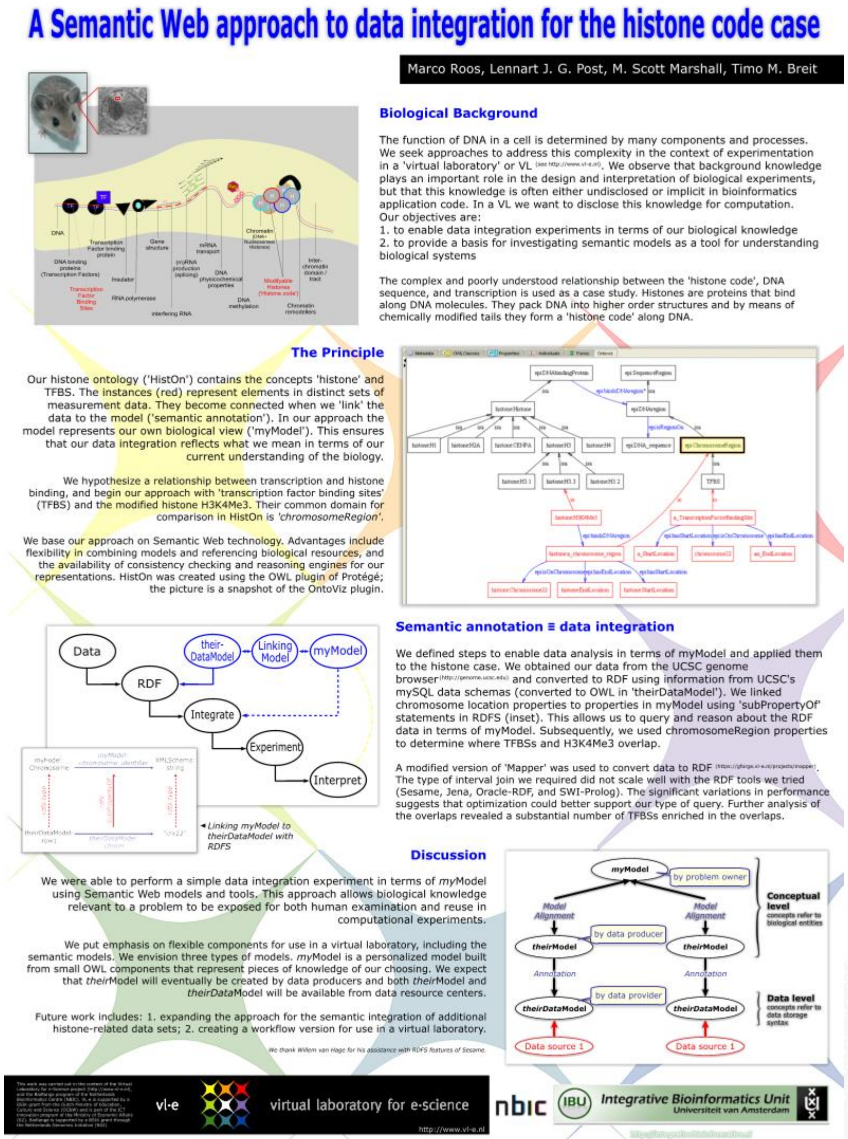
DILS 2006 Poster Flash Presentation

virtual laboratory for e·science



A Semantic Web approach to data integration for the histone code case

Marco Roos, Lennart J. G. Post, M. Scott Marshall, Timo M. Breit

DNA
Transcription Factor binding protein
Gene structure
mRNA transport
Chromatin (DNA+Nucleosome+Histones)
DNA binding proteins (Transcription Factors)
Insulator
(m)RNA production (splicing)
DNA physicochemical properties
Modifyable Histones ("Histone code")
Inter-chromatin domain / tract
Transcription Factor Binding Sites
RNA polymerase
interfering RNA
DNA methylation
Chromatin remodellers

Biological Background

The function of DNA in a cell is determined by many components and processes. We seek approaches to address this complexity in the context of experimentation in a 'virtual laboratory' or VL (see http://www.vl-e.nl). We observe that background knowledge plays an important role in the design and interpretation of biological experiments, but that this knowledge is often either undisclosed or implicit in bioinformatics application code. In a VL we want to disclose this knowledge for computation.

Our objectives are:

1. to enable data integration experiments in terms of our biological knowledge
2. to provide a basis for investigating semantic models as a tool for understanding biological systems

The complex and poorly understood relationship between the 'histone code', DNA sequence, and transcription is used as a case study. Histones are proteins that bind along DNA molecules. They pack DNA into higher order structures and by means of chemically modified tails they form a 'histone code' along DNA.

