A Brief Overview of DataFoundry

Terence Critchlow

Center for Applied Scientific Computing

Lawrence Livermore National Laboratory

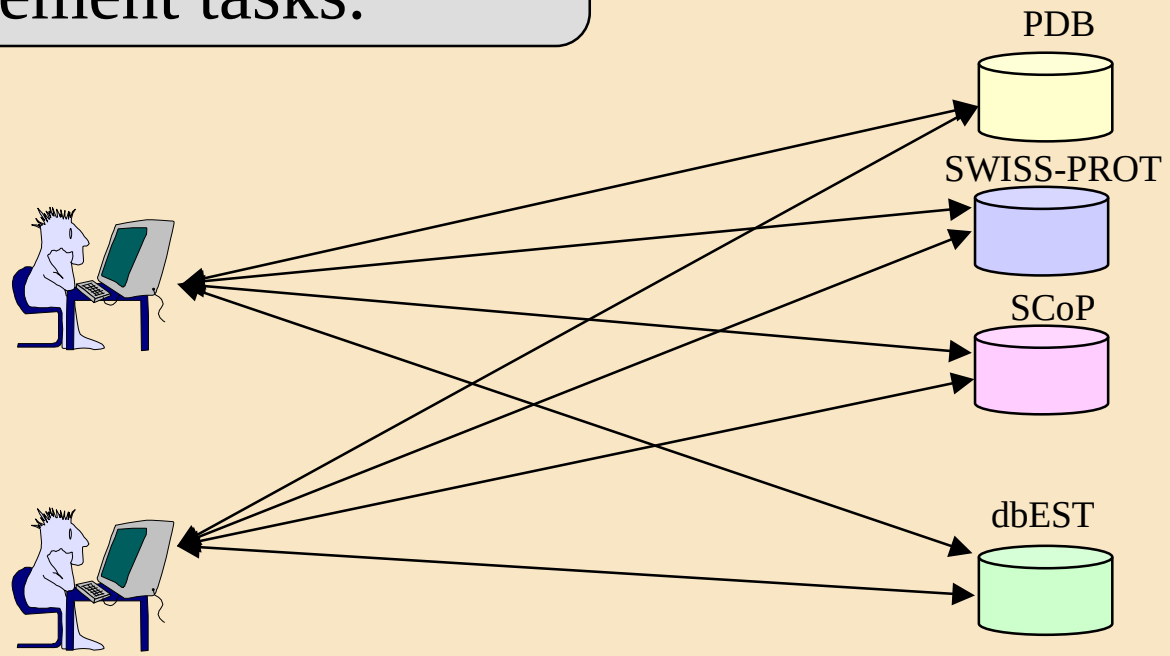
www.llnl.gov/CASC/people/critchlow

UCRL-PRES-144501

Currently, accessing genomics data is challenging.

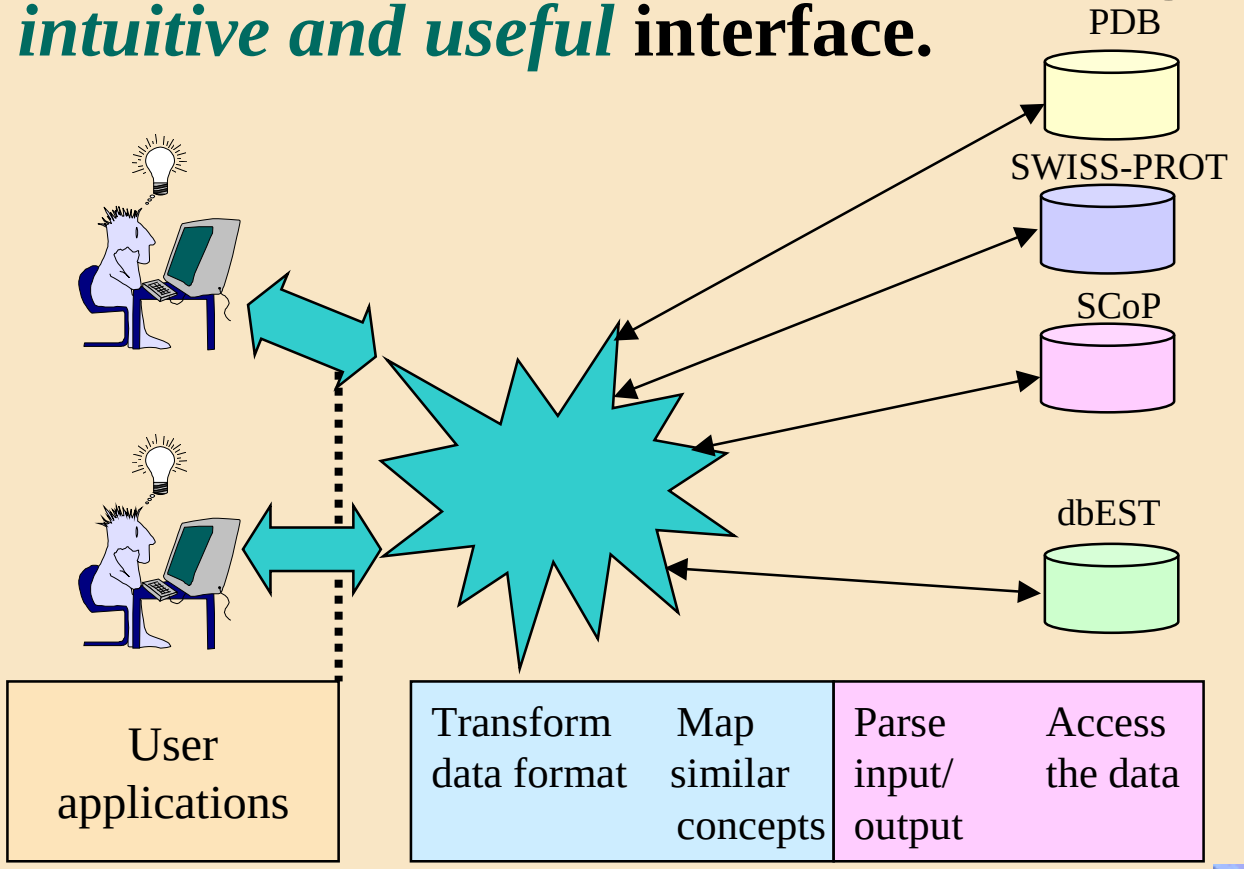
The user is required to perform all data management tasks.

Different users end up doing the same thing.

