

Molecular Biophysics & Biochemistry
447b3 / 747b3

Bioinformatics

Databases II

Mark Gerstein

Class 8, 2/4/98

Yale University

Relational Databases

- Databases make program data **persistent**
- RDB's turn formless data in a number of structured tables
 - ◇ Ways of joining together tables to give various views of the data

Adaptor: An Introduction

171

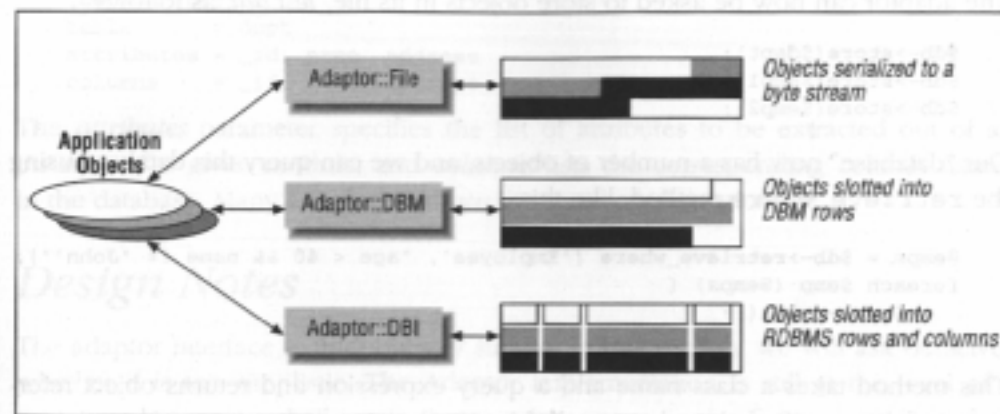


Figure 11-1. Adaptor modules

Unstructured Data

This type of “membership” analysis has been performed previously in terms of the occurrence of sequence motifs, families, functions, and biochemical pathways. Starting from the most basic units, genomes have been compared in terms of the relative frequencies of short oligonucleotide and oligopeptide “words” (Blaisdell et al., 1996; Karlin & Burge, 1995; Karlin et al., 1992; Karlin et al., 1996). The degree of gene duplication in a number of genomes has been ascertained (Brenner et al., 1995; Koonin et al., 1996b; Riley & Labedan, 1997; Wolfe & Shields, 1997; Gerstein, 1997; Tamames et al., 1997). Other analyses have looked at how many highly conserved sequence families in one organism are present in another (Green et al., 1993; Koonin et al., 1995; Tatusov et al., 1997; Ouzounis et al., 1995a,b; Clayton et al., 1997). Finally, if sequences can be related to specific functions and pathways, one can see whether homologous sequences in two organisms truly have the same role (ortholog vs. paralog) and whether particular pathways are present or absent in different organisms (Karp et al., 1996a; Karp et al., 1996b; Koonin et al., 1996a; Mushegian & Koonin, 1996; Tatusov et al., 1996, 1997). This work has yielded many interesting conclusions in terms of pathways that are modified or absent in certain organisms. For instance, the essential citric acid cycle is found to be highly modified in *H. influenzae* (Fleischmann et al., 1995; Tatusov et al., 1996). Furthermore, identifying pathways and proteins unique to certain microbes may prove useful for developing drugs (e.g. antibiotics against bacteria, Tatusov et al., 1997). In some genome annotation systems, attempts have been made to integrate a variety of membership analyses and perform them on a large scale in a highly automated fashion (Bork et al., 1992a; Bork et al., 1992b; Scharf et al., 1994; Casari et al., 1995; Ouzounis et al., 1995a; Gaasterland & Sensen, 1996).

Semi- Structured Data

```

REMARK      8 HET GROUP TRIVIAL NAME: FLAVIN ADENINE DINUCLEOTIDE (FAD)      1FNB   79
REMARK      8 CAS REGISTRY NUMBER: 146-14-5                                1FNB   80
REMARK      8 SEQUENCE NUMBER: 315                                          1FNB   81
REMARK      8 NUMBER OF ATOMS IN GROUP: 53                                   1FNB   82
REMARK      8                                                                    1FNB   83
REMARK      8 HET GROUP TRIVIAL NAME: PHOSPHATE                             1FNB   84
REMARK      8 SEQUENCE NUMBER: 316                                          1FNB   85
REMARK      8 NUMBER OF ATOMS IN GROUP: 5                                    1FNB   86
REMARK      8                                                                    1FNB   87
REMARK      8 HET GROUP TRIVIAL NAME: SULFATE                               1FNB   88
REMARK      8 SEQUENCE NUMBER: 317                                          1FNB   89
REMARK      8 NUMBER OF ATOMS IN GROUP: 5                                    1FNB   90
REMARK      8                                                                    1FNB   91
REMARK      8 HET GROUP TRIVIAL NAME: K2 PT(CN)4                            1FNB   92
REMARK      8 CHARGE: 2- ( PT(CN)4 -- )                                     1FNB   93
REMARK      8 SEQUENCE NUMBER: PT1 - PT7                                    1FNB   94
REMARK      8 NUMBER OF ATOMS IN GROUP: 9                                    1FNB   95
REMARK      8 ADDITIONAL COMMENTS: BINDING SITES USED IN MIR PHASING        1FNB   96
REMARK      8                                                                    1FNB   97
REMARK      8 HEAVY ATOM PARAMETERS ARE AS FOLLOWS:                        1FNB   98
REMARK      8 PT      PT      1      11.832 -8.309 27.027 0.68 33.00          1FNB   99
REMARK      8 PT      PT      2      13.996 -2.135 13.212 0.42 40.00          1FNB  100
REMARK      8 PT      PT      3      33.293 18.752 27.229 0.32 42.00          1FNB  101
REMARK      8 PT      PT      4      19.961 -15.348 -10.328 0.23 28.00        1FNB  102
REMARK      8 PT      PT      5       8.312 14.713 35.679 0.26 31.00          1FNB  103
REMARK      8 PT      PT      6      27.594 -7.790 23.540 0.14 35.00          1FNB  104
REMARK      8 PT      PT      7      15.917 -9.001 12.608 0.30 50.00          1FNB  105
REMARK      8                                                                    1FNB  106
REMARK      8 HET GROUP TRIVIAL NAME: URANYL NITRATE (UO2--)                1FNB  107
REMARK      8 EMPIRICAL FORMULA: UO2 (NO3)2                                 1FNB  108
REMARK      8 CHARGE: 2-                                                    1FNB  109
REMARK      8 SEQUENCE NUMBER: UR1 - UR13                                   1FNB  110
REMARK      8 NUMBER OF ATOMS IN GROUP: 3                                    1FNB  111
REMARK      8 ADDITIONAL COMMENTS: BINDING SITES USED IN MIR PHASING        1FNB  112
REMARK      8                                                                    1FNB  113
REMARK      8 HEAVY ATOM PARAMETERS ARE AS FOLLOWS:                        1FNB  114
REMARK      8  U      UR      1       8.513 16.214 36.081 0.49 27.00          1FNB  115

```

Structured Data

gid_	TrgStrt	TrgStop	did
HI0299	119	135	d193l__
HI0572	180	240	d1aba__
HI0989	56	125	d1aco_1
HI0988	106	458	d1aco_2
HI0154	2	76	d1acp__
HI1633	2	432	d1adea__
HI0349	1	183	d1aky__
HI1309	35	52	d1alo_3
HI0589	8	25	d1alo_3
HI1358	239	444	d1amg_2
HI1358	218	410	d1amy_2
HI0460	20	24	d1ans__
HI1386	139	147	d1ans__
HI0421	11	14	d1ans__
HI0361	285	295	d1ans__
HI0835	100	106	d1ans__

did_	fids
d2rs51_	1.002.007
d1imr__	1.010.002
d1pyib1	1.007.030
d1dxt_	1.001.001
d181l__	1.004.002
d1vmoa_	1.002.044
d2gsq_1	1.001.031
d1etb2_	1.002.003
d1guha1	1.001.031
d1hrc__	1.001.003
d150lc_	1.004.002
d1dmf__	1.007.035
d1l19__	1.004.002
d1yrnc_	1.010.002
d1apld_	1.001.004
d1ndab2	1.003.004
d2rmai_	1.002.036

fid_	bestrep	N_minsp	N_scop	objname
1.001.001	d1flp__	8	340	Globin-like
1.001.002	d1hdj__	4	33	Long alpha-hairpin
1.001.003	d1ctj__	9	78	Cytochrome c
1.001.004	d1enh__	18	76	DNA-binding 3-helical bundle
1.001.005	d1dtr_2	1	3	Diphtheria toxin repressor (DtxR) dimeriz
1.001.006	d1tns__	1	2	Mu transposase, DNA-binding domain
1.001.007	d2spca_	1	2	Spectrin repeat unit
1.001.008	d1bdd__	1	4	Immunoglobulin-binding protein A modules
1.001.009	d1bal__	1	5	Peripheral subunit-binding domain of 2-ox
1.001.010	d2erl__	3	5	Protozoan pheromone proteins

SQL

- SIMPLE Language for Building and Querying Tables
 - CREATE a table
 - INSERT values into it
 - SELECT various entries from it (tuples, rows)
 - UPDATE the values
-
- Example: How Many Globin Folds are there in E. coli versus Yeast?

matches table

gid_	TrgStrt	TrgStop	did	score
HI0299	119	135	d193l__	3.1
HI0572	180	240	d1aba__	0.0032
HI0989	56	125	d1aco_1	0.0049
HI0988	106	458	d1aco_2	4.4e-14
HI0154	2	76	d1acp__	1.2e-23
HI1633	2	432	d1adea_	0
HI0349	1	183	d1aky__	7.6e-36
HI1309	35	52	d1alo_3	1.1
HI0589	8	25	d1alo_3	1.8
HI1358	239	444	d1amg_2	0.002
HI1358	218	410	d1amy_2	0.00037
HI0460	20	24	d1ans__	1.8
HI1386	139	147	d1ans__	3.3
HI0421	11	14	d1ans__	6.4
HI0361	285	295	d1ans__	8.2
HI0835	100	106	d1ans__	9.7

```

create table
matches(
  gid char255,
    # Genome_ID
  TrgStrt int,
    # Start of
    # Match in Gene
  TrgStop int,
    # End of Match
    # in Gene
  did char255,
    # ID Matching
    # Structure
  score real
    # e-value
    # of Match
)

```

matches table 2

gid_	TrgStrt	TrgStop	did	score
HI0299	119	135	d193l__	3.1
HI0572	180	240	d1aba__	0.0032
HI0989	56	125	d1aco_1	0.0049
HI0988	106	458	d1aco_2	4.4e-14
HI0154	2	76	d1acp__	1.2e-23
HI1633	2	432	d1adea_	0
HI0349	1	183	d1aky__	7.6e-36
HI1309	35	52	d1alo_3	1.1
HI0589	8	25	d1alo_3	1.8
HI1358	239	444	d1amg_2	0.002
HI1358	218	410	d1amy_2	0.00037
HI0460	20	24	d1ans__	1.8
HI1386	139	147	d1ans__	3.3
HI0421	11	14	d1ans__	6.4
HI0361	285	295	d1ans__	8.2
HI0835	100	106	d1ans__	9.7

insert into

matches

(gid, TrgStrt,
TrgStop, did,
score)

values

(HI0299, 119, 135,
d193l__, 3.1)

structures table

did_	fid
d2rs51_	1.002.007
d1imr__	1.010.002
d1pyib1	1.007.030
d1dxt_d	1.001.001
d181l__	1.004.002
d1vmoa_	1.002.044
d2gsq_1	1.001.031
d1etb2_	1.002.003
d1guha1	1.001.031
d1hrc__	1.001.003
d150lc_	1.004.002
d1dmf__	1.007.035
d1l19__	1.004.002
d1yrnc_	1.010.002
d1apl_d	1.001.004
d1ndab2	1.003.004
d2rmai_	1.002.036

create table

```
structures(
  did char255,
    # ID Matching
    # Structure
  fid char255,
    # ID of fold that
    # structure has
)
```

folds table

```

create table
folds(
    fid char255,
        # fold ID
    bestrep char255,
    N_hlx int,
    N_beta int,
        # number of helices & sheets
    name char255
        # name of fold
)

```

fid_	bestrep	N_hlx	N_beta	name
1.001.001	d1flp__	8	0	Globin-like
1.001.002	d1hdj__	4	0	Long alpha-hairpin
1.001.003	d1ctj__	9	0	Cytochrome c
1.001.004	d1enh__	2	0	DNA-binding 3-helical bundle
1.001.005	d1dtr_2	1	3	Diphtheria toxin repressor (DtxR) dimeriz
1.001.006	d1tns__	1	2	Mu transposase, DNA-binding domain
1.001.007	d2spca_	0	2	Spectrin repeat unit
1.001.008	d1bdd__	0	4	Immunoglobulin-binding protein A modules
1.001.009	d1bal__	0	5	Peripheral subunit-binding domain of 2-ox
1.001.010	d2erl__	3	5	Protozoan pheromone proteins

Table
Interpretation

HI Gene	1	
HI Gene	2	
HI Gene	3	
HI Gene	4	
HI Gene	5	
HI Gene	6	
HI Gene	7	
HI Gene	8	
HI Gene	9	
HI Gene	10	

Match Table: Ways Structures A, B, and C can match HI Genome

Structures have a limited number of folds, which have various characteristics

Structures

Folds

Str A

Str B

Str C

Str D

Str E



Structure of a Table

- Row
 - ◊ Entity, Tuple, Instance
- Column
 - ◊ Field
 - ◊ Attribute of an Entity
 - ◊ dimension
- Key
 - ◊ Certain Attributes (or combination of attributes) can uniquely identify an object, these are keys
- NULL
 - ◊ Variant Records

	key	key				
Table	attr-a	attr-b	attr-c	attr-d	attr-e	attr-f
tuple-1	a1	b1	c1	d1	e1	f1
tuple-2	a2	b2	c2	d2	e2	f2
tuple-3	a3	b3	c3	d3	e3	f3
tuple-4	a4	b4	c4	d4	e4	f4
tuple-5	a5	b5	c5	d5	e5	f5
tuple-6	a6	b6	c6	d6		
tuple-7	a7	b7	c7	d7		f7
tuple-8	a8	b8	c8	d8	e8	f8
tuple-9	a9	b9	c9	d9	e9	f9
tuple-10	a10	b10	c10	d10		f10
tuple-11	a11	b11	c11	d11	e11	f11
tuple-12	a12	b12	c12	d12	e12	f12
tuple-13	a13	b13	c13	d13	e13	f13
tuple-14	a14	b14	c14	d14	e14	f14