The Principle

Our histone ontology ('HistOn') contains the concepts 'histone' and TFBS. The instances (red) represent elements in distinct sets of measurement data. They become connected when we 'link' the data to the model ('semantic annotation'). In our approach the model represents our own biological view ('myModel'). This ensures that our data integration reflects what we mean in terms of our current understanding of the biology.

We hypothesize a relationship between transcription and histone binding, and begin our approach with 'transcription factor binding sites' (TFBS) and the modified histone H3K4Me3. Their common domain for comparison in HistOn is *'chromosomeRegion'*.

We base our approach on Semantic Web technology. Advantages include flexibility in combining models and referencing biological resources, and the availability of consistency checking and reasoning engines for our representations. HistOn was created using the OWL plugin of Protégé; the picture is a snapshot of the OntoViz plugin.

◂ *Linking myModel to theirDataModel with RDFS*

Semantic annotation ≡ data integration

We defined steps to enable data analysis in terms of myModel and applied them to the histone case. We obtained our data from the UCSC genome browser (http://genome.ucsc.edu) and converted to RDF using information from UCSC's mySQL data schemas (converted to OWL in 'theirDataModel'). We linked chromosome location properties to properties in myModel using 'subPropertyOf' statements in RDFS (inset). This allows us to query and reason about the RDF data in terms of myModel. Subsequently, we used chromosomeRegion properties to determine where TFBSs and H3K4Me3 overlap.

A modified version of 'Mapper' was used to convert data to RDF (https://gforge.vl-e.nl/projects/mapper). The type of interval join we required did not scale well with the RDF tools we tried (Sesame, Jena, Oracle-RDF, and SWI-Prolog). The significant variations in performance suggests that optimization could better support our type of query. Further analysis of the overlaps revealed a substantial number of TFBSs enriched in the overlaps.

Discussion

We were able to perform a simple data integration experiment in terms of *my*Model using Semantic Web models and tools. This approach allows biological knowledge relevant to a problem to be exposed for both human examination and reuse in computational experiments.

We put emphasis on flexible components for use in a virtual laboratory, including the semantic models. We envision three types of models. *my*Model is a personalized model built from small OWL components that represent pieces of knowledge of our choosing. We expect that *their*Model will eventually be created by data producers and both *their*Model and *theirData*Model will be available from data resource centers.

Future work includes: 1. expanding the approach for the semantic integration of additional histone-related data sets; 2. creating a workflow version for use in a virtual laboratory.

We thank Willem van Hage for his assistance with RDFS features of Sesame.

This work was carried out in the context of the Virtual Laboratory for e-Science project (http://www.vl-e.nl), and the BioRange program of the Netherlands Bioinformatics Centre (NBIC). VL-e is supported by a BSIK grant from the Dutch Ministry of Education, Culture and Science (OC&W) and is part of the ICT innovation program of the Ministry of Economic Affairs (EZ). BioRange is supported by a BSIK grant through the Netherlands Genomics Initiative (NGI).

