REDEVELOPING AND OPTIMIZING THE INTERACTIVE CHROMATIN MODELING WEB SERVER [ICM]

Inderbir Sondh

Mentor: Dr. Tom Bishop

DNA

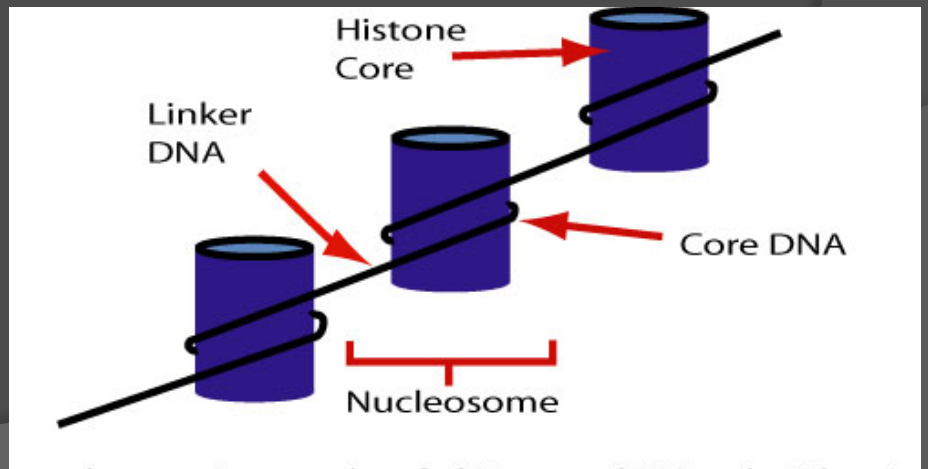
-> 4 Bases- A, C, G, and T

-> Going along a strand gives the sequence

-> A pairs with T, G with C

-> Exists in folded and unfolded forms

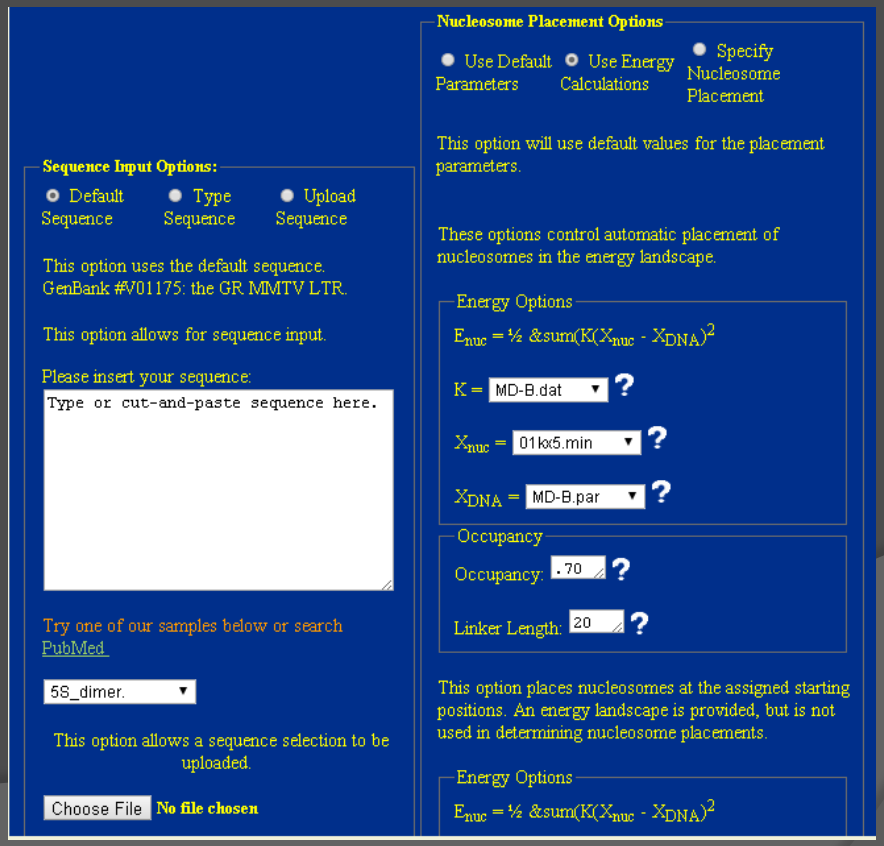
-> Histone = protein that DNA wraps around



