F, Stamatoyannopoulos G (1990): Developmental regulation of human fetal-to-adult globin gene switching in transgenic mice. *Nature* 344:309-313
- Forrester WC, Takegawa S, Papayannopoulou T, Stamatoyannopoulos G, Groudine M (1987): Evidence for a locus activation region: The formation of developmentally stable hypersensitive sites in globin-expressing hybrids. *Nucleic Acids Res* 15:10159-10176
- Gotoh O (1990): Consistency of optimal sequence alignments. *Bull Math Biol* 52:509-525
- Grosveld F, van Assendelft GB, Greaves DR, Kollias G (1987): Position-independent, high-level expression of the human  $\beta$ -globin gene in transgenic mice. *Cell* 51:975-985
- Hall A (1990): ras and GAP—Who's controlling whom? *Cell* 61:921-923
- Hall C, Monfries C, Smith P, Lim HH, Kozman R, Ahmed S, Vanniasingham V, Leung T, Lim L (1990): Novel human brain cDNA encoding a 34,000 Mr protein n-chimaerin, related to both

- the regulatory domain of protein kinase C and BCR, the product of the breakpoint cluster region gene. *J Mol Biol* 211:11-16
- He B, Chen P, Chen SY, Vancura KL, Michaelis S, Powers S (1991): RAM2, an essential gene of yeast, and RAM1 encode the two polypeptide components of the farnesyltransferase that prenylates a-factor and Ras proteins. *Proc Natl Acad Sci USA* 88:11373-11377
- Huang X, Hardison RC, Miller W (1990): A space-efficient algorithm for local similarities. *CABIOS* 6:373-381
- Huang X, Miller W (1991): A time-efficient, linear-space local similarity algorithm. *Adv Appl Math* 12:337-357
- Karlin S, Altschul S (1990): Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci USA* 87:2264-2268
- Kitten GT, Nigg EA (1991): The CaaX motif is required for isoprenylation, carboxyl methylation, and nuclear membrane association of lamin B2. *J Cell Biol* 113:13-23
- Koch CA, Anderson D, Moran MF, Ellis C, Pawson T (1991): SH2 and SH3 domains: Elements that control interactions of cytoplasmic signaling proteins. *Science* 252:668-673
- Kohl NE, Biehl RE, Schaber MD, Rands E, Soderman DD, He B, Moores SL, Pompliano DL, Ferro-Novick S, Powers S, Thomas KA, Gibbs J (1991): Structural homology among mammalian and *Saccharomyces cerevisiae* isoprenyl-protein transferases. *J Biol Chem* 266:18884-18888
- Li Q, Powers PA, Smithies O (1985): Nucleotide sequence of 16-kilobase pairs of DNA 5' to the human  $\epsilon$ -globin gene. *J Biol Chem* 260:14901-14910
- Li Q, Zhou B, Powers P, Enver T, Stamatoyannopoulos G (1990):  $\beta$ -globin locus activation regions: Conservation of organization, structure and function. *Proc Natl Acad Sci USA* 87:8207-8211
- Li Q, Zhou B, Powers P, Enver T, Stamatoyannopoulos G (1991): Primary structure of the goat  $\beta$ -globin locus control region. *Genomics* 9:488-499
- Lipman DJ, Altschul SF, Kececiglu JD (1989): A tool for multiple sequence alignment. *Proc. Natl. Acad. Sci. USA* 86:4412-4415
- McCormick F (1989): ras GTPase activating protein: Signal transmitter and signal terminator. *Cell* 56:5-8
- Marchuk DA, Saulino AM, Tavakkol R, Swaroop M, Wallace MR, Andersen LB, Mitchell AL, Gutmann DII, Boguski MS, Collins FS (1991): cDNA cloning of the type 1 neurofibromatosis gene: Complete sequence of the NF1 gene product. *Genomics* 11:931-940
- Martin GA, Viskochil D, Bollag G, McCabe PC, Crosier WJ, Haubruck II, Conroy L, Clark R, O'Connell P, Cawthon RM, Innis MA, McCormick F (1990): The GAP-related domain of the neurofibromatosis type 1 gene product interacts with ras p21. *Cell* 63:843-849
- Maru Y, Witte ON (1991): The BCR gene encodes a novel serine/threonine kinase activity with a single exon. *Cell* 67:459-468
- Ney PA, Sorrentino BP, McDonagh KT, Nienhuis AW (1990): Tandem AP-1-binding sites within the human  $\beta$ -globin dominant control region function as an inducible enhancer in erythroid cells. *Genes Dev* 4:993-1006
