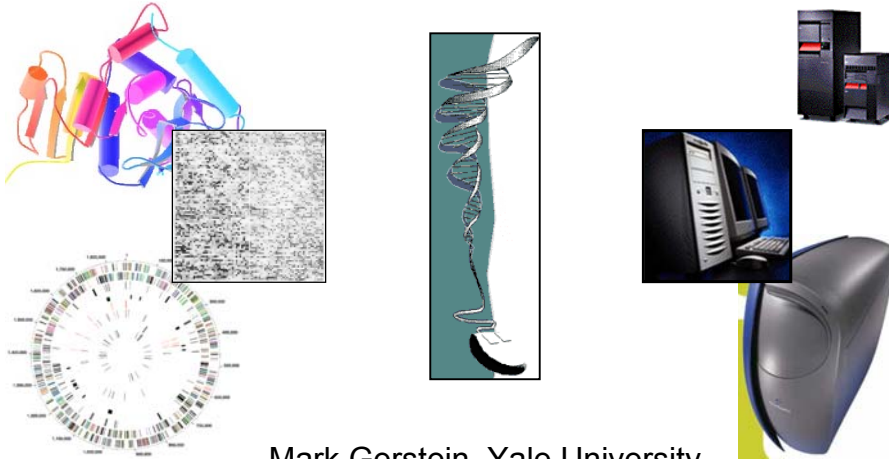


BIOINFORMATICS Databases



Mark Gerstein, Yale University
bioinfo.mbb.yale.edu/mbb452a

Start of class
2002, 11.18 (Bioinfo-11)
[after expression discussion]

Contents: Databases

- Structuring Information in Tables
- Keys and Joins
- Normalization
- Complex RDB encoding
- Indexes and Optimization
- Forms and Reports

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.,

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

```

5 (c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

Structured Data

gid	TrgStrt	TrgStop	did
H10299	119	135	di931__
H10572	180	240	diabs__
H10989	56	125	diaco_1
H10988	106	458	diaco_2
H10154	2	76	diacp__
H11633	2	432	diadea__
H10349	1	183	diaky__
H11309	35	52	dialc_3
H10589	8	25	dialc_3
H11358	239	444	diamg_2
H11358	218	410	diamy_2
H10460	20	24	dians__
H11386	139	147	dians__
H10421	11	14	dians__
H10361	285	295	dians__
H10835	100	106	dians__