What is a Key?

```
table matches(gid, TrgStrt, TrgStop, did, score)
table structures(did, fid)
table folds(fid, bestrep, N_hlx, N_beta, name)
```

gid -> many matches

gid,TrgStrt -> unique match (one tuple)

thus, primary key gid,TrgStrt

gid,TrgStop -> unique match as well

fid -> many did's, but did -> one fid

thus, primary key did

one-to-one between fid and name

SQL Select on a Single Table

	key	key				
Table	attr-a	attr-b	attr-c	attr-d	attr-e	attr-f
tuple-1	a1	b1	c1	d1	e1	f1
tuple-2	a2	b2	c2	d2	e2	f2
tuple-3	a3	b3	c3	d3	e3	f3
tuple-4	a4	b4	c4	d4	e4	f4
tuple-5	a5	b5	c5	d5	e5	f5
tuple-6	a6	b6	c6	d6		
tuple-7	a7	b7	c7	d7		f7
tuple-8	a8	b8	c8	d8	e8	f8
tuple-9	a9	b9	c9	d9	e9	f9
tuple-10	a10	b10	c10	d10		f10
tuple-11	a11	b11	c11	d11	e11	f11
tuple-12	a12	b12	c12	d12	e12	f12
tuple-13	a13	b13	c13	d13	e13	f13
tuple-14	a14	b14	c14	d14	e14	f14

- Select {columns} from {a table}
where {row-selection is true}
- projection of a selection
- Sort result on a attribute

SQL Select on a Single Table, Example

gid_	TrgStrt	TrgStop	did	score
HI0299	119	135	d193l__	3.1
HI0572	180	240	d1aba__	0.0032
HI0989	56	125	d1aco_1	0.0049
HI0349	1	183	d1aky__	7.6e-36
HI1309	35	52	d1alo_3	1.1
HI0589	8	25	d1alo_3	1.8
HI1358	239	444	d1amg_2	0.002
HI0016	1	173	d1dar_2	2e-07
HI0016	179	274	d1dar_1	8.5e-06
HI0016	399	476	d1dar_4	0.00031
HI0460	20	24	d1ans__	1.8
HI1386	139	147	d1ans__	3.3
HI0421	11	14	d1ans__	6.4
HI0361	285	295	d1ans__	8.2
HI0835	100	106	d1ans__	9.7

- Select * from matches where gid= HI0016

HI0016	1	173	d1dar_2	2e-07
HI0016	179	274	d1dar_1	8.5e-06
HI0016	399	476	d1dar_4	0.00031

- Select * from matches where gid= HI0016 and TrgStrt=179

HI0016	179	274	d1dar_1	8.5e-06
--------	-----	-----	---------	---------

SQL Select on a Single Table, Example 2

gid_	TrgStrt	TrgStop	did	score
HI0299	119	135	d193l__	3.1
HI0572	180	240	d1aba__	0.0032
HI0989	56	125	d1aco_1	0.0049
HI0349	1	183	d1aky__	7.6e-36
HI1309	35	52	d1alo_3	1.1
HI0589	8	25	d1alo_3	1.8
HI1358	239	444	d1amg_2	0.002
HI0016	1	173	d1dar_2	2e-07
HI0016	179	274	d1dar_1	8.5e-06
HI0016	399	476	d1dar_4	0.00031
HI0460	20	24	d1ans__	1.8
HI1386	139	147	d1ans__	3.3
HI0421	11	14	d1ans__	6.4
HI0361	285	295	d1ans__	8.2
HI0835	100	106	d1ans__	9.7

- Select did from matches where score < 0.0001

d1aky__, d1dar_2, d1dar_1

HI0349	1	183	d1aky__	7.6e-36
I0016	1	173	d1dar_2	2e-07
HI0016	179	274	d1dar_1	8.5e-06

Joins

gid_	TrgStrt	TrgStop	did	score
HI0299	119	135	d193l__	3.1
HI0572	180	240	d1aba__	0.0032
HI0989	56	125	d1aco_1	0.0049
HI0988	106	458	d1aco_2	4.4e-14
HI0154	2	76	d1acp__	1.2e-23
HI1633	2	432	d1adea__	0
HI0349	1	183	d1aky__	7.6e-36
HI1309	35	52	d1alo_3	1.1
HI0589	8	25	d1alo_3	1.8
HI1358	239	444	d1amg_2	0.002
HI1358	218	410	d1amy_2	0.00037
HI0460	20	24	d1ans__	1.8
HI1386	139	147	d1ans__	3.3
HI0421	11	14	d1ans__	6.4
HI0361	285	295	d1ans__	8.2
HI0835	100	106	d1ans__	9.7

did_	fid
d2rs51_	1.002.007
d1imr__	1.010.002
d1pyib1	1.007.030
d1dxt_d	1.001.001
d181l__	1.004.002
d1vmoa_	1.002.044
d2gsq_1	1.001.031
d1etb2_	1.002.003
d1guha1	1.001.031
d1hrc__	1.001.003
d150lc_	1.004.002
d1dmf__	1.007.035
d1l19__	1.004.002
d1yrnc_	1.010.002
d1ans__	1.007.008
d2rmai_	1.002.036

fid_	bestrep	N_hlx	N_beta	name
1.001.001	d1flp__	8	0	Globin-like
1.001.002	d1hdj__	4	0	Long alpha-hairpin
1.001.003	d1ctj__	9	0	Cytochrome c
1.001.004	d1enh__	2	0	DNA-binding 3-helical bundle
1.001.005	d1dtr_2	1	3	Diphtheria toxin repressor (DtxR) dimeriz
1.001.006	d1tns__	1	2	Mu transposase, DNA-binding domain
1.001.007	d2spca_	0	2	Spectrin repeat unit
1.001.008	d1bdd__	0	4	Immunoglobulin-binding protein A modules
1.007.008	d1qkt__	4	3	Neurotoxin III (ATX III)
1.001.010	d2erl__	3	5	Protozoan pheromone proteins

SQL Select on Multiple Tables

- Select *
from matches, structures, folds
where
matches.gid = HI0361
and matches.did=structures.did
and structures.fid = folds.fid
- Returns
matches | structures | folds
HI0361,285,295,d1ans__,8.2 | d1ans__,1.007.008 | 1.007.008,d1qkt__,4, 3,Neurotoxin III ...
- Select score,name from matches, structures, folds
where gid = HI0361and matches.did=structures.did
and structures.fid = folds.fid
8.2, Neurotoxin III ...

Foreign Key matches

structures

gid_	TrgStrt	TrgStop	did	score
HI0299	119	135	d193l__	3.1
HI0572	180	240	d1aba__	0.0032
HI0989	56	125	d1aco_1	0.0049
HI0988	106	458	d1aco_2	4.4e-14
HI0154	2	76	d1acp__	1.2e-23
HI1633	2	432	d1adea__	0
HI0349	1	183	d1aky__	7.6e-36
HI1309	35	52	d1alo_3	1.1
HI0589	8	25	d1alo_3	1.8
HI1358	239	444	d1amg_2	0.002
HI1358	218	410	d1amy_2	0.00037
HI0460	20	24	d1ans__	1.8
HI1386	139	147	d1ans__	3.3
HI0421	11	14	d1ans__	6.4
HI0361	285	295	d1ans__	8.2
HI0835	100	106	d1ans__	9.7

did_	fid
d2rs51_	1.002.007
d1imr__	1.010.002
d1pyib1	1.007.030
d1dxt_d	1.001.001
d181l__	1.004.002
d1vmoa_	1.002.044
d2gsq_1	1.001.031
d1etb2_	1.002.003
d1guha1	1.001.031
d1hrc__	1.001.003
d150lc_	1.004.002
d1dmf__	1.007.035
d1119__	1.004.002
d1yrnc_	1.010.002
d1ans__	1.007.008
d2rmai_	1.002.036

matches.did is a (foreign) key in the structures table --
i.e. looks up exactly one structure.

Selection as Array Lookup

- Same for a fold identifier from a structure id
 - ◇ `$fid=$structure{$did}`
 - ◇ (perl pseudo-code)
- Same for matches and folds tables, but this time arrays return multiple values and have multiple field keys
 - ◇ `($bestrep, $N_hlx, $N_beta, $name) = $folds{$fid}`
 - ◇ `($TrgStop,$did,$score)=$match{$gid,$TrgStrt}`
- Joining as a double-lookup
 - ◇ `$did = 1mbd__`
`($bestrep, $N_hlx, $N_beta, $name) = $folds{ $structures{$did} }`
 - ◇ Select bestrep,N_hlx,N_beta,name from structures, folds where structures.fid = folds.fid and structures.did = 1mbd__

SQL Select on Multiple Tables

Table 1	key	key			Table 2	did	fid
	gid	TrgStrt	TrgStop	did			
tuple-1	HI001	12	200	d1mbd__	tuple-i	d1lfg_1	1.007.006
tuple-2	HI002	15	231	d1hhba__	tuple-i	d1lfg_1	1.007.006
tuple-3	HI002	100	343	d1lfg_1	tuple-i	d1lfg_1	1.007.006
tuple-4	HI003	12	80	d1lfg_1	tuple-i	d1lfg_1	1.007.006
tuple-5	HI009	200	260	d1mba__	tuple-i	d1lfg_1	1.007.006
tuple-6	HI023	300	450	d2ubx__	tuple-i	d1lfg_1	1.007.006
tuple-7	HI045	2	89	d2lmg__	tuple-i	d1lfg_1	1.007.006
tuple-1	HI001	12	200	d1mbd__	tuple-ii	d1mba__	1.003.002
tuple-2	HI002	15	231	d1hhba__	tuple-ii	d1mba__	1.003.002
tuple-3	HI002	100	343	d1lfg_1	tuple-ii	d1mba__	1.003.002
tuple-4	HI003	12	80	d1lfg_1	tuple-ii	d1mba__	1.003.002
tuple-5	HI009	200	260	d1mba__	tuple-ii	d1mba__	1.003.002
tuple-6	HI023	300	450	d2ubx__	tuple-ii	d1mba__	1.003.002
tuple-7	HI045	2	89	d2lmg__	tuple-ii	d1mba__	1.003.002

- Select {columns} from {huge cross-product of tables} where {row-selection is true}
 - ◊ cross-product $T(1) \times T(2)$ builds a huge virtual table where every row of $T(1)$ is paired with every row of $T(2)$. Then perform selection on this.
- Select fid from matches, structures where $gid=HI009$ and $matches.did = structures.did$

Cross Product $A \times B$

$A(1)$ = Row 1 of Table A

$A(2)$ = Row 2 of Table A

$A(i)$ = Row i of Table A

A has N rows
and C columns

$B(1)$ = Row 1 of Table B

$B(2)$ = Row 2 of Table B

$B(i)$ = Row i of Table B

B has M rows
and K columns

$A \times B =$

$A \times B$ has
 $N \times M$ rows
and
 $C+K$ columns

$A(1)B(1)$

$A(1)B(2)$

$A(1)B(3)$

...

$A(1)B(M)$

$A(2)B(1)$

$A(2)B(2)$

$A(2)B(3)$

...

$A(2)B(M)$

$A(N)B(1)$

$A(N)B(2)$

$A(N)B(3)$

...

$A(N)B(M)$

ER-diagrams

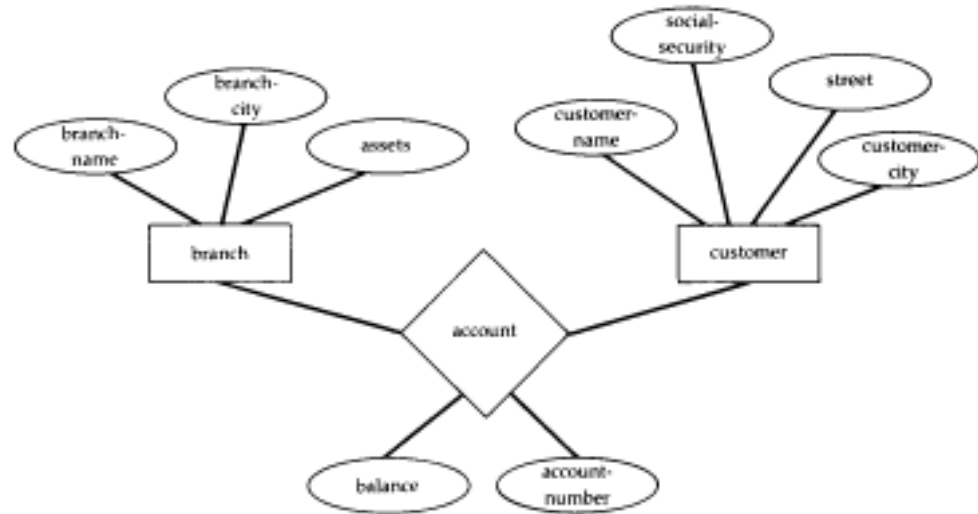


Figure 2.23 E-R diagram with *account* as a relationship set.

