

5/16/10

Do-It-Yourself Bioinformatics with the FMI Galaxy Server

Hans-Rudolf Hotz (hrh@fmi.ch)

Friedrich Miescher Institute for Biomedical Research Basel, Switzerland http://www.fmi.ch/



Friedrich Miescher Institute for Biomedical Research (part of the Novartis Research Foundation)



~ 300 employees (~40 nationalities) 99 PhD students and 86 Post Docs



Bioinformatics / Computational Biology

four FTE providing support for....

the "average" lab scientist, using computers to:



the "modern" lab scientist, using computers to:



Friedrich Miescher Institute for Biomedical Research

the FMI Galaxy Servers:

http://galaxy.fmi.ch

- dedicated, virtualized linux server
- 4 cores, 32GB RAM
- storage via NFS (IBM DS4000)
- changeset: 3545 (March)
- external authentication

http://galaxy-dev.fmi.ch

- IBM x3755
- 2 dual core AMD, 16GB RAM
- storage via NFS (IBM DS4000)
- changeset: 3545

temporary test servers

on desktops/laptops and/or different port numbers



the FMI Galaxy Server (http://galaxy.fmi.ch)

Tools

67 users (have logged in at least once)

- ~ 13 jobs executed / day (Oct. till May)
 - 24 % upload.py

...

- 4.2 % gops_intersect.py
- 4.1 % cutWrapper.py
- 3.3 % fixedValueColumn.py
- 2.8 % motif_search.pl
- 2.6 % data_source.py
- 2.6 % remove_beginning.pl





self-written/added tools

motif search

uploading data

accessing data "outside of Galaxy"

Next Gen Sequencing



motif search

request:

show me all human UniProtKB/Swiss-Prot entries which contain the following motif: "W[ST]XW"



.....what about a pre-installed copy of human UniProtKB/Swiss-Prot entries?
.....can you provide information about subcellular location?
.....provide domain and 3D-structure info
.....and I want to know the sequence to the left and right of the motif



Motif Search

Your motif:

W[ST]XW

see below for examples

protein file:

human (reviewed) UniProtKB

O mouse (reviewed) UniProtKB

C. elegans (reviewed) UniProtKB

O P. falciparum (isolate 3D7) (reviewed) UniProtKB

O "human TSR proteins"

O "pompep" - limited results mark-up

display additional ammino acids to the left and the right (in lower case) of the found motif (max 6 amino acids, not possible for pompep):

4

display proteins without hits:

• "no"

○ "yes" - you could end up with >20,000 lines!

Execute



Result for "W[ST]XW" search in UniProtKB/Swiss-Prot (release 2010_04) for *Homo sapiens*

Accession	Description	start	end	sequence	overlaps with	subcellular location	3D-structure info
<u>P46098</u>	5-hydroxytryptamine receptor 3A	472	475	lvml WSIW qya	-	Cell junction. synapse. postsy	78
Q99758	ATP-binding cassette sub-family A member 3	305	308	swlh WSAW fllf	-	Membrane: Multi-pass membrane	-
<u>P22570</u>	NADPH:adrenodoxin oxidoreductase, mitochondrial	11	14	rwwgWSAWprtr	-	Mitochondrion matrix.	-
<u>Q5T2P9</u>	Arf-GAP with GTPase, ANK repeat and PH domain- containing protein 10	646	649	lltg WTSW pemp	-	-	-0
Q5VTM2	Arf-GAP with GTPase, ANK repeat and PH domain- containing protein 9	691	694	lltg WTSW pemp	-	-	-0
<u>Q8N1P7</u>	Absent in melanoma 1-like protein	194	197	eypd WSHW ggyd	Beta/gamma crystallin 'Greek key' 4 (169 - 211)	a	<u>HSSP</u>
<u>P26442</u>	Autocrine motility factor receptor, isoform 1	153	156	sgvd WTAW gggr	-	Membrane: Single-pass type I m	-
P04424	Argininosuccinate lyase	169	172	qpir WSHW ilsh	-	÷	1AOS 1K62
Q9HBZ2	Aryl hydrocarbon receptor nuclear translocator 2	657	660	psev WSQW qsqh	-	<u>Nucleus</u> (Potential).	-
P15848	Arylsulfatase B	319	322	rgrkWSLWeggv	-	Lysosome.	1FSU
Q9ULK2	Ataxin-7-like protein 1	19	22	lgkpWSSWidaa	5		-

how does it work.....



