# A relational database of transcription factors

David Ghosh
Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA 02115, USA

Received December 20, 1989; Revised and Accepted March 2, 1990

## ABSTRACT

**Recent advances in the understanding of eukaryotic gene regulation have produced an extensive body of transcriptionally-related sequence information in the biological literature (1 – 8 for reviews), and have created a need for computing structures that organize and manage this information. The 'relational model'(9 – 11) represents an approach that is finding increasing application in the design of biological databases(12 – 16). This report describes the compilation of information regarding eukaryotic transcription factors, the organization of this information into five tables, the computational applications of the resultant relational database that are of theoretical as well as experimental interest, and possible avenues of further development.**

## INTRODUCTION

Most current biological databases (17,18) have linear designs that do not easily allow the retention of information at different levels of complexity. These flat-file designs differ fundamentally from relational databases in that the latter are comprised of multiple tables linked by shared information. The latter design makes possible the definition of relationships between different pieces of data in the database in a manner that would be, at best, cumbersome and impractical in the linear design. The database described in this report was organized so as to allow the incorporation of classes of information that exist in the current understanding of the eukaryotic transcriptional machinery. In this design, different aspects of a generic depiction (Figure 1A) were assigned as 'objects' for the purpose of relational database organization. Each 'object' from the generic depiction was mapped to a corresponding table(Figure 1B), such that separate categories of information regarding eukaryotic transcriptional regulation could be contained in separate tables. The ability of relational databases to fully delineate one-to-one versus one-to-many relationships is of some relevance to the present-day problems of the transcription field, in which these types of relations, especially as they apply to factors and their cognate recognition sequences, are often somewhat ill-defined. In the general organization of the database (Figure 2), one-to-one relations are indicated by single arrowheads, and one-to-many relations by double arrowheads. The type of information contained in the tables is indicated by the field names, all of which are included per each corresponding table. Table I presents in detail the lengths and descriptions of each field in each table.

## METHODS

Sequence information contained in the ELEMENTS table, and some entries in the CDNAS table were extracted from the GENBANK database. All other information, including sequences in the SITES and DOMAINS tables, was essentially derived from the literature. Sequence information contained in SITES was derived from sequences reported in publications as having some significance, or from DNA binding or mutagenesis studies presented in the results of experiments. Sequence entries in the SITES and DOMAINS tables were entered largely as they are reported in the literature. Sequence entries in the CDNAS and ELEMENTS tables were broken into 250 or 240 base sequence segments and entered as ~ 1 kilobase entries per record. Data entry in all cases was accomplished through the use of routines written in the 'dBASE' and 'C' languages.

## RESULTS AND DISCUSSION

### Compilation

SITES contains information about sequences recognized by sequence-specific transcription factors, with each field corresponding to a separate parameter of the sequence. DOMAINS contains information regarding amino acid motifs such as the 'zinc finger'(19), 'leucine zipper'(20), and 'POU'(21) domains and contains both actual peptide sequences as well as published consensus motifs. FACTORS contains general information about specific factors, CDNAS contains sequences of transcription factor cDNA clones, and ELEMENTS contains sequences of regulatory elements such as promoters and enhancers. Currently the SITES, DOMAINS, FACTORS, CDNAS, and ELEMENTS tables contain 1440, 201, 320, 18, and 184 records respectively. The aggregate of the SITES, DOMAINS, FACTORS, and CDNAS tables point to roughly 550 publications in the transcription literature.

The chosen fields were ones that might capture information most often directly associated with the entities being represented in the given records in a table. Thus, for an entry in the FACTORS table, which was intended to contain biochemical information regarding particular transcription factors, attributes such as the molecular weight, posttranslation modifications, and

* Present Address: National Center for Biotechnology Information, NLM, NIH, Bethesda, MD 20894, USA