- Korth & Silberschatz
 - ◇ branch $\Leftrightarrow$ matches (gid-start +++ did)
 - ◇ customer $\Leftrightarrow$ folds (fid +++)
 - ◇ linked by
 - account $\Leftrightarrow$ structures (did fid)

Aggregate Functions-- Statistics on Attributes

- Query Statistics
 - ◇ select gid, count (distinct did) from matches
 - ◇ select max(N_hlx) from folds where N_beta = 0
- How many matches to globins in the E. coli genome
- Complex Query by nesting selections
 - ◇ E <= select fid from folds where name contains "globin"
 - ◇ D <= select did from structures where fid in E
 - ◇ N <= select count(distinct gid, TrgStrt) from matches where did in D and score < .01

Joins

gid_	TrgStrt	TrgStop	did	score
HI0299	119	135	d193l__	3.1
HI0572	180	240	d1aba__	0.0032
HI0989	56	125	d1aco_1	0.0049
HI0988	106	458	d1aco_2	4.4e-14
HI0154	2	76	d1acp__	1.2e-23
HI1633	2	432	d1adea__	0
HI0349	1	183	d1aky__	7.6e-36
HI1309	35	52	d1alo_3	1.1
HI0589	8	25	d1alo_3	1.8
HI1358	239	444	d1amg_2	0.002
HI1358	218	410	d1amy_2	0.00037
HI0460	20	24	d1ans__	1.8
HI1386	139	147	d1ans__	3.3
HI0421	11	14	d1ans__	6.4
HI0361	285	295	d1ans__	8.2
HI0835	100	106	d1ans__	9.7

did_	fid
d2rs51_	1.002.007
d1imr__	1.010.002
d1pyib1	1.007.030
d1dxt_d	1.001.001
d181l__	1.004.002
d1vmoa_	1.002.044
d2gsq_1	1.001.031
d1etb2_	1.002.003
d1guha1	1.001.031
d1hrc__	1.001.003
d150lc_	1.004.002
d1dmf__	1.007.035
d1l19__	1.004.002
d1yrnc_	1.010.002
d1ans__	1.007.008
d2rmai_	1.002.036

fid_	bestrep	N_hlx	N_beta	name
1.001.001	d1flp__	8	0	Globin-like
1.001.002	d1hdj__	4	0	Long alpha-hairpin
1.001.003	d1ctj__	9	0	Cytochrome c
1.001.004	d1enh__	2	0	DNA-binding 3-helical bundle
1.001.005	d1dtr_2	1	3	Diphtheria toxin repressor (DtxR) dimeriz
1.001.006	d1tns__	1	2	Mu transposase, DNA-binding domain
1.001.007	d2spca_	0	2	Spectrin repeat unit
1.001.008	d1bdd__	0	4	Immunoglobulin-binding protein A modules
1.007.008	d1qkt__	4	3	Neurotoxin III (ATX III)
1.001.010	d2erl__	3	5	Protozoan pheromone proteins

Join Gives Unnormalized Table

Joining Two or More Tables with a Select Query Gives a New, “Bigger” Table

gid_	TrgStrt	TrgStop	did	score	fid	N_hlx	N_beta	name
HI0299	119	135	d193l__	3.1	1.010.002	0	2	Spectrin repeat unit
HI0572	180	240	d1aba__	0.0032	1.002.045	1	2	Mu transposase, DNA-binding domain
HI0989	56	125	d1aco_1	0.0049	1.001.031	8	0	Globin-like
HI0988	106	458	d1aco_2	4.4e-14	1.001.031	8	0	Globin-like
HI0154	2	76	d1acp__	1.2e-23	1.001.031	8	0	Globin-like
HI1633	2	432	d1adea__	0	1.010.002	0	2	Spectrin repeat unit
HI0349	1	183	d1aky__	7.6e-36	1.001.031	8	0	Globin-like
HI1309	35	52	d1alo_3	1.1	1.007.008	4	3	Neurotoxin III (ATX III)
HI0589	8	25	d1alo_3	1.8	1.002.045	1	2	Mu transposase, DNA-binding domain
HI1358	239	444	d1amg_2	0.002	1.004.002	1	3	Diphtheria toxin repressor (DtxR)
HI1358	218	410	d1amy_2	0.00037	1.002.044	0	4	Immunoglobulin-binding protein A
HI0460	20	24	d1ans__	1.8	1.007.008	4	3	Neurotoxin III (ATX III)
HI1386	139	147	d1ans__	3.3	1.007.008	4	3	Neurotoxin III (ATX III)
HI0421	11	14	d1ans__	6.4	1.007.008	4	3	Neurotoxin III (ATX III)
HI0361	285	295	d1ans__	8.2	1.007.008	4	3	Neurotoxin III (ATX III)
HI0835	100	106	d1ans__	9.7	1.007.008	4	3	Neurotoxin III (ATX III)

Normalization

- What if Want to update Fold 1.007.008 to be “Neurotoxin IV”?
 ◇ Many Updates
- So Good if Previously Normalized into Separate Tables
 ◇ Eliminate Redundancy
 ◇ Allow Consistent Updating

gid_	TrgStrt	TrgStop	did	score	fid	N_hlx	N_beta	name
HI0299	119	135	d193l__	3.1	1.010.002	0	2	Spectrin repeat unit
HI0572	180	240	d1aba__	0.0032	1.002.045	1	2	Mu transposase, DNA-binding domain
HI0989	56	125	d1aco_1	0.0049	1.001.031	8	0	Globin-like
HI0988	106	458	d1aco_2	4.4e-14	1.001.031	8	0	Globin-like
HI0154	2	76	d1acp__	1.2e-23	1.001.031	8	0	Globin-like
HI1633	2	432	d1adea__	0	1.010.002	0	2	Spectrin repeat unit
HI0349	1	183	d1aky__	7.6e-36	1.001.031	8	0	Globin-like
HI1309	35	52	d1alo_3	1.1	1.007.008	4	3	Neurotoxin III (ATX III)
HI0589	8	25	d1alo_3	1.8	1.002.045	1	2	Mu transposase, DNA-binding domain
HI1358	239	444	d1amg_2	0.002	1.004.002	1	3	Diphtheria toxin repressor (DtxR)
HI1358	218	410	d1amy_2	0.00037	1.002.044	0	4	Immunoglobulin-binding protein A
HI0460	20	24	d1ans__	1.8	1.007.008	4	3	Neurotoxin III (ATX III)
HI1386	139	147	d1ans__	3.3	1.007.008	4	3	Neurotoxin III (ATX III)
HI0421	11	14	d1ans__	6.4	1.007.008	4	3	Neurotoxin III (ATX III)
HI0361	285	295	d1ans__	8.2	1.007.008	4	3	Neurotoxin III (ATX III)
HI0835	100	106	d1ans__	9.7	1.007.008	4	3	Neurotoxin III (ATX III)

Normalization Example

Un-normalized



Normalized

Name	City	Area-Code	Phone-Number
Charles	NY	212	345-6789
Mark	SF	415	236-8982
Jane	NY	212	567-2345
Jeff	SF	415	435-3535
Jack	Boston	617	234-9988

Name	City	Phone-Number
Charles	NY	345-6789
Mark	SF	236-8982
Jane	NY	567-2345
Jeff	SF	435-3535
Jack	Boston	234-9988

City	Area-Code
NY	212
SF	415
Boston	617

Normalized Tables

Theory of Normalization

gid_	TrgStrt	TrgStop	did	score
HI0299	119	135	d193l__	3.1
HI0572	180	240	d1aba__	0.0032
HI0989	56	125	d1aco_1	0.0049
HI0988	106	458	d1aco_2	4.4e-14
HI0154	2	76	d1acp__	1.2e-23
HI1633	2	432	d1adea__	0
HI0349	1	183	d1aky__	7.6e-36
HI1309	35	52	d1alo_3	1.1
HI0589	8	25	d1alo_3	1.8
HI1358	239	444	d1amg_2	0.002
HI1358	218	410	d1amy_2	0.00037
HI0460	20	24	d1ans__	1.8
HI1386	139	147	d1ans__	3.3
HI0421	11	14	d1ans__	6.4
HI0361	285	295	d1ans__	8.2
HI0835	100	106	d1ans__	9.7

did_	fid
d2rs51_	1.002.007
d1imr__	1.010.002
d1pyib1	1.007.030
d1dxt_d_	1.001.001
d181l__	1.004.002
d1vmoa_	1.002.044
d2gsq_1	1.001.031
d1etb2_	1.002.003
d1guha1	1.001.031
d1hrc__	1.001.003
d150lc_	1.004.002
d1dmf__	1.007.035
d1l19__	1.004.002
d1yrnc_	1.010.002
d1ans__	1.007.008
d2rmai_	1.002.036

fid_	bestrep	N_hlx	N_beta	name
1.001.001	d1flp__	8	0	Globin-like
1.001.002	d1hdj__	4	0	Long alpha-hairpin
1.001.003	d1ctj__	9	0	Cytochrome c
1.001.004	d1enh__	2	0	DNA-binding 3-helical bundle
1.001.005	d1dtr_2	1	3	Diphtheria toxin repressor (DtxR) dimeriz
1.001.006	d1tns__	1	2	Mu transposase, DNA-binding domain
1.001.007	d2spca_	0	2	Spectrin repeat unit
1.001.008	d1bdd__	0	4	Immunoglobulin-binding protein A modules
1.007.008	d1qkt__	4	3	Neurotoxin III (ATX III)
1.001.010	d2erl__	3	5	Protozoan pheromone proteins

Query Optimization

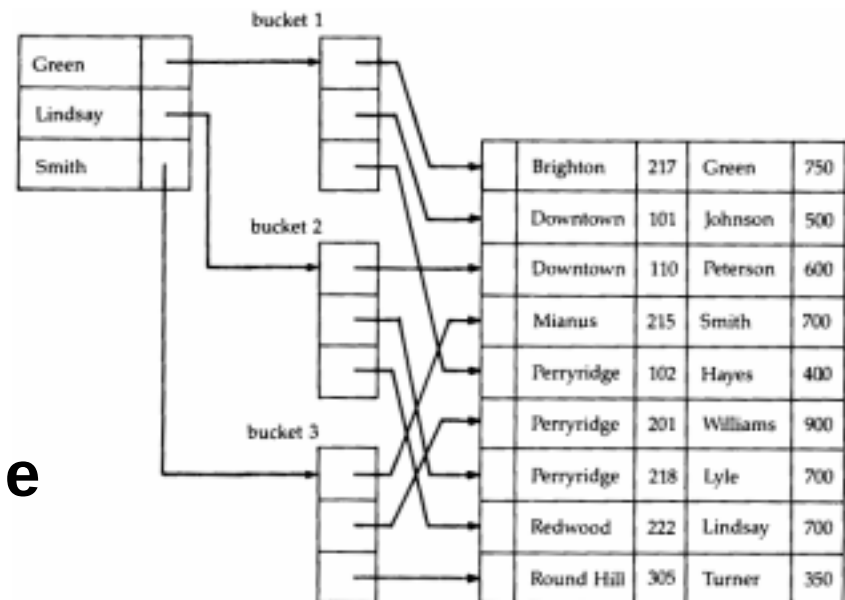
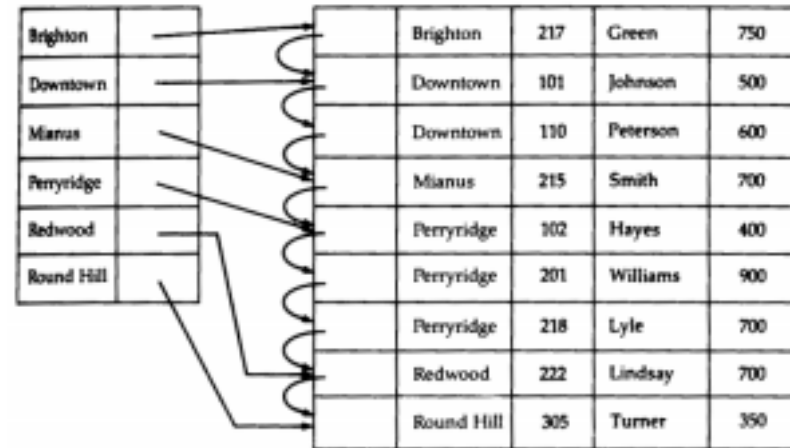
- Get at the Data Quickly!!
- Indexes
- Hash Function Reproduce the Effect of Indexes
 - ◊ Rapidly Associate a Bucket with Each Key
- Joining 10 tables, which to do first?
 - ◊ Joining is slow so store some tables in unnormalized form
 - o Speed vs Memory

Indexes Speed Access

	Brighton	217	Green	750
	Downtown	101	Johnson	500
	Downtown	110	Peterson	600
	Mianus	215	Smith	700
	Perryridge	102	Hayes	400
	Perryridge	201	Williams	900
	Perryridge	218	Lyle	700
	Redwood	222	Lindsay	700
	Round Hill	305	Turner	350

**No
Index**

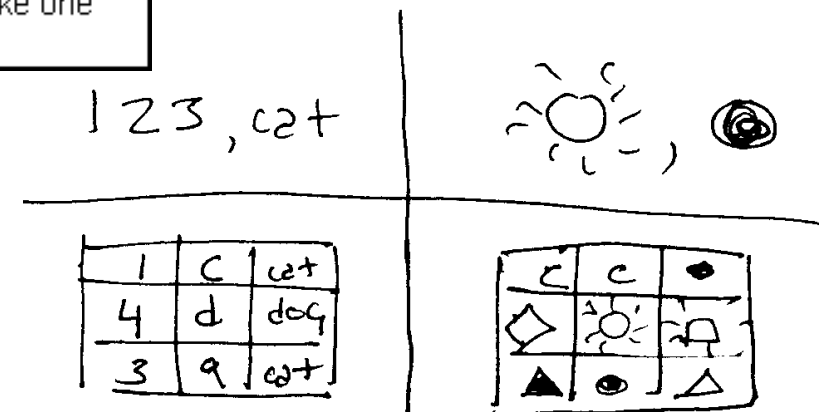
**One
Index**



**Double
Index**

		Simple Data Types	Complex Data Types
Simple Structure	Data	int: 1,2,3	struct A {pointer-list + char}
		chars: hello, text	method Am acts on A
	DB	Simple File	Object DB (OODB)
		File with unstructured text	Complex data and methods stored in a file
	Example	Your .login file	Persistent data from C++ program with an "image" datatype and method for comparing images
Relational Structure	What	Relational Database (RDB)	Object Relational DB (ORDB)
		Rows and columns contain ints and chars	Rows and columns contain complex objects and methods are defined to handle them
	Query Lang.	SQL	OQL
	Ex. Query	A query can ask for all names containing first names stored at 10PM	A query can ask for all images that look like one stored at 10PM

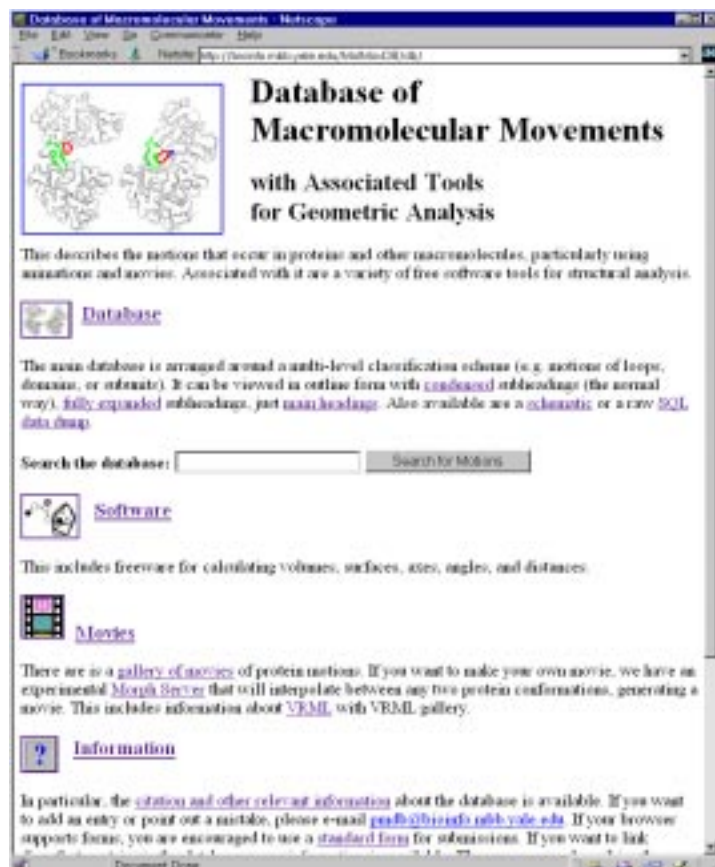
Object Databases



Forms & reports [user views]

- Reports are the result of running a succession of selects queries on a database, joining together a number of tables, and then pasting the results together
- Forms are the same but they are editable
- Forms and Reports represent particular views of the data
 - ◇ For instance, one can be keyed on gene id listing all the structures matching a gene and the other could be keyed on structure id listing all the gene matching a given structure

Example Report: Motions Database



Database of Macromolecular Movements
with Associated Tools
for Geometric Analysis

This describes the motions that occur in proteins and other macromolecules, particularly using animations and movies. Associated with it are a variety of free software tools for structural analysis.