```

did_ fids
d2rs51_ 1.002.007
dlimr_ 1.010.002
dlpyibl 1.007.030
dlldtd_ 1.001.001
di811_ 1.004.002
dlvmoa_ 1.002.044
d2gsq_1 1.001.031
dletb2_ 1.002.003
dlguha1 1.001.031
dlhrc_ 1.001.003
d1501c_ 1.004.002
dlldmf_ 1.007.035
dl119_ 1.004.002
dlyrnc_ 1.010.002
dlapld_ 1.001.004
dlndab2 1.003.004
d2rmai_ 1.002.036

```

fid	bestrep	N_minsp	N_scop	objname
1.001.001	dflp__	8	340	Globin-like
1.001.002	dihdj__	4	33	Long alpha-hairpin
1.001.003	dictj__	9	78	Cytochrome c
1.001.004	dlenh__	18	76	DNA-binding 3-helical bundle
1.001.005	dldtr_2	1	3	Diphtheria toxin repressor (DtxR) dimeriz
1.001.006	ditns__	1	2	Mu transposase, DNA-binding domain
1.001.007	d2spca_	1	2	Spectrin repeat unit
1.001.008	dibdd__	1	4	Immunoglobulin-binding protein A modules
1.001.009	dibal__	1	5	Peripheral subunit-binding domain of 2-ox
1.001.010	d2erl__	3	5	Protozoan pheromone proteins

6 (c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

Extra

Turn the Survey into a Table (I)

0	Person Number	5	1	20	8	13	22	9	21	7	25	11
1	First Name	john	jason	josh	jerry	jesse	jennifer	jill	mark	martin	nurmy	net
8	Major Field?	biophysics	MS&B	MS&B	Molecular Biophysics and Biochemistry	mib	m&B	mib	Molecular Biophysics & Biochemistry	MS&B	MS&B	MS&B
17	Are you combining this half-course module with another one?	y	y	y	n	n	y	n	n	n	n	n
18	If so, which one?	macromolecular crystallography	Topics in Nucleic Acids	not decided yet	n		macromolecular crystallography		NA	NA		
7.5	Comment on if taking for credit											
53	Are there specific topics that you want to cover? (use words above)		BLAST searching, Dynamic Programming	protein alignment algorithms, joining together two databases tables	grof, a recursive descent parser, hashing function, recursion, no because I	linkage and sib pair analysis, experiment at binary structures/terminations		none		chemokines	robotics	neural nets
50.5	Comment on if over/undersubscribed											
4	Status	O	O	O	U	U	U	O	U	U	O	U
7.5	Are you taking this for credit?	y	y	y	y	y	y	y	y	y	y	y
17	Are you combining this half-course module with another one?	y	y	y	n	n	y	n	n	n	n	n
16	Do you think a bioinformatics course should be offered again?	y	y	y	y	y	y	y	y	y	y	y
16	If course is ever subscribed this year, would you want to take it next year?	n	y	n	n	n	n		y	y	n	y
87	Is time change to Mon. & Wed. 9:05-10:20 & NOT Fri. Ok?	y	y	y	y	y	y	y	y	y	n	y
87	Is time change to Mon. & Wed. 9:30-10:45 & NOT Fri. Ok?	y	y	y	y	y	y	y	y	n	y	y
87	Is time change to Mon. & Wed. 9:20-10:35 & NOT Fri. Ok?	n	y	n	n	n	n	n	y	n	y	n
10	Can you program in perl?	n	y	n	n	n	n	n	y	n	n	n

Unique Identifier for Person?



7 (c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

Extra

Turn the Survey into a Table (II)

87	Is time change to Mon. & Wed. 9:30-10:45 & NOT Fri. Ok?	y	y	y	y	n	y	y	y	n	y	y
87	Is time change to Mon. & Wed. 9:20-10:35 & NOT Fri. Ok?	y	y	y	y	y	y	y	y	n	y	y
10	Can you program in perl?	n	y	n	n	n	n	n	y	n	n	n
12	Have you taken single-variable calculus?	y	y	y	y	y	y	y	y	y	y	y
13	Do you have a Parthenon Account?	y	y	y	y	y	y	y	y	y	y	y
14	Do you have an easy ability to read and create web pages?	y	y	n	y	y	y	n	y	y	y	n
15.5	Do you have a web home page?	y	y	n	y	y	y	n	y	y	y	y
20	Did not fill in survey but was at class	n	n	n	n	n	n	n	n	n	n	n
80	First Class Attendance	y	y	y	n	n	y	y	y	y	y	y
19	Familiarity with 'Genetic code'	3	3	2	3	3	2	1	3	1	2	2