Friedrich Miescher Institute for Biomedical Research

motif search

similar tool for genomes

- using EMBOSS:fuzznuc
- connecting to public ensembl database to get neighboring genes, exon/intron, etc



too slow!

- currently re-writing using the Bioconductor package : Biostring



individual 'public' sequences

Fetch a sequence	
Accession:	_
P51003	
see below for examples	4
database:	
💽 UniProtKB	
🔘 RefSeq	
🔘 EMBL/Genbank	
format:	
💽 fasta	
🔘 full entry	
Execute	
Excedite	



Tools

Get Data

- <u>Upload File</u> from your computer
- Load local abi reads from the FMI sequencing facility
- · Access Libraries stored locally
- <u>UCSC Main</u> table browser
- BioMart Central server
- Fetch a sequence from UniProtKB, RefSeq, EMBL
- Text Manipulation FASTA manipulation FASTQ manipulation Filter and Sort Join, Subtract and Group

>sp|P51003|PAPOA_HUMAN Poly(A) polymerase alpha OS=Homo sapiens GN=P MPFPVTTQGSQQTQPPQKHYGITSPISLAAPKETDCVLTQKLIETLKPFGVFEEEEELQR RILILGKLNNLVKEWIREISESKNLPQSVIENVGGKIFTFGSYRLGVHTKGADIDALCVA PRHVDRSDFFTSFYDKLKLQEEVKDLRAVEEAFVPVIKLCFDGIEIDILFARLALQTIPE DLDLRDDSLLKNLDIRCIRSLNGCRVTDEILHLVPNIDNFRLTLRAIKLWAKRHNIYSNI LGFLGGVSWAMLVARTCQLYPNAIASTLVHKFFLVFSKWEWPNPVLLKQPEECNLNLPVW DPRVNPSDRYHLMPIITPAYPQQNSTYNVSVSTRMVMVEEFKQGLAITDEILLSKAEWSK LFEAPNFFQKYKHYIVLLASAPTEKQRLEWVGLVESKIRILVGSLEKNEFITLAHVNPQS FPAPKENPDKEEFRTMWVIGLVFKKTENSENLSVDLTYDIQSFTDTVYRQAINSKMFEVD MKIAAMHVKRKQLHQLLPNHVLQKKKKHSTEGVKLTALNDSSLDLSMDSDNSVSPSTS ATKTSPLNSSGSSQGRNSPAPAVTAASVTNIQATEVSVPQVNSSESSGGTSSESIPQTAT QPAISPPPKPTVSRVVSSTRLVNPPPRSSGNAATSGNAATKIPTPIVGVKRTSSPHKEES PKKTKTEEDETSEDANCLALSGHDKTEAKEQLDTETSTTQSETIQTAASLLASQKTSSTD LSDIPALPANPIPVIKNSIKLRLNR





'ABI PRISM 3730' sequencer data (yes, they are still in use.....)

 everybody (ie. each desktop) has access to the reads from his/her research group

drwxr-x+	2	dnaseq	gpielage	3484	2010-05-05	13:30	pielage.j
drwxr-x+	2	dnaseq	grijli	1134	2010-04-19	13:30	rijli.f
drwxr-x+	11	dnaseq	groska	10201	2010-03-26	09:00	roska.b
drwxr-x+	16	dnaseq	gschub	9712	2010-05-04	14:30	schubeler.d

```
02620_D03_193434SCD_013.ab1
02620_D03_193434SCD_013.seq
02620_D04_193435SCD_013.ab1
02620_D04_193435SCD_013.seq
02620_D05_193436SCD_021.seq
02620_D05_193436SCD_021.ab1
```



individual file uploads possible, but not very efficient



request:

- keep the directory structure
- maintain security

problem:

- using 'Data libraries' and assigning 'roles' to each file seemed un-doable
- providing different tools (ie. GUI) for each user is not possible

solution:

- give "Galaxy" access to all the reads
- use *\$userEmail* to provide security



```
<tool id="load_reads" name="Load local abi reads" version="1">
  <description>from the FMI sequencing facility</description>
  <command interpreter="perl">load_abi_reads.pl $reads $dir $userEmail
$output
```

```
</command>
```

```
<inputs>
```

```
<page>
```

```
<param name="dir" type="select" label="reads directories"
help="Use tickboxes to select a directory with reads"
display="radio"</pre>
```

```
dynamic options="ds readDirOptions()"/>
```

```
</page>
```

```
<page>
```

```
<param name="reads" type="select" label="available reads"
help="Use tickboxes to select reads"
display="checkboxes" multiple="true"</pre>
```

```
dynamic options="ds readsOptions(dir)"/>
```

```
</page>
```

```
</inputs>
```

```
<outputs>
```

```
<data format="fasta" name="output" label="Reads" />
```

</outputs>

```
<code file="more_code_for_load_abi_reads.py" />
```

<help>

</help>

</tool>



Load local abi reads (step 1 of 2)

reads directo	ri	e	S:
---------------	----	---	----

🔘 alcedo.j
🔘 arber.s
O bentires.m
🔘 buehler.m
🔘 caroni.p
🔘 chiquet.r
⊖ ciosk.r
🔘 filipowicz.w
⊖ friedrich.r
🕘 gasser.s
🔘 grosshans.h
O hemmings.b
🔘 hofsteenge.j
O hynes.n
🔘 luthi.a
🔘 matthias.p
🔘 matus.a
🔘 meins.f
🔘 monard.d
🔘 nagamine.y
🔘 oertner.t
🔘 peters.a
🔘 pielage.j
🔘 rijli.f
O roska.b

• schubeler.d

🔾 sun.f

O thoma.n

Use tickboxes to select a directory with reads

Next step

Load local abi reads (step 2 of 2)

available readss:

Select All Unselect All 01906_D08_136031SCD_029.seq 01906_D09_136032SCD_037.seq 01906_D10_136033SCD_037.seq 01906_D11_136034SCD_045.seq 01906_D12_136035SCD_045.seq 01906_E01_136036SCD_004.seq 01906_E02_136037SCD_004.seq 01906_E03_136038SCD_012.seq

02620_D02_193433SCD_005.seq
 02620_D03_193434SCD_013.seq
 02620_D04_193435SCD_013.seq
 02620_D05_193436SCD_021.seq
 02620_D06_193437SCD_021.seq
 02620_D07_193438SCD_029.seq
 Use tickboxes to select reads to be stored in a multi sequence fasta file

Execute



>02620_D03_193434SCD_013.ab1	History
>02620_D03_193434SCD_013.ab1 NNNNNNNNNNNNNNGCTGNNGGCGNNTNNGTTGGGTAACGCCAGGGTT GACGTTGTAAAACGACGGCCAGTGAATTGTAATACGACTCACTATAGGG CTGCAGGGGATAACTTCGTATAATGTATACTATAC	History History CGGATTGGGGGG CGGATCCA CTGTAGGGCAA CCAGGATGGAC CCAGGAAGAAAG CTCCCGGGTTTA TGTGTCTGCAG CAGGACGTGGC CAGGACGTGGC CAGGACGAGGAA CTGTGTCTGCAG CAGGACGTGGC CAGGACAAGGGAA
AGAAAATCCCTTCAGAATCCCTTTTAAGACGGTGGCAGAAGGACAATAC CTCGGCCATCATTGCTTTCCCTGAGCTCCAGCGTTCCTAATGCGAGGTC TCTCCTGCACTCAAATGAGAAGCTTATCNN >02620_D04_193435SCD_013.ab1 NNNNNNNNNNNNNNNTTTCNCACAGGAAACAGCTATGACCATGA TCGAANNNNNNCTCACTAAAGGGAACAAAAGCTGGTACCGGGCCCCCC ACTTCGTATAATGTATACTATAC	CAGATGCCTTT CGTCCCCCACC TTACGCCAAGC CTCGAGCCAAG CCTCGAGCCATA TGTAGGGCAAG CAGGTGAGGCAAG CAGGGTGAGCA CAGGAAGAAAGG CCCGGGTTTAC CGTGTCTGCAGC TGGACGTGGCC AACAAGGGAAA CAGATGCCTTTC CGCCCCCCCCCC