Database

The main database is arranged around a multi-level classification scheme (e.g. motions of loops, domains, or subunits). It can be viewed in outline form with [condensed subheadings](#) (the normal way), [fully expanded subheadings](#), just [main headings](#). Also available are a [schematic](#) or a raw [SQL data dump](#).

Search the database:

Software

This includes freeware for calculating volumes, surfaces, area, angles, and distances.

Movies

There are a [gallery of movies](#) of protein motions. If you want to make your own movie, we have an experimental [Movie Server](#) that will interpolate between any two protein conformations, generating a movie. This includes information about [VRML](#), with VRML gallery.

Information

In particular, the [citation and other relevant information](#) about the database is available. If you want to add an entry or point out a mistake, please e-mail pubs@bioinfo.mbb.yale.edu. If your browser supports frames, you are encouraged to use a [standard form](#) for submissions. If you want to link:

Report on Calmodulin




Motion in Calmodulin [cm]

Excerpt from: Motions, Hinge Mechanisms (12-6-11)

Classification

- Classified in **2004**, 1st, 2nd, 3rd, 4th, 5th, 6th, 7th, 8th, 9th, 10th, 11th, 12th, 13th, 14th, 15th, 16th, 17th, 18th, 19th, 20th, 21st, 22nd, 23rd, 24th, 25th, 26th, 27th, 28th, 29th, 30th, 31st, 32nd, 33rd, 34th, 35th, 36th, 37th, 38th, 39th, 40th, 41st, 42nd, 43rd, 44th, 45th, 46th, 47th, 48th, 49th, 50th, 51st, 52nd, 53rd, 54th, 55th, 56th, 57th, 58th, 59th, 60th, 61st, 62nd, 63rd, 64th, 65th, 66th, 67th, 68th, 69th, 70th, 71st, 72nd, 73rd, 74th, 75th, 76th, 77th, 78th, 79th, 80th, 81st, 82nd, 83rd, 84th, 85th, 86th, 87th, 88th, 89th, 90th, 91st, 92nd, 93rd, 94th, 95th, 96th, 97th, 98th, 99th, 100th, 101st, 102nd, 103rd, 104th, 105th, 106th, 107th, 108th, 109th, 110th, 111th, 112th, 113th, 114th, 115th, 116th, 117th, 118th, 119th, 120th, 121st, 122nd, 123rd, 124th, 125th, 126th, 127th, 128th, 129th, 130th, 131st, 132nd, 133rd, 134th, 135th, 136th, 137th, 138th, 139th, 140th, 141st, 142nd, 143rd, 144th, 145th, 146th, 147th, 148th, 149th, 150th, 151st, 152nd, 153rd, 154th, 155th, 156th, 157th, 158th, 159th, 160th, 161st, 162nd, 163rd, 164th, 165th, 166th, 167th, 168th, 169th, 170th, 171st, 172nd, 173rd, 174th, 175th, 176th, 177th, 178th, 179th, 180th, 181st, 182nd, 183rd, 184th, 185th, 186th, 187th, 188th, 189th, 190th, 191st, 192nd, 193rd, 194th, 195th, 196th, 197th, 198th, 199th, 200th, 201st, 202nd, 203rd, 204th, 205th, 206th, 207th, 208th, 209th, 210th, 211st, 212th, 213th, 214th, 215th, 216th, 217th, 218th, 219th, 220th, 221st, 222nd, 223rd, 224th, 225th, 226th, 227th, 228th, 229th, 230th, 231st, 232nd, 233rd, 234th, 235th, 236th, 237th, 238th, 239th, 240th, 241st, 242nd, 243rd, 244th, 245th, 246th, 247th, 248th, 249th, 250th, 251st, 252nd, 253rd, 254th, 255th, 256th, 257th, 258th, 259th, 260th, 261st, 262nd, 263rd, 264th, 265th, 266th, 267th, 268th, 269th, 270th, 271st, 272nd, 273rd, 274th, 275th, 276th, 277th, 278th, 279th, 280th, 281st, 282nd, 283rd, 284th, 285th, 286th, 287th, 288th, 289th, 290th, 291st, 292nd, 293rd, 294th, 295th, 296th, 297th, 298th, 299th, 300th, 301st, 302nd, 303rd, 304th, 305th, 306th, 307th, 308th, 309th, 310th, 311st, 312th, 313th, 314th, 315th, 316th, 317th, 318th, 319th, 320th, 321st, 322nd, 323rd, 324th, 325th, 326th, 327th, 328th, 329th, 330th, 331st, 332nd, 333rd, 334th, 335th, 336th, 337th, 338th, 339th, 340th, 341st, 342nd, 343rd, 344th, 345th, 346th, 347th, 348th, 349th, 350th, 351st, 352nd, 353rd, 354th, 355th, 356th, 357th, 358th, 359th, 360th, 361st, 362nd, 363rd, 364th, 365th, 366th, 367th, 368th, 369th, 370th, 371st, 372nd, 373rd, 374th, 375th, 376th, 377th, 378th, 379th, 380th, 381st, 382nd, 383rd, 384th, 385th, 386th, 387th, 388th, 389th, 390th, 391st, 392nd, 393rd, 394th, 395th, 396th, 397th, 398th, 399th, 400th, 401st, 402nd, 403rd, 404th, 405th, 406th, 407th, 408th, 409th, 410th, 411st, 412th, 413th, 414th, 415th, 416th, 417th, 418th, 419th, 420th, 421st, 422nd, 423rd, 424th, 425th, 426th, 427th, 428th, 429th, 430th, 431st, 432nd, 433rd, 434th, 435th, 436th, 437th, 438th, 439th, 440th, 441st, 442nd, 443rd, 444th, 445th, 446th, 447th, 448th, 449th, 450th, 451st, 452nd, 453rd, 454th, 455th, 456th, 457th, 458th, 459th, 460th, 461st, 462nd, 463rd, 464th, 465th, 466th, 467th, 468th, 469th, 470th, 471st, 472nd, 473rd, 474th, 475th, 476th, 477th, 478th, 479th, 480th, 481st, 482nd, 483rd, 484th, 485th, 486th, 487th, 488th, 489th, 490th, 491st, 492nd, 493rd, 494th, 495th, 496th, 497th, 498th, 499th, 500th, 501st, 502nd, 503rd, 504th, 505th, 506th, 507th, 508th, 509th, 510th, 511st, 512th, 513th, 514th, 515th, 516th, 517th, 518th, 519th, 520th, 521st, 522nd, 523rd, 524th, 525th, 526th, 527th, 528th, 529th, 530th, 531st, 532nd, 533rd, 534th, 535th, 536th, 537th, 538th, 539th, 540th, 541st, 542nd, 543rd, 544th, 545th, 546th, 547th, 548th, 549th, 550th, 551st, 552nd, 553rd, 554th, 555th, 556th, 557th, 558th, 559th, 560th, 561st, 562nd, 563rd, 564th, 565th, 566th, 567th, 568th, 569th, 570th, 571st, 572nd, 573rd, 574th, 575th, 576th, 577th, 578th, 579th, 580th, 581st, 582nd, 583rd, 584th, 585th, 586th, 587th, 588th, 589th, 590th, 591st, 592nd, 593rd, 594th, 595th, 596th, 597th, 598th, 599th, 600th, 601st, 602nd, 603rd, 604th, 605th, 606th, 607th, 608th, 609th, 610th, 611st, 612th, 613th, 614th, 615th, 616th, 617th, 618th, 619th, 620th, 621st, 622nd, 623rd, 624th, 625th, 626th, 627th, 628th, 629th, 630th, 631st, 632nd, 633rd, 634th, 635th, 636th, 637th, 638th, 639th, 640th, 641st, 642nd, 643rd, 644th, 645th, 646th, 647th, 648th, 649th, 650th, 651st, 652nd, 653rd, 654th, 655th, 656th, 657th, 658th, 659th, 660th, 661st, 662nd, 663rd, 664th, 665th, 666th, 667th, 668th, 669th, 670th, 671st, 672nd, 673rd, 674th, 675th, 676th, 677th, 678th, 679th, 680th, 681st, 682nd, 683rd, 684th, 685th, 686th, 687th, 688th, 689th, 690th, 691st, 692nd, 693rd, 694th, 695th, 696th, 697th, 698th, 699th, 700th, 701st, 702nd, 703rd, 704th, 705th, 706th, 707th, 708th, 709th, 710th, 711st, 712th, 713th, 714th, 715th, 716th, 717th, 718th, 719th, 720th, 721st, 722nd, 723rd, 724th, 725th, 726th, 727th, 728th, 729th, 730th, 731st, 732nd, 733rd, 734th, 735th, 736th, 737th, 738th, 739th, 740th, 741st, 742nd, 743rd, 744th, 745th, 746th, 747th, 748th, 749th, 750th, 751st, 752nd, 753rd, 754th, 755th, 756th, 757th, 758th, 759th, 760th, 761st, 762nd, 763rd, 764th, 765th, 766th, 767th, 768th, 769th, 770th, 771st, 772nd, 773rd, 774th, 775th, 776th, 777th, 778th, 779th, 780th, 781st, 782nd, 783rd, 784th, 785th, 786th, 787th, 788th, 789th, 790th, 791st, 792nd, 793rd, 794th, 795th, 796th, 797th, 798th, 799th, 800th, 801st, 802nd, 803rd, 804th, 805th, 806th, 807th, 808th, 809th, 810th, 811st, 812th, 813th, 814th, 815th, 816th, 817th, 818th, 819th, 820th, 821st, 822nd, 823rd, 824th, 825th, 826th, 827th, 828th, 829th, 830th, 831st, 832nd, 833rd, 834th, 835th, 836th, 837th, 838th, 839th, 840th, 841st, 842nd, 843rd, 844th, 845th, 846th, 847th, 848th, 849th, 850th, 851st, 852nd, 853rd, 854th, 855th, 856th, 857th, 858th, 859th, 860th, 861st, 862nd, 863rd, 864th, 865th, 866th, 867th, 868th, 869th, 870th, 871st, 872nd, 873rd, 874th, 875th, 876th, 877th, 878th, 879th, 880th, 881st, 882nd, 883rd, 884th, 885th, 886th, 887th, 888th, 889th, 890th, 891st, 892nd, 893rd, 894th, 895th, 896th, 897th, 898th, 899th, 900th, 901st, 902nd, 903rd, 904th, 905th, 906th, 907th, 908th, 909th, 910th, 911st, 912th, 913th, 914th, 915th, 916th, 917th, 918th, 919th, 920th, 921st, 922nd, 923rd, 924th, 925th, 926th, 927th, 928th, 929th, 930th, 931st, 932nd, 933rd, 934th, 935th, 936th, 937th, 938th, 939th, 940th, 941st, 942nd, 943rd, 944th, 945th, 946th, 947th, 948th, 949th, 950th, 951st, 952nd, 953rd, 954th, 955th, 956th, 957th, 958th, 959th, 960th, 961st, 962nd, 963rd, 964th, 965th, 966th, 967th, 968th, 969th, 970th, 971st, 972nd, 973rd, 974th, 975th, 976th, 977th, 978th, 979th, 980th, 981st, 982nd, 983rd, 984th, 985th, 986th, 987th, 988th, 989th, 990th, 991st, 992nd, 993rd, 994th, 995th, 996th, 997th, 998th, 999th, 1000th, 1001st, 1002nd, 1003rd, 1004th, 1005th, 1006th, 1007th, 1008th, 1009th, 1010th, 1011st, 1012th, 1013th, 1014th, 1015th, 1016th, 1017th, 1018th, 1019th, 1020th, 1021st, 1022nd, 1023rd, 1024th, 1025th, 1026th, 1027th, 1028th, 1029th, 1030th, 1031st, 1032nd, 1033rd, 1034th, 1035th, 1036th, 1037th, 1038th, 1039th, 1040th, 1041st, 1042nd, 1043rd, 1044th, 1045th, 1046th, 1047th, 1048th, 1049th, 1050th, 1051st, 1052nd, 1053rd, 1054th, 1055th, 1056th, 1057th, 1058th, 1059th, 1060th, 1061st, 1062nd, 1063rd, 1064th, 1065th, 1066th, 1067th, 1068th, 1069th, 1070th, 1071st, 1072nd, 1073rd, 1074th, 1075th, 1076th, 1077th, 1078th, 1079th, 1080th, 1081st, 1082nd, 1083rd, 1084th, 1085th, 1086th, 1087th, 1088th, 1089th, 1090th, 1091st, 1092nd, 1093rd, 1094th, 1095th, 1096th, 1097th, 1098th, 1099th, 1100th, 1101st, 1102nd, 1103rd, 1104th, 1105th, 1106th, 1107th, 1108th, 1109th, 1110th, 1111st, 1112th, 1113th, 1114th, 1115th, 1116th, 1117th, 1118th, 1119th, 1120th, 1121st, 1122nd, 1123rd, 1124th, 1125th, 1126th, 1127th, 1128th, 1129th, 1130th, 1131st, 1132nd, 1133rd, 1134th, 1135th, 1136th, 1137th, 1138th, 1139th, 1140th, 1141st, 1142nd, 1143rd, 1144th, 1145th, 1146th, 1147th, 1148th, 1149th, 1150th, 1151st, 1152nd, 1153rd, 1154th, 1155th, 1156th, 1157th, 1158th, 1159th, 1160th, 1161st, 1162nd, 1163rd, 1164th, 1165th, 1166th, 1167th, 1168th, 1169th, 1170th, 1171st, 1172nd, 1173rd, 1174th, 1175th, 1176th, 1177th, 1178th, 1179th, 1180th, 1181st, 1182nd, 1183rd, 1184th, 1185th, 1186th, 1187th, 1188th, 1189th, 1190th, 1191st, 1192nd, 1193rd, 1194th, 1195th, 1196th, 1197th, 1198th, 1199th, 1200th, 1201st, 1202nd, 1203rd, 1204th, 1205th, 1206th, 1207th, 1208th, 1209th, 1210th, 1211st, 1212nd, 1213rd, 1214th, 1215th, 1216th, 1217th, 1218th, 1219th, 1220th, 1221st, 1222nd, 1223rd, 1224th, 1225th, 1226th, 1227th, 1228th, 1229th, 1230th, 1231st, 1232nd, 1233rd, 1234th, 1235th, 1236th, 1237th, 1238th, 1239th, 1240th, 1241st, 1242nd, 1243rd, 1244th, 1245th, 1246th, 1247th, 1248th, 1249th, 1250th, 1251st, 1252nd, 1253rd, 1254th, 1255th, 1256th, 1257th, 1258th, 1259th, 1260th, 1261st, 1262nd, 1263rd, 1264th, 1265th, 1266th, 1267th, 1268th, 1269th, 1270th, 1271st, 1272nd, 1273rd, 1274th, 1275th, 1276th, 1277th, 1278th, 1279th, 1280th, 1281st, 1282nd, 1283rd, 1284th, 1285th, 1286th, 1287th, 1288th, 1289th, 1290th, 1291st, 1292nd, 1293rd, 1294th, 1295th, 1296th, 1297th, 1298th, 1299th, 1300th, 1301st, 1302nd, 1303rd, 1304th, 1305th, 1306th, 1307th, 1308th, 1309th, 1310th, 1311st, 1312nd, 1313rd, 1314th, 1315th, 1316th, 1317th, 1318th, 1319th, 1320th, 1321st, 1322nd, 1323rd, 1324th, 1325th, 1326th, 1327th, 1328th, 1329th, 1330th, 1331st, 1332nd, 1333rd, 1334th, 1335th, 1336th, 1337th, 1338th, 1339th, 1340th, 1341st, 1342nd, 1343rd, 1344th, 1345th, 1346th, 1347th, 1348th, 1349th, 1350th, 1351st, 1352nd, 1353rd, 1354th, 1355th, 1356th, 1357th, 1358th, 1359th, 1360th, 1361st, 1362nd, 1363rd, 1364th, 1365th, 1366th, 1367th, 1368th, 1369th, 1370th, 1371st, 1372nd, 1373rd, 1374th, 1375th, 1376th, 1377th, 1378th, 1379th, 1380th, 1381st, 1382nd, 1383rd, 1384th, 1385th, 1386th, 1387th, 1388th, 1389th, 1390th, 1391st, 1392nd, 1393rd, 1394th, 1395th, 1396th, 1397th, 1398th, 1399th, 1400th, 1401st, 1402nd, 1403rd, 1404th, 1405th, 1406th, 1407th, 1408th, 1409th, 1410th, 1411st, 1412nd, 1413rd, 1414th, 1415th, 1416th, 1417th, 1418th, 1419th, 1420th, 1421st, 1422nd, 1423rd, 1424th, 1425th, 1426th, 1427th, 1428th, 1429th, 1430th, 1431st, 1432nd, 1433rd, 1434th, 1435th, 1436th, 1437th, 1438th, 1439th, 1440th, 1441st, 1442nd, 1443rd, 1444th, 1445th, 1446th, 1447th, 1448th, 1449th, 1450th, 1451st, 1452nd, 1453rd, 1454th, 1455th, 1456th, 1457th, 1458th, 1459th, 1460th, 1461st, 1462nd, 1463rd, 1464th, 1465th, 1466th, 1467th, 1468th, 1469th, 1470th, 1471st, 1472nd, 1473rd, 1474th, 1475th, 1476th, 1477th, 1478th, 1479th, 1480th, 1481st, 1482nd, 1483rd, 1484th, 1485th, 1486th, 1487th, 1488th, 1489th, 1490th, 1491st, 1492nd, 1493rd, 1494th, 1495th, 1496th, 1497th, 1498th, 1499th, 1500th, 1501st, 1502nd, 1503rd, 1504th, 1505th, 1506th, 1507th, 1508th, 1509th, 1510th, 1511st, 1512nd, 1513rd, 1514th, 1515th, 1516th, 1517th, 1518th, 1519th, 1520th, 1521st, 1522nd, 1523rd, 1524th, 1525th, 1526th, 1527th, 1528th, 1529th, 1530th, 1531st, 1532nd, 1533rd, 1534th, 1535th, 1536th, 1537th, 1538th, 1539th, 1540th, 1541st, 1542nd, 1543rd, 1544th, 1545th, 1546th, 1547th, 1548th, 1549th, 1550th, 1551st, 1552nd, 1553rd, 1554th, 1555th, 1556th, 1557th, 1558th, 1559th, 1560th, 1561st, 1562nd, 1563rd, 1564th, 1565th, 1566th, 1567th, 1568th, 1569th, 1570th, 1571st, 1572nd, 1573rd, 1574th, 1575th, 1576th, 1577th, 1578th, 1579th, 1580th, 1581st, 1582nd, 1583rd, 1584th, 1585th, 1586th, 1587th, 1588th, 1589th, 1590th, 1591st, 1592nd, 1593rd, 1594th, 1595th, 1596th, 1597th, 1598th, 1599th, 1600th, 1601st, 1602nd, 1603rd, 1604th, 1605th, 1606th, 1607th, 1608th, 1609th, 1610th, 1611st, 1612nd, 1613rd, 1614th, 1615th, 1616th, 1617th, 1618th, 1619th, 1620th, 1621st, 1622nd, 1623rd, 1624th, 1625th, 1626th, 1627th, 1628th, 1629th, 1630th, 1631st, 1632nd, 1633rd, 1634th, 1635th, 1636th, 1637th, 1638th, 1639th, 1640th, 1641st, 1642nd, 1643rd, 1644th, 1645th, 1646th, 1647th, 1648th, 1649th, 1650th, 1651st, 1652nd, 1653rd, 1654th, 1655th, 1656th, 1657th, 1658th, 1659th, 1660th, 1661st, 1662nd, 1663rd, 1664th, 1665th, 1666th, 1667th, 1668th, 1669th, 1670th, 1671st, 1672nd, 1673rd, 1674th, 1675th, 1676th, 1677th, 1678th, 1679th, 1680th, 1681st, 1682nd, 1683rd, 1684th, 1685th, 1686th, 1687th, 1688th, 1689th, 1690th, 1691st, 1692nd, 1693rd, 1694th, 1695th, 1696th, 1697th, 1698th, 1699th, 1700th, 1701st, 1702nd, 1703rd, 1704th, 1705th, 1706th, 1707th, 1708th, 1709th, 1710th, 1711st, 1712nd, 1713rd, 1714th, 1715th, 1716th, 1717th, 1718th, 1719th, 1720th, 1721st, 1722nd, 1723rd, 1724th, 1725th, 1726th, 1727th, 1728th, 1729th, 1730th, 1731st, 1732nd, 1733rd, 1734th, 1735th, 1736th, 1737th, 1738th, 1739th, 1740th, 1741st, 1742nd, 1743rd, 1744th, 1745th, 1746th, 1747th, 1748th, 1749th, 1750th, 1751st, 1752nd, 1753rd, 1754th, 1755th, 1756th, 1757th, 1758th, 1759th, 1760th, 1761st, 1762nd, 1763rd, 1764th, 1765th, 1766th, 1767th, 1768th, 1769th, 1770th, 1771st, 1772nd, 1773rd, 1774th, 1775th, 1776th, 1777th, 1778th, 1779th, 1780th, 1781st, 1782nd, 1783rd, 1784th, 1785th, 1786th, 1787th, 1788th, 1789th, 1790th, 1791st, 1792nd, 1793rd, 1794th, 1795th, 1796th, 1797th, 1798th, 1799th, 1800th, 1801st, 1802nd, 1803rd, 1