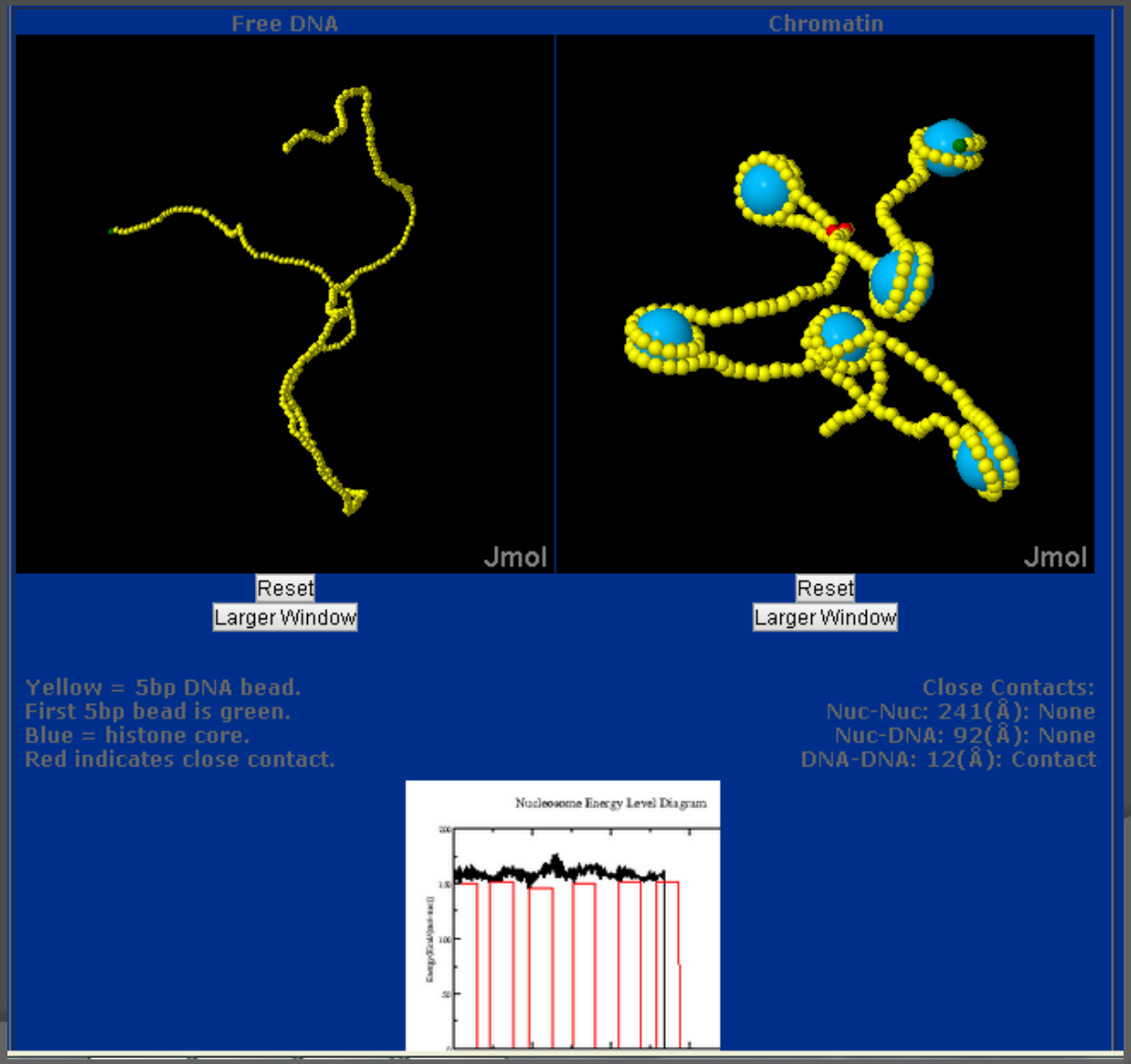
a nucleosome is a complex of a histone and 146 nucleotide pairs

What is the ICM?

- Software that generates a 3-D model of a given DNA sequence.
- Initial Inputs:



ICM Final Output



Needed Improvements to ICM

- Currently can only efficiently handle sequences around 10,000-20,000 base pairs long.
- There are many sequences that are much longer (human genome is billions long!)
- Interface should be easier to use and integrated with other DNA research tools.