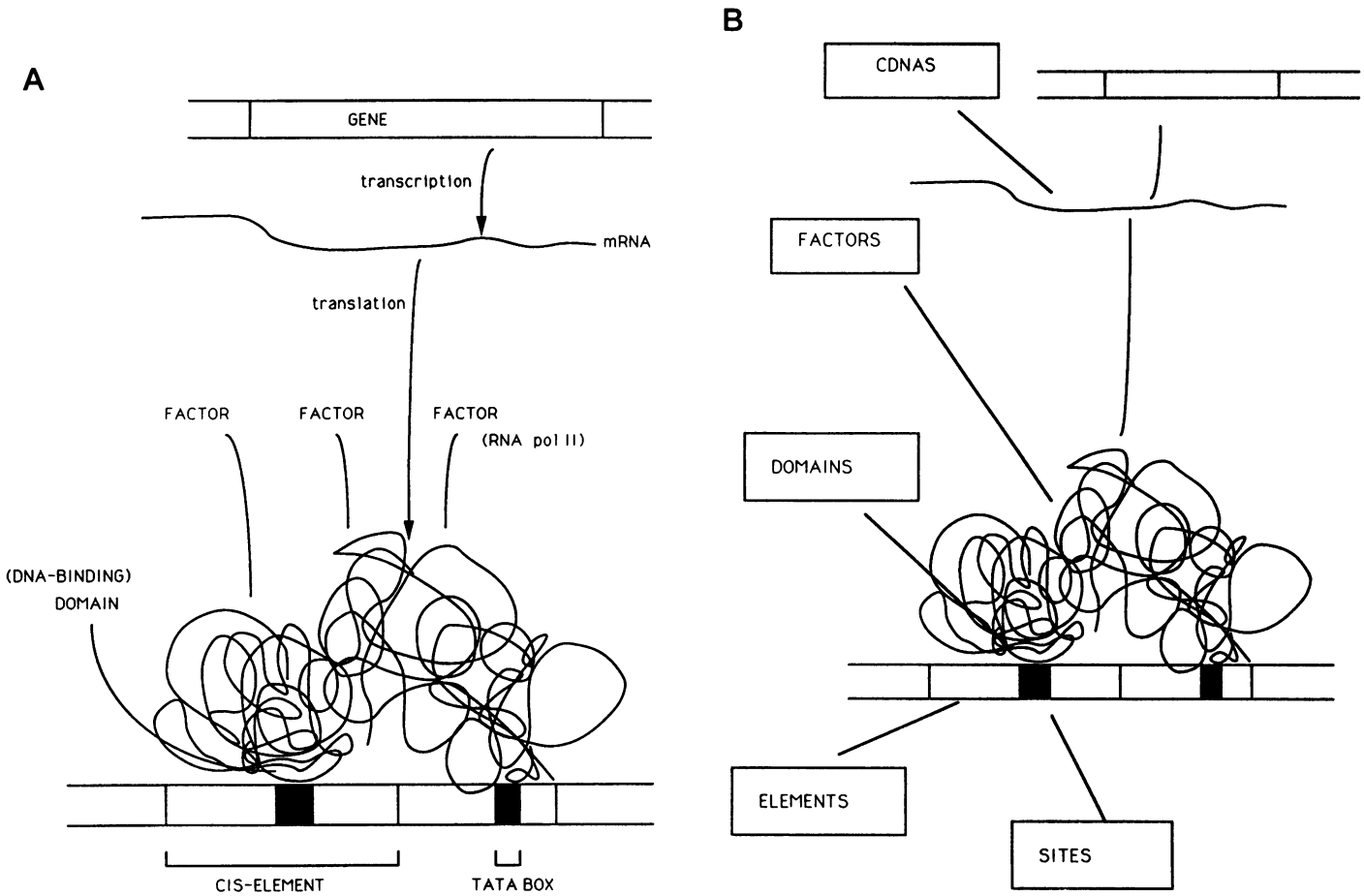



**Figure 1:** Conceptual Background. A. Generic depiction of molecular basis of transcriptional activation. Genes encoding transcription factors are transcribed into RNA, then processed and translated into proteins that participate in the transcriptional machinery. B. Mapping of entities contained in panel A, to specific tables. Each table thus contains a separate category of information pertaining to the transcription machinery. This database is primarily organized around the transcriptional phenomena depicted in the lower part of this figure.

| Table | Field | Length | Description |
|---|---|---|---|
| cdnas | fac_name | 15 | Name of factor |
| cdnas | clone_type | 4 | Type of clone(cdna, etc) |
| cdnas | source | 10 | Source of clone |
| cdnas | clone_name | 10 | Name given to clone in paper |
| cdnas | seq_clas | 2 | AA or NA (in sequence fields) |
| cdnas | accession | 6 | Genbank accession number |
| cdnas | sequence1 | 250 | Base pairs 1-250 |
| cdnas | sequence2 | 250 | Base pairs 251-500 |
| cdnas | sequence3 | 250 | Base pairs 501-750 |
| cdnas | sequence4 | 250 | Base pairs 751-1000 |
| cdnas | comments | 50 | Comments |
| cdnas | main_ref | 40 | Primary literature reference |
| cdnas | ref_n | 15 | A reference number |
| domains | fac_name | 20 | Name of factor |
| domains | domain | 15 | Name of domain(ZF for zinc finger, etc) |
| domains | seq_type | 1 | Individual or consensus sequence |
| domains | aa_start | 4 | Start position in protein |
| domains | aa_end | 4 | Stop position |
| domains | aa_seq | 60 | Amino acid sequence entry (one-letter code) |
| domains | function | 5 | Function of domain, if known |
| domains | comments | 30 | Comments |
| domains | main_ref | 40 | Primary reference |
| domains | ref_n | 20 | A reference number |
| elements | gene | 15 | Name of associated gene |
| elements | organism | 15 | Organism from which gene is derived |
| elements | locus | 10 | Genbank locus name |
| elements | accession | 10 | Genbank accession number |
| elements | type | 10 | Type of element(enhancer or promoter) |
| elements | sequence1 | 240 | Base pairs 1-240 |
| elements | sequence2 | 240 | Base pairs 241-480 |
| elements | sequence3 | 240 | Base pairs 481-720 |
| elements | sequence4 | 240 | Base pairs 721-960 |
| elements | block | 1 | Segment of sequence(1 for 1st, 2 for 2nd, etc) |
| elements | startsite | 4 | RNA start site in sequence |
| elements | comments | 50 | Comments |
| elements | main_ref | 40 | Primary reference |

| Table | Field | Length | Description |
|---|---|---|---|
| factors | fac_name | 15 | Name of factor |
| factors | distrib | 5 | Tissue distribution of factor |
| factors | system | 5 | System or organism |
| factors | cdna | 1 | Existing cdna? |
| factors | seq_spec | 1 | Sequence-specific? |
| factors | dna_bindin | 1 | DNA-binding? |
| factors | modifs | 15 | Modifications, if any |
| factors | function | 25 | Function, if known |
| factors | comments | 70 | Comments |
| factors | main_ref | 70 | Primary reference |
| factors | source | 21 | Source for isolation of factor |
| factors | gene | 21 | Gene, if known |
| factors | mw | 6 | Molecular weight of protein |
| factors | f_s | 6 | Type of finger structure (if any) |
| factors | syns | 16 | Synonymous or related factors |
| factors | ref_n | 22 | A reference number |
| sites | fac_name | 16 | Name of factor |
| sites | seq_name | 15 | Name of sequence or element |
| sites | sequence | 40 | Nucleic acid sequence |
| sites | seq_type | 1 | Individual or consensus sequence |
| sites | system | 10 | System or organism |
| sites | genome | 1 | Viral or cellular genome |
| sites | promoter | 15 | Promoter where site is found |
| sites | function | 3 | Function of site |
| sites | comments | 50 | Comments |
| sites | main_ref | 70 | Primary reference |
| sites | gene | 20 | Associated gene |
| sites | fac_source | 15 | Source of factor used to map site |
| sites | location | 20 | Location relative to mRNA start |
| sites | method | 11 | Method used to map site |
| sites | ref_n | 10 | A reference number |
| sites | code | 1 | Sequence code (IUPAC in all cases) |
| sites | keyword | 40 | Keyword to be used with other databases |
| sites | locus | 41 | Locus name in other databases |

**Table I.** Data dictionary for transcription factor database. Leftmost column contains the name of each table, second column the field name (or column of the table, third column the (maximum) length of data contained in that field, and the fourth column a brief description of the contents. For sake of convenience, all entries in this database are as characters. Fields which have a description ending in a question mark (i.e. FACTORS.Seq_spec) allow a single character yes/no (Y or N) answer.