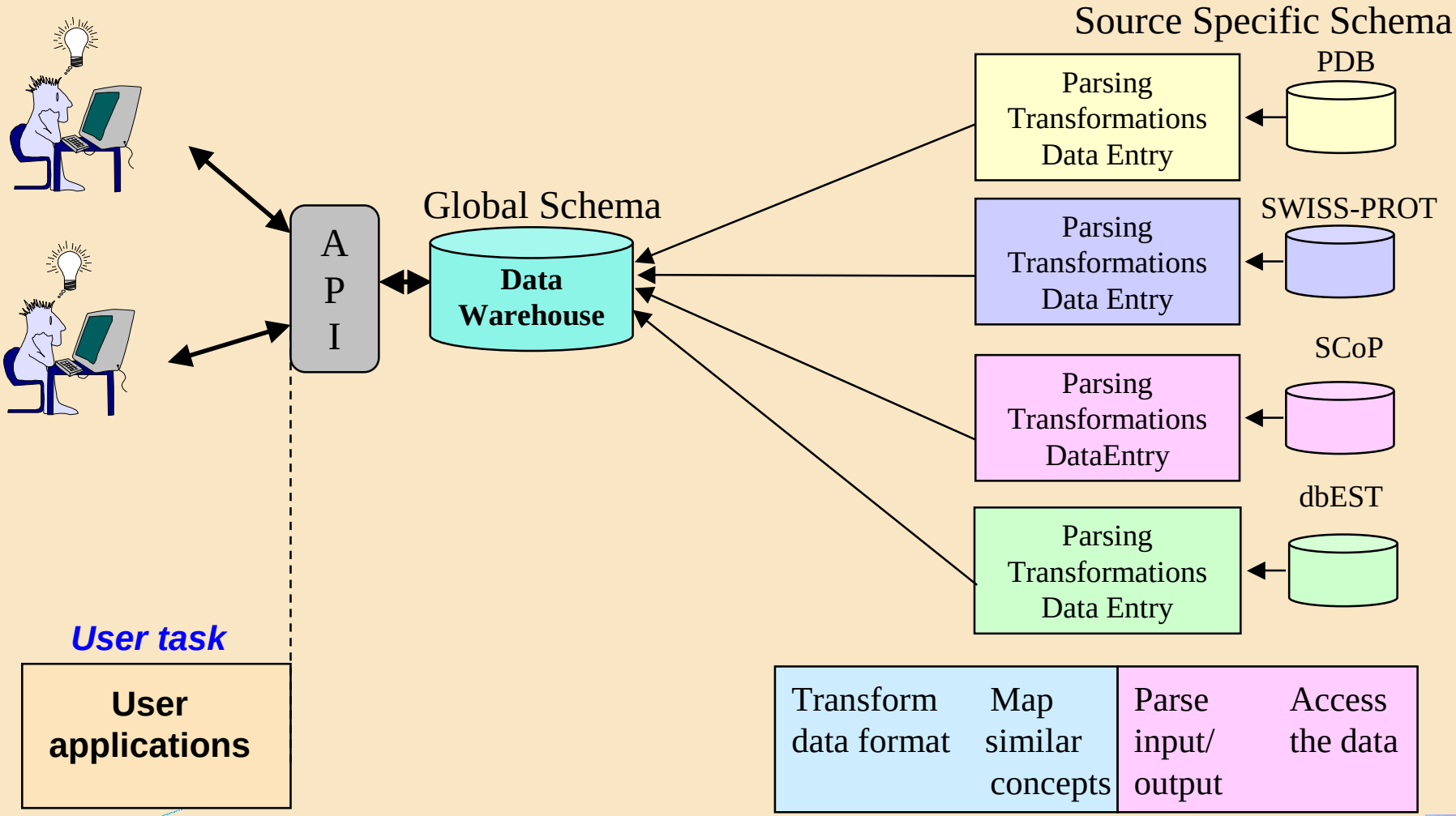


What is our ideal environment?

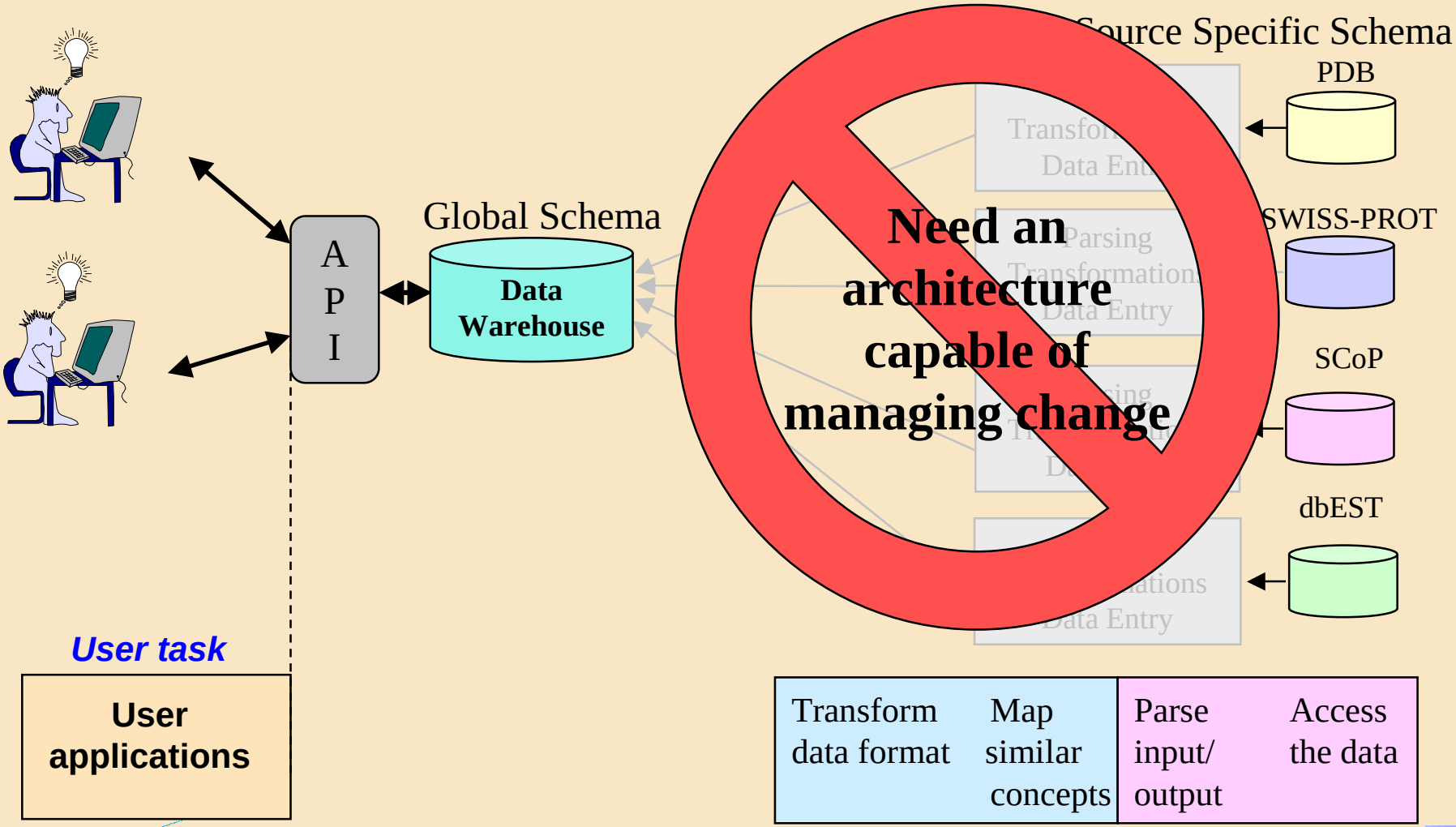
A *single* location that provides *effective* access to a *consistent* view of data from *many* sources through an *intuitive and useful* interface.