20	Familiarity with 'Protein alignment algorithms'	0	1	0	1	1	1.5	0	1	0	0	0
21	Familiarity with 'BLAST search'	1	1	0	2	2	0	0	2	0	1.5	0
22	Familiarity with 'Bioinformatics'	0	1	3	0	2	0	0	0	0	1	0
23	Familiarity with '3D rotations, translations'	2	1	3	1	1	3	1	0	3	0	0
24	Familiarity with 'Consistent Subfactor'	1	1	0	0	0	2.5	0	0	2	0	0
25	Familiarity with 'Bayesian probability'	0	0	0	0	0	0	0	0	0	0	0
26	Familiarity with 'Belief nets'	0	0	0	0	0	0	0	0	1	0	0
27	Familiarity with 'Neural nets'	0	0	1	0	1	0	0	0	1	0	0
28	Familiarity with 'Genetic algorithms'	0	0	0	0	1	0	0	0	0	0	0
29	Familiarity with 'Simulated annealing'	1	2	0	0	2	2	1	0	0	1	0
30	Familiarity with 'Decision trees'	0	1	1	0	2	1	1	2	2	0	0
31	Familiarity with 'Artificial Intelligence'	1	0	0	1	2	0	0	1	0	0	0
32	Familiarity with 'Calculation of Standard Deviation'	3	2	3	3	3	2	3	3	3	3	1
33	Familiarity with 'a Bell-shaped Distribution (as of test scores)'	3	2	3	3	3	3	2	2	3	2	1
34	Familiarity with 'DNA, RNA'	3	3	3	3	3	2	3	3	3	2	3
35	Familiarity with 'Dynamic Programming'	1	0	0	0	1	1	0	3	1	0	0
36	Familiarity with 'alpha-helix'	3	3	3	3	3	3	1	3	3	2	2
37	Familiarity with 'Cell nucleus'	3	3	2	3	2	3	3	3	3	2	2
38	Familiarity with 'ATP, NAD'	3	2	2	3	3	2	3	3	3	2	3
39	Familiarity with 'Force as the Derivative (grad) of Energy'	3	2	3	1	1	3	3	3	2	2	0

Standardized Values



8 (c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

Extra

Turn the Survey into a Table (III)

- Dependencies between Values (dates)
- Unstructured Text



46) Familiarity with 'What GroEL does'	3	3	0	0	1	2	1	0	1	2	0	
46) Familiarity with 'A worm is a metazoan'	1	2	0	0	1	2	0	0	0	2	1	
50) Familiarity with 'E. coli is gram negative'	1	2	0	0	2	2	1	3	1	2	1	
51) Familiarity with 'What chemokines are'	3	2	0	0	3	3	1	0	0	0	2	
Familiarity with 'Joining together two database tables'	0	2	0	2	1	2.5	0	0	0	1	0	
52) Light'	0	2	0	2	1	2.5	0	0	0	1	0	
54) Favorite Fruit (response used for database class)	orange	orange	tangerine	pear	orange	mango	banana	watermelo	in	leaw	honeymelo	nectarine
Favorite Color (response used for database class)	O	O	O	O	B	R	B	B	B	B	B	W
55) class)												
			I don't know if anyone wants to know that I'm really hungry while I am writing this text	I wasn't able to attend Monday because I don't know actually have the time to take it again	I am very interested in taking this class, and since I am a senior with MB&B major, I will not get the chance to take it again	I would really like to take this class		none	#	rope	music	?
56) Any other random thoughts												
61) day												
62) month	Mon	Mon	Thu	Mon	Jan	Jan	Jan	Tue	Jan	Jan	Jan	Tue
63) date	12	12	18	13	14	16	17	18	19	18	19	20
64) Year	1998	1998	1998	1998	1998	1998	1998	1998	1998	1998	1998	1998
65) year	1998	1998	1998	1998	1998	1998	1998	1998	1998	1998	1998	1998



9 (c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

Extra

Statistics are only Possible on Standardized Values

Familiarity with 'DNA, RNA'	0-3	3.0	3	3	3	3	3	3	3	3	3	3	3	3	3
Familiarity with 'alpha-helix'	0-3	2.9	3	3	3	3	3	3	3	3	3	3	3	3	3
Familiarity with 'Cell nucleus'	0-3	2.8	3	3	3	3	2	2	3	3	3	3	3	3	3
Familiarity with 'ATP, NAD'	0-3	2.6	3	3	3	3	2	2	2	2	3	3	3	3	3
Familiarity with 'Genetic code'	0-3	2.6	3	3	3	3	3	2	3	2	3	3	3	3	3
Familiarity with 'Bell-shaped Distribution (as of test scores)'	0-3	2.6	3	3	3	2	2	2	2	3	2	3	3	3	3
