Algorithms and Analysis for Combining Sequences and Maps: Application to the Malaria Parasite *P. falciparum*⋆

Marco Antoniotti[1], Thomas Anantharaman[2], Chiung-Wen Chang[3], Juliette Colinas[3], Salvatore Paxia[1], David C. Schwartz[4], and Bud Mishra[1]

[1] Courant Bioinformatics Group, New York University {marcoxa,paxia,mishra}@cs.nyu.edu
[2] Biostatistics and Medical Informatics Department, University of Wisconsin
[3] Department of Biology, New York University
[4] Laboratory for Molecular and Computational Genomics,
Departments of Genetics and Chemistry, Madison, WI, U.S.A.

1 Extended Abstract

The study of genetics relies on complete nucleotide sequences of the organism together with a description of the transcription units. While this information at its finest level is not often available, or when available, may suffer from various errors due to sequencing or assembly, one can garner much information from significantly coarser descriptions that are easily available in genomic maps. Such maps with high resolution and accuracy as well as partially assembled sequences at various degrees of completion exist for many of the microbial organisms, yeasts, worms, flies and now humans. In general, genetically or physically mapped collections of objects derived from the genome under study are still of immense utility, and require robust bioinformatics tools to validate their mutual consistency and integration. The integrated genomic databases derived from all available sources are likely to prove useful even at an early stage for annotation, gap detection (in sequences) and targeted gap closing, sequence contig phasing and map assisted sequence assembly. In this extended abstract, we focus on these issues with an example involving ordered restriction map data based on optical mapping and partially assembled sequences for the malaria parasite, *Plasmodium falciparum*.

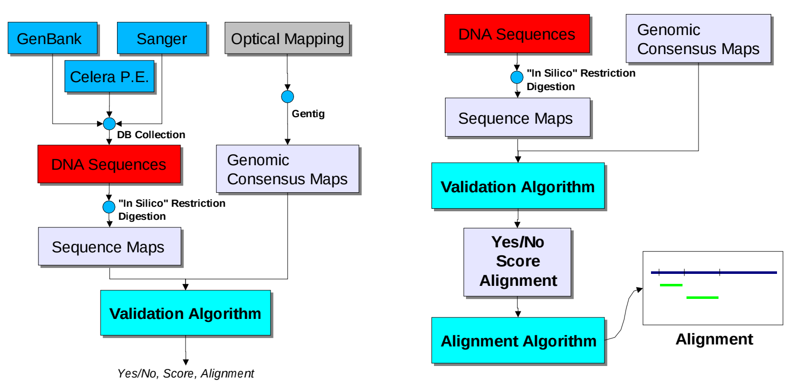

Fig. 1. The validation and alignment processes overall flow.

⋆ This research was conducted under the Department of Energy Grant DoE-25-74100-F1799, the National Cancer Institute Grant NCI 5 RO1 CA79063-03, and the NSF Career Grant IRI-9702071.

2 Description of the Problems

In the following we describe the two problems we studied: the *Map Based Validation Problem* and the *Map Based Alignment Problem*. The first problem characterizes whether a sequence contig (possibly assembled by shotgun sequence assembly) is "correct". The second problem computes an alignment of several sequence contigs along an ordered restriction map, while identifiyng gaps and possible overlaps.

Definition: Map Based Validation Problem. Given a genome wide *ordered restriction map* CM with M restriction fragments and an "in silico" map SM with N fragments (usually $N < M$) obtained from the known sequence of the same organism and with respect to the same restriction enzyme, we compute the best match of the SM against the CM with respect to a score function based on a *likelihood estimate*. The likelihood function models the following error sources:

1. *sizing errors* (espressed as a *standard deviation* σ_i, with $0 \leq i < M$),
2. *false positives*, i.e. restriction enzyme cuts which are reported incorrectly, and
3. *false negatives*, i.e. missing restriction enzyme cuts. □

The score function is based on a Maximum Likelihood Estimate (MLE) of the error model, computed via a Minimization of a Weighted Sum-of-Squares (MWSS). The result of the computation is a set S_i of triples $\langle s_i, x_{(i,j)}, v_{(i,j)} \rangle$ (with $0 \leq j < M$), where s_i is an identifier for the given SM, x_i is a position on CM and v_i is the value associated to this positioning.