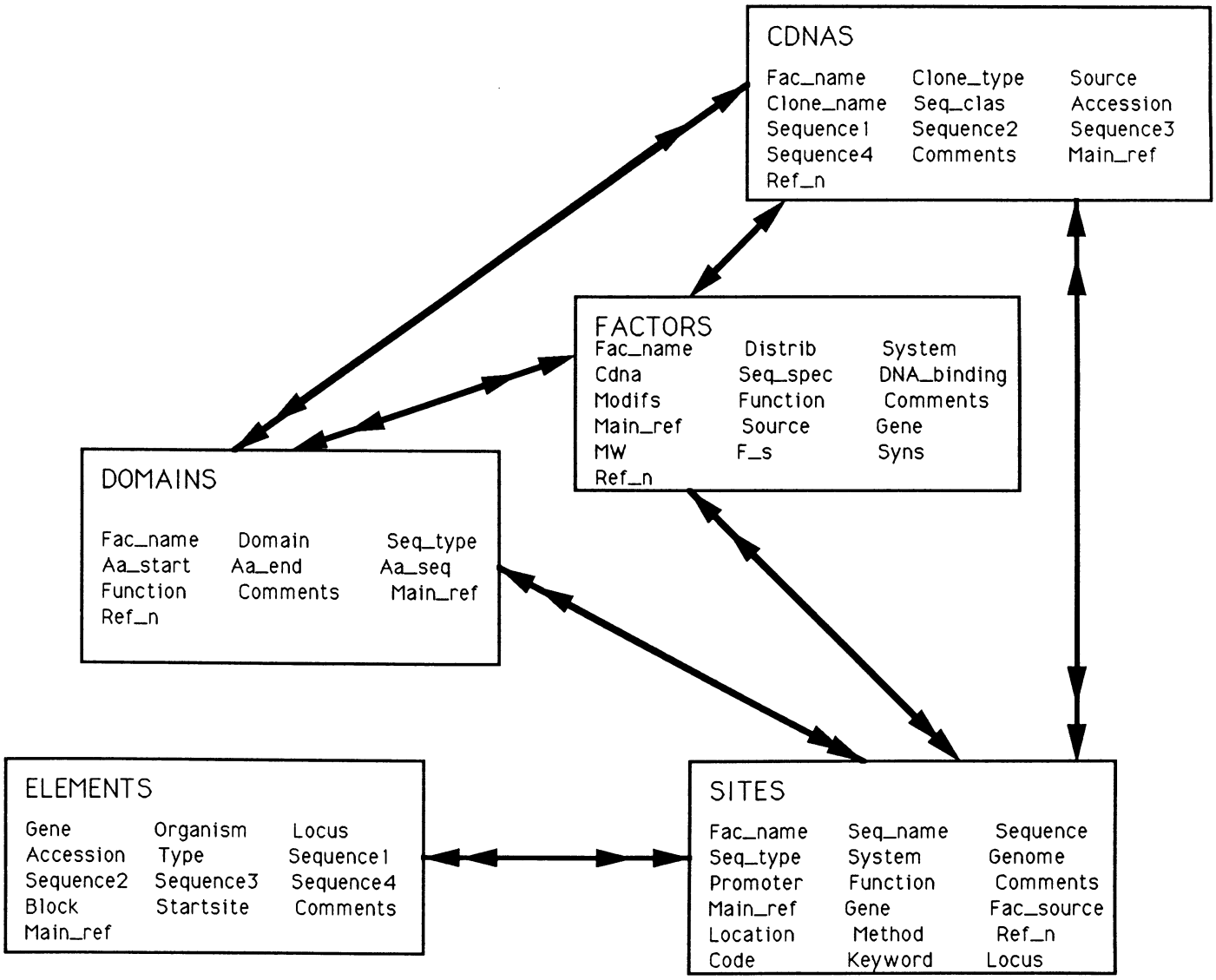



**Figure 2:** Organization of database. General organization of database, including names and contents of tables. Arrows between boxes indicate one-to-many (double arrow) or one-to-one (single arrow) relations which are usable in relational database management systems. Names of fields comprising each table are included (see Table I for descriptions of fields). Biosequence information is contained in the SITES.Sequence, DOMAINS.Aa_seq, ELEMENTS.Sequence1,2,3,4, and CDNAS.Sequence1,2,3,4 fields.

tissue distribution of specific proteins were entered into the Mw, Modifs, and Distrib fields. Whether a factor is sequence-specific or a DNA-binding protein was captured by the Seq_spec and Dna_binding fields. The experimental system, i.e. cell line or organism, used to isolate and study the factor was captured into the System and Source fields. For the SITES table, an effort was made to contain, in addition to the specific reported squence, information regarding what promoter or gene the sequence originated from (Promoter, Gene), whether the sequence is a consensus or an actual individual sequence(Seq_type), whether it comes from a viral or cellular genome(Genome), and where it is located relative to the RNA initiation site(Location). Some effort was made to capture information regarding the source of protein or extract that was used to map the site (Fac_source) and the technology used(Method). Each site was give a unique name(Seq_name) which corresponded, when possible, to a name given by the authors. The DOMAINS table captured information regarding the start and stop codons of the peptide sequence in the total protein (Aa_start, Aa_end), the name given to the domain (Domain) and its presumed function (Function) which was in most cases clearly defined in the publication from which the amino acid sequence was derived. The CDNAS table, in addition to primary nucleotide or amino acid sequence information, contains information corresponding to the cell line from which the cDNA was derived (Source) and name (Clone_name) given to the cDNA clone. In the CDNAS table, clones which are essentially bacterial fusion proteins were essentially entered starting from the first codon after the fusion protein-transcription factor junction. The ELEMENTS table, in addition to containing sequence information, contains fields for entry of information corresponding to the name of the gene (Gene), the organism from which the gene is derived (Organism), the Genbank locus and accession numbers (Locus, Accession), and the RNA start site relative to the first nucleotide of the sequence entered (Startsite). In each table, a primary literature reference field (Main_ref) was used to capture the published

**A**

```
SITES==|=====Fac_name=====|=====Seq_name=====|=======Sequence========|=====Seq_
       :  x               :                   :{ ..ACGTCA..          :
       :                  :                   :                      :
       :                  :                   :                      :

DOMAINS|=====Fac_name=====|======Domain======|=====Seq_type=====|=====Aa_start=
       :{ x               :                   :                  :
       :                  :                   :                  :
       :                  :                   :                  :

    |======Aa_end======|======Aa_seq======|=====Function=====|=====Comments====
    :                  :{                  :                  :
    :                  :                   :                  :
    :                  :                   :                  :
```

**B**

```
select fac_name, aa_seq, main_ref from domains
where fac_name in
(select fac_name from sites where sequence like "%ACGTCA%")
```

**C**

```
ANSWER1|==Sequence==|Fac_name|============================Aa_seq==============
     1 :  GACGTCA   :  CREB   :  EAARKREVRLMKNREAARECRRKKKE
     2 :  GACGTCA   :  CREB   :  LENRVAVLENQNKTLIEELKAL
     3 :  TGACGTCA  :  CREB   :  EAARKREVRLMKNREAARECRRKKKE
     4 :  TGACGTCA  :  CREB   :  LENRVAVLENQNKTLIEELKAL
```