vl·e virtual laboratory for e·science

http://www.vl-e.nl

nbic

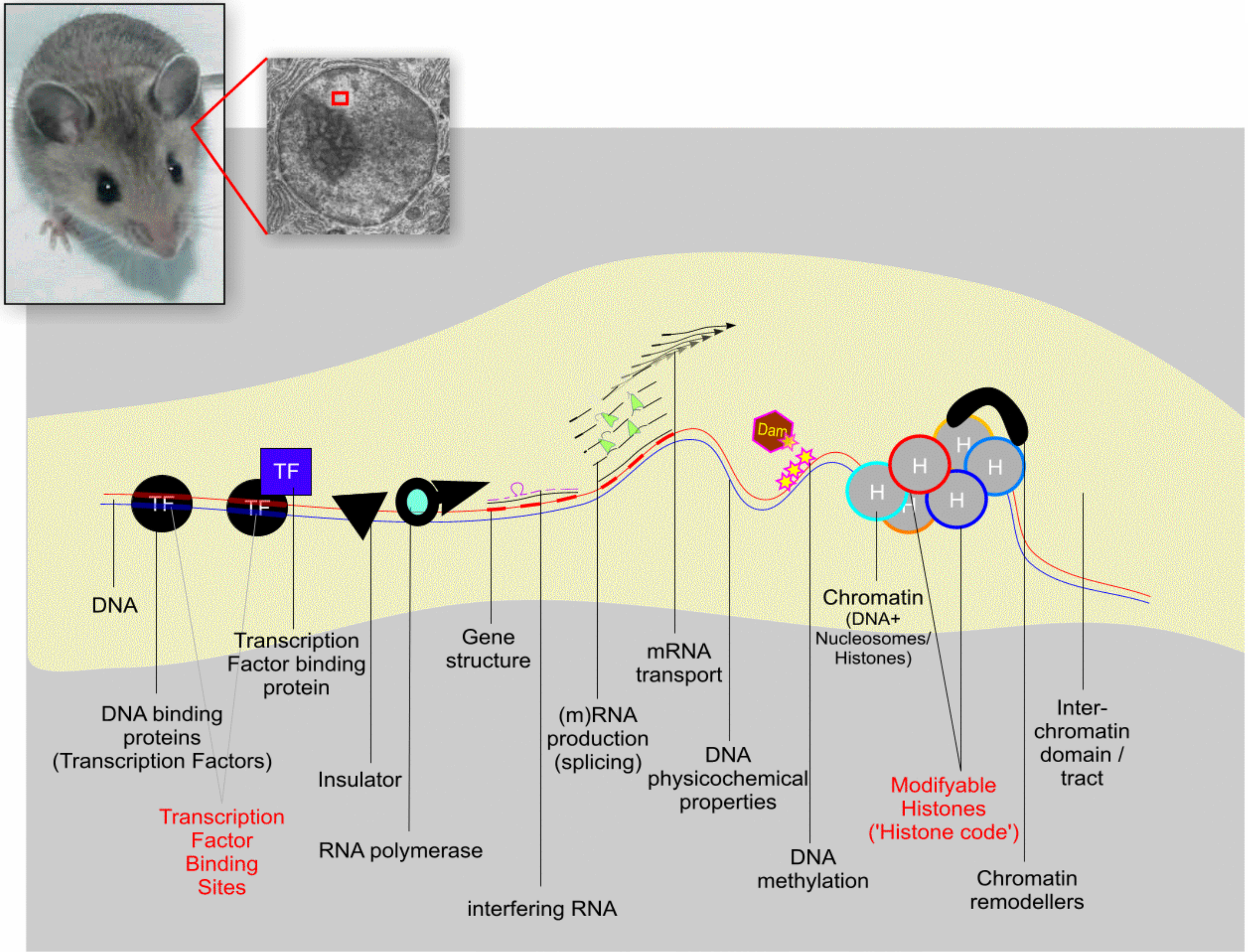
Integrative Bioinformatics Unit
Universiteit van Amsterdam

A Semantic Web approach to data integration for the histone code case

M. Scott Marshall
Marco Roos
Lennart JG Post
Timo M. Breit

Acknowledgements
Roel van Driel
Paul Fransz
Willem van Hage
Machiel Jansen

Biological background



DNA

Transcription Factor binding protein

Gene structure

mRNA transport

Chromatin (DNA+ Nucleosomes/ Histones)

DNA binding proteins (Transcription Factors)

(m)RNA production (splicing)

DNA physicochemical properties

Inter-chromatin domain / tract

Insulator

Modifyable Histones ('Histone code')

Transcription Factor Binding Sites

RNA polymerase

DNA methylation

Chromatin remodellers

interfering RNA

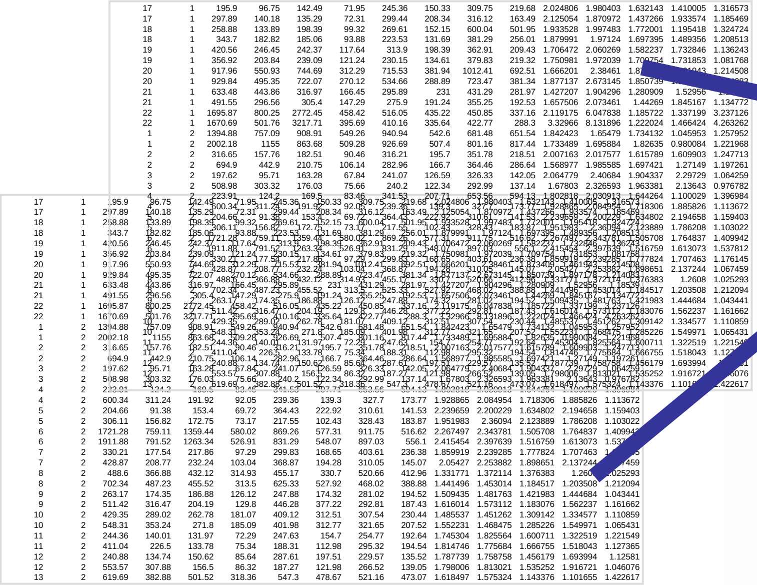
Objectives

Our objectives

1. to enable data integration experiments in terms of our biological knowledge with Semantic Web formats and tools
2. to provide a basis for investigating semantic models as a tool for understanding biological systems