Goals

- Redesign ICM with an object oriented approach including steps to increase efficiency (enough to handle 1 Mil base pairs)
- Integrate with an existing genome browser for more intuitive usage and increased functionality.

Helical Parameters

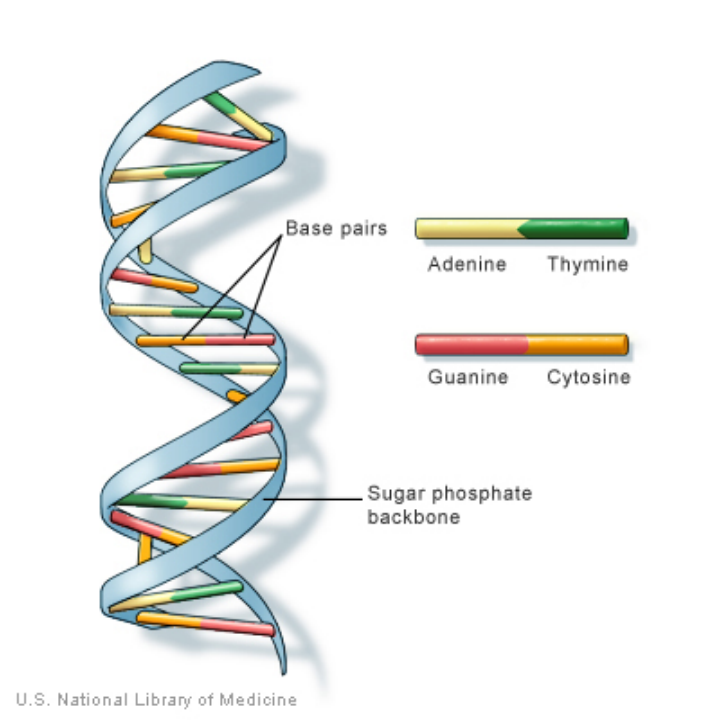
- Used to describe each DNA base pair relative to an adjacent base pair

- Translations across XYZ:

  Shift, Slide Rise

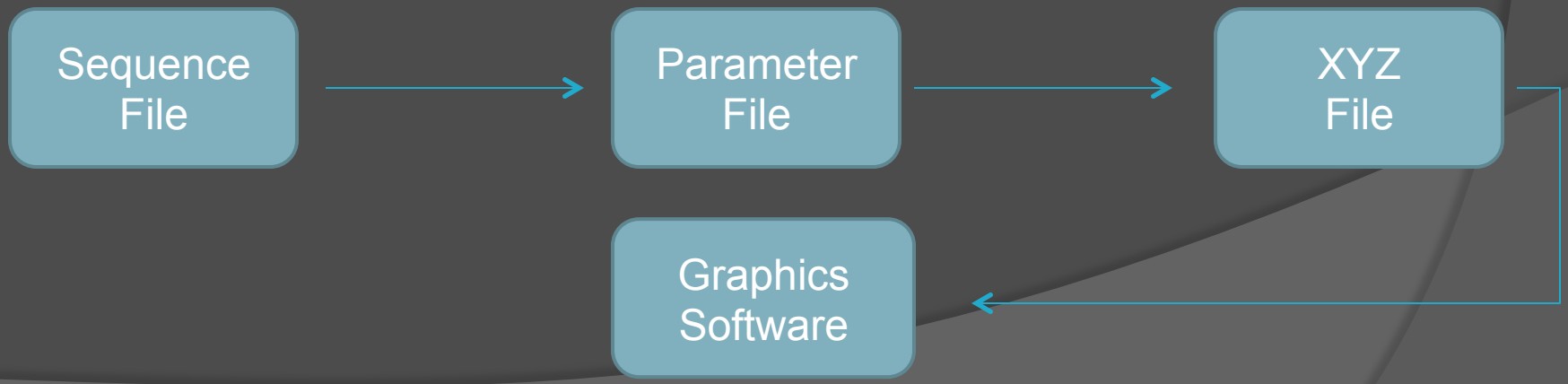
  Rotations around XYZ:

  Tilt, Roll, Twist



ICM Workflow/Logistics

- **Inputs**: DNA Sequence, Energy Models, Temperature, Nucleosome Placement
- **Outputs**: XYZ File, 3-D Rendering (using Jmol)