- Ohya Y, Goebl M, Goodman LE, Peterson-Bjorn S, Friesen JD, Tamanoi F, Anraku Y (1991): Yeast CAL1 is a structural and functional homologue to the DPR1 (RAM) gene involved in ras processing. *J Biol Chem* 266:12356-12360
- Orkin SH (1990): Globin gene regulation and switching: Circa 1990. *Cell* 63:665-672
- Otsu M, Hiles I, Gout I, Fry MJ, Ruiz-Larrea F, Panayotou G, Thompson A, Dhand R, Hsuan J, Totty N, Smith AD, Morgan SJ, Courtneidge SA, Parker PJ, Waterfield MD (1991): Characterization of two 85 kd proteins that associate with receptor tyrosine kinases, middle-T/pp60c-src complexes, and PI3-kinase. *Cell* 65:91-104
- Pearson WR, Lipman D (1988): Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA* 85:2444-2448
- Philipsen S, Talbot D, Fraser P, Grosveld F (1990): The  $\beta$ -globin dominant control region hypersensitive site 2. *EMBO J* 9:2159-2167
- Reiss Y, Scabra MC, Armstrong SA, Slaughter CA, Goldstein JL, Brown MS (1991): Nonidentical subunits of p21<sup>ras</sup> farnesyltransferase. *J Biol Chem* 266:10672-10677
- Reitman M, Lee E, Westphal H, Felsenfeld G (1990): Site-independent expression of the chicken  $\beta^A$ -globin gene in transgenic mice. *Nature* 348:749-752
- Rine J, Kim SII (1990): A role for isoprenoid lipids in the localization and function of an oncoprotein. *New Biol* 2:219-226
- Rossi G, Jiang Y, Newman AP, Ferro-Novick S (1991): Dependence of Ypt1 and Sec4 membrane attachment on Bet2. *Nature* 351:158-161
- Rubinfield B, Munemitsu S, Clark R, Conroy L, Watt K, Crosier WJ, McCormick F, Polakis P (1991): Molecular cloning of a GTPase activating protein specific for the Krev-1 protein p21<sup>ras</sup>. *Cell* 65:1033-1042
- Ryan TM, Behringer RR, Martin NC, Townes TM, Palmiter RD, Brinster RL (1989): A single erythroid-specific DNase I super-hypersensitive site activates high levels of human  $\beta$ -globin gene expression in transgenic mice. *Genes Dev* 3:314-323
- Schuler GD, Altschul SF, Lipman DJ (1991): A workbench for multiple alignment construction and analysis. *Proteins: Struct Funct Genet* 9:180-190
- Schwartz S, Yang CM, Hardison RC, Miller W (1991): Software tools for analyzing pairwise sequence alignments. *Nucleic Acids Res* 19:4663-4667
- Skolnik EY, Margolis B, Mohammadi M, Lowenstein E, Fischer R, Drepps A, Ullrich A, Schlessinger J (1991): Cloning of PI3 kinase-associated p85 utilizing a novel method for expression/cloning of target proteins for receptor tyrosine kinases. *Cell* 65:83-90
- States DJ, Boguski MS (1991): Similarity and homology. In Devereux J, Gribskov M (eds) *Sequence Analysis Primer*, New York, Freeman, p 89
- Talbot D, Philipsen S, Fraser P, Grosveld F (1990): Detailed analysis of the site 3 region of the human  $\beta$ -globin dominant control region. *EMBO J* 9:2169-2178
- Townes TM, Behringer RR (1990): Human globin locus activation region (LAR): Role in temporal control. *Trends Genet* 7:219-223
- Tuan D, Solomon W, Li Q, London IM (1985): The " $\beta$ -like globin" gene domain in human erythroid cells. *Proc. Natl. Acad. Sci. USA* 82:6384-6388
- Vingron M, Argos P (1991): Motif recognition and alignment for many sequences by comparison of dot-matrices. *J Mol Biol* 218:33-43
- Wang Y, Boguski MS, Riggs M, Rodgers L, Wigler M (1991): Sar1, a gene from *Schizosaccharomyces pombe* encoding a protein that regulates ras1. *Cell Regul* 2:453-465
- Wigler MH (1990): GAPs in understanding Ras. *Nature* 346:696-697
- Xu G, Lin K, Tanaka D, Dunn D, Wood D, Gesteland R, White R, Weiss R, Tamanoi R (1990): The catalytic domain of the neurofibromatosis type 1 gene product stimulates ras GTPase and complements ira mutants of *S. cerevisiae*. *Cell* 63:835-841