Definition: Map Based Alignment Problem. Given a list of sets $S_i \mid_{i=0}^{\ell}$ obtained from the validation procedure (i.e. $S_i = \{\langle s_i, x_{(i,j)}, v_{(i,j)} \rangle \mid 1 \leq j \leq k\}$, we want to choose at most one triple from each S_i, while satisfying the following global conditions:

1. The chosen s_i's do not *overlap* (although this requirement may be relaxed);
2. $\sum_i \left(I_i \times v_{(i,j)}\right)$ is minimized (over each j in a set S_i);
3. $n - \sum_i I_i$ is minimized;

where I_i is an *indicator* variable assuming a value 1 if a triplet from S_i is included in the chosen set, and 0 otherwise. □

It should be immediately clear that objectives (2) and (3) conflict: the minimum of objective (2) is achieved when no sequence is chosen, while (3) requires to choose as many sequences as possible, irrespective of the score values. We resolve this conflict by a weighting scheme involving a Lagrangian-like term linearly combining the two contradictory objectives.

We considered two approximation algorithms to solve this problem. We started by considering the problem for the special case when $k = 1$ and devises an efficient algorithm. Next, we considered the general case when $k > 1$ and devise good approximation heuristics. Finally we devised two different solutions for the problem: a *Greedy* solution, a (yet another) *Dynamic Programming* solution, and analyzed their applicability. We conjecture that the problem is $\mathcal{NP}$-complete in the general setting ($k > 1$), since the problem is assimilable to INDEPENDENT SET.

3 Experimental Results

We used the NYU-BiG ApplE (Bioinformatics Group Application Environment) software to run several experiments involving *P. falciparum* genome. We give two representative examples below (one for validation and one for alignment). The complete results with explanations are viewable at `http://bioinformatics.cat.nyu.edu/valis/projects/PFalciparum/`.

We obtained the sequences for the *P. falciparum*'s 14 chromosomes from the `www.plasmodb.org` site. Our experiment cut the sequences "in silico" using the `BamHI` and `Nhe-I` restriction enzymes. The resulting maps are input to the validation program along with appropriate optical ordered restriction maps.

3.1 Validation of Chromosomes 2 and 3

We produce two "in silico" maps for the chromosome 2 and chromosome 3 sequences with the enzyme `BamH I`. The optical ordered restriction maps we used were published in [4, 5]. Since the published maps omitted all the statistically relevant information, we also used maps generated subsequent to the publication, by an improved version of the gentig program.

We produced the results reported in Tables 1, 2, and 3. Table 1 shows the match of the sequence maps for chromosomes 2 and 3 against the consensus maps generated by gentig. Tables 2 and 3 show the match of the sequence maps against the consensus map published in [3].

The table results are to be read in the following way. First the results are grouped by $\mathcal{P}$-score and then they are ordered by their effective MLE/MWSS score. It turns out (as expected) that the best match has both a high $\mathcal{P}$-score and a low MLE/MWSS score. The matches with 0.0 $\mathcal{P}$-score are reported to show how they are not influencing the result.

The results are a summary referring to maps obtained using the `Bam-HI` restriction enzyme. We also have the results for maps produced with the `Nhe-I` restriction enzyme.