Example Report: Motions Database

Report shows information, merging together many tables with variable amounts of information. Form same but allows entry.



Schema

```
CREATE TABLE relations (
  id_ CHAR(15),
  id_to_ CHAR(15),
  type CHAR(30),
  comment CHAR(512)
)

CREATE TABLE single_vals (
  id_ CHAR(10),
  name_ CHAR(30),
  val CHAR(30),
  comment CHAR(500)
)

CREATE TABLE structures (
  id_ CHAR(10),
  pdb_id_ CHAR(8),
  name_short CHAR(50),
  chain CHAR(1),
  name_long CHAR(100)
)

CREATE TABLE value_names (
  abbrev_ CHAR(15),
  name CHAR(50)
)

CREATE TABLE endnote_refs (
  num_I INT,
  name CHAR(512)
)
```

```
CREATE TABLE classes (
  class_num_ CHAR(10),
  new CHAR(10),
  class_name CHAR(80)
)

CREATE TABLE classifications (
  id_ CHAR(10),
  class_num CHAR(10)
)

CREATE TABLE links (
  id_ CHAR(10),
  url_ CHAR(150),
  hilit_text CHAR(100),
  other_text CHAR(500),
  flag CHAR(5)
)

CREATE TABLE names (
  id_ CHAR(10),
  seq_num_n INT,
  name CHAR(255)
)

CREATE TABLE refs (
  id_ CHAR(10),
  medline_I INT,
  endnote_I INT,
  flag_n INT
)

CREATE TABLE descriptions (
  id_ CHAR(10),
  num_I INT,
  prose CHAR(5000)
)
```

Example Report: Motions Database

Motion in Calmodulin [cm]

Classification

Known Domain Motion, Hinge Mechanism [D-h-2]

Structures

- Closed is **2BBM**; fly, NMR, closed with peptide
(Links to [PDB](#), [Entrez](#), [SCOP](#), [Core-Structures](#), [VRML-lines](#), and [VRML-tubes](#)).
- Closed is **1CTR**
(Links to [PDB](#), [Entrez](#), [SCOP](#), [Core-Structures](#), [VRML-lines](#), and [VRML-tubes](#)).
- Closed is **1CDL**; mammalian, recomb., X-ray
(Links to [PDB](#), [Entrez](#), [SCOP](#), [Core-Structures](#), [VRML-lines](#), and [VRML-tubes](#)).
- Closed (conf. 3) is **2BBN**; fly, NMR, closed with 2nd peptide
(Links to [PDB](#), [Entrez](#), [SCOP](#), [Core-Structures](#), [VRML-lines](#), and [VRML-tubes](#)).
- Open is **1CLL**; human, X-ray, refined
(Links to [PDB](#), [Entrez](#), [SCOP](#), [Core-Structures](#), [VRML-lines](#), and [VRML-tubes](#)).
- Open is **4CLN**; fly, X-ray
(Links to [PDB](#), [Entrez](#), [SCOP](#), [Core-Structures](#), [VRML-lines](#), and [VRML-tubes](#)).

Description

- Basically, this hinge motion involves long helix splitting into 2 helices (inclined at ~100 degrees) with strand in between.
- The unligated form of calmodulin contains two globular domains, connected by a long helix. NMR and X-ray structures of ligated calmodulin show the molecule binding to peptide helices with different sequences and the two domains closing around the peptide far enough to make contact with each other. In this motion, the long interdomain helix, which is known to have only marginal stability in solution, partly unfolds to break into two helical segments connected by a 4-residue hinge region in an extended conformation. The angle between the axes of the two helical segments is ~100 degrees. As there is an additional twist around the helix axes, the total rotation of one domain relative to the other is upwards of 150 degrees. Calmodulin can bind peptides with different sequences because of flexibility in the side



Structures: Variable Number Per ID (Var. Num. of Phone Num. per Person), Foreign Key into PDB

Example Report: Motions Database

Particular values describing motion

- Annotation Level (1..10) = 7
- Domain 1 (residue selection) = 2 - 80
- Domain 2 (residue selection) = 81 - 147
- Location of a Hinge (residue selection) = 72 - 82 (4cln v. 2bbm)
- Maximum CA displacement (Å) = 60 (After sieve-fitting on domain-1)
- Maximum Rotation (degrees) = 148.02
- Number of Inter-domain connections = 1
- Number of Significant Torsion Angle Changes = 18 (Greater than 20 degrees)
- Number of hinges = 1



```
$sth = $dbh->query("SELECT value_names.name,
                    single_vals.val, single_vals.comment ".
                    "FROM value_names, single_vals ".
                    "WHERE single_vals.id_ = '$id' AND
                    single_vals.name_ = value_names.abbrev_ ".
                    "ORDER BY value_names.name");

$rows = $sth->numrows;

if ($rows > 0) {
    &PrintHead("Particular values describing motion");
    for ($i=0; $i<$rows; $i++) {
        @values = $sth->fetchrow;
        PrintSingleVals(@values);
    }
}
```

Single Values: Joining Two Tables and Iterating in Perl

Example Report: Motions Database

References

- W E Meador, A R Means and F A Quirocho (1992). Target enzyme recognition by Calmodulin: 2.4 structure of a Calmodulin-Peptide Complex. Science. 257: 1251-1255.
(Medline-UI 92390716: Report or Entrez export)
- M Ikura, G M Clore, A M Gronenborn, G Zhu, C B Klee and A Bax (1992). Solution structure of a Calmodulin-Target peptide complex by multidimensional NMR. Science. 256: 632-644.
(Medline-UI 92263094: Report or Entrez export)
- W E Meador, A R Means and F A Quirocho (1993). Modulation of calmodulin plasticity in molecular recognition on the basis of x-ray structures. Science. 262: 1718-1721.
(Medline-UI 94082290: Report or Entrez export)
- L Stryer (1995). Biochemistry. New York, W H Freeman and Company.