History	Options 💌
🖏 🗖 Talk	0 🖻
4: Reads 3 sequences, format ? Info:	● Ø X : fasta, database: ⊘ 📄
>02620_D03_193434SCD NNNNNNNNNNNNNNNGCTGN GACGTTGTAAAACGACGGCC CTGCAGGGGGATAACTTCGTA CTAGTAACGGCCGCCAGTGT GATTGTACAGACCAACCTTT	_013.ab1 NGGCGNNTNNGTTGGGTA AGTGAATTGTAATACGAC TAATGTATACTATAC
<u>3: P51003</u>	• / ×
2: Motif Search	• / %
<u>1: Motif Search</u>	• / %

FMI pipeline for tag based applications (Illumina GA II)



Friedrich Miescher Institute for Biomedical Research

FMI: DeepSeqRepository

- <u>Remove adaptor sequences</u> from 5'/3' ends of reads
- Import sample into deepSeq repository
- <u>Trigger annoation (alignment)</u> of a sample against reference sequences
- <u>Select samples</u> from the deepSeq repository
- <u>Create quality control plots</u> (repeatedness, error rate and aligned fraction)
- <u>Create full report</u> that classifies reads into biological types
- <u>Extract data</u> from the repository (expression levels, etc.)

the pipeline 'starts' with a fasta file:

'upload file'

'Access Libraries'





data is inside "Galaxy"





Remove adaptor sequences

Clean-up these sequences:

5: DM_100000_RNAseq.fa 🗘

Sequence to be removed from 5'-end of the reads:

If empty, nothing will be removed from the 5'-end of reads.

Sequence to be removed from 3'-end of the reads:

ACGTACGTACGT

If empty, nothing will be removed from the 3'-end of reads.

Execute



data is inside "Galaxy"

1004					
#	SUMMARY:				
#	Number of sequences:				
#	input sequences	:	100000	(100.0%)
#	unique sequences	:	89948	i	89.9%)
#				0.50	
#	Adaptor removal:				
#	full matches to 5'-adaptor	:	0	(0.0%)
#	inexact suffix matches to 5'-adaptor	:	0	(0.0%)
ŧ	exact suffix matches to 5'-adaptor	:	0	(0.0%)
ŧ	full matches to 3'-adaptor	:	23	(0.0%)
ŧ	inexact prefix matches to 3'-adaptor	:	116	(0.1%)
ŧ	exact prefix matches to 3'-adaptor	:	32615	(32.6%)
ŧ					
ŧ	total matches to 5'-adaptor	:	0	(0.0%)
ŧ	total matches to 3'-adaptor	:	32754	(32.8%)
ŧ					
ŧ	unique sequences (w/o adaptors)	:	89946	(89.9%)
¥					
¥	Sequence filtering:				
ŧ	5'-adaptor alignment too short	:	0	(0.0%)
ŧ	3'-adaptor alignment too short	:	0	(0.0%)
ŧ	too short	:	16	(0.0%)
ŧ	too many non-base characters	:	223	(0.2%)
ŧ	low entropy	:	44	(0.0%)
ŧ					
έ.	Final sequences:				
¥.	magaad		99717	1	99.7%)

History	Options 👻
🔁 🗖 Talk	4 🖻
<u>7:</u> DM 100000 RN/	■ Ø X Aseq.fa log
6: DM 100000 RN/ 89,673 sequence database: <u>?</u> Info: 	Aseq.fa cleaned s, format: fasta, O
>-1350550920:7:1: AAAAGTCGGGCAAGCAG >-1350550920:7:1: GAGAAGCCGCAGGCCTG >-1350550920:7:1: GTACAGCCATTGACCG	TTTT:945 1 SAACCCAAAAACAGTCG 1082:464 1 CTGCCTTGGATGCCACGCT T1:17 1 TAGCTGTCCAGCTCGAAG
<u>5:</u> DM 100000 RN/ 5.9 Mb, format: f	● Ø X Aseq.fa asta, database: <u>?</u> sta file