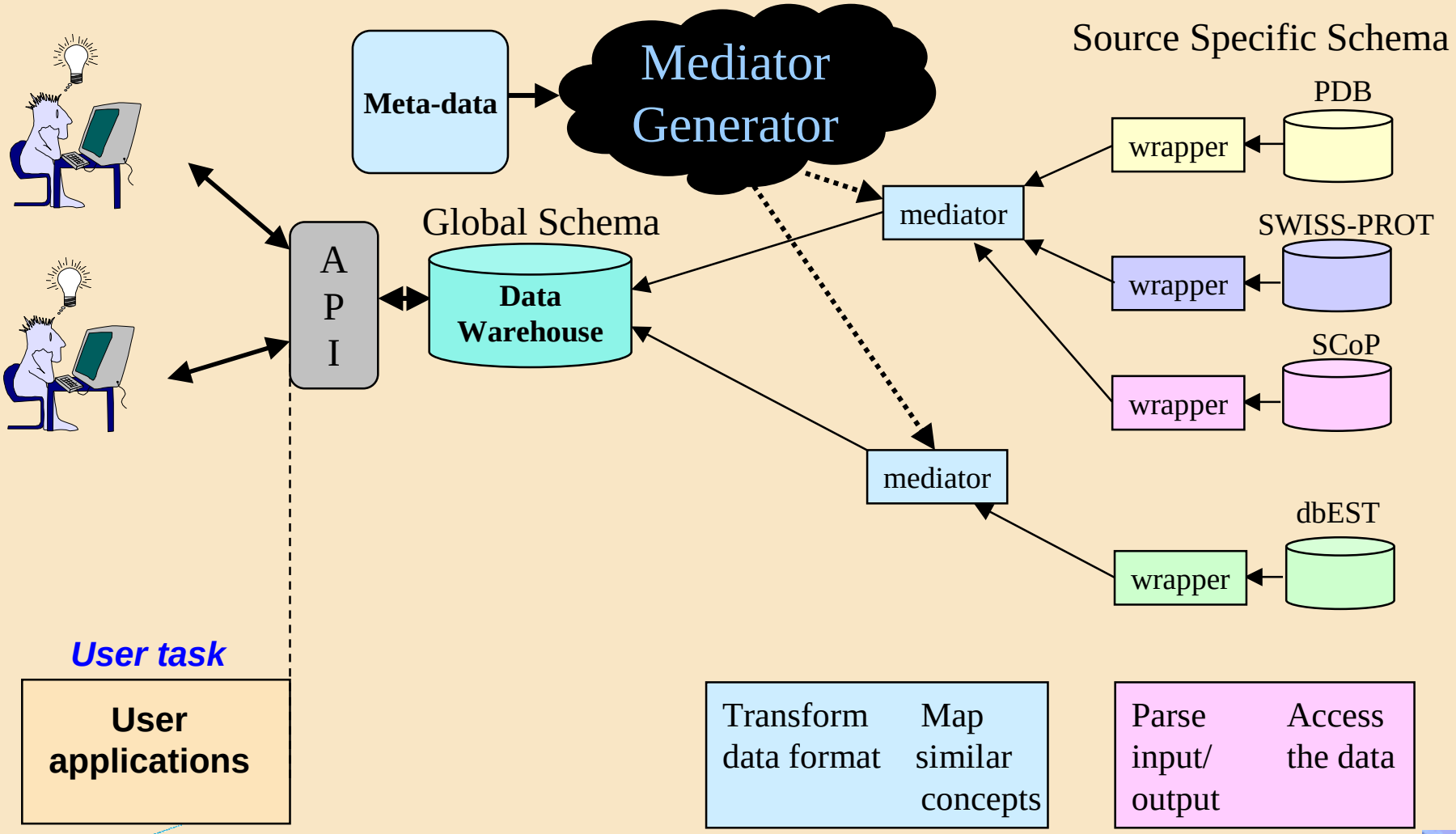
Typically data warehouses can provide that functionality.



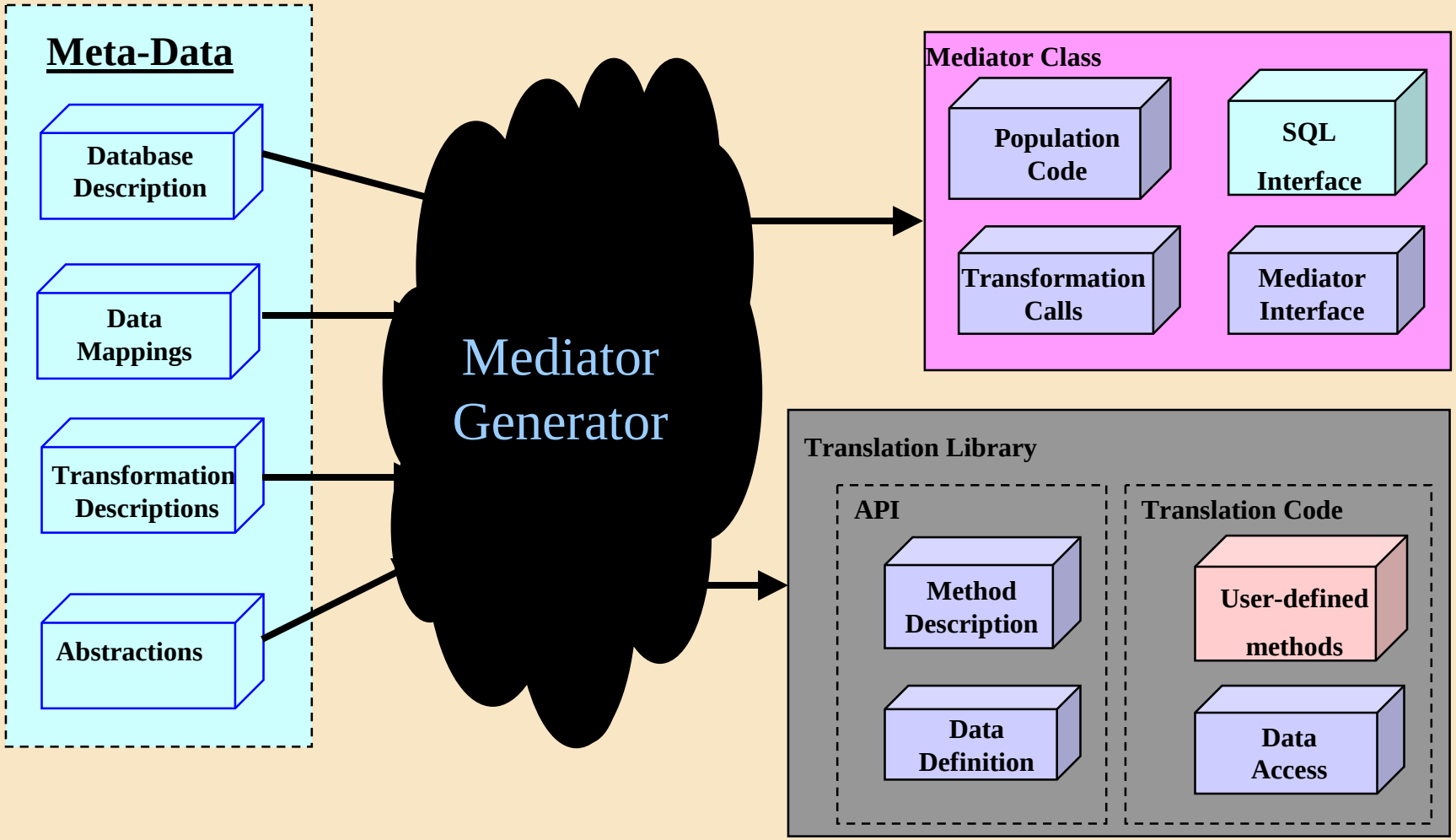
When a source changes, propagating the change can be time consuming.



DataFoundry separates these tasks.



The mediator generator translates the meta-data into C++ code .



Results:

Integrating SCoP into warehouse that already contains PDB and SWISS-PROT.