**D**

```
create procedure find_sites_domains @seqtry varchar(30)
as
select fac_name, aa_seq, main_ref from domains
where fac_name in
(select fac_name from sites where sequence like @seqtry)
```

**Figure 3:** Multi-table query. Sample query utlizing multiple tables. Panel (A) shows query implemented in PARADOX according to the QBE or Query-by-Example (33) method. The character 'x' indicates a link between the SITES and DOMAINS tables by the Fac_name field. The '{' character indicates fields to display in the ANSWER table(SITES.Fac_name, SITES.Sequence, DOMAINS.Aa_seq). This query initially searches for all entries in the SITES.Sequence column that contain ACGTCA (..ACGTCA..). Middle panel (B) shows the same query as in panel (A) structured in Structured Query Language or SQL(34) and implemented in SYBASE. Panel (C) presents the answer to the query of panels A and B. The bottom panel (D) presents the query structured as a general procedure implementable in SYBASE and which can be applied to any DNA sequence.

citation containing the information for that particular entry. A comments field (Comments) was included in each table to capture novel or significant unusual information pertaining to an entry that could not be contained in any of the other fields.

Thus, each entry in each table corresponds to information that was presented as a result in a specific published report. An effort

was made to limit data entered into this database to those sequences and attributes that are in the published literature. Information was entered whenever it was presented or readily discernable from the data or text in a publication. In all cases an effort was made to avoid subjective interpretations of the data or conclusions of the authors, but rather to enter information in accordance with conclusions made by the investigators. Certain fields, such as Function in the FACTORS and SITES tables, or Location in the SITES table, are problematic since the corresponding information is not always easily discernable from a given report. Other problematic fields include the CDNAS.Accession, FACTORS.Distrib, FACTORS.Mw, ELEMENTS.Startsite, SITES.Fac_name, SITES.Seq_name, and SITES.Promoter fields. A recurring problem that occurs in the maintenance of this database lies in the nomenclature used to describe transcription factors. In many cases multiple names exist for the same factor, such as EBP20 and C/EBP (22–24), and one term is used to refer to a family of proteins as in the cases of AP-1(25–27) and NF-I(28–32). Minor nomenclature differences such as 'HSF' and 'HSTF', or 'NF-I', 'NF-1', and 'NFI' can grossly affect the later computational accessibility of the pertinent information that, as in these cases, refers to the same protein or family of proteins. The primary limitation of the database at this time, however, is the non-separation of relationships from fixed data. In an improved design, as is currently being explored, correlation tables would separate the large number of many-to-many relations of the current entity-relation design (Figure 2). A redistribution of fields might then occur such that fields that are more descriptive or relationships (such as SITES.Fac_source) would be moved to the linking table(such as a SITES-FACTORS table).

## Relational Queries

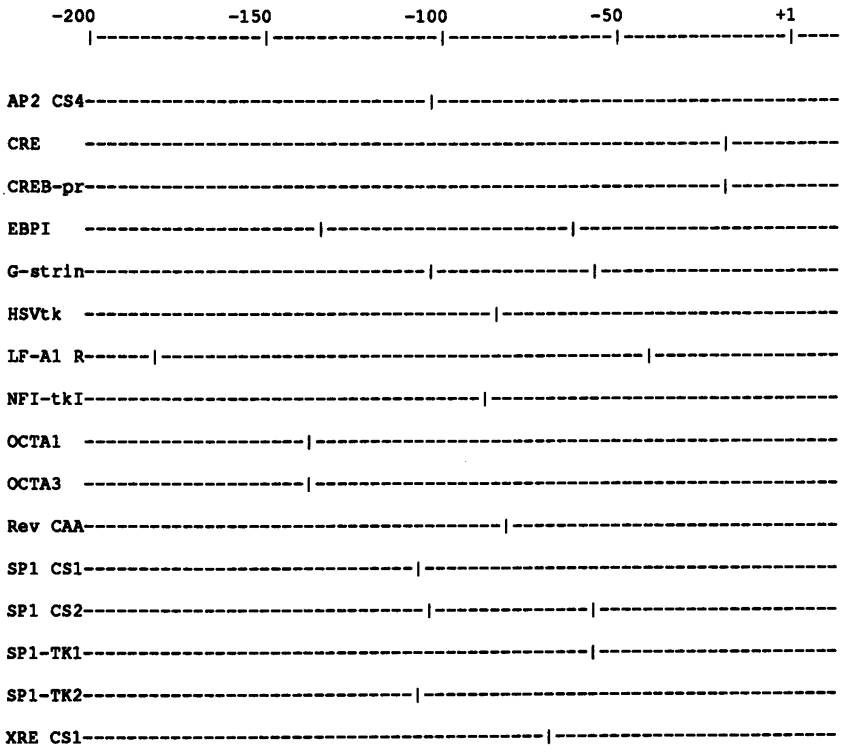
Each table except for ELEMENTS has the shared field Fac_name, which can be used in defining relations and performing multi-table queries, in the context of a relational database management system. Figure 3A shows a typical query which may be performed in the relational database management system PARADOX, and the corresponding results. This query is structured according to the QBE (Query-By-Example) method(33), and asks the question 'What known peptide domains are found in proteins that recognize the DNA sequence 'ACGTCA'?' In this case the query is performed in one step by linking the SITES and DOMAINS tables by their Fac_name fields, by searching for all occurrences of ACGTCA in the SITES.Sequence field, and by listing the SITES.Sequence, DOMAINS.Fac_name, and DOMAINS.Aa_seq entries in the ANSWER table. The same question is structured as an SQL (Structured Query Language) query(34) as implementable in SYBASE, in Figure 3B. The Aa_seq entries listed in the ANSWER in Figure 3C were derived by the QBE method(Figure 3A), but are esssentially identical to the results of the SQL query, and correspond to the CREB leucine zipper(35), and the adjacent charged domain(36). The query may be generalized as an SQL procedure as presented in Figure 3D, which may be used to apply the above question 'What known peptide domains....' to any DNA sequence. In principle, this database may be implemented in any commonly available commercial or non-commercial database management system. To date, the tables have been implemented in SYBASE, INGRES, dBASE, and PARADOX. Through database conversion routines, they can be provided as ASCII files, or in any of sixteen possible database or spreadsheet formats.

