Deconvoluting BAC-gene Relationships Using a Physical Map

Y. Wu[1], L. Liu[1], T. Close[2], S. Lonardi[1]

[1]Department of Computer Science & Engineering

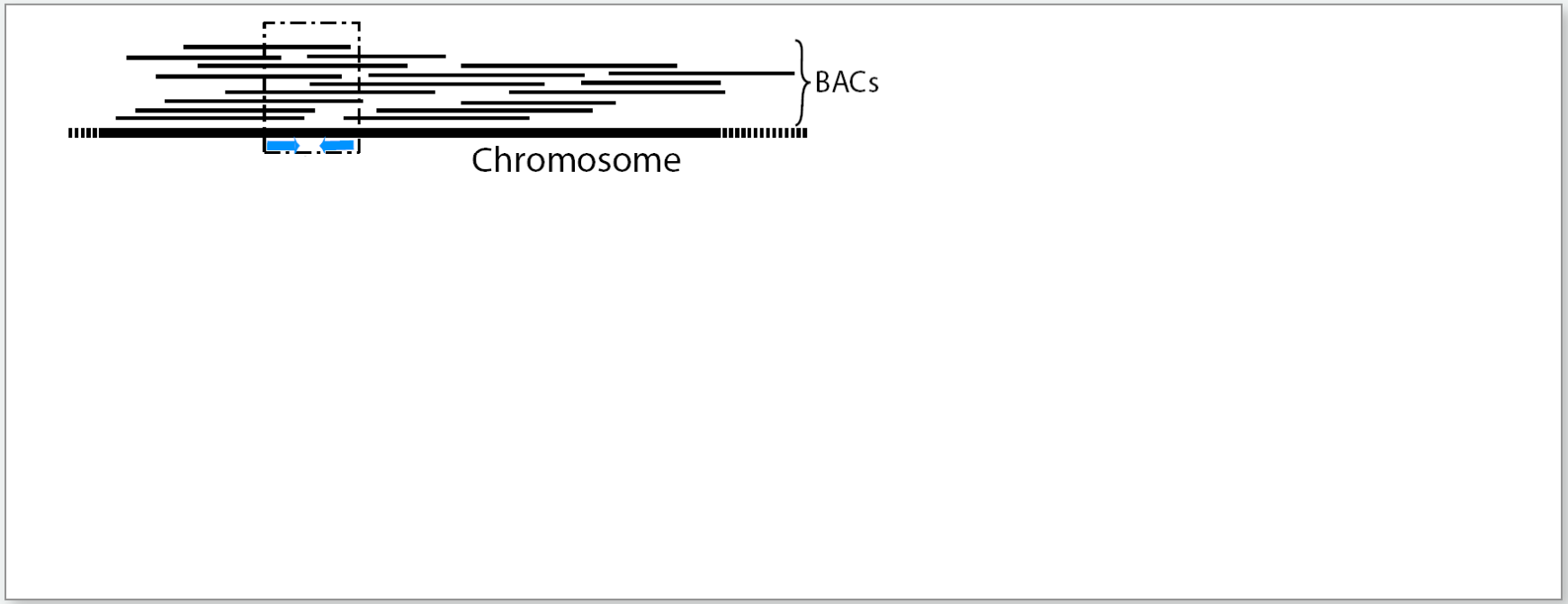
[2]Department of Botany & Plant Sciences