Familiarity with 'Calculation of Standard Deviation'	0-3	2.4	2	3	2	2	1	2	2	3	1	2	3	3	3
Familiarity with 'Proteins are tightly packed'	0-3	2.3	2	2	3	3	2	2	3	3	3	3	3	3	3
Familiarity with 'E. coli is gram negative'	0-3	2.2	2	1	3	2	2	2	3	1	3	3	3	3	3
Familiarity with 'Force as the Derivative (grad) of Energy'	0-3	2.0	2	3	2	2	3	2	2	3	3	3	3	3	3
Familiarity with 'Protein families'	0-3	1.9	3	2	3	2	2	2	1	1	1	3	3	3	2
Familiarity with 'What GroEL does'	0-3	1.9	3	3	3	3	2	3	3	3	3	3	3	1	2
Familiarity with 'A worm is a metazoan'	0-3	1.8	3	1	2	3	2	2	3	2	3	3	3	1	2
Familiarity with 'What chemokines are'	0-3	1.8	3	3	1	2	2	2	2	3	2	3	3	3	3
Familiarity with 'BLAST search'	0-2	1.4	3	1	1	1	2	1	2	1	2	2	2	2	2
Familiarity with 'A P-value of .01'	0-3	1.3	2	3	1	2	2	1	1	1	2	2	3	2	2
Familiarity with '3D rotations, translations'	0-3	1.2	2	2	2	1	1	1	3	1	1	2	1	1	2
Familiarity with 'Poisson-Boltzman Equation'	0-3	1.1	2	2	2	1	2	2	1	1	1	2	1	1	2
Familiarity with 'Protein alignment algorithms'	0-3	1.0	2	1	1	2	2	2	1	2	2	1	2	1	2
Familiarity with 'Simulated annealing'	0-3	0.9	1	2	2	1	1	1	1	1	1	1	1	1	2
Familiarity with 'An Extreme Value Distribution'	0-3	0.7	3	1	1	1	1	1	1	1	1	1	2	1	2
Familiarity with 'Joining together two database tables'	0-3	0.7	1	1	1	2	1	1	1	1	1	1	1	1	2
Familiarity with 'Artificial Intelligence'	0-3	0.7	1	1	1	1	1	1	1	1	1	1	1	1	1
Familiarity with 'Sequence homology twilight zone'	0-3	0.6	2	2	2	2	2	2	2	2	2	2	1	2	2
Familiarity with 'Decision trees'	0-3	0.6	1	1	1	1	1	1	1	1	1	1	1	1	1
Familiarity with 'Constraint Satisfaction'	0-3	0.6	1	1	1	1	1	1	1	1	1	1	1	1	1
Familiarity with 'Genetic algorithms'	0-3	0.5	1	1	1	1	1	1	1	1	1	1	1	1	1
Familiarity with 'Robotics'	0-3	0.5	1	1	1	1	1	3	1	1	1	1	1	1	1
Familiarity with 'Dynamic Programming'	0-3	0.5	1	1	1	1	1	1	1	1	1	1	1	1	1
Familiarity with 'Bayesian probability'	0-3	0.4	1	1	1	1	1	1	1	1	1	1	1	1	2
Familiarity with 'Neural nets'	0-3	0.4	1	1	1	1	1	1	1	1	1	2	1	1	1
Familiarity with 'A Recursive Descent Parser'	0-3	0.4	1	1	1	1	1	1	1	1	1	1	1	1	1
Familiarity with 'A Hashing Function'	0-3	0.1	1	1	1	1	1	1	1	1	1	1	1	1	1
Familiarity with 'Belief nets'	0-3	0.0	1	1	1	1	1	1	1	1	1	1	1	1	1
Average		1.4	1.6	1.6	1.5	1.4	1.3	1.2	1.2	1.0	1.0	1.0	1.0	1.0	1.0

10 (c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

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

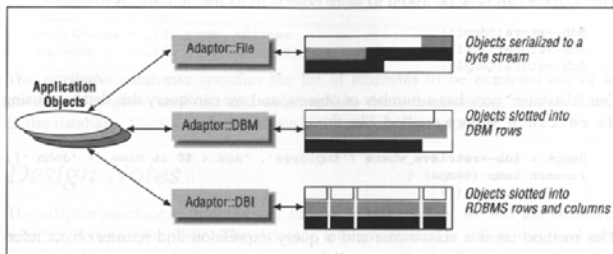


Figure 11-1. Adaptor modules

SQL

Core

- 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	d1931__	3.1
HI0572	180	240	dlaba__	0.0032
HI0989	56	125	dlaco_1	0.0049
HI0988	106	458	dlaco_2	4.4e-14
HI0154	2	76	dlacp__	1.2e-23
HI1633	2	432	dladea__	0
HI0349	1	183	dlaky__	7.6e-36
HI1309	35	52	dlalo_3	1.1
HI0589	8	25	dlalo_3	1.8
HI1358	239	444	dlamg_2	0.002
HI1358	218	410	dlamy_2	0.00037
HI0460	20	24	dlans__	1.8
HI1386	139	147	dlans__	3.3
HI0421	11	14	dlans__	6.4
HI0361	285	295	dlans__	8.2
HI0835	100	106	dlans__	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	d1931__	3.1
HI0572	180	240	dlaba__	0.0032
HI0989	56	125	dlaco_1	0.0049
HI0988	106	458	dlaco_2	4.4e-14
HI0154	2	76	dlacp__	1.2e-23
HI1633	2	432	dladea__	0
HI0349	1	183	dlaky__	7.6e-36
HI1309	35	52	dlalo_3	1.1
HI0589	8	25	dlalo_3	1.8
HI1358	239	444	dlamg_2	0.002
HI1358	218	410	dlamy_2	0.00037
HI0460	20	24	dlans__	1.8
HI1386	139	147	dlans__	3.3
HI0421	11	14	dlans__	6.4
HI0361	285	295	dlans__	8.2
HI0835	100	106	dlans__	9.7