Data integration by scripts

virtual laboratory for e·science



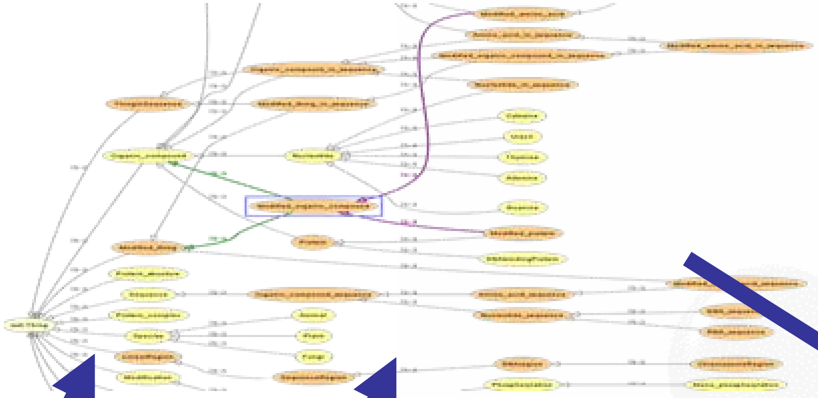
17	1	195.9	96.75	142.49	71.95	245.36	150.33	309.75	219.68	2.024806	1.980403	1.632143	1.410005	1.316573
17	1	297.89	140.18	135.29	72.31	299.44	208.34	316.12	163.49	2.125054	1.870972	1.437266	1.933574	1.185469
18	1	258.88	133.89	198.39	99.32	269.61	152.15	600.04	501.95	1.933528	1.997483	1.772001	1.195418	1.324724
18	1	343.7	182.82	185.06	93.88	223.53	131.69	381.29	256.01	1.879991	1.97124	1.697395	1.489356	1.208513
19	1	420.56	246.45	242.37	117.64	313.9	198.39	362.91	209.43	1.706472	2.060269	1.582237	1.732846	1.136243
19	1	356.92	203.84	239.09	121.24	230.15	134.61	379.83	219.32	1.750981	1.972039	1.709754	1.731853	1.081768
20	1	917.96	550.93	744.69	312.29	715.53	381.94	1012.41	692.51	1.666201	2.38461	1.873409	1.461943	1.214508
20	1	929.84	495.35	722.07	270.12	534.66	288.89	723.47	381.34	1.877137	2.673145	1.850739	1.897178	1.214083
21	1	633.48	443.86	316.97	166.45	295.89	231	431.29	281.97	1.427207	1.904296	1.280909	1.52956	1.18539
21	1	491.55	296.56	305.4	147.29	275.9	191.24	355.25	192.53	1.657506	2.073461	1.44269	1.845167	1.134772
22	1	1695.87	800.25	2772.45	458.42	516.05	435.22	450.85	337.16	2.119175	6.047838	1.185722	1.337199	3.237126
22	1	1670.69	501.76	3217.71	395.69	410.16	335.64	422.77	288.3	3.32966	8.131896	1.222024	1.466424	4.263262
1	2	1394.88	757.09	908.91	549.26	940.94	542.6	681.48	651.54	1.842423	1.65479	1.734132	1.045953	1.257952
1	2	2002.18	1155	863.68	509.28	926.69	507.4	801.16	817.44	1.733489	1.695884	1.82635	0.980084	1.221968
2	2	316.65	157.76	182.51	90.46	316.21	195.7	351.78	218.51	2.007163	2.017577	1.615789	1.609903	1.247713
2	2	694.9	442.9	210.75	106.14	282.96	166.7	364.46	286.64	1.568977	1.985585	1.697421	1.27149	1.197261
3	2	197.62	95.71	163.28	67.84	241.07	126.59	326.33	142.05	2.064779	2.40684	1.904337	2.29729	1.064259
3	2	508.98	303.32	176.03	75.66	240.2	122.34	292.99	137.14	1.67803	2.326593	1.963381	2.13643	0.976782
4	2	223.91	124.2	169.5	83.46	341.53	207.71	653.56	594.13	1.802818	2.030913	1.644264	1.100029	1.396984
4	2	600.34	311.24	191.92	92.05	239.36	139.3	327.7	173.77	1.928865	2.084954	1.718306	1.885826	1.113672
5	2	204.66	91.38	153.4	69.72	364.43	222.92	310.61	141.53	2.239659	2.200229	1.634802	2.194658	1.159403
5	2	306.11	156.82	172.75	73.17	217.55	102.43	328.43	183.87	1.951983	2.36094	2.123889	1.786208	1.103022
6	2	1721.28	759.11	1359.44	580.02	869.26	577.31	911.75	516.62	2.267497	2.343781	1.505708	1.764837	1.409942
6	2	1911.88	791.52	1263.34	526.91	831.29	548.07	897.03	556.1	2.415454	2.397639	1.516759	1.613073	1.537812
7	2	330.21	177.54	217.86	97.29	299.83	168.65	403.61	236.38	1.859919	2.239285	1.777824	1.707463	1.176145
7	2	428.87	208.77	232.24	103.04	368.87	194.28	310.05	145.07	2.05427	2.253882	1.898651	2.137244	1.067459
8	2	488.6	366.88	432.12	314.93	455.17	330.7	520.66	412.96	1.331771	1.372114	1.376383	1.2608	1.025293
8	2	702.34	487.23	455.52	313.5	625.33	527.92	468.02	388.88	1.441496	1.453014	1.184517	1.203508	1.212094
9	2	263.17	174.35	186.88	126.12	247.88	174.32	281.02	194.52	1.509435	1.481763	1.421983	1.444684	1.043441
9	2	511.42	316.47	204.19	129.8	446.28	377.22	292.81	187.43	1.616014	1.573112	1.183076	1.562237	1.161662
10	2	429.35	289.02	262.78	181.07	409.12	312.51	307.54	230.44	1.485537	1.451262	1.309142	1.334577	1.110859
10	2	548.31	353.24	271.8	185.09	401.98	312.77	321.65	207.52	1.552231	1.468475	1.285226	1.549971	1.065431
11	2	244.36	140.01	131.97	72.29	247.63	154.7	254.77	192.64	1.745304	1.825564	1.600711	1.322519	1.221549
11	2	411.04	226.5	133.78	75.34	188.31	112.98	295.32	194.54	1.814746	1.775684	1.666755	1.518043	1.127365
12	2	240.88	134.74	150.62	85.64	287.61	197.51	229.57	135.52	1.787739	1.758758	1.456179	1.693994	1.12581
12	2	553.57	307.88	156.5	86.32	187.27	121.98	266.52	139.05	1.798006	1.813021	1.535252	1.916721	1.046076
13	2	619.69	382.88	501.52	318.36	547.3	478.67	521.16	473.07	1.618497	1.575324	1.143376	1.101655	1.422617