Sequence File

sequence.txt

1	A
2	C
3	G
4	A
5	C
6	C
7	C
8	T
9	C
10	G
11	A
12	A

Reference File

16 base-pairs

0 ***local base-pair & step parameters***

	Shear	Stretch	Stagger	Buckle	Prop-Tw	Opening	Shift	Slide	Rise	Tilt	Roll	Twist
A-A	-0.06	-0.02	-0.03	0.14	-6.91	0.44	-0.06	-0.03	3.17	-1.49	1.32	31.92
A-C	0.03	-0.02	-0.02	-1.43	-7.77	0.20	-0.05	0.04	3.19	0.27	2.14	32.00
A-G	0.03	-0.02	-0.02	-1.43	-7.76	0.21	0.10	-0.25	3.22	-0.58	3.16	28.49
A-T	-0.06	-0.02	-0.03	0.13	-6.91	0.43	-0.00	-0.08	3.12	0.00	2.01	30.18
C-A	-0.06	-0.02	-0.03	0.12	-6.92	0.42	0.02	0.25	3.12	0.21	9.19	27.86
C-C	0.03	-0.02	-0.02	-1.44	-7.79	0.20	0.15	-0.28	3.34	0.15	5.68	29.57
C-G	0.03	-0.02	-0.02	-1.42	-7.77	0.20	0.00	0.30	3.07	0.00	8.07	27.24
C-T	-0.06	-0.02	-0.03	0.13	-6.90	0.42	-0.10	-0.25	3.22	+0.58	3.15	28.50
G-A	-0.06	-0.02	-0.03	0.13	-6.91	0.42	-0.05	0.22	3.23	-0.30	3.72	32.99
G-C	0.03	-0.02	-0.02	-1.41	-7.76	0.20	-0.00	0.24	3.23	0.00	1.65	34.74
G-G	0.03	-0.02	-0.02	-1.44	-7.76	0.21	-0.15	-0.28	3.34	-0.16	5.68	29.57
G-T	-0.06	-0.02	-0.03	0.13	-6.92	0.43	+0.05	0.04	3.19	-0.27	2.13	32.00
T-A	-0.06	-0.02	-0.03	0.13	-6.90	0.43	-0.00	0.24	3.17	0.00	10.30	28.82
T-C	0.03	-0.02	-0.02	-1.43	-7.78	0.20	+0.05	0.22	3.23	0.30	3.71	32.99
T-G	0.03	-0.02	-0.02	-1.42	-7.77	0.20	-0.02	0.25	3.12	-0.21	9.19	27.85
T-T	-0.06	-0.02	-0.03	0.11	-6.90	0.43	+0.06	-0.03	3.17	+1.50	1.31	31.92

Parameter File

6197 base_pairs

0 ***local base-pair & step parameters***

	Shear	Stretch	Stagger	Buckle	Prop-Tw	Opening	Shift	Slide	Rise	Tilt	Roll	Twist
G-C	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A-T	-0.06	-0.02	-0.03	0.13	-6.91	0.42	-0.05	0.22	3.23	-0.30	3.72	32.99
T-A	-0.06	-0.02	-0.03	0.13	-6.91	0.43	-0.00	-0.08	3.12	0.00	2.01	30.18
C-G	0.03	-0.02	-0.02	-1.43	-7.78	0.20	0.05	0.22	3.23	0.30	3.71	32.99
C-G	0.03	-0.02	-0.02	-1.44	-7.79	0.20	0.15	-0.28	3.34	0.15	5.68	29.57
G-C	0.03	-0.02	-0.02	-1.42	-7.77	0.20	0.00	0.30	3.07	0.00	8.07	27.24
T-A	-0.06	-0.02	-0.03	0.13	-6.92	0.43	0.05	0.04	3.19	-0.27	2.13	32.00
C-G	0.03	-0.02	-0.02	-1.43	-7.78	0.20	0.05	0.22	3.23	0.30	3.71	32.99
G-C	0.03	-0.02	-0.02	-1.42	-7.77	0.20	0.00	0.30	3.07	0.00	8.07	27.24

El Hassan's Algorithm

- Process used to convert a .par file to a .xyz file (par -> xyz)
- Involves a series of rotation matrix multiplications.