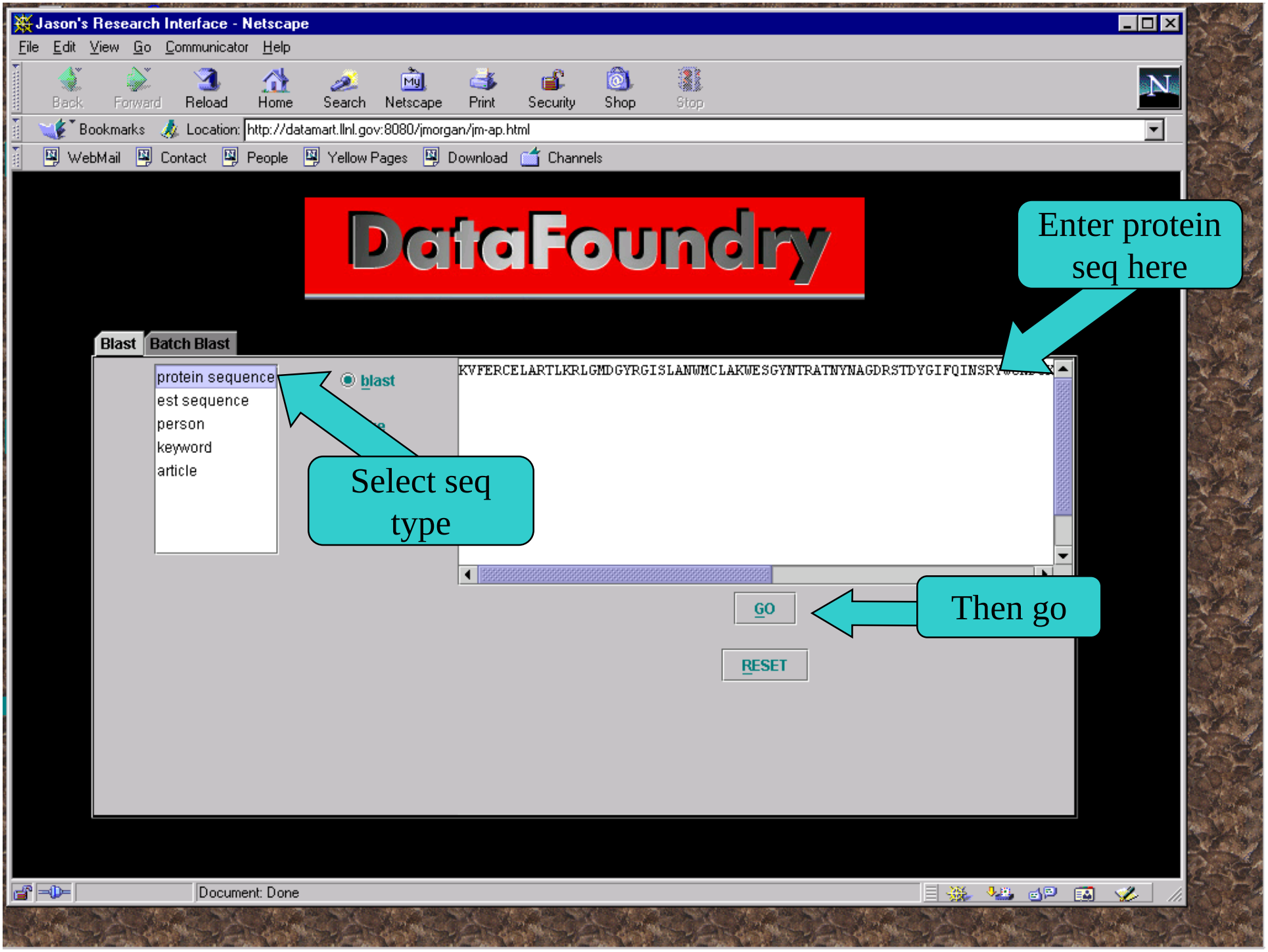
Activity/ integration style	manual	meta-data	diff	%diff
understanding SCOP	2.0	2.0	0.0	0
writing wrapper	4.5	2.5	2.0	44%
modifying schema	0.5	0.5	0.0	0
writing mediator	4.0	0.0	4.0	---
modifying meta-data	0.0	1.0	(1.0)	---
total time in days	11.0	6.0	5.0	45%

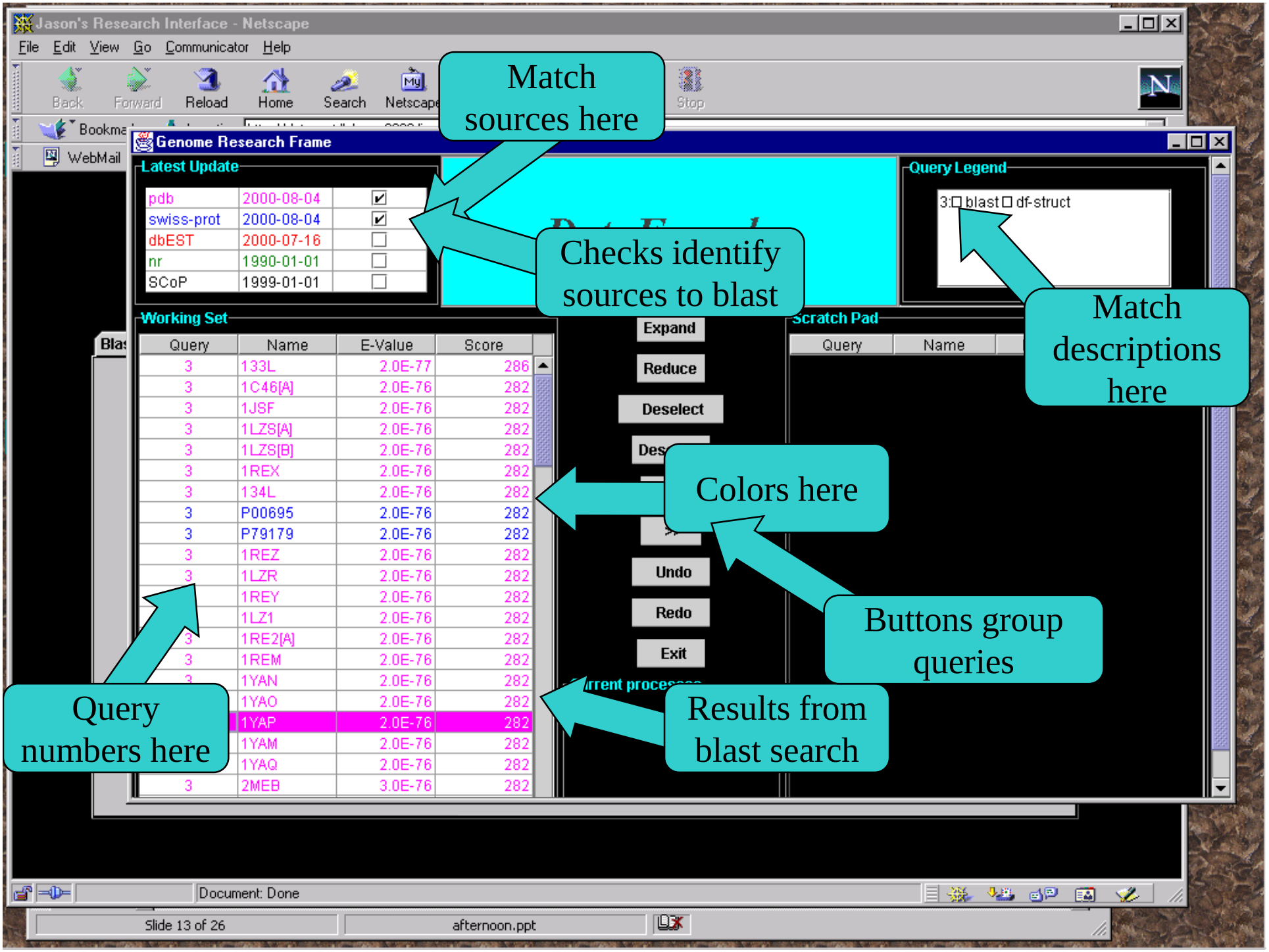
Technology has been licensed.

