Molecular Biophysics & Biochemistry 447b3 / 747b3

Bioinformatics

Databases

Mark Gerstein

Class 8, 2/2/98

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

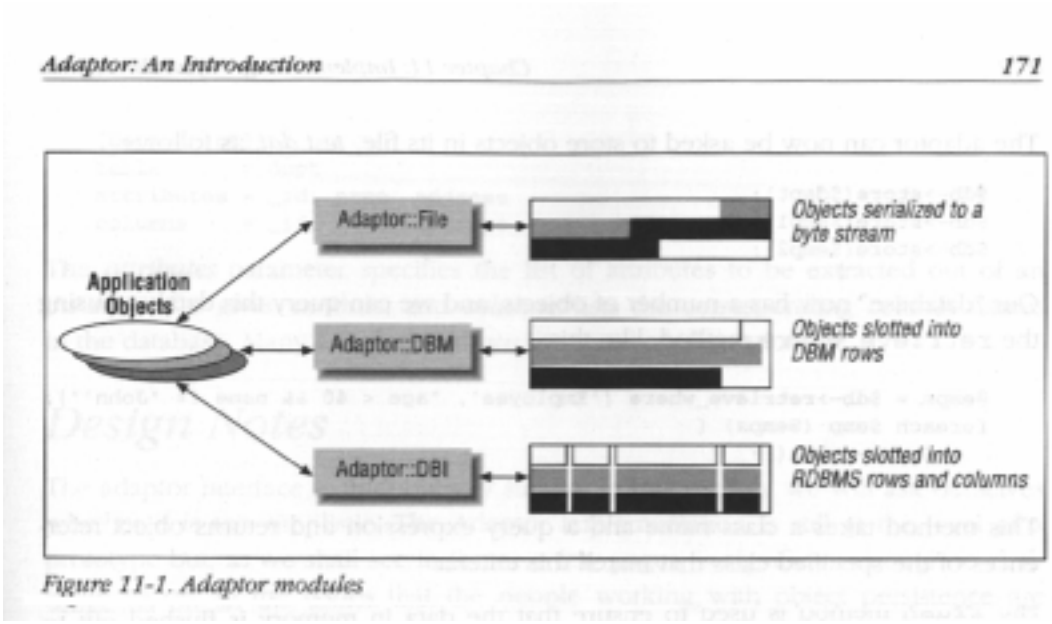


Figure 11-1. Adaptor modules

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
d1dxtd_	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

Survey 1

*** First-Name
<1> (free-text) ::
Joe
*** Last-Name
<2> (free-text) ::
Bone
*** e-mail Address
<3> (free-text) ::
Joe.Bone@yale.edu

*** Status
<4> ([U]ndergraduate/[G]raduate/[O]ther) :: G

*** Class? (e.g. first year grad. student, senior undergrad.)
<5> (free-text) :: first year

*** Major Field?
<6> (free-text) :: mbb

*** Are you taking this for credit?
<7> (y/n) ::
y
*** Is time change to Mon. & Wed. 9:30-10:45 & NOT Fri. OK?
<8> (y/n) :: y