$$R_X(\theta) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & Cos(\theta) & -Sin(\theta) \\ 0 & Sin(\theta) & Cos(\theta) \end{bmatrix}$$

$$R_Y(\theta) = \begin{bmatrix} Cos(\theta) & 0 & -Sin(\theta) \\ 0 & 1 & 0 \\ Sin(\theta) & 0 & Cos(\theta) \end{bmatrix}$$

$$R_Z(\theta) = \begin{bmatrix} Cos(\theta) & Sin(\theta) & 0 \\ -Sin(\theta) & Cos(\theta) & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\mathbf{T}_{i+1} = \left[\mathbf{R}_z\left(\frac{\Omega}{2} - \phi\right) \mathbf{R}_y(\Gamma) \mathbf{R}_z\left(\frac{\Omega}{2} + \phi\right) \right] \mathbf{T}_i \qquad (9)$$

$$\mathbf{T}_{mst} = \left[\mathbf{R}_z\left(\frac{\Omega}{2} - \phi\right) \mathbf{R}_y\left(\frac{\Gamma}{2}\right) \mathbf{R}_z(\phi) \right] \mathbf{T}_i \qquad (10)$$

$$\mathbf{r}^o_{i+1} = \mathbf{r}^o_i + D_x \mathbf{x}_{mst} + D_y \mathbf{y}_{mst} + D_z \mathbf{z}_{mst} \qquad (11)$$

XYZ FILE

```
COMMENT  TcB par2xyz
CA         0.00000        0.00000        0.00000
H1         1.00000        0.00000        0.00000
H2         0.00000        1.00000        0.00000
H3         0.00000        0.00000        1.00000
CA        -0.01227        0.23463        3.22933
H1         0.82456        0.77837        3.16564
H2        -0.55635        1.07355        3.24274
H3         0.04846        0.25807        4.22721
```

- **CA-**Central Atom Coordinates
- **H-**Director / Pointer for an axis

Efficiency of par -> xyz

- One of the most time consuming process of the program
- Since test files are up to millions of base pairs long, process must be optimized.
- Application runs on a webpage, quickness is a necessity.

Current Running Times

Running Time Comparison of New and Old Code				
	C++		FORTRAN	
# of Base Pairs	Running Time (s)	CPU Usage (%)	Running Time (s)	CPU Usage (%)
1000000	68.3	74.5	25.69	100
100000	7.3	74.8	2.54	99.6
10000	0.83	66.2	0.25	100
1000	0.09	55.5	0.03	66.6
100	0.02	50	0.003	0
10	0	0	0	0

New code is slower…..File I/O issues

C++ code was also writing files at the same time

Ways to Improve Running Time

- Introduce new data structure to hold all .par data and prevent opening and closing files many times.
- Execute El Hassan's Algorithm while expanding matrices beforehand so that multiple matrix multiplications are not needed.

Ways to Improve Running Time (2)

- We can parameterize the rotations using unit quaternions.
- Quaternion algebra is especially practical for rotation calculations, very likely to increase efficiency.