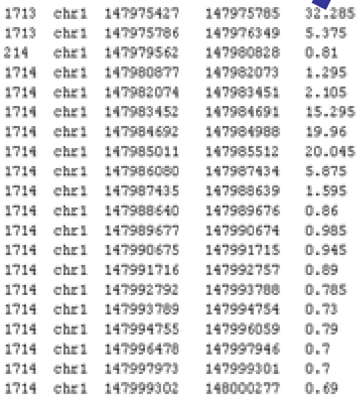
‘Computational experiment’

```
# (c) 2006, SARA, Bart Heupers
# Example use :
# python get_overlap.py TFBSConsSites_chr22.txt encodeSangerChipH3K4me3_chr22.txt
> overlaps.txt
import sys, os

def read_line(f):
    '''Read a line from the file and check whether the third argument is
numerical'''
    line_ok = False
    while not line_ok:
        line = f.readline()
        if line == "":
            return ""
        li = line.split()
        if li[2].isdigit():
            line_ok = True
    li[2] = int(li[2])
    li[3] = int(li[3])
    return li

def main():
    list1 = [];
    if len(sys.argv) < 3:
       print "Use : get_overlap <file1> <file2>"
       os._exit(1)
    file1 = open(sys.argv[1])
    file2 = open(sys.argv[2])

    list1.append(read_line(file1))
    line2 = read_line(file2)
    while list1[-1] != "" and line2 != "":
        if line2[3] <= list1[0][2]:
            # end element 2 smaller then start smallest element 1
            # read new element 2
            line2 = read_line(file2)
        elif line2[2] > list1[-1][3]:
            # start element 2 bigger than end last element 1
            # read new element 1, remove all existing elements 1
            list1 = []
            list1.append(read_line(file1))
        else:
            # There is some overlap
            # check all elements in list in for overlap
            for line1 in list1:
                if (line2[3] >= line1[2] and line2[3] <= line1[3]) or (line1[3]
>= line2[2] and line1[3] <= line2[3]):
                    # Overlap
                    print "Overlappping elements :\n", line1, "\n", line2
            line2 = read_line(file2)
    file1.close()
    file2.close()


if __name__ == '__main__':
    main()
```