*** Is time change to Mon. & Wed. 9:05-10:20 & NOT Fri. OK?
<9> (y/n) :: y

*** Can you program in perl?
<10> (y/n) :: n

*** Can you program in C?
<11> (y/n) :: y

*** Have you taken single-variable calculus?
<12> (y/n) :: y

*** Do you have a Pantheon Account?
<13> (y/n) :: y

*** Do you have easy ability to read and create web pages?
<14> (y/n) :: y

*** URL of Your Web Home Page, if you have one
<15> (free-text) :: http://pantheon.yale.edu/~mg269

*** Do you think a bioinformatics course should be offered again?
<16> (y/n) :: y

*** Are you combining this half-course module with another one?
<17> (y/n) :: y

*** If so, which one?
<18> (free-text) :: mbb460b4

*** Level of Familiarity with 'Genetic code'
<19> (0-3) :: 3

*** Level of Familiarity with 'Protein alignment algorithms'
<20> (0-3) :: 1

*** Level of Familiarity with 'BLAST search'
<21> (0-3) :: 2

*** Level of Familiarity with 'Robotics'
<22> (0-3) :: 0

*** Level of Familiarity with '3D rotations, translations'
<23> (0-3) :: 1

*** Level of Familiarity with 'Constraint Satisfaction'
<24> (0-3) :: 0

*** Level of Familiarity with 'Bayesian probability'
<25> (0-3) :: 0

*** Level of Familiarity with 'Belief nets'
<26> (0-3) :: 0

*** Level of Familiarity with 'Neural nets'
<27> (0-3) :: 1

*** Level of Familiarity with 'Genetic algorithms'
<28> (0-3) :: 1

Survey 2

*** Level of Familiarity with 'Simulated annealing'
<29> (0-3) :: 1

*** Level of Familiarity with 'Decision trees'
<30> (0-3) :: 1

*** Level of Familiarity with 'Artificial Intelligence'
<31> (0-3) :: 2

*** Level of Familiarity with 'Calculation of Standard Deviation'
<32> (0-3) :: 3

*** Level of Familiarity with 'a Bell-shaped Distribution (of test scores)'
<33> (0-3) :: 3

*** Level of Familiarity with 'DNA, RNA'
<34> (0-3) :: 3

*** Level of Familiarity with 'Dynamic Programming'
<35> (0-3) :: 1

*** Level of Familiarity with 'alpha-helix'
<36> (0-3) :: 3

*** Level of Familiarity with 'Cell nucleus'
<37> (0-3) :: 3

*** Level of Familiarity with 'ATP, NAD'
<38> (0-3) :: 3

*** Level of Familiarity with 'Force as the Derivative (grad) of Energy'
<39> (0-3) :: 1

*** Level of Familiarity with 'A P-value of .01'
<40> (0-3) :: 2

*** Level of Familiarity with 'An Extreme Value Distribution'
<41> (0-3) :: 0

*** Level of Familiarity with 'Proteins are tightly packed' http://bioinfo.mbb.yale.edu/course
<42> (0-3) :: 1

*** Level of Familiarity with 'Sequence homology twilight zone'
<43> (0-3) :: 1

*** Level of Familiarity with 'Protein families'
<44> (0-3) :: 1

*** Level of Familiarity with 'Poisson-Boltzman Equation'
<45> (0-3) :: 0

*** Level of Familiarity with 'A Hashing Function'
<46> (0-3) :: 0

*** Level of Familiarity with 'A Recursive Descent Parser'
<47> (0-3) :: 0

*** Level of Familiarity with 'What GroEL does'
<48> (0-3) :: 1

*** Level of Familiarity with 'A worm is a metazoa'
<49> (0-3) :: 1

*** Level of Familiarity with 'E. coli is gram negative'
<50> (0-3) :: 1

*** Level of Familiarity with 'What chemokines are'
<51> (0-3) :: 0

*** Level of Familiarity with 'Joining together two database tables'
<52> (0-3) :: 1

*** Are there specific topics that you want to cover? (use words above)
<53> (free-text) ::
using blast and understanding its output,
molecular modeling, genomes