| SITE | POSITION | GAPS | SCORE | OCCURRENCE |
|---|---|---|---|---|
| NFkB-HIV(Len=11) | 647 | 0 gaps) | 11 | 1 |
| NFkB-IgkLC(Len=11) | 647 | 0 gaps) | 11 | 1 |
| NFkB_CS1(Len=11) | 647 | 0 gaps) | 11 | 1 |
| NFkB_CS2(Len=11) | 647 | 0 gaps) | 11 | 1 |
| EBP1_RS(Len=10) | 648 | 0 gaps) | 10 | 1 |
| HIV_enh(Len=10) | 648 | 0 gaps) | 10 | 1 |
| TEF1-GTI*'(Len=9) | 650 | 0 gaps) | 9 | 1 |
| IE1.9(Len=6) | 652 | 0 gaps) | 6 | 1 |
| SV40.13'(Len=7) | 652 | 0 gaps) | 7 | 1 |
| SV40.16'(Len=7) | 652 | 0 gaps) | 7 | 1 |
| SV40.6'(Len=7) | 652 | 0 gaps) | 7 | 1 |
| conalb_US1'(Len=5) | 653 | 0 gaps) | 5 | 1 |
| T-Ag-SV40.2(Len=5) | 661 | 0 gaps) | 5 | 1 |
| T-Ag-polyo.4(Len=5) | 661 | 0 gaps) | 5 | 1 |
| Adh1_US1'(Len=5) | 665 | 0 gaps) | 5 | 1 |
| Adh1_US3(Len=5) | 665 | 0 gaps) | 5 | 1 |
| SV40.11'(Len=5) | 671 | 0 gaps) | 5 | 4 |
| SP1_CS1(Len=10) | 672 | 0 gaps) | 10 | 1 |
| Sp1-HIV-1.2(Len=10) | 672 | 0 gaps) | 10 | 1 |
| G-string(Len=6) | 673 | 0 gaps) | 6 | 1 |
| HSV.tk.1'(Len=6) | 673 | 0 gaps) | 6 | 1 |
| HSV.tk.2(Len=6) | 673 | 0 gaps) | 6 | 1 |
| HVRP7(Len=8) | 673 | 0 gaps) | 8 | 1 |
| JCVprom.1(Len=8) | 673 | 0 gaps) | 8 | 1 |
| NEG_RS1'(Len=6) | 673 | 0 gaps) | 6 | 1 |
| SP1?-U2snR.3(Len=6) | 673 | 0 gaps) | 6 | 1 |
| SP1_CS2(Len=6) | 673 | 0 gaps) | 6 | 1 |
| SP1_RS1(Len=6) | 673 | 0 gaps) | 6 | 1 |
| Sp1-IE-3.1'(Len=6) | 673 | 0 gaps) | 6 | 1 |
| Sp1-IE-3.2(Len=6) | 673 | 0 gaps) | 6 | 1 |
| Sp1-IE-3.4'(Len=6) | 673 | 0 gaps) | 6 | 1 |
| Sp1-IE-3.5'(Len=6) | 673 | 0 gaps) | 6 | 1 |
| Sp1-IE-4/5'(Len=6) | 673 | 0 gaps) | 6 | 1 |
| Sp1-SV40.1(Len=6) | 673 | 0 gaps) | 6 | 1 |
| Sp1-SV40.5(Len=6) | 673 | 0 gaps) | 6 | 1 |
| Sp1?-U2snR.1(Len=6) | 673 | 0 gaps) | 6 | 1 |
| SUP-DIS.1(Len=6) | 674 | 0 gaps) | 6 | 1 |
| Sp1-IE-3.3'(Len=6) | 674 | 0 gaps) | 6 | 1 |
| Sp1-IE-4/5'(Len=6) | 674 | 0 gaps) | 6 | 1 |
| Sp1-hsp70(Len=6) | 674 | 0 gaps) | 6 | 1 |
| hsp70.2(Len=6) | 674 | 0 gaps) | 6 | 1 |
| LBP-1_RS(Len=5) | 680 | 0 gaps) | 5 | 8 |
| EBPII'(Len=7) | 681 | 0 gaps) | 7 | 1 |
| SV40.11'(Len=5) | 681 | 0 gaps) | 5 | 3 |
| CAP_CS(Len=11) | 682 | 0 gaps) | 11 | 1 |
| Sp1-HIV-1.3(Len=10) | 683 | 0 gaps) | 10 | 1 |
| ZESTE_RS(Len=6) | 692 | 0 gaps) | 6 | 2 |
| hist.gene.1(Len=5) | 702 | 0 gaps) | 5 | 2 |
| LVc-Mo-MuL(Len=5) | 705 | 0 gaps) | 5 | 1 |
| LVc_RS(Len=5) | 705 | 0 gaps) | 5 | 1 |
| EBPI'(Len=5) | 707 | 0 gaps) | 5 | 1 |
| EFII-RSV'(Len=6) | 707 | 0 gaps) | 6 | 1 |
| hist.gene.3(Len=5) | 711 | 0 gaps) | 5 | 1 |
| LBP-1_RS(Len=5) | 735 | 0 gaps) | 5 | 7 |
| SV40.11'(Len=5) | 736 | 0 gaps) | 5 | 2 |
| ZESTE_RS'(Len=6) | 740 | 0 gaps) | 6 | 1 |
| UBP1_RS(Len=8) | 742 | 0 gaps) | 8 | 2 |
| LBP-1_RS(Len=5) | 745 | 0 gaps) | 5 | 6 |
| MAJ-SPERM.1'(Len=6) | 757 | 0 gaps) | 6 | 1 |
| MAJ-SPERM.1(Len=6) | 757 | 0 gaps) | 6 | 1 |
| CRE.0(Len=10) | 760 | 0 gaps) | 10 | 2 |
| ZESTE_RS(Len=6) | 762 | 0 gaps) | 6 | 1 |
| SV40.11'(Len=5) | 767 | 0 gaps) | 5 | 1 |
| UBP1_RS(Len=8) | 774 | 0 gaps) | 8 | 1 |

**Figure 4:** Analysis of proviral HIV 3' LTR RNA initiation region. Typical primary analysis output from DYNAMIC. This analysis was performed against the HIV (HXB2 isolate) 3' LTR (query sequence) using the sequence entries from the SITES.Sequence field (search set). The output shows matching locus name(derived from SITES.Seq_name), length of matching sequence, position of match, number of gaps, relative score, and occurrence of numbers for each match. Since this particular analysis allowed no mismatches or gaps, the scre was always equal to the sequence length. Occurrence numbers are listed in descending order so that the first occurrence of a matching SITES sequence will indicate the number of total occurences of that particular matching sequence, in the query sequence.