```
insert into
matches
(gid, TrgStrt,
  TrgStop, did,
  score)
values
(HI0299, 119,
  135, d1931__,
  3.1)
```

End of class 2002,11.18 (Bioinfo-11) [after expression discussion]

structures table

```
did_    fid
d2rs51_ 1.002.007
d1imr_  1.010.002
dlpyibl 1.007.030
didxtd_ 1.001.001
d1811_  1.004.002
dlvmoa_ 1.002.044
d2gsq_1 1.001.031
d1etb2_ 1.002.003
dlguhal 1.001.031
dlhrc_  1.001.003
d1501c_ 1.004.002
dldmf_  1.007.035
d1119_  1.004.002
d1yrnc_ 1.010.002
dlapld_ 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
)
```

**10 K domain
structure IDs (did)
vs. 300 fold IDs
(fid)**

folders 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	dflp__	8	0	Globin-like
1.001.002	dihdj__	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

The “problem” databased in the example here: Matching structures to a genome

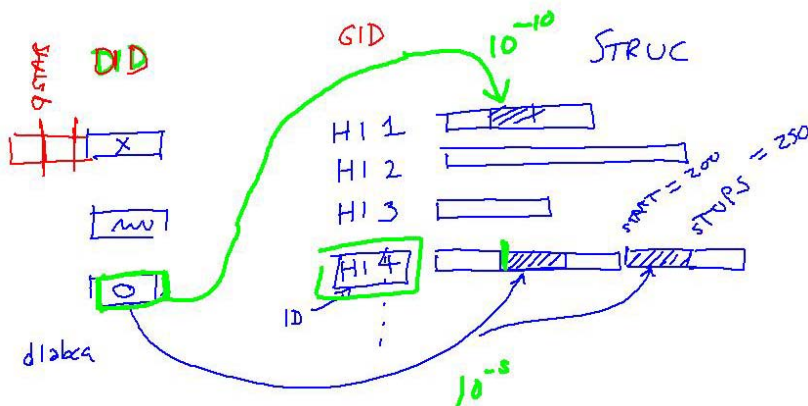
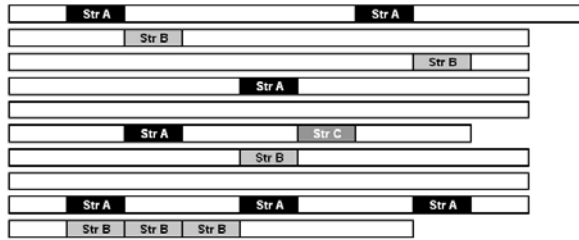


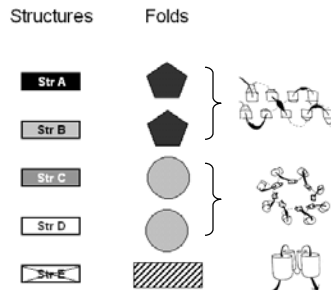
Table Interpretation

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



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

Table	key attr-a	key 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



1<->1
1->many
many->1

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_	TrgStprt	TrgStop	did	score
HI0299	119	135	d1931__	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 TrgStprt=179

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

SQL Select on a Single Table, Example 2

gid_	TrgStprt	TrgStop	did	score
HI0299	119	135	d1931__	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
HI0016	1	173	d1dar_2	2e-07
HI0016	179	274	d1dar_1	8.5e-06

Joins

Core

Foreign Key

Folds

Matches				Structures	
gid	TrgStrt	TrgStop	did	score	
HI0299	119	135	d1931__	3.1	
HI0572	180	240	diaba__	0.0032	
HI0989	56	125	diaco_1	0.0049	
HI0988	106	458	diaco_2	4.4e-14	
HI0154	2	76	diacp__	1.2e-23	
HI1633	2	432	diadea__	0	
HI0349	1	183	diaky__	7.6e-36	
HI1309	35	52	dialo_3	1.1	
HI0589	8	25	dialo_3	1.8	
HI1358	239	444	diamg_2	0.002	
HI1358	218	410	diamy_2	0.00037	
HI0460	20	24	dians__	1.8	
HI1386	139	147	dians__	3.3	
HI0421	11	14	dians__	6.4	
HI0361	285	295	dians__	8.2	
HI0835	100	106	dians__	9.7	

did	fid
d2rs51__	1.002.007
dlimr__	1.010.002
d1pyib1	1.007.030
didxtd__	1.001.001
d1811__	1.004.002
d1vmoa__	1.002.044
d2gsq_1	1.001.031
dletb2__	1.002.003
d1guhal	1.001.031
dlhrc__	1.001.003
d1501c__	1.004.002
dldmf__	1.007.035
d1119__	1.004.002
dlyrnc__	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	d1lenh__	2	0	DNA-binding 3-helical bundle
1.001.005	didtr_2	1	3	Diphtheria toxin repressor (DtxR) dimeriz
1.001.006	d1tms__	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

25 (c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

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 = HI0361 and matches.did=structures.did
 and structures.fid = folds.fid

 8.2, Neurotoxin III ...

26 (c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

Foreign Key matches

