ALGORITHMS FOR COMBINING SEQUENCE DATA WITH OPTICAL MAPS
Examples from *P. falciparum* and *E. coli*

Marco Antoniotti[1], Thomas Anantharaman[2], Chiung-Wen Chang[1], David C. Schwartz[2] and Bud Mishra[1,3]

[1]New York University Courant Bioinformatics Group
[2]Laboratory for Molecular and Computational Genomics
University of Wisconsin, Madison
[3]Cold Spring Harbor Laboratory

Current, genome wide shotgun sequence assembly methods have produced a very large amount of data that can eventually answer many important biological questions. Concurrently, our groups at NYU and University of Wisconsin have developed a Single Molecule approach to produce High-Resolution Ordered Restriction Maps (Optical Mapping) that can be integrated with the sequence data for the purposes described below:

1 - speed up the sequence assembly by providing an appropriate scaffolding and
2 - establish the correctness of a completed sequence assembly by aligning in silico restriction maps from sequences with the optical maps
3 - align, orient and phase sequence contigs and thus, facilitate the gap closing and finishing steps.

We ran our validation and alignment algorithms on the *P. falciparum* and validated the assembly of chromosomes 2 and 3, while phasing the known contigs of other chromosomes.

We also ran our algorithms on the *E. coli* K12 data and are able to suggest possible local improvements on the assembled sequence.

The Algorithms

We developed three alignment algorithms based on a Maximum Likelihood Estimation approach

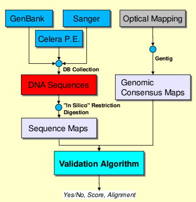
Validation: used to establish the correctness of a sequence assembly against an (Optical) Ordered Restriction Map
Alignment: used to suggest anchoring of sequence contigs against an Ordered Restriction Map
Restriction Fragments Translocation (RFT) Detection: used to find RFTs and to propose a correction of the map

The Statistical Error Model

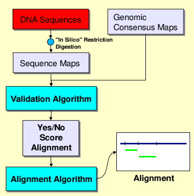
$$p_c \times \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} e^{\frac{\left(\sum_{k=i-u}^{i} l_k^2 - \sum_{h=j-v}^{j} a_h^2\right)}{2\hat{\sigma}^2}} \times (1-p_c) \times p_f$$

The core of probabilistic model of the sources of error used in the Likelihood Estimate of an alignment: essentially a combination of a Gaussian, a Poisson and a Bernoulli process. The sources of errors considered are
1 - *sizing* errors
2 - *missing cuts*
3 - *false cuts* (false positives)

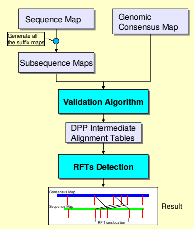
The **Validation** algorithm is computed via a specialized *Dynamic Programming Procedure* (DPP). The **Alignment** algorithm uses an iterative 1D DPP to approximate a (conjectured) *NP*-complete problem.



Validation of Sequence Data Architecture

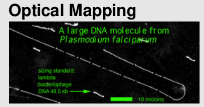


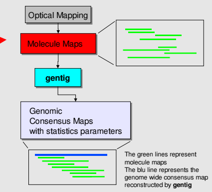
Multiple ***Alignments*** against Consensus Map Scaffold



Restriction Fragments Translocations Detection and Correction Architecture

Optical Mapping





Optical Mapping is a technology that it used to reconstructs Genome wide Ordered Restriction Maps of a given organism by analyzing images of Single DNA Molecules.
The diagram shows the flow of information is the current setup, where the **gentig** program is used to produce the resulting *Consensus Map.*

Sample Application

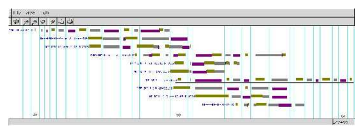
P. Falciparum Chromosome 2 Validation

rank	MLE/MWSS score	*P*-score	# Missing Cuts	# False Cuts
1	77.308	1.000	1	0
2	385.146	1.000	12	2
3	454.146	1.000	10	3
4	149.037	0.980	10	4
5	155.710	0.100	13	6
6	125.088	0.000	8	2

P. Falciparum Chromosome 3 Validation

rank	MLE/MWSS score	*P*-score	# Missing Cuts	# False Cuts
1	87.020	1.000	1	0
2	184.232	1.000	15	3
3	290.579	1.000	13	5
4	231.384	0.998	14	2
5	237.722	0.002	13	5
6	130.866	0.000	15	4

These two tables report the **Validation** results for chromosomes 2 and 3 of the Malaria parasite *P. falciparum*. Each row represents an alignment of the chromosome *in silico* ordered restriction map against the published chromosome optical consensus map. The **rank** column lists the order of the alignment as produced by the Validation algorithm. The **MLE/MWSS score** column contains the Maximum Likelihood Estimate values computed for the alignment (as a *Minimum Weighted Sum of Squares*): low values are better. The ***P*-score** column contains a number (between 0 and 1) characterizing the relevance of the alignment in comparison with all the other ones (the greyed rows show alignments with a 0 *P*-score, hence discardable). The **# Missing Cuts** column and the **# False Cuts** column are self explanatory. As expected the Validation algorithm optimizes for low values in both columns.



Multiple Alignment

The picture to the left shows a screenshot of the **NYU VALIS** environment displaying the result of the **Alignment** algorithm. The published contigs of chromosome 12 of *P. falciparum* are anchored along the published (Optical) Consensus Map. This application aids in finishing assembly and gap closing.

Future Developments

The algorithms developed for **Validation**, **Alignment** and **Restriction Fragment Translocation** Detection and Correction have a variety of applications. The algorithm core and implementation can be tuned for other types of genomic maps. Other immediate applications (beyond Validation of published sequence data and Sequence Assembly aid) include karyotyping, especially in conjunction with population analysis and cancer evolution.

Please visit our web sites **`http://bioinformatics.cat.nyu.edu`** for more information, publications, and references

In cooperation with: CSH UNIVERSITY OF WISCONSIN MADISON