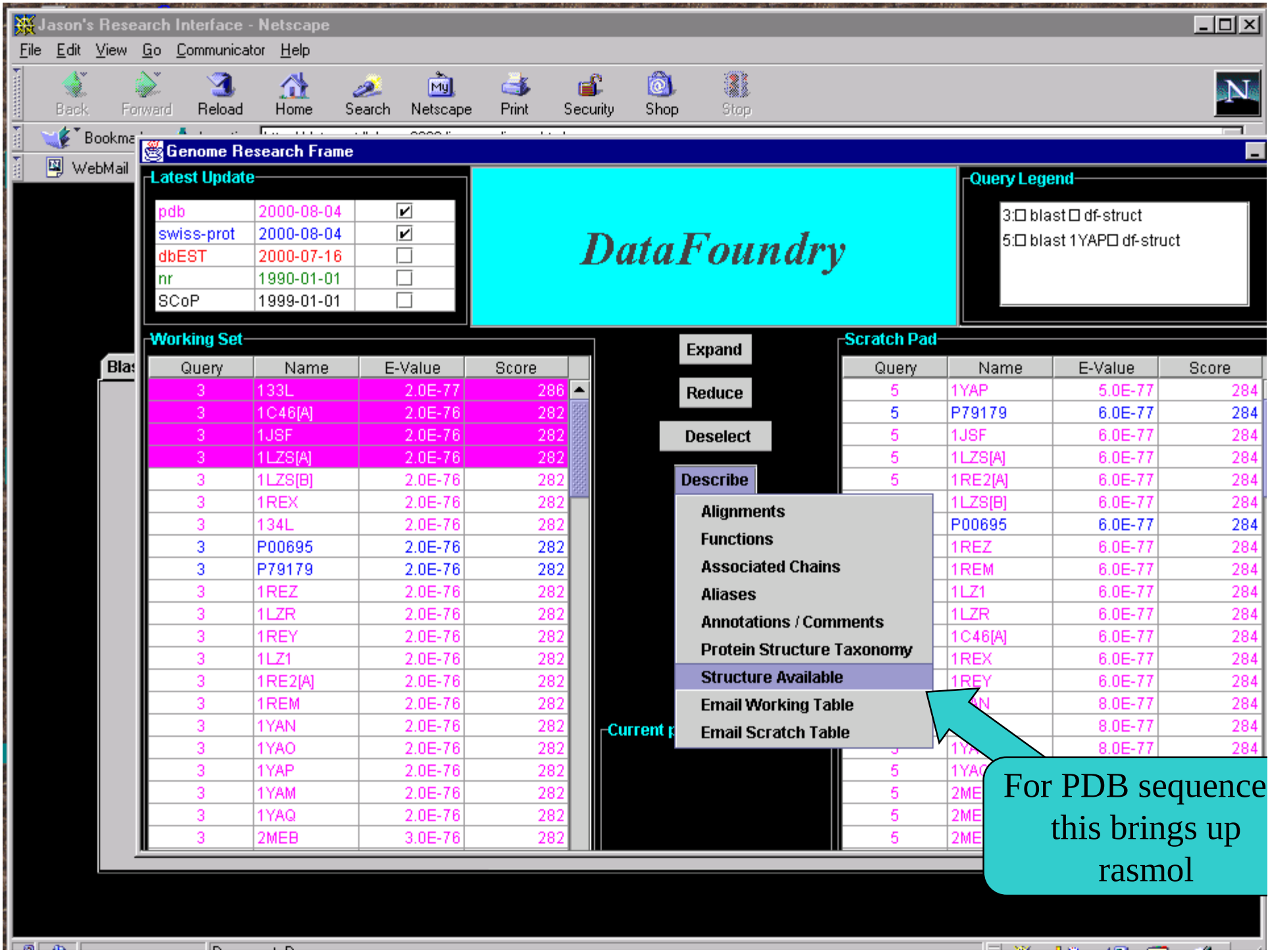
Once the back-end warehouse was in place, we developed a useful interface to interact with it.

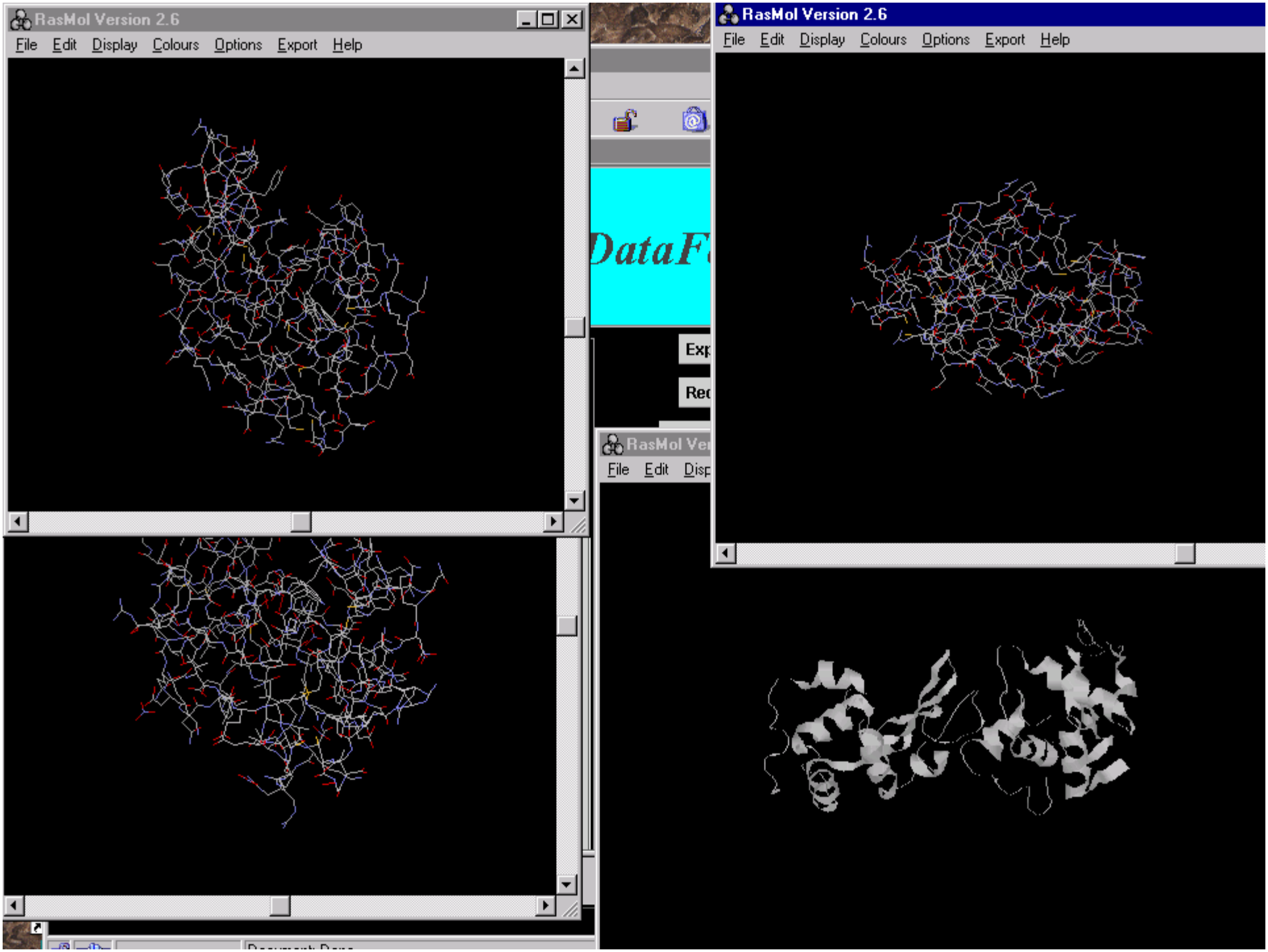
- **Original applet interface allows exploration of a single sequence.**
  - **Blasts against several sources at once**
  - **Complex and iterative queries**
  - **Connections to Rasmol (other programs can be added)**