Import sample	
Import these sequences: 6: DM_100000_RNAseq.fa cleaned	
Sample identifier:	
galaxy_talk	
The sample identifier has to be unique.	
Sample description:	
a small example	
You can enter any description information into t	his free text field.
Execute	



data is outside "Galaxy"

loading sequences		History	Options 👻
filtering sequences sample ID: description unique sequences unique sq after filter total counts The new sample galaxy_t.	galaxy_talk hansrudolf.hotz@fmi.ch: a small example 89673 89673 99717 alk was successfully imported	 Talk <u>B:</u> <u>DM 100000 RN/</u> <u>import log</u> <u>7:</u> DM 100000 RN/ 	Aseq fa log
		<u>6:</u> DM 100000 RN/ <u>5:</u> DM 100000 RN/	 ● Ø X Aseq.fa cleaned ● Ø X Aseq.fa



data is outside "Galaxy"

galaxy	gbioinfo	4096	2010-05-11	14:50	galaxy_talk
galaxy	gbioinfo	4096	2010-05-06	13:46	galaxy_workshop
galaxy	gbioinfo	4096	2009-11-18	17:46	GM_8_com_H3K9me3
galaxy	gbioinfo	4096	2010-05-04	15:55	ME_K27me3
haruhotz	gbioinfo	4096	2010-05-05	09:03	ME_K27me3_2
lienertf	gbioinfo	4096	2009-12-08	09:08	CA_H3K9me2a_quant
•••					

cd galaxy_talk

. . .

galaxy gbioinfo402010-05-0613:46description.txtgalaxy gbioinfo40962010-05-0613:48errorLogsgalaxy gbioinfo40962010-05-0613:48fullReportsgalaxy gbioinfo40962010-05-0613:48mappingsgalaxy gbioinfo63049772010-05-0613:46seqs.tabgalaxy gbioinfo63049772010-05-0613:46seqs_unfi.tab

head description.txt

hansrudolf.hotz@fmi.ch: a small example



data is outside "Galaxy"

Sample identifier:	27
galaxy_talk	÷
The sample identifier to be annotated	d.
Genome assembly:	
dm3 (D.melanogaster Apr. 200	6, BDGP F
Select the reference genome for the	sample.
Annotation database:	
dmV01 (D.melanogaster v1) 🗘	
Select the reference annotation datab	ase for the sample



data is outside "Galaxy"





Sel	lect	sam	pl	es

Available samples:

Select All Unselect All

100205_433DMAAXX_6_b10 (ce6, sPombe :

sPombeV01, ceV01)

100205_433DMAAXX_6_b11 (sPombe :

sPombeV01)

100205_433DMAAXX_6_b12 (sPombe :

sPombeV01)

100205_433DMAAXX_6_b13 (sPombe :

sPombeV01)

100205_433DMAAXX_6_b14 (sPombe : sPombeV01)

fly_mRNA (dm3, ce6 : dmV01, ceV01)

galaxy_talk (dm3 : dmV01)

galaxy_workshop (dm3 : dmV01)

mouseTest (mm9, mm8 : mmV02)

Use tickboxes to select samples to be stored in a sample set









data is outside "Galaxy"

Create quality control plots
Sample selection:
Sample list
Do you want to work with a single sample or a sample list?
Sample list:
10: sample list
The list with sample identifiers to be used. If you don't see any items, use 'Select samples' first.
Genome assembly:
ce6 (C.elegans May 2008, WS190)
Select the reference genome assembly.
Annotation database:
ceV01 (C.elegans v1)
Select the reference annotation database.
Read length:
36
Analyze up to this many bases in each read alignment.
Maximum number of alignment errors:
2
This is the maxium number of errors to be expected in alignments.
Execute



data is outside "Galaxy"

Sample selection:	
Single sample 🗘	
Do you want to work with a single sample or a sample list?	
Sample identifier:	
galaxy_talk	
The sample identifier to be used.	
Genome assembly:	
dm3 (D.melanogaster Apr. 2006, BDGP I	
Select the reference genome assembly.	
Select the reference genome assembly. Annotation database:	
Select the reference genome assembly. Annotation database: dmV01 (D.melanogaster v1)	
Select the reference genome assembly. Annotation database: dmV01 (D.melanogaster v1) Select the reference annotation database.	
Select the reference genome assembly. Annotation database: dmV01 (D.melanogaster v1) Select the reference annotation database. Read length:	
Select the reference genome assembly. Annotation database: (dmV01 (D.melanogaster v1) Select the reference annotation database. Read length: 36	
Select the reference genome assembly. Annotation database: (dmV01 (D.melanogaster v1) Select the reference annotation database. Read length: 36 Analyze up to this many bases in each read alignment.	
Select the reference genome assembly. Annotation database: (dmV01 (D.melanogaster v1) Select the reference annotation database. Read length: 36 Analyze up to this many bases in each read alignment. Maximum number of alignment errors:	
Select the reference genome assembly. Annotation database: (mV01 (D.melanogaster v1) Select the reference annotation database. Read length: 36 Analyze up to this many bases in each read alignment. Maximum number of alignment errors: 2	



data (result) is inside "Galaxy"