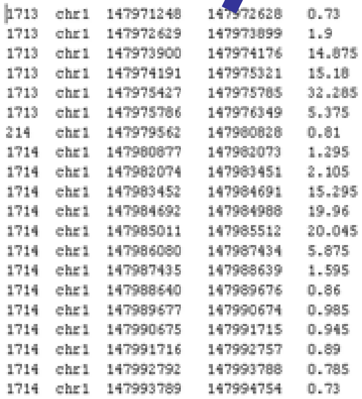
Semantic integration and workflow

HistOn (Histone Ontology)

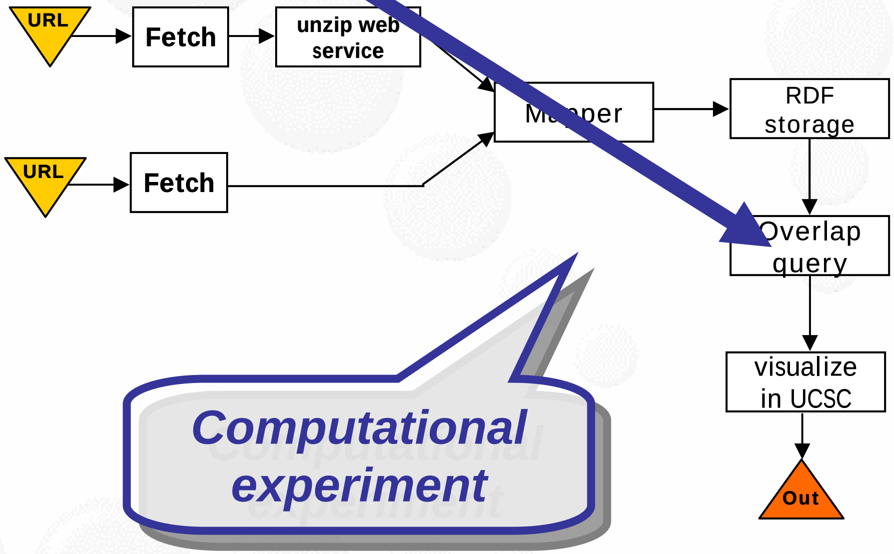




```
1713  chr1  147975427  147975785  32.285
1713  chr1  147975786  147976349  5.375
214   chr1  147979562  147980828  0.81
1714  chr1  147980877  147982073  1.295
1714  chr1  147982074  147983451  2.105
1714  chr1  147983452  147984691  15.295
1714  chr1  147984692  147984988  19.96
1714  chr1  147985011  147985512  20.045
1714  chr1  147986080  147987434  5.875
1714  chr1  147987435  147988639  1.595
1714  chr1  147988640  147989676  0.86
1714  chr1  147989677  147990674  0.985
1714  chr1  147990675  147991715  0.945
1714  chr1  147991716  147992757  0.89
1714  chr1  147992792  147993788  0.785
1714  chr1  147993789  147994754  0.73
1714  chr1  147994755  147996059  0.79
1714  chr1  147996478  147997946  0.7
1714  chr1  147997973  147999301  0.7
1714  chr1  147999302  148000277  0.69
```