## Sequence Analysis

In the types of analyses presented in Figure 3, sequence searches are done within a specific relational database management system and are dependent on the pattern-matching algorithm used by that database system. Generally, these types of searches consider sequences as simple character strings that are handled as 'regular expressions' as they are in most database management, word processing, or spreadsheet software. Consequently, these analyses do not handle nucleotide ambiguity codes, mismatches, or gaps as do those performed by many molecular biology sequence analysis programs. Sequence information contained in SITES and




| SEQUENCE | P/250 | N/HSV GENOME |
|---|---|---|
| YCSCCMNSSS | .05 | 536 |
| CGTCA | .25 | 322 |
| GCGTCA | .06 | 90 |
| ATGCA | .25 | 125 |
| GGGCGG | .06 | 684 |
| ATTGGCGAAT | .0002 | 1 |
| TGRMCC | .24 | 499 |
| GCCAATGA | .004 | 3 |
| ATTTGCAT | .004 | 4 |
| ATGCAAAT | .004 | 4 |
| ATTGG | .25 | 126 |
| KGGGCGGRRY | .08 | 63 |
| GGGCGG | .06 | 684 |
| GGGGCGGCGC | .0002 | 9 |
| TGGGCGGGGT | .0002 | 6 |
| CACGW | .10 | 199 |

**Figure 5:** Promoter searches. Analysis of HSV tk promoter(52) using a SITES-derived search file. Analysis was performed using the program MICROGENIE. P/250 refers to the probability of occurrence of the matching sequence in 250 base pairs of random DNA sequence, and N/HSV genome refers to the number of occurrences of the sequence in the 156 kilobase pair sequence of the HSV strain 17 genome. Listed SITES sequences are presented as they are contained in the SITES table, and ambiguity codes are according to the IUPAC(45) convention.

in DOMAINS can, however, be placed into formats for use by sequence analysis software, and thus far have been used through the MICROGENIE(37), UWGCG(38), and FASTA(39) series of programs. Tests of the DOMAINS sequence information search set suggest that the FASTA method is accurately predictive of the locations of common motifs found in known transcription factors. DOMAINS, in conjunction with FASTA, can be used to rapidly analyze new transcription factor clones isolated using recently developed technologies (40–42).

A sequence analysis tool developed at the MBCRR(43) named DYNAMIC(44) has been especially adapted and proven useful for analyses of DNA sequences with derivatives from the SITES table. This program produces output similar to that produced by commercial restriction enzyme recognition site mapping programs, but provides the user with greater control over the analysis parameters. The scoring matrices, thresholds, and gap penalties can be designed for a specific application, in this case transcription factor recognition sequence searches. DYNAMIC uses a similarity matrix whose entries are in accordance with the IUPAC ambiguity codes(45) for nucleotides, in computing scores. Further software can use the output from DYNAMIC to extract sequence and literature reference information from the original SITES table, corresponding to each listed matching SITES entry.

Figure 4 presents the results of an analysis of the HIV-1 (HXB2 isolate) 3′ long terminal repeat (LTR) (46) using a nucleic acid sequence search file derived from SITES. The proviral sequence was analyzed using this search file with DYNAMIC. The analysis was performed using parameters that allow no gaps or mismatches and produces an output which lists each perfect match to an entry in the SITES table. The predicted interactions with Sp1 and NF-kB have been previously reported(47–49). In this region of the HIV-1 LTR immediately encompassing the RNA start site, certain other interactions such as ones with the murine leukemia virus LTR-binding LVc(50), and the Rous sarcoma virus LTR-binding EFII(51) are predicted.

Figure 5 presents the results of an analysis of the Herpes Simplex virus thymidine kinase promoter-regulatory element(52) using MICROGENIE. Matching sequences are listed under the SEQUENCE columns, probability of occurrence in the P/250 column, and numbers of total occurrences in the HSV-1 strain 17(53) genome in the N/HSV column. The probability of occurrence of the matching SITES sequences in an equivalent (250 bp) length of random DNA sequence was intended as an initial assessment of the significance of each match. This probability analysis made the simplest assumption that the four nucleotides are equally distributed in this genome, and did not take into consideration the decreased distribution of CpG's in eukaryotic DNA or the high GC content of this genome(54). While the Sp1 and NF-I recognition sequences are known to bind factors and to have activating ability by DNA-binding, mutagenesis, and *in vivo* functional assays(55–58), the AP2, CREB/ATF, and Octa sequences have not been well-characterized. Though the HSV-tk promoter is an extensively characterized one, the existing data in the literature do not rule out a functional relevance for the newly predicted sequences. In addition, the predicted AP2 site lies in close proximity to the Sp1 site as it does in the SV40 early promoter(59), and the CREB/ATF site occurs in proximity to the TATA box as it does in a number of promoters(60).

Computationally-derived predictions similar to those presented

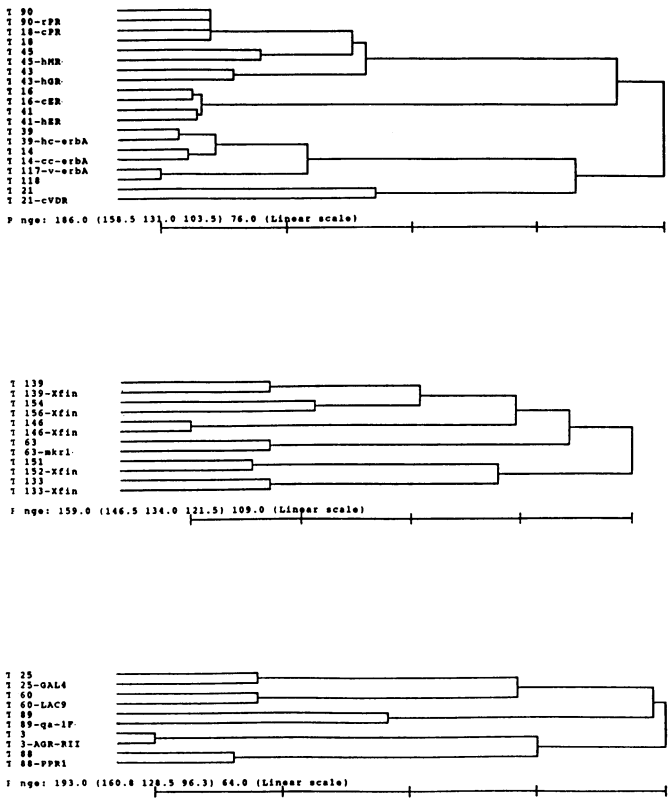



**Figure 6:** Domains clusters. Cluster analysis of peptide sequences contained in the DOMAINS table. The amino acid sequences were extracted from the DOMAINS.Aa_seq field and placed into a FASTA-compatible format. Each individual sequence was analyzed against the total set through the FASTA pairwise similarity method, and scores above 27 saved. From this output each pair and their score was extracted, and assembled into a format suitable for cluster analysis. The largest groups were plotted as trees, with the horizonital axis corresponding to pairwise distance. The three trees shown, as described in text, correspond to (A) zinc fingers of proteins in the steroid receptor superfamily, to (B) a series of zinc fingers homologous to Xfin-31, and to (C) domains with high homology to the yeast GAL4 zinc finger.

above have been tested experimentally(61). Although the predictive methods are found to have some validity as tested in experimental assays, improved computational methods for the prediction of sequence-specific interactions are under development. These include consideration of the probabilities of occurrence, the use of position-dependent weighting methods, and the correlation of matching SITES sequences derived from more than one type of search algorithm.