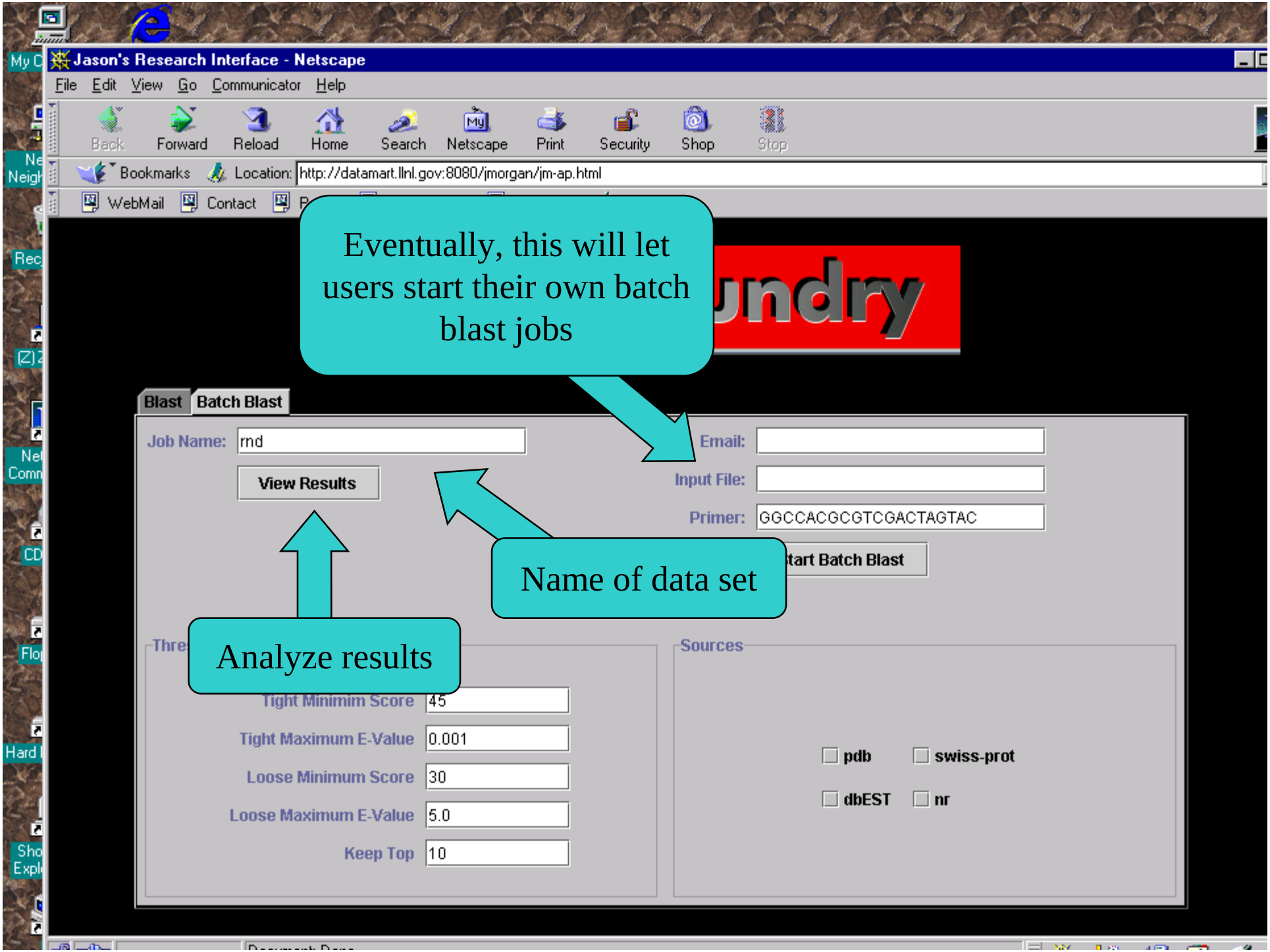
- **New interface allows user to interact with the results of a batch blast.**
  - **Developed in conjunction with Matt and Allen**
  - **Goal: easily identify those clones whose function is known**