Chromosome 2 Validation Summary A

rank	MLE/MWSS score	$\mathcal{P}$-score	map id	# missing cuts	# false cuts
1	80.869	1.000	1302	0	1
2	126.835	1.000	1326	12	4
3	132.890	1.000	1414	12	2
4	127.488	0.980	1305	8	4
5	105.861	0.000	1302	2	1

Table 1. The data reported shows the best "matches" found by the validation checker in the case of *P. falciparum* chromosome 2. The "in silico" sequence map was obtained from the TIGR database sequence. The `Bam-HI` sequence map (as well as its reversed) was checked against 75 (optical) consensus maps produced by gentig. The 75 optical maps cover the entire *P. falciparum* genome. The validity checker found its best matches against the map tagged 1302.

Chromosome 2 Validation Summary B

rank	MLE/MWSS score	$\mathcal{P}$-score	map id	# missing cuts	# false cuts
1	77.308	1.000	NYU-WISC	1	0
2	385.146	1.000	NYU-WISC	12	2
3	454.146	1.000	NYU-WISC	10	3
4	149.037	0.980	NYU-WISC	10	4
5	155.710	0.100	NYU-WISC	13	6
6	125.088	0.000	NYU-WISC	8	2
7	130.866	0.000	NYU-WISC	8	4

Table 2. The data reported shows the best "matches" found by the validation checker in the case of *P. falciparum* chromosome 2. The "in silico" sequence map was obtained from the TIGR database sequence. The `Bam-HI` sequence map (as well as its reversed) was checked against the map published in [3].

Chromosome 3 Validation Summary

rank	MLE/MWSS score	$\mathcal{P}$-score	map id	# missing cuts	# false cuts
1	87.020	1.000	NYU-WISC	1	0
2	184.232	1.000	NYU-WISC	15	3
3	290.579	1.000	NYU-WISC	13	5
4	231.384	0.998	NYU-WISC	14	2
5	237.722	0.002	NYU-WISC	13	5
6	192.926	0.000	NYU-WISC	15	5

Table 3. The data reported shows the best "matches" found by the validation checker in the case of *P. falciparum* chromosome 3. The "in silico" sequence map was obtained from the TIGR database sequence. The `Bam-HI` sequence map (as well as its reversed) was checked against the map published in [3].

3.2 Alignment Experiment of Contigs for Chromosome 12

We ran the alignment tool for the contigs assigned to each chromosome in the `www.plasmodb.org` database. In this section we show, as an example, part of the alignment results for chromosome 12. The complete results of our validation and alignment experiments is available at `http://bioinformatics.cat.nyu.edu/projects/PFalciparum/`. The alignment tool proposed an alignment of the contigs of length greater that 20 Kbps. We are limited to this number by the resolution of the particular optical maps used. Parts of the alignment are shown in Figure 2. The image shows contigs 11 and 13 from Chromosome 12 anchored in positions 39 and 50 of the optical map. In both cases the map overlap constraint was turned off.

3.3 Result Summary for Chromosome 12.

Restriction Enzyme `Bam-HI`**.**
Putative Sequence Anchors

Position	Sequence Contig	Reversal	Missing Cuts	False Cuts
17	`chr12_05`		0	0
54	`chr12_13`		0	0
26	`chr12_06`		1	0
41	`chr12_11`		1	0
23	`chr12_15`	RM	1	0
11	`chr12_04`		1	0
15	`chr12_07`	RM	1	0
77	`chr12_14`	RM	1	0
32	`chr12_09`	RM	2	0
23	`chr12_04`	RM	3	0
22	`chr12_11`	RM	4	0
36	`chr12_06`	RM	4	0
58	`chr12_15`		4	0
54	`chr12_15`	RM	6	0
16	`chr12_02`		6	0

Joins
`chr12_15` (R, RM) + `chr12_06`
`chr12_11` + `chr12_13`
`chr12_05` + `chr12_06`

Restriction Enzyme `Nhe-I`.
Putative Sequence Anchors

Position	Sequence Contig	Reversal	Missing Cuts	False Cuts
10	`chr12_16`	RM	1	0
49	`chr12_06`	RM	1	0
35	`chr12_16`		2	0
38	`chr12_10`	RM	2	0
26	`chr12_11`	RM	5	0

Joins: none.