### Cluster Analysis

As an example of a more general use of the sequence information contained in DOMAINS, a FASTA-compatible library file was derived from the Aa_seq (amino acid sequence) entries of this table. This collection of peptide sequences was run against itself using the FASTA method, and scores above a threshold(see legend) saved. These scores were then clustered into binary dendograms (62) that graphically depict the pairwise similarities within groups of sequences. Among the largest clusters in these results are those (Figure 6) that suggest the existence of structurally-related peptide domain families. The steroid receptor cluster contains proteins having the closely related GRE/ERE/TRE specificity which mediated by residues essentially within the zinc-finger(63,64), the GAL4 cluster contains proteins whose sequence-specificity has been shown to be mediated by residues adjacent to the finger domain(65), and the Xfin cluster includes the DOMAIN enpry TF152 or Xfin-31(66) whose DNA-binding activity is reported to be nonspecific. The suggested structural homologies of these peptide sequences may imply similarities of function, such as common mechanisms for DNA sequence recognition. This method provides one computational framework for the classification of transcription factors that may complement other(8,67) such efforts.

### Future Development

A number of areas exist for the further development of this database and its associated computational tools. First, sufficient information is contained in the database to derive consensus matrices(68) from the SITES and DOMAINS tables. Algorithms exist for performing search analyses using consensus matrices(69,70), but it is likely that these as well as more standard search algorithms(71) can be modified and improved upon, to develop sequence analysis methods that are more accurately predictive of factor interactions. Second, a series of relation matrices(or correlation tables) can be developed which directly link specific records from different tables to one another. This step, in addition to fundamentally separating fixed data from their interrelations, may allow the inclusion of further information regarding these relations such as, for example, the relative sequence-specific affinity of a factor for a certain binding site. Third, other tables such as ones dedicated to genes(GENES), publications(REFERENCES), and nomenclature(NAMES), may be constructed. Fourth, pointers which provide interfaces to other databases can be developed; these databases may include ones under development, or structural databases such as the Brookhaven crystallographic database(72), bibliographic databases such as MEDLINE(73), chromosomal and linkage databases(74), and sequence databases such as GENBANK(75), EMBL(76), and PIR(77). These interfaces, along with the above other areas of development, may further extend the utility of this database as an information resource as well as a tool for biological research.

## ACKNOWLEDGEMENTS

I thank Mark Boguski, Michael Lenardo, David Landsman, James Ostell, and Temple Smith for critically reading this manuscript. This work was partially supported by grants from NIH to Temple Smith at the MBCRR/DFCI.

## REFERENCES

1. Maniatis, T., B. Goodbourn, and J.A. Fischer (1987) Science **236**, 1237–1245.
2. Jones, N.C., P.W.J. Rigby, and E.B. Ziff (1988) Genes Dev **2**, 267–287.
3. Ptashne, M. (1988) Nature **335**, 683–689.
4. Wasylyk, B. (1988) Biochim Biophys Acta **951**, 17–35.
5. Johnson, P.F., and S.L. McKnight (1989) Ann Rev Bioch **58**, 799–839.
6. Mitchell, P.J., and R. Tjian (1989) Science **245**, 371–378.
7. Salzman, A.G., and R. Weinmann (1989) FASEB J **3**, 1723–1733.
8. Struhl, K. (1989) Trends Bioch Sci **14**, 137–140.
9. Codd, E.F. (1970) Commun ACM **13**, 377–378.
10. Codd, E.F. (1979) ACM Trans Database Systems **4**, 397–434.
11. Date, C.J. (1986) Relational Databases: Selected Writings Addison-Wesley Publishing Company, Reading, MA.
12. Kanehisa, M., J.W. Fickett, and W.B. Good (1984) Nucleic Acids Res **12**, 149–158.
13. Kanehisa, M., P. Klein, and C. DeLisi (1984) Nucleic Acids Res **12**, 417–428.