*** Favorite Fruit (response used for database class)
<54> (free-text) ::
apple

*** Favorite Color (response used for database class)
<55> ([R]ed/[G]reen/[B]lue/[Y]ellow/[W]hite/[O]ther) :: Y

*** Any other random thoughts
<56> (free-text) ::
I hope you made it here :)

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	jessie	jennifer	jill	mark	martin	murrey	mel
6	Major Field?	biophysics	MB&B	MB&B	Molecular Biophysics and Biochemistry	mbb	mb&b	mbb	Molecular Biophysics &Biochemistry	MB&B	MB&B	MB&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 database tables	groel, a recursive descent parser, hashing function, poisson-no because I	linkage and sib pair analysis, experimental tertiary structure determination		none		chemokines	robotics	neural nets
58.5	Comment on if oversubscribed						n (I will not be here)					
4	Status	O	O	O	U	U	U	O	U	U	O	U
7	Are you taking this for credit?	y	y	y	y	y	y	n	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
58	If course is oversubscribed this year, would you want to take it next year?	n		y	n	n	n		y	y	n	y
9	Is time change to Mon. & Wed. 9:05-10:20 & NOT Fri. OK?	y	y	y	y	y	y	y	y	y	n	y
8	Is time change to Mon. & Wed. 9:30-10:45 & NOT Fri. OK?	y	y	y	y	n	y	y	y	n	y	y
57	Is time change to Mon. & Wed. 9:20-10:35 & NOT Fri. OK?	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

Unique Identifier for Person?

Turn the Survey into a Table (II)

8	Is time change to Mon. & Wed. 9:30-10:45 & NOT Fri. OK?	y	y	y	y	n	y	y	y	n	y	y
57	Is time change to Mon. & Wed. 9:20-10:35 & NOT Fri. OK?	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
11	Can you program in C?	n	n	y	n	y	n	n	y	y	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 Pantheon Account?	y	y	y	y	y	y	n	y	y	y	y
14	Do you have 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	?	n	n	y	y	y	?
70	Did not fill in survey but was at class	n	n	n	n	n	n	n	n	n	n	n
60	First Class Attendence	y	y	y	n			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
21	Familiarity with 'BLAST search'	1	1	0	2	3	0	0	2	0	1	0
22	Familiarity with 'Robotics'	0	1	3	0	2	0	0	0	1	0	0
23	Familiarity with '3D rotations, translations'	2	1	3	1	1	3	1	0	3	0	0
24	Familiarity with 'Constraint Satisfaction'	1	1	0	0	0	2.5	0	0	2	0	0
25	Familiarity with 'Bayesian probability'	0	0	0	0	1	0	0	0	0	1	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	2	0	1	2	0	0
32	Familiarity with 'Calculation of Standard Deviation'	3	2	3	3	3	3	2	3	3	2	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	3	2	3	3	3	3
35	Familiarity with 'Dynamic Programming'	1	0	0	0	1	1	0	3	0	1	0
36	Familiarity with 'alpha-helix'	3	3	3	3	3	3	2	3	3	2	2
37	Familiarity with 'Cell nucleus'	3	3	2	3	3	2	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

Standard-ized Values

Turn the Survey into a Table (III)

- Dependencies between Values (dates)
- Unstructured Text

48 Familiarity with 'What GroEL does'	3	3	0	0	1	2	1	0	1	2	0
49 Familiarity with 'A worm is a metazoa'	1	3	0	3	1	2	0	0	0	2	1
50 Familiarity with 'E. coli is gram negative'	1	2	1	3	2	2	1	3	1	2	1
51 Familiarity with 'What chemokines are'	3	2	0	3	3	1	0	0	0	2	0
52 Familiarity with 'Joining together two database tables'	0	2	0	2	1	2.5	0	0	0	1	0
54 Favorite Fruit (response used for database class)	organe	orange	tangerine	pear	orange	mango	banana	watermelon	kiwi	honeymelon	nectarine
55 Favorite Color (response used for database class)	G	G	O	O	B	R	B	B	B	B	W
56 Any other random thoughts			I don't know know if anyone wants to know that I am really hungry while I am writing this text	I wasn't able to attend class on Monday because I didn't know I would actually have the time slot free for this class. I hope that's all right.	I am very interested in taking this class, and since I am a senior in the MB&B major, I will not get the chance to take it again. Myprogram	I would really like to take this class	none	p	nope	music	?
61 day	Mon	Mon	Thu	Tue	Wed	Fri	Tue	Fri	Tue	Fri	Tue
62 month	Jan	Jan	Jan	Jan	Jan	Jan	Jan	Jan	Jan	Jan	Jan
63 date	12	12	15	13	14	16	13	16	13	16	13
64 hhmmss	13:43:18	11:14:10	21:00:41	13:15:19	14:20:28	1:08:01	15:16:25	1:07:59	11:08:16	14:49:31	16:37:24
65 year	1998	1998	1998	1998	1998	1998	1998	1998	1998	1998	1998