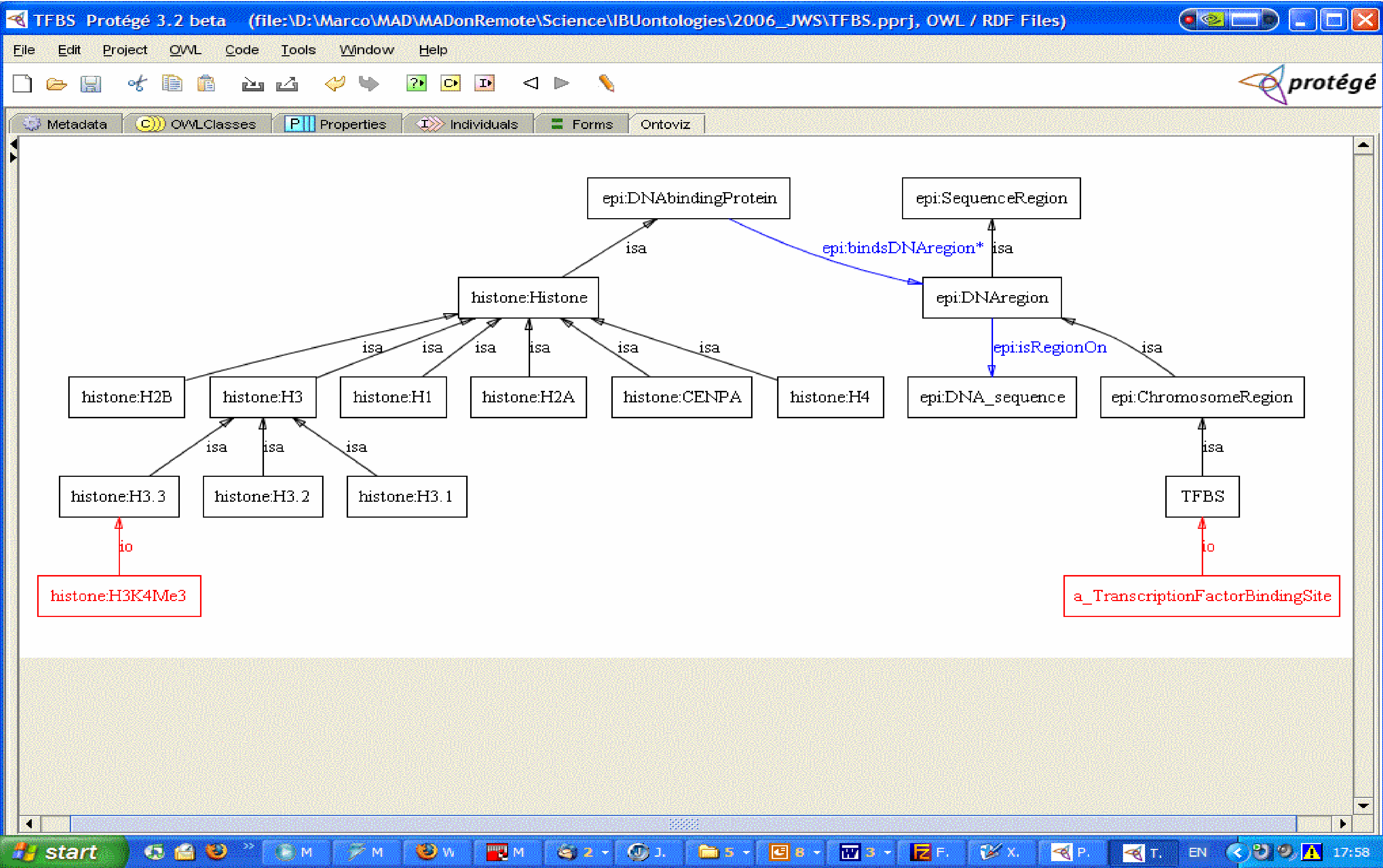
```
1713  chr1  147971248  147972628  0.73
1713  chr1  147972629  147973899  1.9
1713  chr1  147973900  147974176  14.875
1713  chr1  147974191  147975321  15.18
1713  chr1  147975427  147975785  32.285
1713  chr1  147975786  147976349  5.375
214   chr1  147979562  147980828  0.81
1714  chr1  147980877  147982073  1.295
1714  chr1  147982074  147983451  2.105
1714  chr1  147983452  147984691  15.295
1714  chr1  147984692  147984988  19.96
1714  chr1  147985011  147985512  20.045
1714  chr1  147986080  147987434  5.875
1714  chr1  147987435  147988639  1.595
1714  chr1  147988640  147989676  0.86
1714  chr1  147989677  147990674  0.985
1714  chr1  147990675  147991715  0.945
1714  chr1  147991716  147992757  0.89
1714  chr1  147992792  147993788  0.785
1714  chr1  147993789  147994754  0.73
```



Computational experiment

Data integration through Histone Ontology: ‘HistOn’