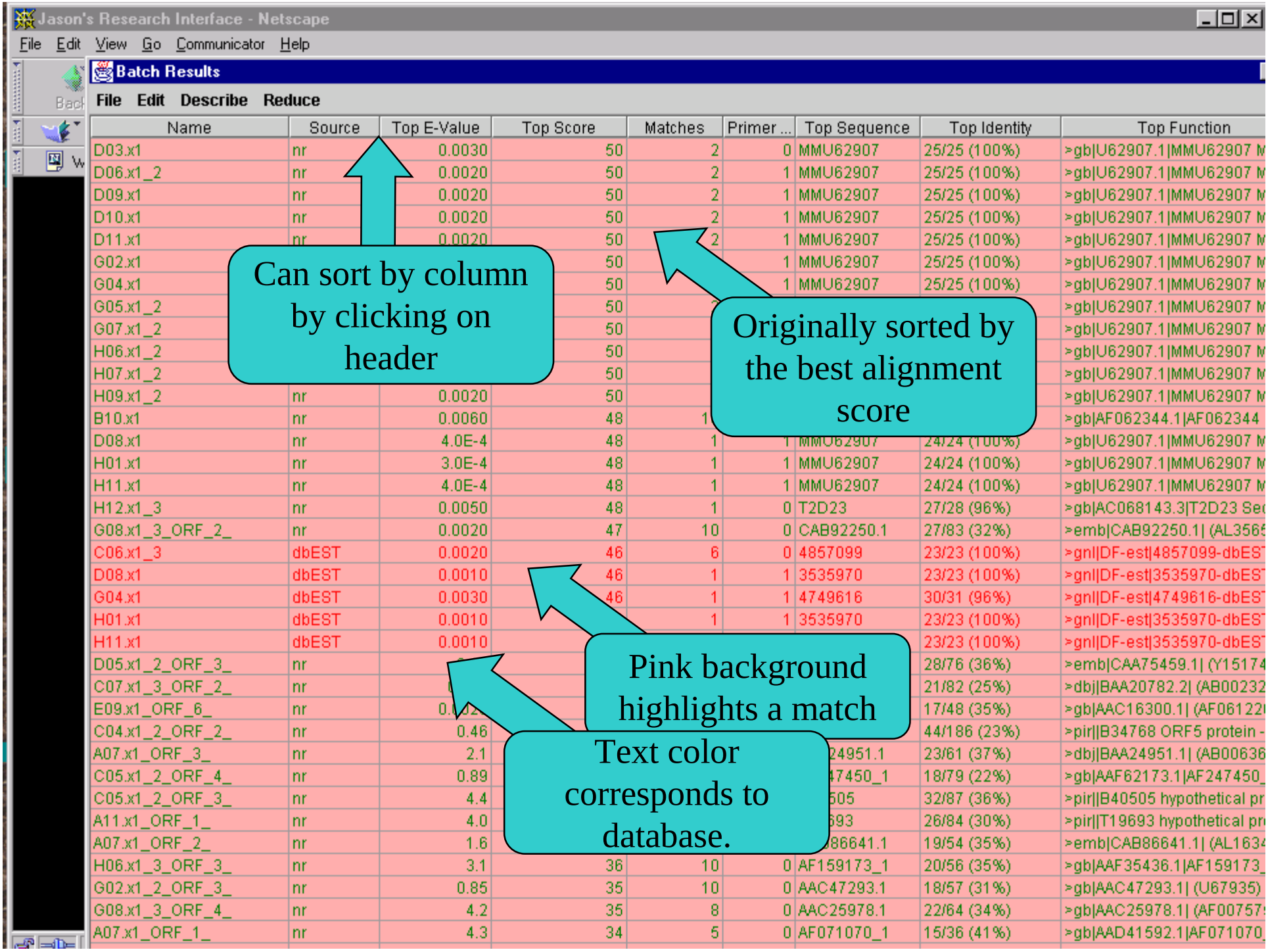

Jason's Research Interface - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: http://datamart.llnl.gov:8080/jmorgan/jm-ap.html
WebMail Contact People Yellow Pages Download Channels
DataFoundry
Enter protein seq here
Blast
Batch Blast
protein sequence
est sequence
person
keyword
article
blast
KVFERCELARTLKRLGMDGYRGISLANWMCLAKWESGYNTRATNYNAGDRSTDYGIFQINSRY
Select seq type
GO
Then go
RESET
Document: Done


Jason's Research Interface - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Stop
Bookmarks
WebMail
Genome Research Frame
Latest Update
pdb 2000-08-04
swiss-prot 2000-08-04
dbEST 2000-07-16
nr 1990-01-01
SCoP 1999-01-01
Query Legend
3: blast df-struct
Working Set
Query Name E-Value Score
3 133L 2.0E-77 286
3 1C46[A] 2.0E-76 282
3 1JSF 2.0E-76 282
3 1LZS[A] 2.0E-76 282
3 1LZS[B] 2.0E-76 282
3 1REX 2.0E-76 282
3 134L 2.0E-76 282
3 P00695 2.0E-76 282
3 P79179 2.0E-76 282
3 1REZ 2.0E-76 282
3 1LZR 2.0E-76 282
1REY 2.0E-76 282
1LZ1 2.0E-76 282
1RE2[A] 2.0E-76 282
1REM 2.0E-76 282
1YAN 2.0E-76 282
1YAO 2.0E-76 282
1YAP 2.0E-76 282
1YAM 2.0E-76 282
1YAQ 2.0E-76 282
3 2MEB 3.0E-76 282
Expand
Reduce
Deselect
Undo
Redo
Exit
Scratch Pad
Query Name
Match sources here
Checks identify sources to blast
Match descriptions here
Colors here
Buttons group queries
Query numbers here
Results from blast search
Document: Done
Slide 13 of 26
afternoon.ppt


Jason's Research Interface - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Genome Research Frame
Latest Update
pdb 2000-08-04
swiss-prot 2000-08-04
dbEST 2000-07-16
nr 1990-01-01
SCoP 1999-01-01
DataFoundry
Query Legend
3: blast df-struct
5: blast 1YAP df-struct
Working Set
Query Name E-Value Score
3 133L 2.0E-77 286
3 1C46[A] 2.0E-76 282
3 1JSF 2.0E-76 282
3 1LZS[A] 2.0E-76 282
3 1LZS[B] 2.0E-76 282
3 1REX 2.0E-76 282
3 134L 2.0E-76 282
3 P00695 2.0E-76 282
3 P79179 2.0E-76 282
3 1REZ 2.0E-76 282
3 1LZR 2.0E-76 282
3 1REY 2.0E-76 282
3 1LZ1 2.0E-76 282
3 1RE2[A] 2.0E-76 282
3 1REM 2.0E-76 282
3 1YAN 2.0E-76 282
3 1YAO 2.0E-76 282
3 1YAP 2.0E-76 282
3 1YAM 2.0E-76 282
3 1YAQ 2.0E-76 282
3 2MEB 3.0E-76 282
Expand
Reduce
Deselect
Describe
Alignments
Functions
Associated Chains
Aliases
Annotations / Comments
Protein Structure Taxonomy
Structure Available
Email Working Table
Email Scratch Table
Scratch Pad
Query Name E-Value Score
5 1YAP 5.0E-77 284
5 P79179 6.0E-77 284
5 1JSF 6.0E-77 284
5 1LZS[A] 6.0E-77 284
5 1RE2[A] 6.0E-77 284
1LZS[B] 6.0E-77 284
P00695 6.0E-77 284
1REZ 6.0E-77 284
1REM 6.0E-77 284
1LZ1 6.0E-77 284
1LZR 6.0E-77 284
1C46[A] 6.0E-77 284
1REX 6.0E-77 284
1REY 6.0E-77 284
8.0E-77 284
8.0E-77 284
8.0E-77 284
5 1YAQ
5 2ME
5 2ME
5 2ME
For PDB sequence this brings up rasmol