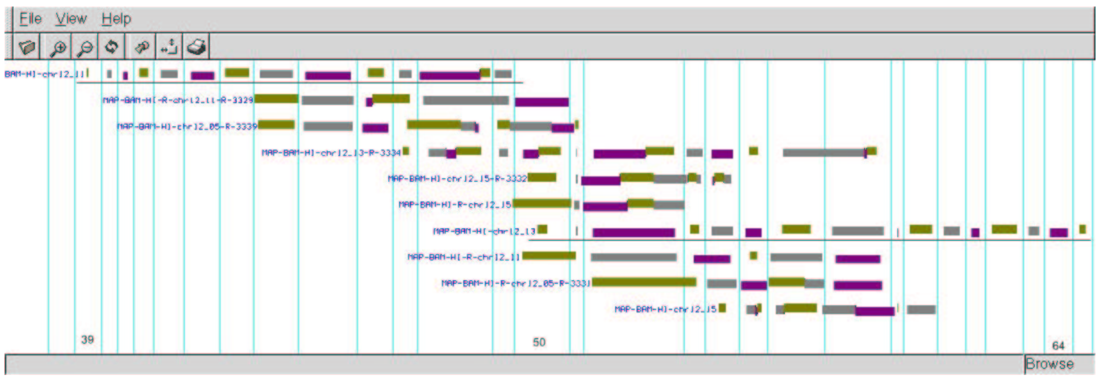

Fig. 2. A region of the alignment of Chromosome 12. The two underlined maps in position 39 and 50 of the map represent very good anchoring of contigs 11 and 13 to the (optical) map. The alignment was obtained without the overlap filter. The overlapping maps are candidate for a BLAST run.

References

1. M. Antoniotti, T. Anantharaman, S. Paxia, and B. Mishra. Genomics via Optical Mapping IV: Sequence Validation via Optical Map Matching. Technical Report CIMS-TR-811, NYU Courant Bioinformatics Group, 719 Broadway 12th Floor, New York, NY, 10003, U.S.A., 2001.
2. C. Aston, B. Mishra, and D. C. Schwartz. Optical Mapping and Its Potential for Large-Scale Sequencing Projects. *Trends in Biotechnology*, 17:297–302, 1999.
3. M. J. Gardner et al. Chromosome 2 sequence of the human malaria parasite *Plasmodium Falciparum*. *Science*, 282:1126–1132, 1998.
4. J. Jing, Z. Lai, C. Aston, J. Lin, D. J. Carucci, M. J. Gardner, B. Mishra, T. Anantharaman, H. Tettelin, L. M. Cummings, S. L. Hoffman, J. C. Venter, and D. C. Schwartz. Optical Mapping of *Plasmodium Falciparum* Chromosome 2. *Genome Research*, 9:175–181, 1999.
5. Z. Lai, J. Jing, C. Aston, V. Clarke, J. Apodaca, E. T. Dimalanta, D. J. Carucci, M. J. Gardner, B. Mishra, T. Anantharaman, S. Paxia, S. L. Hoffman, J. C. Venter, E. Huff, and D. C. Schwartz. A shotgun optical map of the entire *Plasmodium Falciparum* genome. *Nature Genetics*, 23:309–313, 1999.
6. X. Su, M. T. Ferdig, Y. Huang, C Q. Huynh, A. Liu, J You, J. C. Wootton, and T. E. Wellems. A Genetic Map and Recombination Parameters of the Human Malaria Parasite *Plasmodium falciparum*. *Science*, 286, 1999.
7. X.-Z. Su and T. E. Wellems. Genome Discovery and Malaria Research: Current Status and Promise. In I. W. Sherman, editor, *Malaria: Parasite Biology, Pathogenesis, and Protection*. ASM Press, Washington, D.C., U.S.A., 1998.