virtual laboratory for e-science



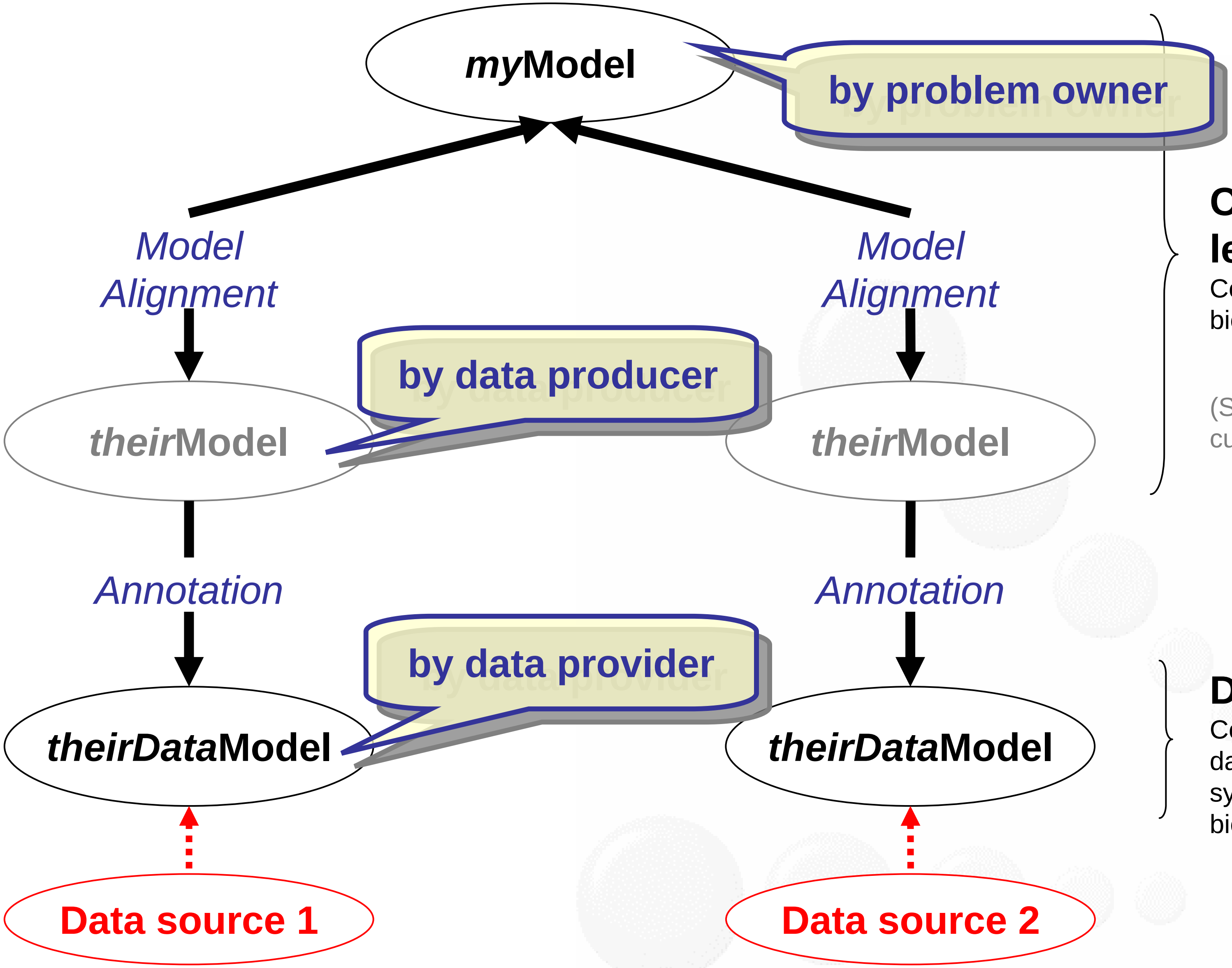
End of presentation

Please visit

http://integrativebioinformatics.nl/semanticdataintegration.html

Our concept

Manuscript in preparation



*my*Model

by problem owner

Model Alignment

Model Alignment

by data producer

*their*Model

*their*Model

Annotation

Annotation

by data provider

*theirData*Model

*theirData*Model

Data source 1

Data source 2

Conceptual level

Concepts refer to biological entities

(Skipped in current scenario)

Data level

Concepts refer to data storage syntax, not biological entities

Query in terms of myModel (toy example)

```
SELECT * FROM {x}
       myModel:chromosomeIdentifier {y}
       USING NAMESPACE
       myModel = <http://www.somewhere.org/myModel.owl#>
```

Table 1: Sesame query results (also available in RDF)

x	y
http://www.encodeSangerChipH3K4me3.xml/row1	"chr1"
http://www.encodeSangerChipH3K4me3.xml/row2	"chr1"

2 results found in 1 ms.

The histone code case

‘Elucidate the relationship between the Histone code, DNA sequence, and transcriptional activity’

Histone code case characteristics

The histone code case

Multifaceted problem, no 1:1 relationships
Requires explorative analysis (phenomenon discovery)
Requires ‘large-scale’ data integration

⇓

Semantic modelling approach?

Histone code case characteristics

virtual laboratory for e·science

The histone code case

Multifaceted
no 1:1 relationships
Requires data integration and modelling

⇓

Semantic modelling approach?