NAMES

id_	seq_num_n	name
aat	7	Aspartate Amino Transferase (AAT)
acetyl	1005	Acetylcholinesterase
br	97	Bacteriorhodopsin (br)
cm	23	Calmodulin

REFS

id_	medline_I	endnote_I
acetyl	0	1007
br	90294303	893
br	93154310	313
cm	92263094	648
cm	92390716	647
cm	94082290	673

ENDNOTE_REFS

num_I	name
313	S Subramaniam, M Gerstein, D Oesterhelt and R H Hender
893	R Henderson, J M Baldwin, T A Ceska, F Zemlin, E Beckm
1007	M K Gilson, T P Straatsma, JA A McCammon, D R Ripoll,
647	W E Meador, A R Means and F A Quirocho (1992). Target e
648	M Ikura, G M Clore, A M Gronenborn, G Zhu, C B Klee an
649	B-H Oh, J Pandit, C-H Kang, K Nikaido, S Gokcen, G F-L



References:

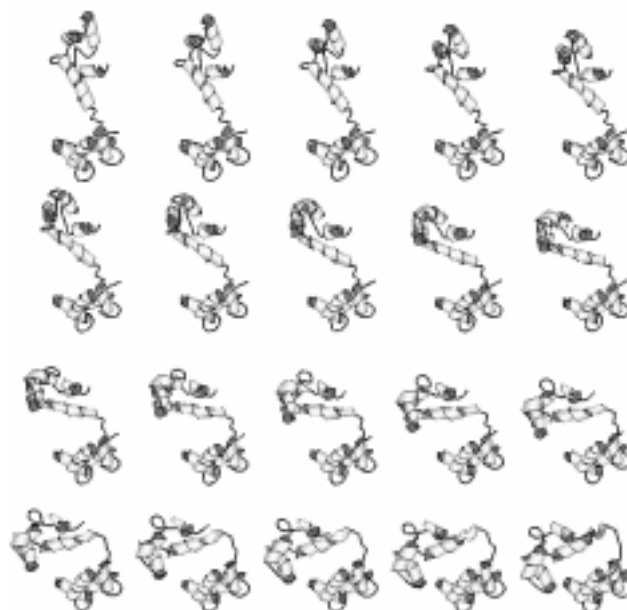
Join Two Lists (Protein Names and References) with a Table Containing Key for each List (a Relation: protein has reference.)

```
SELECT endnote_refs.name, refs.medline_I
FROM endnote_refs,refs WHERE refs.id_ =
'cm' AND refs.endnote_I =
endnote_refs.num_I
```

Example Report: Motions Database

Data and Graphics

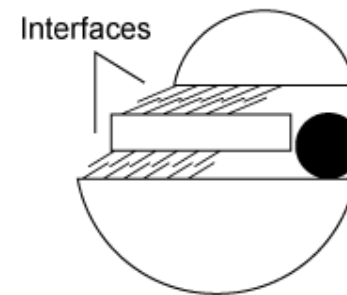
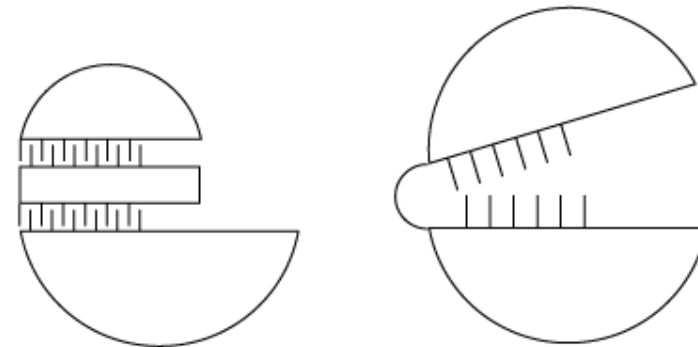
- o 4x4 Transformation Matrix. The following 4x4 matrix [1 .. 16] orients the opened form so that the axis of rotation is along the z-axis and the origin is at the molecular centroid.
- o Closed Form. Adapted from Biochemistry, Copyright 1995, Lubert Stryer.
- o MOVIES. A page giving pointers to movies of the motion.
- o Open Form. Adapted from Biochemistry, Copyright 1995, Lubert Stryer.
- o The closing of calmodulin in 3 steps. From another viewpoint, showing open, closed, and partially open.
- o Torsion Changes + Atom Deviations. Columns are, respectively: residue, phi-C, psi-C, rotamer-C, dphi, dpsi, dCA (after doing a fit) (O=open-form, C=closed-form).



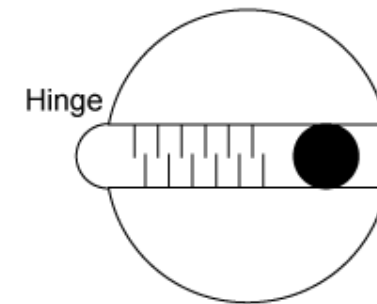
**Graphics:
How to Store
Complex Data?
(File Pointers,
BLOBS, OODB)**

Packing Based Classification

	Number Known Forms	Size of Motion	Mechanism of Motion	Examples	#
Motion	2 forms	Fragment	Hinge	TIM, LDH, TGL	14
			Shear	Insulin	3
			Unclassifiable	MS2 Coat	3
		Domain	Hinge	LF, ADK, CM	16
			Shear	CS, TrpR, AAT	8
			Refold	Serpin, RT	3
			Special	Ig elbow	1
			Unclassifiable	TBP, EF-tu	3
	1 form	Subunit	Allosteric	PFK, Hb, GP	4
			Non-allosteric	Ig VL-VH	2
		Domain	Unclassifiable		
			Hinge		
			Shear		
	1 form	Fragment	Unclassifiable	bR	1
			Refold		
			Hinge	LF~TF, SBP	10
		Domain	Shear	HK~PGK, HSP	4
			Special		
			Unclassifiable	Myosin	4
	1 form	Subunit	Allosteric		
			Non-allosteric		
			Unclassifiable	PCNA, GroEL	3



Shear Motion

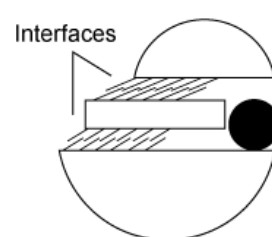
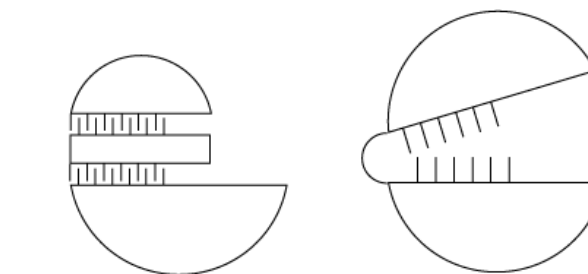
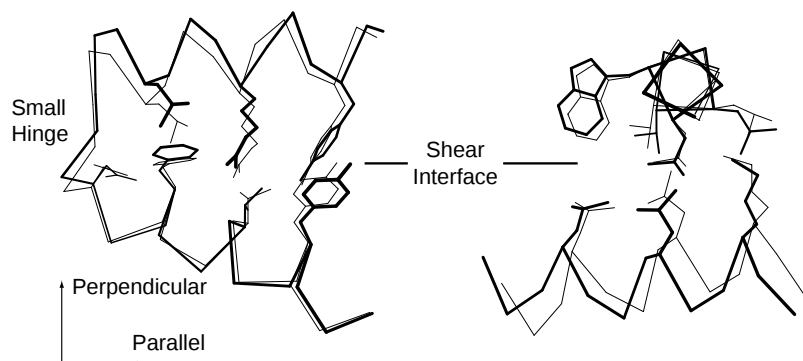
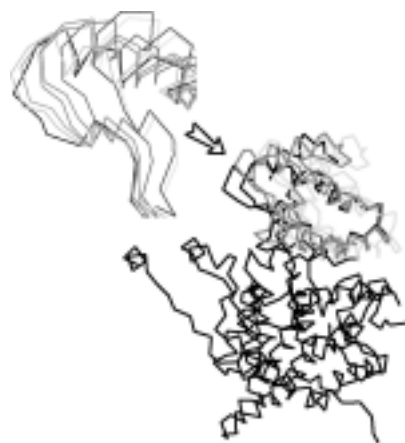


Hinge Motion

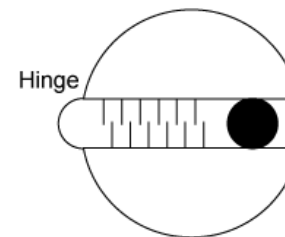
Packing Based Classification: Hinge v Shear

Shear

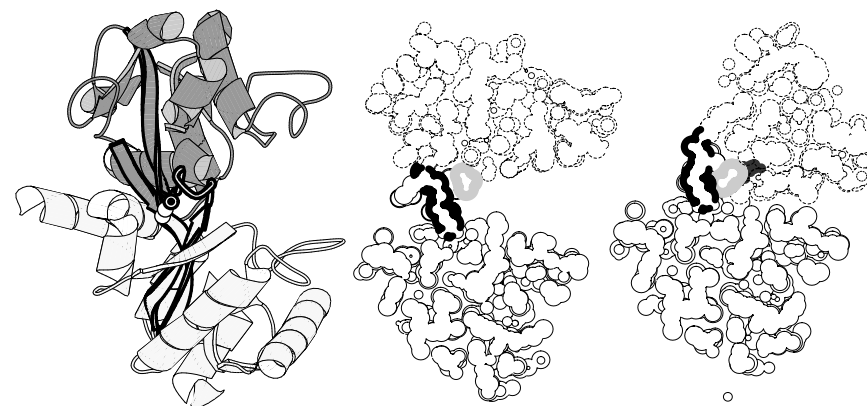
Mechanism
Involves Many
Small Motions
across a
Continuously
Maintained
Interface



Shear Motion



Hinge Motion



Hinge Mechanism involves absence of steric constraints, esp. at hinge

Simple Forms Example: Links Page

Add your own link:

<http://bioinfo.mbb.yale.edu/ius/?Mlval=links&page=course>

Aspects of Forms: Transactions and Security

- Transactions
 - ◇ Genome Centers and United Airlines!
 - ◇ Log each entry and enable **UNDO**
- Security
 - ◇ Only certain users can modify certain fields

Large-scale Example: Census DB

- 9 Genome Comparison
- 1437 Relational Tables
- 442 Mb
- Simple ASCII Layout

Molecular Biology Information: Whole Genomes

- The Revolution Driving Everything

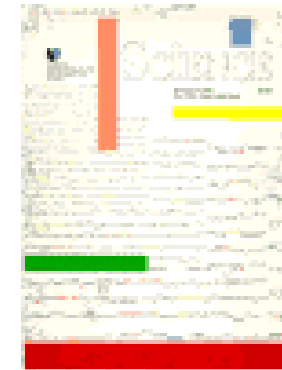
Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M., McKenney, K., Sutton, G., Fitzhugh, W., Fields, C., Gocayne, J. D., Scott, J., Shirley, R., Liu, L. I., Glodek, A., Kelley, J. M., Weidman, J. F., Phillips, C. A., Spriggs, T., Hedblom, E., Cotton, M. D., Utterback, T. R., Hanna, M. C., Nguyen, D. T., Saudek, D. M., Brandon, R. C., Fine, L. D., Fritchman, J. L., Fuhrmann, J. L., Geoghagen, N. S. M., Gnehm, C. L., McDonald, L. A., Small,

K. V., Fraser, C. M., Smith, H. O. & **Venter**, J. C. (1995). "Whole-genome

random sequencing and assembly of *Haemophilus influenzae* rd."

Science 269: 496-512.

(Picture adapted from TIGR website,
<http://www.tigr.org>)



- Integrative Data

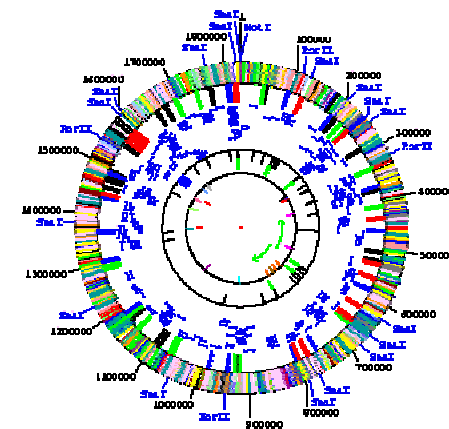
1995, HI (bacteria): 1.6 Mb & 1600 genes done

1997, yeast: 13 Mb & ~6000 genes for yeast

1998: 14 completed genomes!

1998, worm: 75 of 100 Mb done
with 13 K genes so far

2003, human: 3 Gb & 100 K genes...