Create full report Sample selection: Single sample Do you want to work with a single sample or a sample list? Sample identifier: \$ galaxy_talk The sample identifier to be used. Genome assembly: dm3 (D.melanogaster Apr. 2006, BDGP I Select the reference genome assembly. Annotation database: dmV01 (D.melanogaster v1) 🗘 Select the reference annotation database. Execute



data (result) is inside "Galaxy"

	#dm3-dm	V01-aln2 normsk	galaxy talk	History	Options 👻
	rRNA+	1088			
	rRNA-	536			/2 📑 🚺
	tRNA+	0			~ =
	tRNA-	0		Talk	
	snRNA+	1			
	snRNA-	0		14: dm3-dm	1V01-aln2 👁 🖉 💥
	snoRNA+	0		full report	
	snoRNA-	0			
	miRNA+	0		12.00	6
	miRNA-	0		13: QC-plot	Traction I V X
	piRNA+	1		aligned (dm	<u>3-dmV01-aln2)</u>
	piRNA-	0			
	mRNA+	40375		12: QC-plot	:error 👁 🖉 🕱 🗋
	mRNA-	36623		profiles (dm	3-dmV01-aln2)
1	unknown	unmapped	5850	> promes (um	5 unitor unity
	unknown	overmapped	689	11.00	- 0.00
	unknown	non-repeat	14554	11: QC-plot	• • • • ×
	unknown	repeat 0		repeatednes	is (dm3-dmV01-aln2)
	000000000000000000000000000000000000000				-



Extract data (step 1 of 2)	
Sample selection:	
Single sample	
Do you want to work with a single sample or a sample list?	
Sample identifier:	
galaxy_talk	
The sample identifier to be used.	
Genome assembly:	
dm3 (D.melanogaster Apr. 2006, BDGP	
Select the reference genome assembly.	
Annotation database:	
dmV01 (D.melanogaster v1) ≑	
Select the reference annotation database for the sample.	







data (result) is inside "Galaxy"

#targetId;dm3-	dmV01-aln2;maxhits=100;norm;type=mF	History	Options -
NM 001007095	2.91756		
NM_001007096	2.99338		120
NM 001014452	9.41581	00	~ 🖻
NM 001014453	7.19727	Talk	
NM 001014454	9.32627		
NM 001014455	9.64373	15: annotation summar	v O O X
NM 001014456	0.00000		+
NM 001014457	0.00000	14. dm2_dm1/01_sln2	@ D W
NM 001014458	0.00000	<u>14: um3-umv01-ain2</u>	
NM 001014459	3.76299	<u>full report</u>	
NM 001014460	12.59257		
NM 001014461	0.00000	13: QC-plot: fraction	00%
NM 001014462	0.00000	aligned (dm3-dmV01-	aln2)
NM 001014463	0.00000		1
NM 001014464	0.60443	12: OC platt arrar	⊕ D ∾
NM 001014465	0.00000	12. QC-plot. error	
NM 001014466	0.00000	profiles (dm3-dmV01-	ain2)
NM 001014467	0.00000		2527/250
NM 001014468	0.00000	11: QC-plot:	• 0 ×
NM 001014469	0.00000	repeatedness (dm3-dm	V01-aln2)
NM 001014474	0.00000		
NM 001014475	5.39321	10. comple list	
NM 001014476	7.11771	10: sample list	
NM 001014477	7.24376		
NM 001014478	0.00000	9: galaxy talk annotation	on • 0 🕱
NM 001014479	0.00000	log	
NM 001014480	0.00000	1 lines format: txt_data	hase 7

....just a bunch of Perl scripts!



Friedrich Miescher Institute for Biomedical Research

Acknowledgment

Computational Biology

- Michael Stadler
- Dimos Gaidatzis

Functional Genomics

- Tim Roloff

IT Support

- Thomas Übermeier
- Jan Welker
- Dean Flanders

....and all the people from the Galaxy