Statistics are only Possible on Standarized Values

Familiarity with 'DNA, RNA'	0-3	3.0	3	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	2	2	3	3				3	3	3	3	3
Familiarity with 'Cell nucleus'	0-3	2.8	3	3	3	3	2	2	3	2	3				3	3	3	3	3
Familiarity with 'ATP, NAD'	0-3	2.6	3	3	3	2	2	2		2	3				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	3
Familiarity with 'a Bell-shaped Distribution (as of test scores)'	0-3	2.6	3	3	3	2	2	2	2	3	2				3	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	3
Familiarity with 'Proteins are tightly packed'	0-3	2.3	2	2	3	3	3	2		2	3				3	2	3	3	2
Familiarity with 'E. coli is gram negative'	0-3	2.2	2	1	3	2	2	2	3	1	3				3	3	2	2	3
Familiarity with 'Force as the Derivative (grad) of Energy'	0-3	2.0	2	3	2	2	3	2	2	3					2	2	3	3	2
Familiarity with 'Protein families'	0-3	1.9	3	2	3	2	2	2	1	1	1				3	2	3	2	2
Familiarity with 'What GroEL does'	0-3	1.9	3	3	3	3	3	2	3		3				3	3	1	2	2
Familiarity with 'A worm is a metazoa'	0-3	1.8	3	1	2	3	2	2	3		2				3	3	1	2	2
Familiarity with 'What chemokines are'	0-3	1.8	2	3	1	2	2	2			2				3	3	3	2	3
Familiarity with 'BLAST search'	0-3	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	2	3	2	
Familiarity with '3D rotations, translations'	0-3	1.2		2	2	1				3	1				2	1	1	2	2
Familiarity with 'Poisson-Boltzman Equation'	0-3	1.1	2	2	2	1	2	2		1					2	2	1	2	
Familiarity with 'Protein alignment algorithms'	0-3	1.0	2		1	1	2		2						1	2	1	2	2
Familiarity with 'Simulated annealing'	0-3	0.9		1	2	2	1	1	1						1	1	1	2	2
Familiarity with 'An Extreme Value Distribution'	0-3	0.7		3	1			1	1	1					1	2	3		
Familiarity with 'Joining together two database tables'	0-3	0.7			1	2		1								1	1	2	
Familiarity with 'Artificial Intelligence'	0-3	0.7	1	1											2	1			
Familiarity with 'Sequence homology twilight zone'	0-3	0.6	2		2			2	2						1		2		
Familiarity with 'Decision trees'	0-3	0.6	1			1				1					1		1		
Familiarity with 'Constraint Satisfaction'	0-3	0.6	1	1		1									1	1	1	2	
Familiarity with 'Genetic algorithms'	0-3	0.5	1						1						1	2	2	1	2
Familiarity with 'Robotics'	0-3	0.5	1			1				3					1	1		1	
Familiarity with 'Dynamic Programming'	0-3	0.5	1	1				1							1			1	
Familiarity with 'Bayesian probability'	0-3	0.4	1		1			1								1	1	1	2
Familiarity with 'Neural nets'	0-3	0.4	1		1					1						2		1	
Familiarity with 'A Recursive Descent Parser'	0-3	0.1		1					2										
Familiarity with 'A Hashing Function'	0-3	0.1						1									1		
Familiarity with 'Belief nets'	0-3	0.0																	
Average		1.4	1.6	1.6	1.5	1.4	1.3	1.2	1.2	1.0	1.0				1.7	1.8	1.7	1.7	1.4