14. Blundell, T.L., B.L. Sibanda, M.J.E. Sternberg, and J.M. Thornton (1987) Nature **326**, 347–352.
15. McGregor, M.J., S.A. Islam, and M.J.E. Sternberg (1987) J Mol Biol **198**, 295–310.
16. Islam, S.A., and M.J.E. Sternberg (1989) Prot Engineer **2**, 431–442.
17. Bishop, M.J., B. Ginsburg, C.J. Rawlins, and R. Wakeford (1987) In Bishop, M.J., and Rawlins, C.J. (eds.), Nucleic acid and protein sequence analysis, a practical approach. IRL Press, Oxford, England, pps. 83–113.
18. Lawton, J.R., F.A. Martinez, and C. Burks (1989) Nucleic Acids Res **17**, 5885–5899.
19. Miller, J., A.D. McLachlan, and A. Klug (1985) EMBO J **4**, 1609–1614.
20. Landschultz, W.H., P.F. Johnson, and S.L. McKnight (1988) Science **240**, 1759–1764.
21. Herr, W., R.A. Sturm, R.G. Clerc, L.M. Corcoran, D. Baltimore, P.A. Sharp, H.A. Ingraham, M.G. Rosenfeld, M. Finney, G. Ruvkun, and H.R. Horvitz (1988) Genes Dev **2**, 1513–1516.
22. Carlberg, K., T.A. Ryden, and K. Beemon (1988) J Virol **88**, 1617–1624.
23. Landschultz, W.H., P.F. Johnson, E.Y. Adashi, and B.J. Graves (1988) Genes Dev **2**, 786–800.
24. Ryden, T.A., and K. Beemon (1989) Mol Cell Biol **9**, 1155–1164.
25. Nakabeppu, Y., K. Ryder, and D. Nathans (1988) Cell **55**, 907–915.
26. Chin, R., P. Angel, and M. Karin (1989) Cell **59**, 979–989.
27. Hirai, S.-I., R.-P. Ryseck, F. Mechta, R. Bravo, and M. Yaniv (1989) EMBO J **8**, 143–1439.
28. Meisterernst, M., L. Rogge, C. Donath, I. Gender, F. Lottspeich, R. Mertz, T. Dobner, R. Fockler, G. Steltzer, and E.L. Winnacker (1988) FEBS Lett **236**, 27–32.
29. Ristiniemi, J., and J. Oikarinen (1989) J Biol Chem **264**, 2164–2174.
30. Rupp, R.A.W., and A.E. Sippel (1987) Nucleic Acids Res **15**, 9707–9726.
31. Paonessa, G., F. Gounari, R. Frank, and R. Cortese (1988) EMBO J **7**, 3115–3123.
32. Gil, G., J.R. Smith, J.L. Goldstein, C.A. Slaughter, K. Orth, M.S. Brown, and T.F. Osborne (1988) Proc Natl Acad Sci **85**, 8963–8967.
33. Zloof, M.M. (1977) IBM Systems J **16**, 324–343.
34. Date, C.J. (1987) A Guide to the SQL Standard. Addison-Wesley Publishing Company, Reading, MA.
35. Hoeffler, J.P., T.E. Meyer, Y. Yun, J.L. Jameson, and J.F. Habener (1988) Science **242**, 1430–1433.
36. Turner, R. and R. Tjian (1989) Science **243**, 1689–1694.
37. Queen, C., and L.J. Korn (1984) Nucleic Acids Res **12**, 581–599.
38. Devereux, J., P. Haberli, and O. Smithies (1984) Nucleic Acids Res **12**, 387–395.
39. Pearson, W.R., and D.J. Lipman (1988) Proc Natl Acad Sci **85**, 2444–2448.
40. Kadonaga, J.T., K.R. Carner, F.R. Masiarz, and R. Tjian (1987) Cell **51**, 1079–1090.
41. Singh, H., J.H. LeBowitz, A.S. Baldwin, and P.A. Sharp (1988) Cell **52**, 415–423.
42. Vinson, C.R., K.L. LaMarco, P.F. Johnson, W.H. Landschultz, and S.L. McKnight (1988) Genes Dev **2**, 801–806.
43. Smith, T.F., K.G. Gruskin, S. Tolman, and D. Faulkner (1986) Nucleic Acids Res **14**, 25–29.
44. algorithm for this program was derived from, Smith, T.F., and M.S. Waterman (1981) J Mol Biol **147**, 195–197. Modifications to the program were by D.V. Faulkner and T.F. Smith (1988) and by D. Ghosh (1989).
45. Nomenclature Committee of the International Union of Biochemists (1985) Eur J Biochem **150**, 1–5.
46. Starcich, B., L.Ratner, S.F. Josephs, T. Okamato, R.C. Gallo, and F. Wong-Staal (1985) Science **227**, 538–540.
47. Jones, K.A., J.T. Kadonaga, P.A. Luciw, and R. Tjian (1986) Science **232**, 755–759.
48. Wu, F.K., J.A. Garcia, D. Harrish, and R.B. Gaynor (1988) EMBO J **7**, 2117–2130.
49. Nabel, G., and D. Baltimore (1987) Nature **326**, 711–713.
50. Speck, N.A., and D. Baltimore (1987) Mol Cell Biol **7**, 1101–1110.
51. Sealy, L., and R. Chalkley (1987) Mol Cell Biol **7**, 787–798.
52. Coen, D.M., S.P. Weinheimer, and S.L. McKnight (1986) Science **234**, 53–59.
53. McGeoch, D.J., M.A. Dalrymple, A.J. Davidson, A. Dolan, M.C. Frame, D. McNab, L.J. Perry, J.E. Scott, and P. Taylor (1988) J Gen Virol **69**, 1531–1574.
54. Honess, R.W., U.A. Gompals, B.G. Barell, M. Gaxton, K.R. Cameron, R. Staden, Y.N. Chang, and G.S. Hayward (1989) J Gen Virol **70**, 837–855.
55. Jones, K.A., K.R. Yamamoto, and R. Tjian (1985) Cell **42**, 559–572.
56. Jones, K.A., J.T. Kadonaga, P.J. Rosenfeld, T.J. Kelly, and R. Tjian (1987) Cell **48**, 79–89.
57. Ben-Hattar, J., P. Beard, and J. Jiricny (1989) Nucleic Acids Res **17**, 10179–10190.
58. Boni, J., and D.M. Coen (1989) J Virol **63**, 4088–4092.
59. Mitchell, P.J., C. Wang, and R. Tjian (1987) Cell **50**, 847–861.
60. Hurst, H.C., and N.C. Jones (1987) Genes Dev **1**, 1132–1146.
61. Ghosh, D. (1989) submitted.
62. Sneath, P.H., and R.R. Sokal (1973) Numerical Taxonomy. W.H. Freeman, San Fransisco, CA.
63. Beato, M. (1989) Cell **56**, 335–344.
64. Mader, S., V. Kumar, H. deVernen, and P. Chambon (1989) Nature **338**, 271–274.
65. Corton, J.C., and S.A. Johnston (1989) Nature **340**, 724–727.
66. Lee, M.S., G.P. Gippert, K.V. Soma, D.A. Case, and P.E. Wright (1989) Science **245**, 635–637.
67. Evans, R.M., and S.M. Hollenberg (1988) Cell **52**, 1–3.
68. Stormo, G.F. (1988) Annu Rev Biophys Biophys Chem **19**, 241–263.
69. Staden, R. (1986) Nucleic Acids Res **14**, 217–231.
70. Gribskov, M., A.D. McLachlan, and D. Eisenberg (1987) Proc Natl Acad Sci **84**, 4355–4358.
71. Smith, T.S., and M.S. Waterman (1981) Adv Appl Math **2**, 482–485.
72. Lesk, A.M., D.R. Boswell, V.I. Lesk, V.E. Lesk, and A. Bairoch (1989) Prot Seq Data Anal **2**, 295–308.
73. Bicknell, F.J., R. Rada, S. Davidson, and R. Stands (1989) Nucleic Acids Res **16**, 1667–1686.
74. Pericak-vance, M.A., W.Y. Hung, L Yanoka, C. Haynes, R.J. Bertlett, J.M. Vance, J. Lu, J. Siddiqui, P.C. Gaskell, J. Starich, et.al. (1988) Aust Paediatr J **24** Suppl 1, 87–89.
75. Burks, C., and L. Tomlinson (1989) Proc Natl Acad Sci **86**, 408.
76. Hamm, G.H., and G.N. Cameron (1986) Nucleic Acids Res **14**, 5–9.
77. Sidman, K.E., D.G. George, W.C. Barker, and L.T. Hunt (1988) Nucleic Acids Res **16**, 1869–1871.