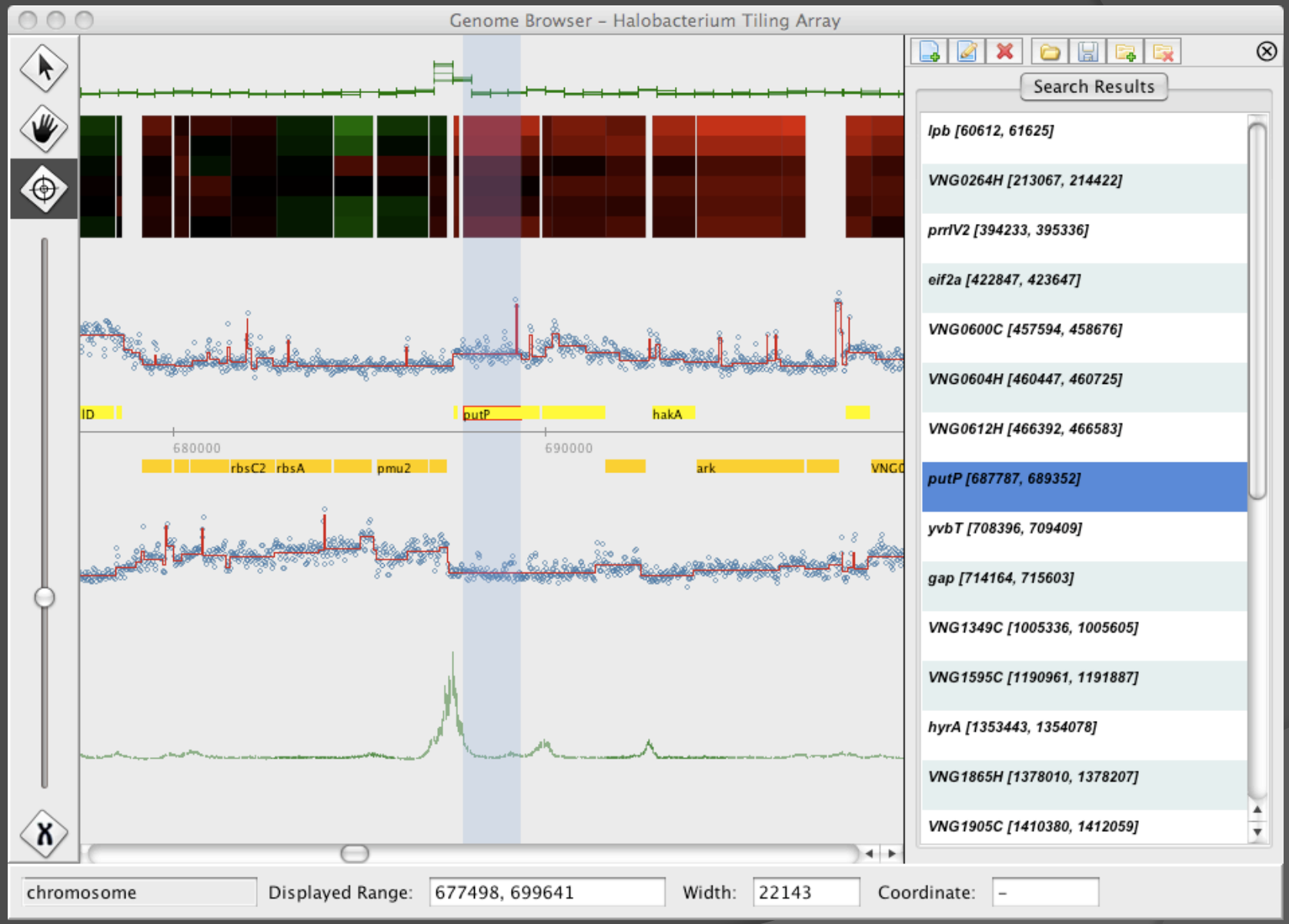
Future Plans: The Big Picture

- **Integration with a Genome Browser.**

This allows a user to input a sequence directly from a DNA database.

Also saves calculation time, as the genome browser can tell exactly where nucleosomes should be placed.

More intuitive and aesthetically pleasing user interface.


Genome Browser – Halobacterium Tiling Array
Search Results
lpb [60612, 61625]
VNG0264H [213067, 214422]
prrIV2 [394233, 395336]
eif2a [422847, 423647]
VNG0600C [457594, 458676]
VNG0604H [460447, 460725]
VNG0612H [466392, 466583]
putP [687787, 689352]
yvbT [708396, 709409]
gap [714164, 715603]
VNG1349C [1005336, 1005605]
VNG1595C [1190961, 1191887]
hyrA [1353443, 1354078]
VNG1865H [1378010, 1378207]
VNG1905C [1410380, 1412059]
ID
putP
hakA
680000
690000
rbsC2
rbsA
pmu2
ark
chromosome
Displayed Range:
677498, 699641
Width:
22143
Coordinate:
–

References

- El Hassan, M.A. and Calladine C.R., 1995, The Assessment of the Geometry of Dinucleotide Steps in Double-Helical DNA; a New Local Calculation Scheme, *J. Mol. Biol.*, Vol. 251, p. 648-664
- Bishop, T.C. and Stolz, R.C., 2010, ICM Web: the interactive chromatin modeling web server, *Nucleic Acids Research*, Vol. 38, Web Server Issue. DOI: 10.1093/nar/gkq496
- http://ghr.nlm.nih.gov/handbook/illustrations/dnastructure.jpg
- http://gaggle.systemsbiology.net/docs/geese/genomebrowser/genome_browser.png