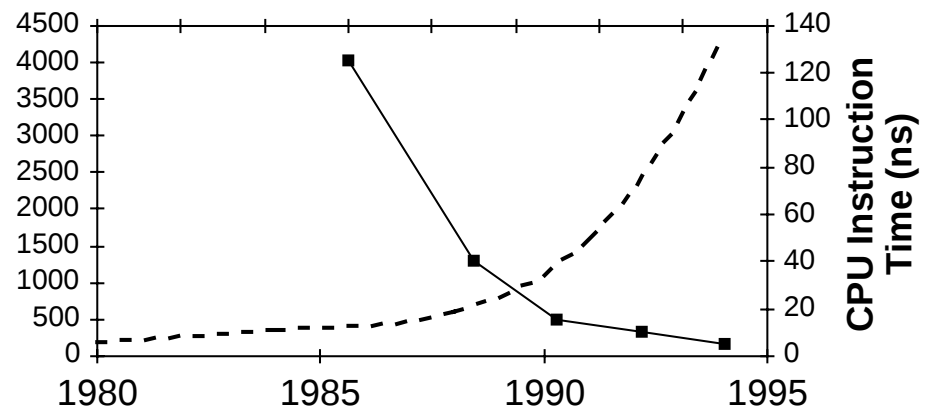
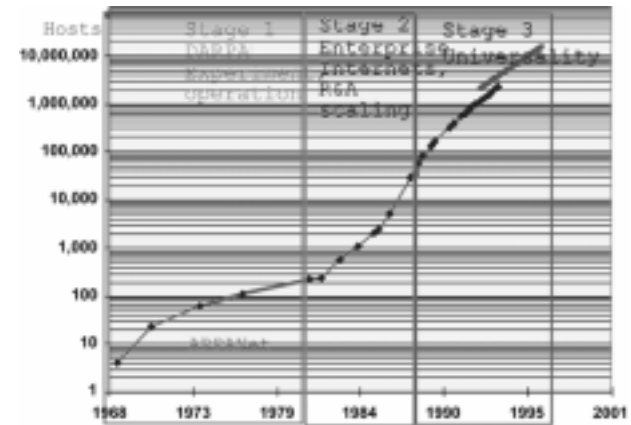
Exponential Growth of Data Matched by Development of Computer Technology

- CPU vs Disk & Net
 - ◇ As important as the increase in computer speed has been, the ability to store large amounts of information on computers is even more crucial
- Driving Force in Bioinformatics

(Internet picture adapted from D Brutlag, Stanford)

Num.
Protein
Domain
Structures

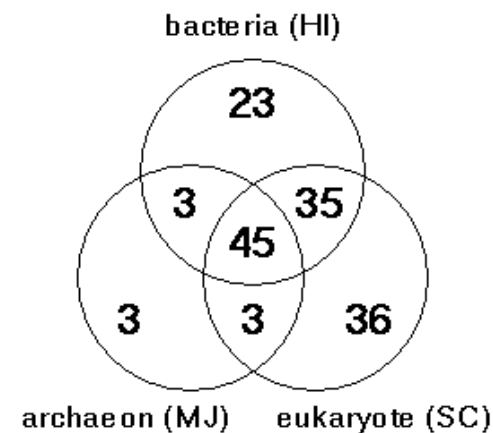
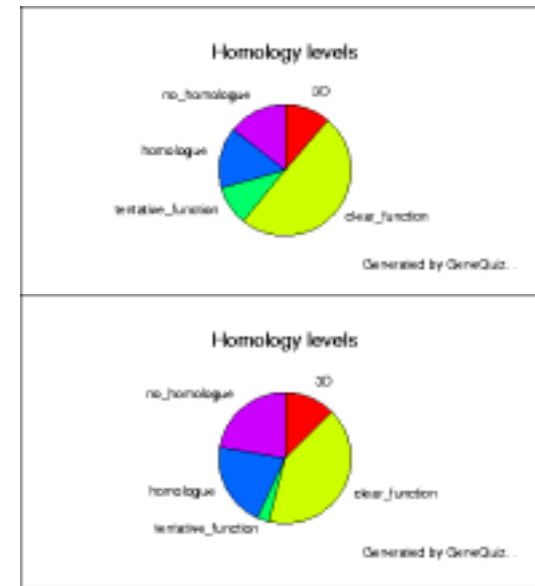
Internet Hosts



Major Application II: Overall Genome Characterization

- Overall Occurrence of a Certain Feature in the Genome
 - ◇ e.g. how many kinases in Yeast
- Compare Organisms and Tissues
 - ◇ Expression levels in Cancerous vs Normal Tissues
- Databases, Statistics

(Clock figures, yeast v. Synechocystis, adapted from GeneQuiz Web Page, Sander Group, EBI)



Fold Library

Biology (~1000) -- Limited Number of Folds

(Chothia's 1000 folds, Surveys of the "Parts List of Shapes", Which folds in Which Organisms?, Statistical rather than mathematical conclusions, Database)

Physics (~10)

(12 matter particles, 18 SM const., CODATA, mathematical relationships, in the head)

Fundamental Constants

Symbol	Value	Unit
c	299 792 458	m s^{-1}
μ_0	$4\pi \times 10^{-7}$	N A^{-2}
ϵ_0	$\approx 12.566 370 614 \dots$	$10^{-17} \text{ N A}^{-2}$
e	8.854 187 817...	$10^{-12} \text{ F m}^{-1}$
G	6.672 59(85)	$10^{-11} \text{ m}^2 \text{ kg}^{-1} \text{ s}^{-2}$
$\hbar$	6.625 075 5(40)	10^{-34} J s
h	1.054 572 65(63)	10^{-34} J s
e	1.602 177 33(49)	10^{-19} C
ϕ_0	2.067 834 61(61)	10^{-15} Wb
m_e	9.109 389 7(54)	10^{-31} kg
m_p	1.672 623 1(10)	10^{-27} kg
m_p/m_e	1 836.152 701(37)	
α	7.297 353 08(33)	10^{-3}
α^{-1}	137.035 989 5(61)	
R_∞	10 973 731.534(13)	m^{-1}
N_A, L	6.022 135 7(36)	10^{23} mol^{-1}
F	96 485.339(29)	C mol^{-1}
R	8.314 510(70)	$\text{J mol}^{-1} \text{ K}^{-1}$
k	1.380 658(12)	$10^{-23} \text{ J K}^{-1}$
σ	5.673 51(19)	$10^{-8} \text{ W m}^{-2} \text{ K}^{-4}$



Chemistry (~100)

(112 elements, <http://mwana1.lanl.gov/CST/imagemap/periodic/periodic.html>, trends, on paper)

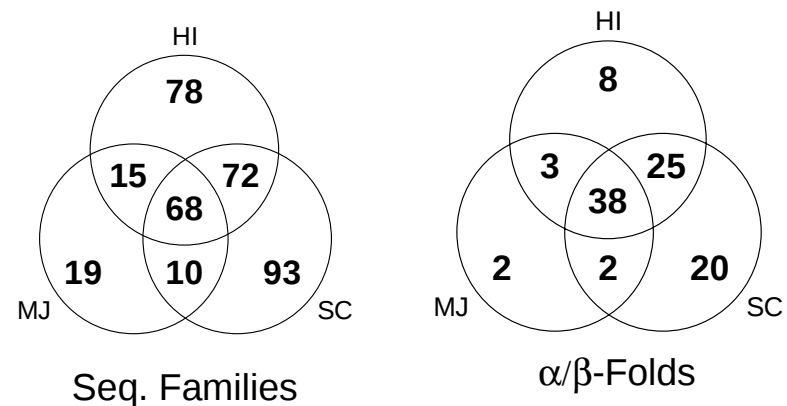
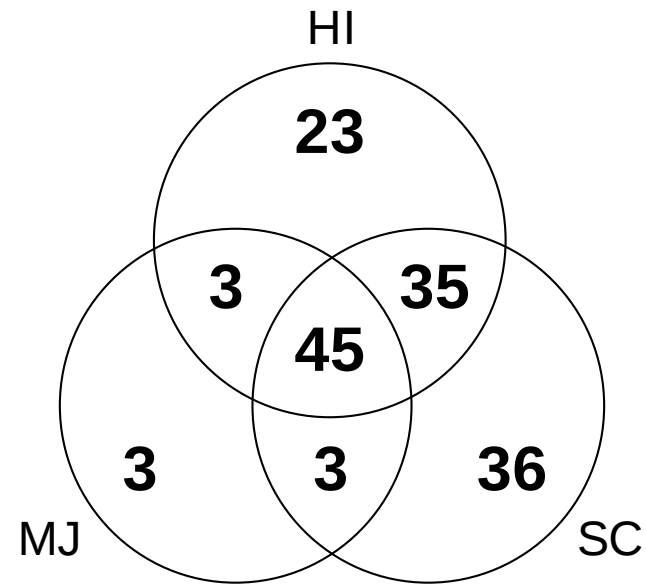
Periodic Table of the Elements																	
1	2															10	11
3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48
51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68
71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88
89	90	91	92	93	94	95	96	97	98	99	100	101	102	103	104	105	106
107	108	109	110	111	112												

* Lanthanide Series	58	59	60	61	62	63	64	65	66	67	68	69	70	71
+ Actinide Series	90	91	92	93	94	95	96	97	98	99	100	101	102	103

Shared Folds in Genomes are α/β

- Compare Genomes against Fold Library
 - ◊ HI (bacteria, ~1700 ORFs)
 - ◊ MJ (archeon, ~1700)
 - ◊ SC (eukaryote, ~6200)
 - ◊ EC, MP, MG, HP, SS as well.
- 135 of ~300 known folds in all 3 (106 HI vs. 54 MJ, PDB bias)
- **45** shared by all, and **38** of these mixed- $\alpha\beta$

Folds



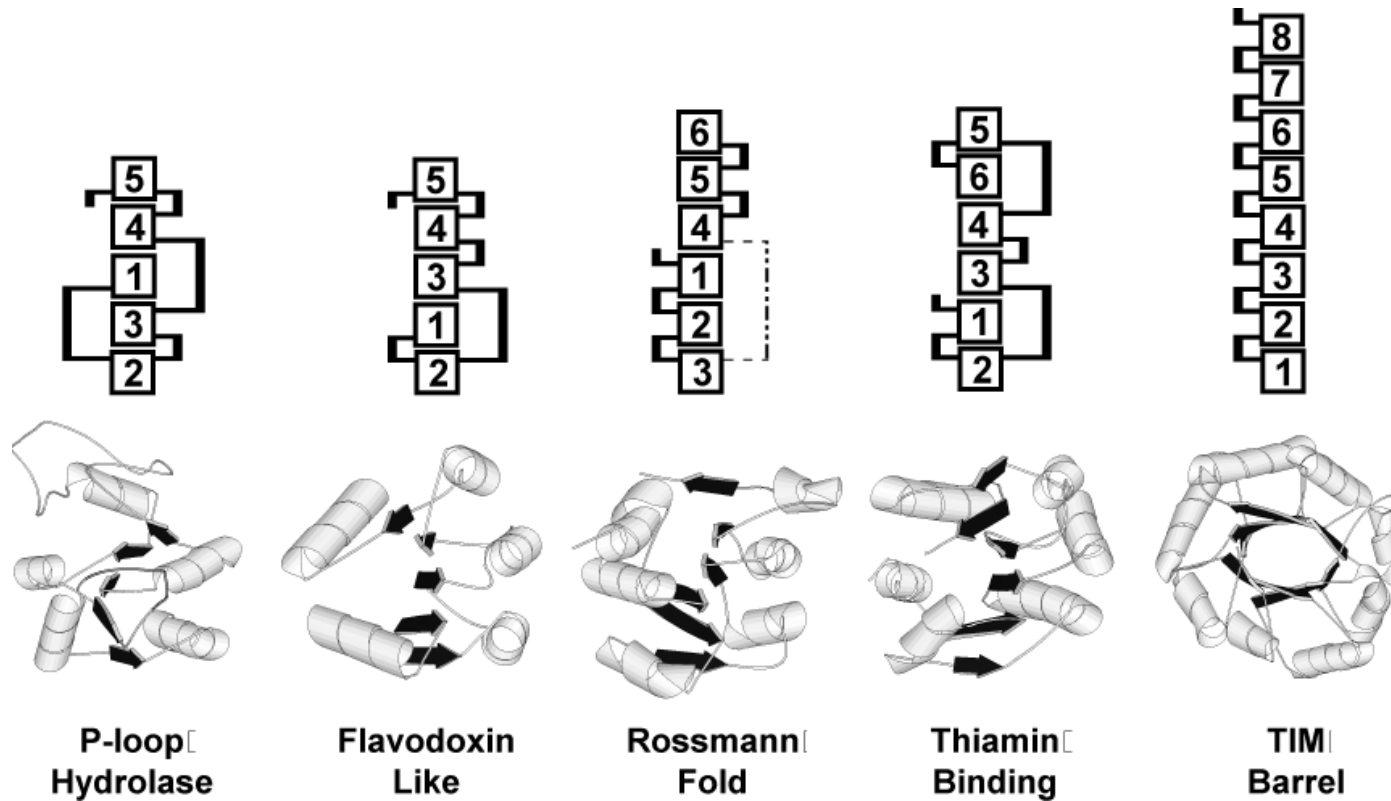
What are the Most Common Folds in the Genomes?

How many of these are
shared?

Num. in genome	Class	Fold Name	Representative Structure (PDB selection)
Top-10 in a eukaryotic genome (SC)			
84	$\alpha+\beta$	Protein kinases (catalytic core)	1irk
49	α/β	<u>P-loop containing NTP hydrolases</u>	1gky
35	α/β	<u>Rossmann Fold</u>	2ohx A:175-324
31	α/β	<u>TIM barrel</u>	1tim A:
25	α/β	Ribonuclease H-like	2rn2
18	S	Classic zinc finger	1zaa C:
14	$\alpha+\beta$	Ubiquitin conjugating enzyme	1aak
12	β	GroES-like	1acy L:109-211
10	α/β	Thioredoxin-like	1trx
9	α/β	<u>Thiamin-binding Fold</u>	1pvd A:2-181
5 x 8	...	...	...
7	α/β	<u>Flavodoxin-like</u>	3chy
Top-11 in a eubacterial genome (HI)			
18	α/β	<u>Rossmann Fold</u>	2ohx A:175-324
13	α/β	<u>P-loop containing NTP hydrolases</u>	1gky
12	α/β	<u>Flavodoxin-like</u>	3chy
10	α/β	<u>TIM barrel</u>	1tim A:
10	$\alpha+\beta$	Ferredoxin-like	1fxd
10	α/β	Ribonuclease H-like	2rn2
6	α/β	Periplasmic binding protein-like II	1sbp
5	α/β	Periplasmic binding protein-like I	2dri
5	$\alpha+\beta$	Like Class II aaRS synthetases	1sry A:111-421
4	β	OB-fold	1pyp
4	α/β	<u>Thiamin-binding Fold</u>	1pvd A:2-181
Top-11 in an archaeal genome (MJ)			
19	$\alpha+\beta$	Ferredoxin-like	1fxd
10	α/β	<u>P-loop containing NTP hydrolases</u>	1gky
7	α/β	<u>TIM barrel</u>	1tim A:
6	α/β	<u>Rossmann Fold</u>	2ohx A:175-324
5	α	Histone-fold	1ntx
4	α/β	<u>Thiamin-binding Fold</u>	1pvd A:2-181
4	α/β	<u>Flavodoxin-like</u>	3chy
4	β	Reductase/elongation factor common	1efg A:283-403
3	$\alpha+\beta$	ATP-grasp	1bnc A:115-330
3	α/β	PLP-dependent transferases	1dka
3	α/β	ATP pyrophosphatases	1gpm A:208-404