gid	TrgStrt	TrgStop	did	score
HI0299	119	135	d1931__	3.1
HI0572	180	240	dlaba__	0.0032
HI0989	56	125	diaco_1	0.0049
HI0988	106	458	diaco_2	4.4e-14
HI0154	2	76	diacp__	1.2e-23
HI1633	2	432	diadea__	0
HI0349	1	183	dlaky__	7.6e-36
HI1309	35	52	dialo_3	1.1
HI0589	8	25	dialo_3	1.8
HI1358	239	444	diamg_2	0.002
HI1358	218	410	diamy_2	0.00037
HI0460	20	24	dians__	1.8
HI1386	139	147	dians__	3.3
HI0421	11	14	dians__	6.4
HI0361	285	295	dians__	8.2
HI0835	100	106	dians__	9.7

structures

did	fid
d2rs51__	1.002.007
dlimr__	1.010.002
dipyib1	1.007.030
dldxtd__	1.001.001
d1811__	1.004.002
dlvmoa__	1.002.044
d2gsq_1	1.001.031
dletb2__	1.002.003
dlguhal	1.001.031
dlhrc__	1.001.003
d1501c__	1.004.002
dldmf__	1.007.035
d1119__	1.004.002
dlyrnc__	.010.002
dians__	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

Matches					Structures		
Table 1	key gid	key TrgStrt	TrgStop	did	Table 2	did	fid
tuple-1	HI001	12	200	d1mbd_	tuple-i	d1fg_1	1.007.006
tuple-2	HI002	15	231	d1hhba_	tuple-i	d1fg_1	1.007.006
tuple-3	HI002	100	343	d1fg_1	tuple-i	d1fg_1	1.007.006
tuple-4	HI003	12	80	d1fg_1	tuple-i	d1fg_1	1.007.006
tuple-5	HI009	200	260	d1mba_	tuple-i	d1fg_1	1.007.006
tuple-6	HI023	300	450	d2ubx_	tuple-i	d1fg_1	1.007.006
tuple-7	HI045	2	89	d2lmg_	tuple-i	d1fg_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	d1fg_1	tuple-ii	d1mba	1.003.002
tuple-4	HI003	12	80	d1fg_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) x 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 x 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 x B =

A x B has
 N x 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

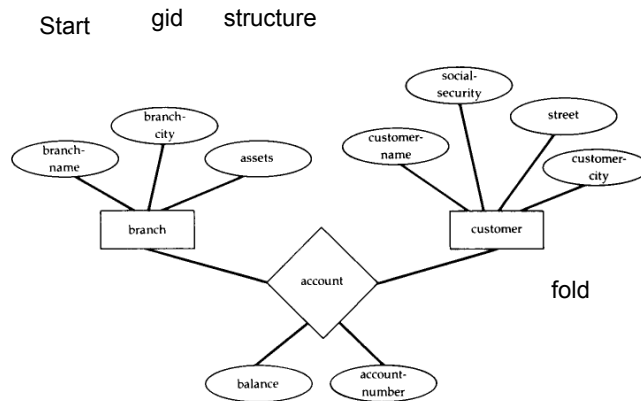


Figure 2.23 E-R diagram with *account* as a relationship set.

- Korth & Silberschatz
 - ◇ branch <=> matches (gid-start +++ did)
 - ◇ customer <=> folds (fid +++)
 - ◇ linked by
 - account <=> 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
 - ◇ F <= select fid from folds where name contains "globin"
 - ◇ D <= select did from structures where fid in F
 - ◇ N <= select count(distinct gid, TrgStrt) from matches where did in D and score < .01

Joins

gid	TrgStprt	TrgStop	did	score
HI0299	119	135	d1931__	3.1
HI0572	180	240	dlaba__	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
d1dxtd__	1.001.001
d1811__	1.004.002
d1vmoa__	1.002.044
d2gsq_1	1.001.031
d1etb2__	1.002.003
d1guhal	1.001.031
d1hrc__	1.001.003
d1501c__	1.004.002
d1dmf__	1.007.035
d1119__	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	d1lenh__	2	0	DNA-binding 3-helical bundle
1.001.005	d1dtr_2	1	3	Diphtheria toxin repressor (DtxR) dimeriz
1.001.006	d1tms__	1	2	Mu transposase, DNA-binding domain
1.001.007	d2spca__	0	2	Spectrin repeat unit
1.001.008	d1lbd__	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

Core

gid	TrgStprt	TrgStop	did	score	fid	N_hlx	N_beta	name
HI0299	119	135	d1931__	3.1	1.010.002	0	2	Spectrin repeat unit
HI0572	180	240	dlaba__	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	d1931__	3.1	1.010.002	0	2	Spectrin repeat unit