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
d1dxtd_	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

```
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

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
d1dxtd_	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

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**

structures

did_	fid
d2rs51_	1.002.007
d1imr__	1.010.002
d1pyib1	1.007.030
d1dxtd_	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

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

	key	key						
Table 1	gid	TrgStrt	TrgStop	did		Table 2	did	fid
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) 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

ER-diagrams

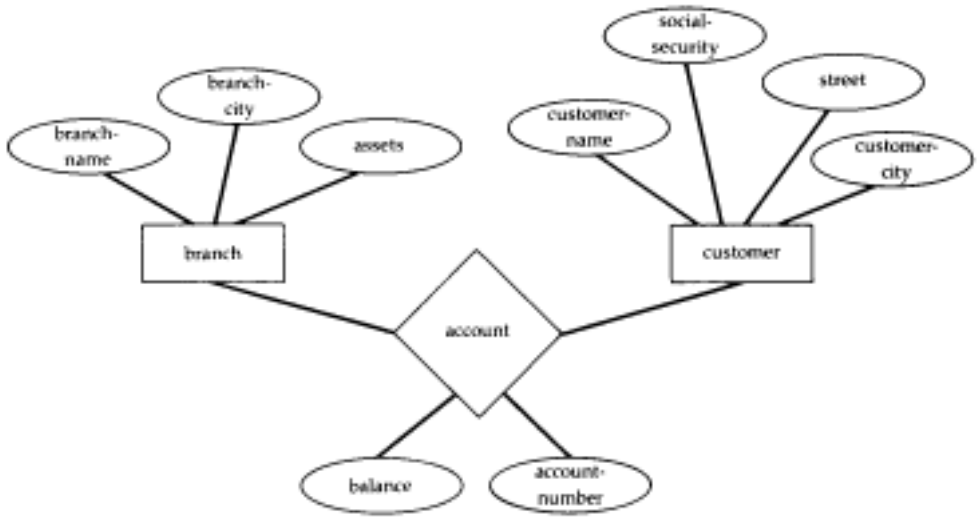


Figure 2.23 E-R diagram with *account* as a relationship set.

- Korth & Silberschatz
  - ◊ branch <=> matches
  - ◊ customer <=> folds
  - ◊ linked by
    account <=> structures

Aggregate Functions--Statistics on Attributes

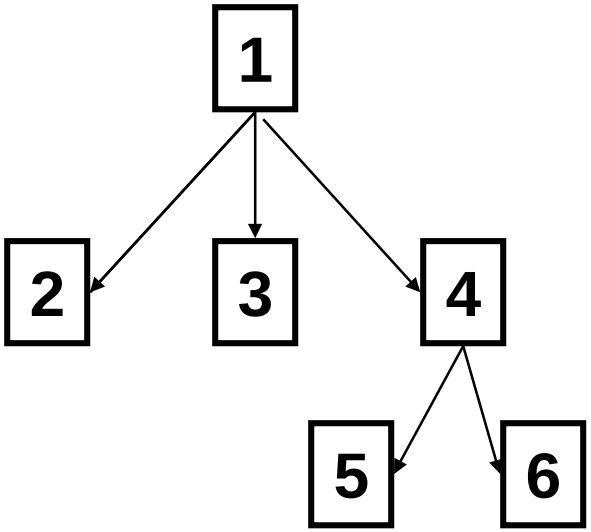
- Query Statistics
  - ◊ select gid, count (distinct did) from matches
  - ◊ select max(N_hlx) from folds where N_beta = 0
- How many matches to globins in the E. coli genome
- Complex Query by nesting selections
  - ◊ F <= select fid from folds where name contains “globin”
  - ◊ D <= select did from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

RDB's can encode complex data: Encoding Trees in RDBs

- Consider table below

N	P
1	0
2	1
3	1
4	1
5	3
6	3

N = node, P=parent



Further topics

- Transactions
  - ◊ Genome Centers and United Airlines!
- Security
- Object Databases
- Indexes and hash tables
- Query Optimization

Forms & reports [user views]

Normalization

- Eliminate Redundancy
- Allow Consistent Updating