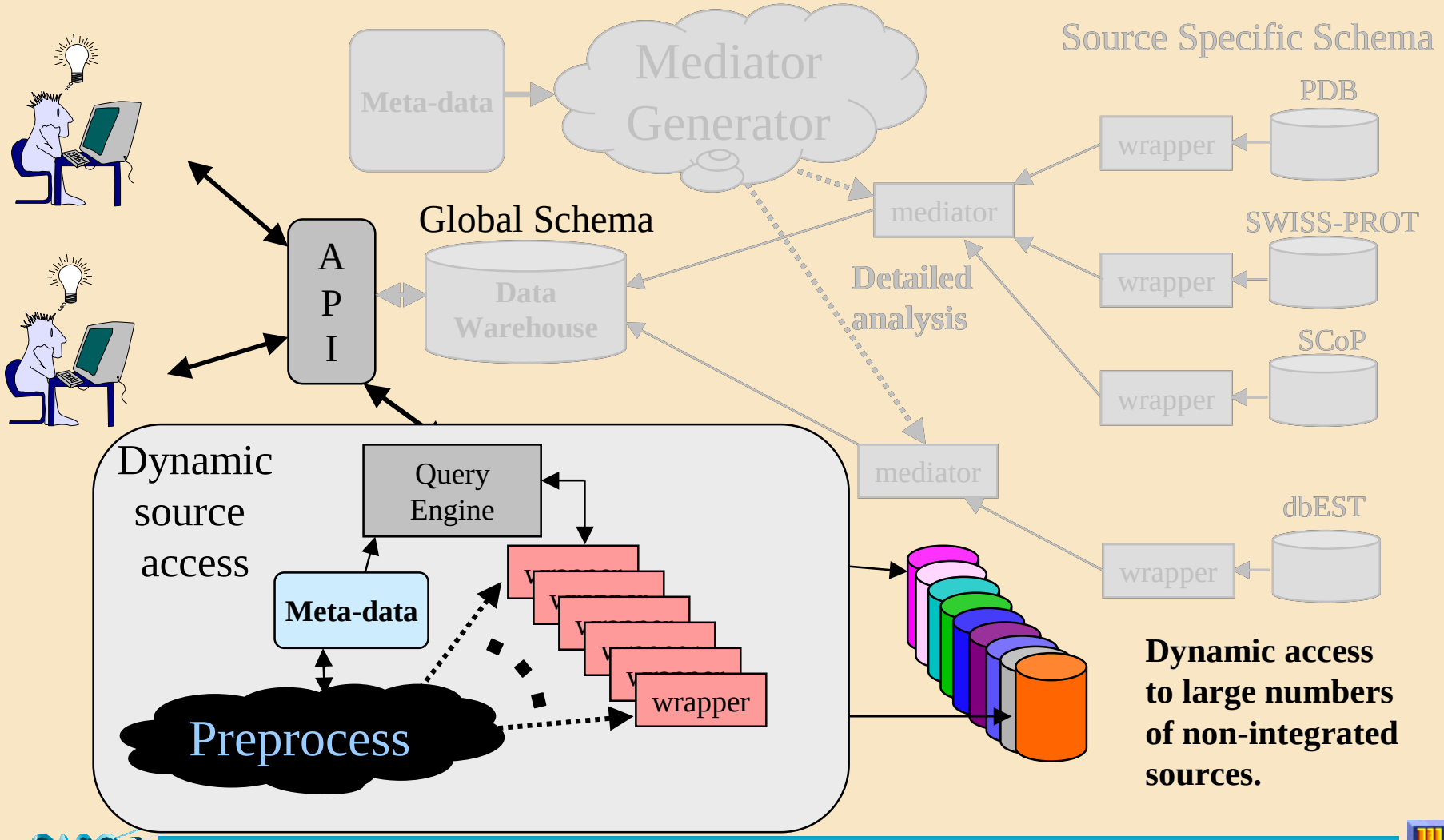

RasMol Version 2.6
File Edit Display Colours Options Export Help
RasMol Version 2.6
File Edit Display Colours Options Export Help
RasMol Ver
File Edit Disp


Jason's Research Interface - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: http://datamart.llnl.gov:8080/jmorgan/jm-ap.html
WebMail Contact
Eventually, this will let users start their own batch blast jobs
Blast
Batch Blast
Job Name: rnd
View Results
Email:
Input File:
Primer: GGCCACGCGTCGACTAGTAC
Name of data set
Analyze results
Sources
Tight Minimim Score 45
Tight Maximum E-Value 0.001
Loose Minimum Score 30
Loose Maximum E-Value 5.0
Keep Top 10
pdb
swiss-prot
dbEST
nr



Name	Source	Top E-Value	Top Score	Matches	Primer ...	Top Sequence	Top Identity	Top Function
D03.x1	nr	0.0030	50	2	0	MMU62907	25/25 (100%)	>gb|U62907.1|MMU62907 M
D06.x1_2	nr	0.0020	50	2	1	MMU62907	25/25 (100%)	>gb|U62907.1|MMU62907 M
D09.x1	nr	0.0020	50	2	1	MMU62907	25/25 (100%)	>gb|U62907.1|MMU62907 M
D10.x1	nr	0.0020	50	2	1	MMU62907	25/25 (100%)	>gb|U62907.1|MMU62907 M
D11.x1	nr	0.0020	50	2	1	MMU62907	25/25 (100%)	>gb|U62907.1|MMU62907 M
G02.x1			50		1	MMU62907	25/25 (100%)	>gb|U62907.1|MMU62907 M
G04.x1			50		1	MMU62907	25/25 (100%)	>gb|U62907.1|MMU62907 M
G05.x1_2			50					>gb|U62907.1|MMU62907 M
G07.x1_2			50					>gb|U62907.1|MMU62907 M
H06.x1_2			50					>gb|U62907.1|MMU62907 M
H07.x1_2			50					>gb|U62907.1|MMU62907 M
H09.x1_2	nr	0.0020	50					>gb|U62907.1|MMU62907 M
B10.x1	nr	0.0060	48	1				>gb|AF062344.1|AF062344
D08.x1	nr	4.0E-4	48	1	1	MMU62907	24/24 (100%)	>gb|U62907.1|MMU62907 M
H01.x1	nr	3.0E-4	48	1	1	MMU62907	24/24 (100%)	>gb|U62907.1|MMU62907 M
H11.x1	nr	4.0E-4	48	1	1	MMU62907	24/24 (100%)	>gb|U62907.1|MMU62907 M
H12.x1_3	nr	0.0050	48	1	0	T2D23	27/28 (96%)	>gb|AC068143.3|T2D23 Se
G08.x1_3_ORF_2_	nr	0.0020	47	10	0	CAB92250.1	27/83 (32%)	>emb|CAB92250.1| (AL356
C06.x1_3	dbEST	0.0020	46	6	0	4857099	23/23 (100%)	>gnl|DF-est|4857099-dbES
D08.x1	dbEST	0.0010	46	1	1	3535970	23/23 (100%)	>gnl|DF-est|3535970-dbES
G04.x1	dbEST	0.0030	46	1	1	4749616	30/31 (96%)	>gnl|DF-est|4749616-dbES
H01.x1	dbEST	0.0010		1	1	3535970	23/23 (100%)	>gnl|DF-est|3535970-dbES
H11.x1	dbEST	0.0010					23/23 (100%)	>gnl|DF-est|3535970-dbES
D05.x1_2_ORF_3_	nr						28/76 (36%)	>emb|CAA75459.1| (Y15174
C07.x1_3_ORF_2_	nr						21/82 (25%)	>dbj|BAA20782.2| (AB00232
E09.x1_ORF_6_	nr	0.0					17/48 (35%)	>gb|AAC16300.1| (AF06122
C04.x1_2_ORF_2_	nr	0.46					44/186 (23%)	>pir||B34768 ORF5 protein -
A07.x1_ORF_3_	nr	2.1				24951.1	23/61 (37%)	>dbj|BAA24951.1| (AB00636
C05.x1_2_ORF_4_	nr	0.89				47450_1	18/79 (22%)	>gb|AAF62173.1|AF247450_
C05.x1_2_ORF_3_	nr	4.4				505	32/87 (36%)	>pir||B40505 hypothetical pr
A11.x1_ORF_1_	nr	4.0				693	26/84 (30%)	>pir||T19693 hypothetical pr
A07.x1_ORF_2_	nr	1.6				86641.1	19/54 (35%)	>emb|CAB86641.1| (AL163
H06.x1_3_ORF_3_	nr	3.1	36	10	0	AF159173_1	20/56 (35%)	>gb|AAF35436.1|AF159173_
G02.x1_2_ORF_3_	nr	0.85	35	10	0	AAC47293.1	18/57 (31%)	>gb|AAC47293.1| (U67935)
G08.x1_3_ORF_4_	nr	4.2	35	8	0	AAC25978.1	22/64 (34%)	>gb|AAC25978.1| (AF00757
A07.x1_ORF_1_	nr	4.3	34	5	0	AF071070_1	15/36 (41%)	>gb|AAD41592.1|AF071070_

We are moving towards a hybrid approach to information access.



Questions?

www.llnl.gov/CASC/people/critchlow