HI0572	180	240	diaba__	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	d1931__	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
d1dxtd__	1.001.001
d1811__	1.004.002
d1vmoa__	1.002.044
d2gsq_1	1.001.031
d1etb2__	1.002.003
d1gihal	1.001.031
d1hrc__	1.001.003
d1501c__	1.004.002
d1dmf__	1.007.035
d1119__	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

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

One Index

Green	750
Lindsay	700
Smith	700

bucket 1

bucket 2

bucket 3

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

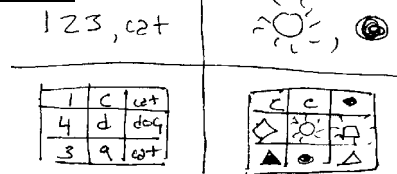
Double Index

		Simple Data Types	Complex Data Types
Simple Structure	Data	int: 1,2,3 chars: hello, text	struct A {pointer-list + char} method Am acts on A
	DB	Simple File File with unstructured text	Object DB (OODB) 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)
	Arrays of Query Lang.	Rows and columns contain ints and chars SQL	Rows and columns contain complex objects and methods are defined to handle them 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

Extra

C, fortran vs. C++



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

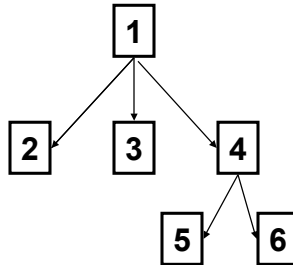
Extra

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

Complex Data Example: Encoding Trees in RDBs

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



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

RDBs Everywhere: Internet Mail

Search Messages

Search for items in **course** where

the sender contains gerstein

and the subject contains survey

and the date is before 2/4/1998

Found 2 matches

Subject: Filled-out survey
Date: Mon, 1/12/98
From: mbb
To: mbb

Please answer the questions below and e-mail this form back to Mark Gerstein@yale.edu with a subject line containing the text "Filled-out Survey". This is not meant to be a quiz or test in any way! You do not need to know the answer to these questions to take the course. Rather, the point behind the questions is so that I can better understand everyone's current level of knowledge. In particular, for the questions beginning with "Level of Familiarity with", answer from 0 to 3 to indicate "never heard of" to "know about in detail".

You do not need to worry about getting the spacing right. Just make sure your response is between the two delimiters, "!" and "****".

This "form" will provide a good example to keep in mind when we talk

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:PTP User:/home/ftp:
nobody*:99:99:Nobody:/:

mlml:cw5ZrAmNBaxvU:106:100:Michael Levitt (linux):/u1/mlml:/bin/tcsh
dabushne:ErR3hu4q0tO7Y: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**

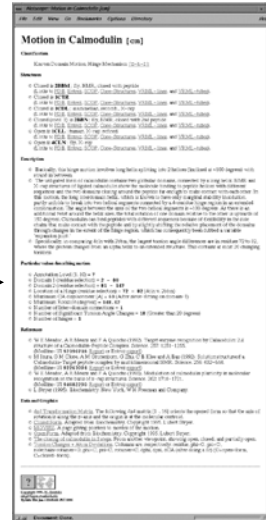


**From here to end
of Databases all is
“extra” unless
otherwise marked.**

Example Report: Motions Database



Report on Calmodulin



Example Report: Motions Database

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



Schema