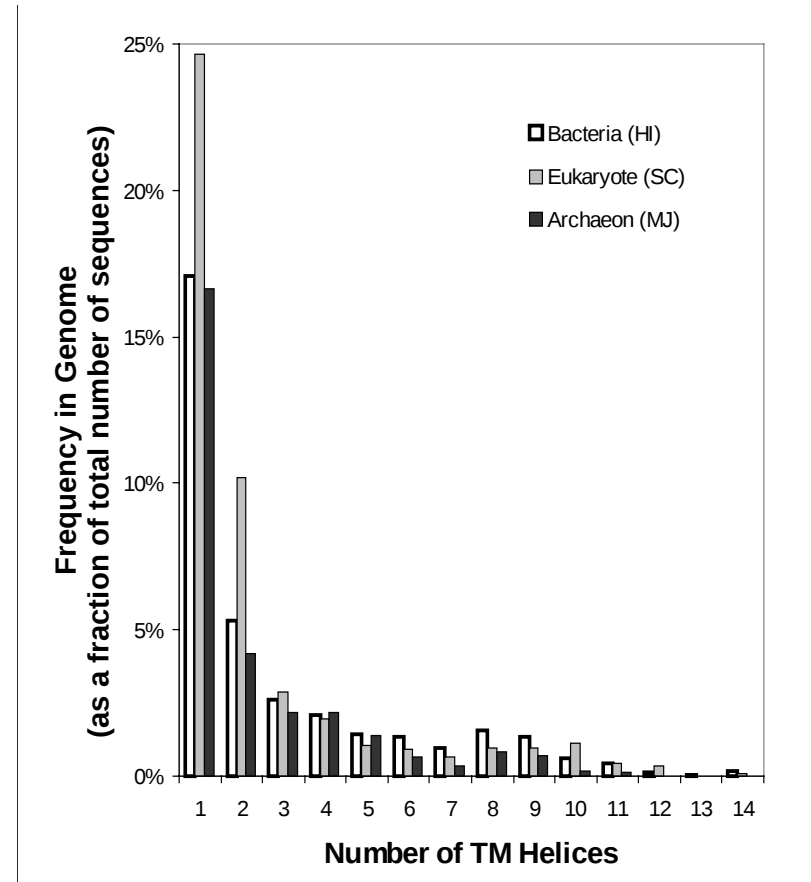
5 Folds in Top-10 All 3 Genomes

All share α/β structure with repeated R.H. $\beta\alpha\beta$ units connecting adjacent strands or nearly so (18+4+2 of 24)



TM-helix “prediction”

- TM prediction (KD, GES).
Count number with 2 peaks,
3 peaks, &c.
- Yeast has more mem. prots.,
esp. 2-TMs
- No preference for particular
supersecondary structures:
7-TM's



	Propensity (kcal/mole)		
	TM	α	β
	helix	helix	strand
Q	+4.1	-1.3	-0.4
S	-0.6	-1.1	-0.9
K	+8.8	-1.5	-0.4
I	-3.1	-1.2	-1.3
W	-1.9	-1.1	-1.0
H	+3.0	-1.1	-0.4
A	-1.6	-1.9	0.0
T	-1.2	-0.6	-1.4
Y	+0.7	-1.2	-1.6
E	+8.2	-1.2	-0.2
G	-1.0	0.0	+1.2
P	+0.2	+3.0	>3.0
C	-2.0	-1.1	-0.8
N	+4.8	-1.0	-0.5
V	-2.6	-0.8	-0.9
R	+12.3	-1.9	-0.4
D	+9.2	-1.0	+0.9
M	-3.4	-1.4	-0.9
L	-2.8	-1.6	-0.5
F	-3.7	-1.0	-1.1

← Each a.a. has different propensity for local structure
 ⇒ Different Compositions
 ⇒ Different Local Structure

However, bulk prediction of 2° struc. gives same fraction of α and β in the genomes (by element, half α , half β) →

2° Structure Prediction

Fraction of residues Predicted to be in...	strand	helix
Avg	17%	39%
SD	1%	2%
EC	17%	39%
HI	16%	41%
HP	15%	42%
MG	17%	39%
MJ	19%	37%
MP	17%	39%
SC	17%	34%
SS	16%	38%

Supersecondary structure words

- Look at super-secondary patterns (“words” such as $\alpha\alpha$ or $\beta\alpha\beta$) in predictions
- Compare observed freq. with expected freq.

odds = $f(\alpha\beta)/f(\alpha)f(\beta)$
(Freq. Words, Karlin)

- Find

HI more $\alpha\alpha$, $\alpha\alpha\alpha$, $\alpha\alpha\alpha\alpha$...

SC more $\beta\beta$, $\beta\beta\beta$, $\beta\beta\beta\beta$...

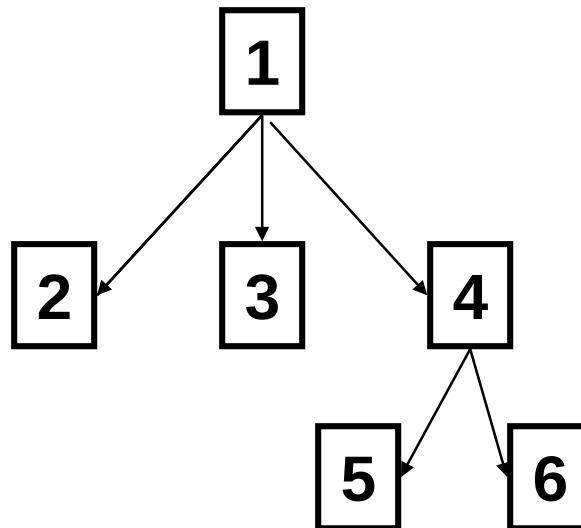
MJ more $\alpha\beta\alpha\beta$, $\beta\alpha\beta\alpha$...

Super-Secondary Structure "Word"	Maximum Difference between 3 Genomes	Relative Abundance (Odds Ratio)			
		HI	MJ	SC	PDB
$\beta\beta$	26%	0.96	1.06	1.24	1.22
$\alpha\alpha$	15%	0.97	0.85	0.83	0.85
$\alpha\beta$	10%	1.09	1.09	0.99	0.95
$\beta\alpha$	7%	0.98	1.00	0.93	0.99
$\beta\beta\beta$	41%	0.96	1.15	1.46	1.62
$\alpha\alpha\alpha$	19%	1.01	0.83	0.84	0.92
$\alpha\beta\alpha$	18%	1.04	1.03	0.87	1.16
$\alpha\alpha\beta$	15%	1.03	0.97	0.89	0.70
$\beta\alpha\beta$	12%	1.15	1.24	1.10	1.19
$\beta\alpha\alpha$	11%	0.93	0.87	0.83	0.78
$\beta\beta\alpha$	9%	0.90	0.94	0.99	0.82
$\alpha\beta\beta$	6%	0.97	0.98	1.03	0.80
$\beta\beta\beta\beta$	54%	1.03	1.35	1.78	2.28
$\alpha\alpha\alpha\alpha$	29%	1.10	0.82	0.89	1.18
$\beta\beta\beta\alpha$	25%	0.85	0.94	1.10	0.98
$\beta\alpha\beta\alpha$	23%	1.11	1.18	0.94	1.48
$\alpha\beta\alpha\beta$	21%	1.21	1.23	0.99	1.39
$\alpha\beta\alpha\alpha$	21%	1.00	0.95	0.81	1.00
...	...	...	...	...	...

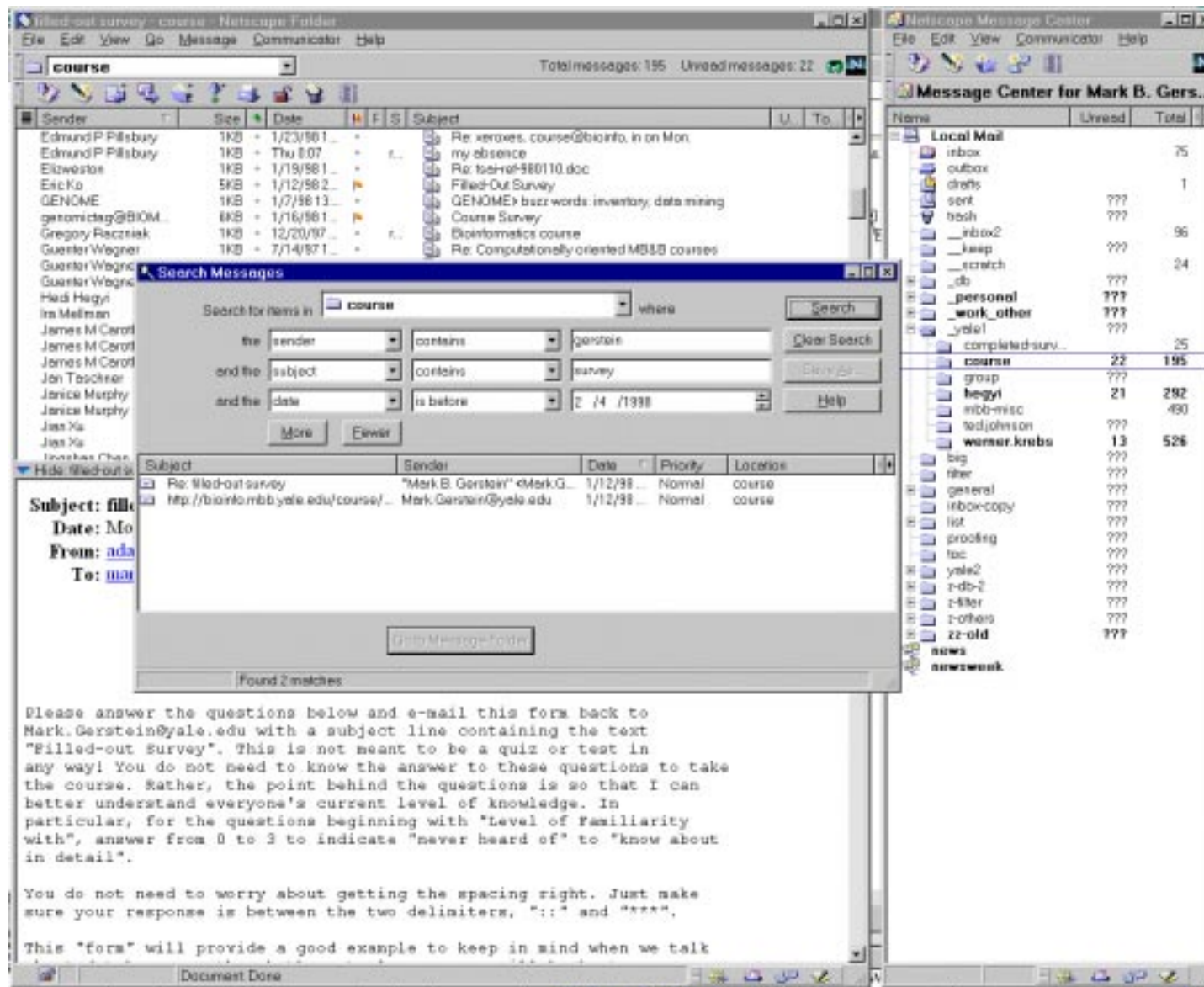
Complex Data Example: Encoding Trees in RDBs

Node	Parent
1	0
2	1
3	1
4	1
5	3
6	3

Node	Name
1	Organism
2	Bacteria
3	Archea
4	Eukarya
5	Metazoa
6	Plants



RDBs Everywhere: Internet Mail



RDBs Everywhere: File System

<u>INODE</u>	SIZE	PERMISSION	USER	GROUP	BYTES	MMM-DD--	YEAR	NAME
120462	1	drwxr-xr-x	10 mbg	gerstein	1024	Feb 12	1997	.
120463	1	drwxr-xr-x	2 mbg	gerstein	1024	Jan 30	1997	./hi-tbl
120464	514	-rw-r--r--	1 mbg	gerstein	525335	Nov 10	1996	./hi-tbl/id_gorss.tbl
120465	19	-rw-r--r--	1 mbg	gerstein	18469	Nov 10	1996	./hi-tbl/id_kytedool.tbl
120466	514	-rw-r--r--	1 mbg	gerstein	525372	Nov 10	1996	./hi-tbl/id_seq.tbl
108224	507	-rw-r--r--	1 mbg	gerstein	518822	Nov 10	1996	./mj-tbl/id_gorss.tbl
108227	54	-rw-r--r--	1 mbg	gerstein	54775	Jan 30	1997	./mj-tbl/id_abcode.tbl
108228	19	-rw-r--r--	1 mbg	gerstein	19131	Nov 11	1996	./mj-tbl/id_kytedool.tbl
108229	106	-rw-r--r--	1 mbg	gerstein	108345	Nov 16	1996	./mj-tbl/word_stats.tbl.bak
108230	106	-rw-r--r--	1 mbg	gerstein	108354	Jan 28	1997	./mj-tbl/word_stats.tbl
108231	7	-rw-r--r--	1 mbg	gerstein	6962	Jan 30	1997	./mj-tbl/hist_seqlen.tbl
108232	7	-rw-r--r--	1 mbg	gerstein	6967	Jan 30	1997	./mj-tbl/hist_num_H_res.tbl
91903	1	drwxr-xr-x	2 mbg	gerstein	1024	Nov 19	1996	./po-tbl

USER:PASSWD:UID:GID:COMMENT:DIR:SHELL

```
ftp:*:14:50:FTP User:/home/ftp:
nobody:*:99:99:Nobody:/:
mlml:cw5ZrAmNBAXvU:106:100:Michael Levitt (linux):/u1/mlml:/bin/tcsh
dabushne:ErR3hu4q0t07Y:108:100:Dave:/u1/dabushne:/bin/tcsh
mbg:V9CPWXAG.mo3E:5514:165:Mark Gerstein,432A, BASS,2-6105,:/u0/mbg:/bin/tcsh
mbgmbg:V9CPWXAG.mo3E:5515:165:logs into mbg,,,,:/u0/mbg:/bin/tcsh
mbg10:V9CPWXAG.mo3E:5516:165:alternate account for mbg:/home/mbg10:/bin/tcsh
local::502:20:Local Installed Packages:/u1/local:/bin/tcsh
login::503:20:Hyper Login:/u0/login:/u0/login/hyper-login.pl
```

find -ls
/etc/passwd