```

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 relations (
  id CHAR(15),
  id CHAR(15),
  type CHAR(30),
  comment CHAR(512)
)
CREATE TABLE links (
  id CHAR(10),
  url CHAR(150),
  hit_text CHAR(100),
  other_text CHAR(500),
  flag CHAR(5)
)
CREATE TABLE single_vals (
  id CHAR(10),
  name CHAR(30),
  val CHAR(30),
  comment CHAR(500)
)
CREATE TABLE names (
  id CHAR(10),
  seq_num INT,
  name CHAR(255)
)
CREATE TABLE structures (
  id CHAR(10),
  pdb_id CHAR(8),
  name_short CHAR(50),
  chain CHAR(1),
  name_long CHAR(100)
)
CREATE TABLE refs (
  id CHAR(10),
  medicine INT,
  endnote INT,
  flag INT
)
CREATE TABLE value_names (
  abbrev CHAR(15),
  name CHAR(50)
)
CREATE TABLE descriptions (
  id CHAR(10),
  num INT,
  prose CHAR(5000)
)
CREATE TABLE endnote_refs (
  num INT,
  name CHAR(512)
)
    
```


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

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 (A) = **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 Quiocho (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 92269094: Report or Entrez export)
- W E Meador, A R Means and F A Quiocho (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. WH Freeman and Company.



NAMES	id	seq	num	n	name
aas					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	649
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 Quiocho (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:

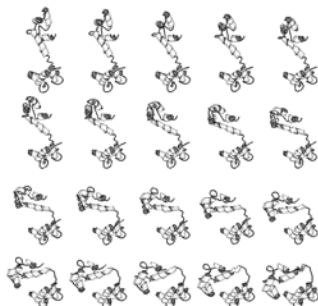
Join Two Lists (Protein Names and References) with a Table Containing Key for each List (a Relation: protein has reference.)

```
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

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

Practical Database Systems

- Oracle, MySQL, Informix, Sybase, IBM DB2

Moving into Surveys