UNIVERSITY OF CALIFORNIA UCRIVERSIDE

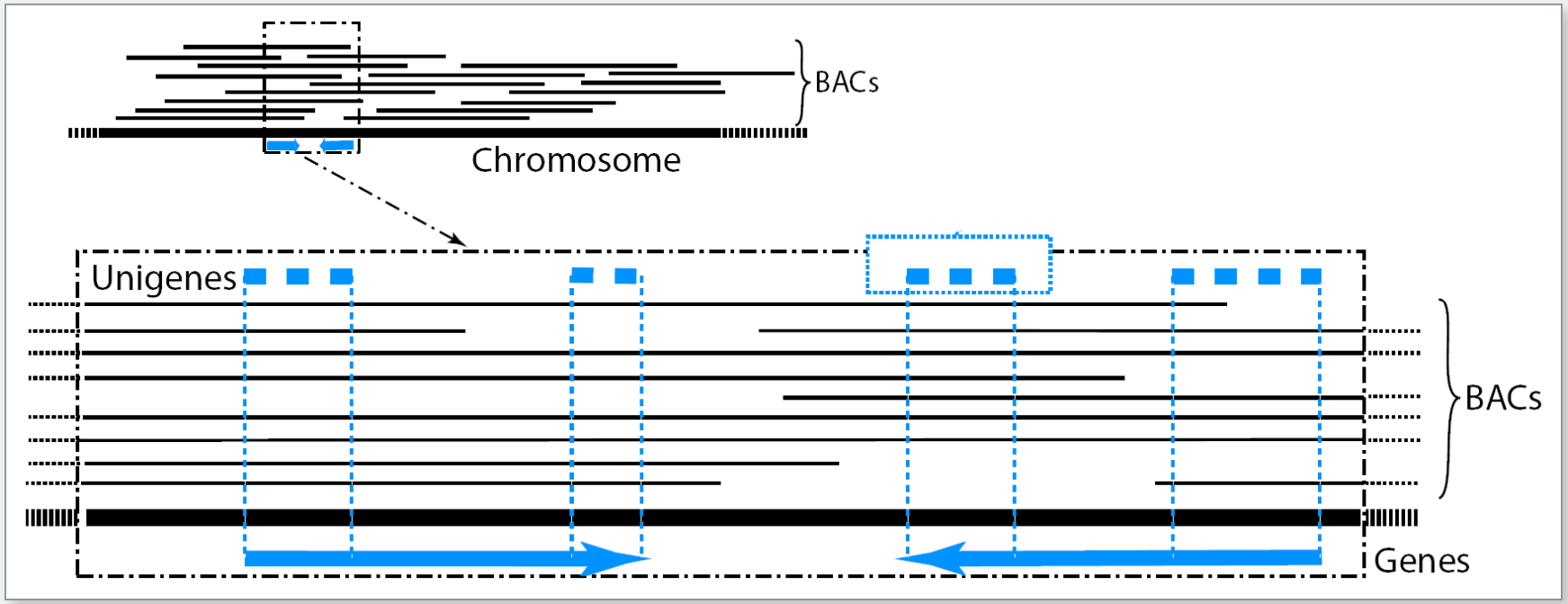
Selective sequencing

- Many organisms are unlikely to be sequenced in the near future due to the large size and highly repetitive content of their genomes
- *Selective sequencing:* obtain the sequence of a small set of BAC clones that contain a specific set of genes of interest
- How do we identify these BAC clones? *BAC-gene deconvolution problem*

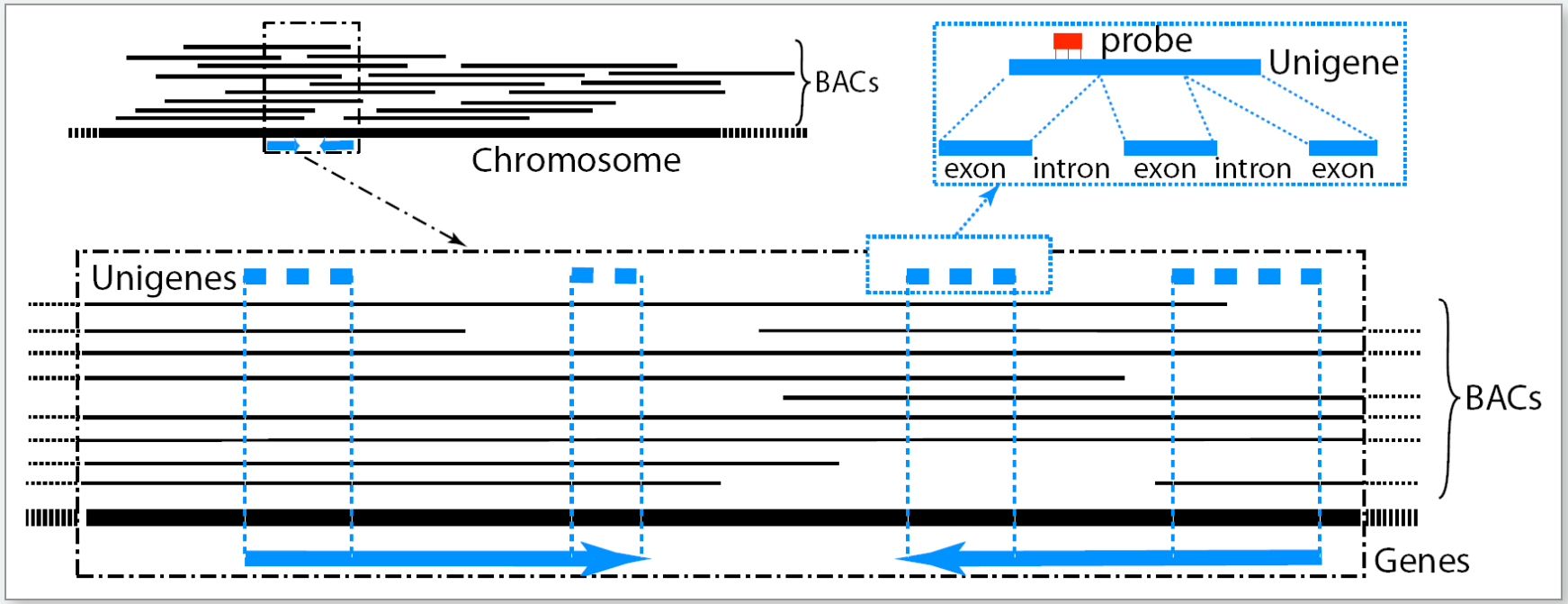
An illustration of the problem



An illustration of the problem



An illustration of the problem



Hybridization with probes

- The presence of a gene in a BAC can be determined by an hybridization experiment (e.g., using a unique probe designed from it)
- Given that typically BAC clones and probes could be in the order of tens of thousands, carrying out an experiment for each pair (BAC,probe) is usually unfeasible
- *Group testing* (or *pooling*) has to be used

Hybridization with pools of probes

- Probes can be arranged into pools for group testing. However, in order to achieve exact deconvolution this strategy could be still unfeasible due to the large number of pools
- *Question*: Can we use a small number of pools (e.g., 1- or 2-decodable pool design) and still achieve accurate deconvolution?

Dealing with the limitations of pooling

- *Answer:* Yes, if one compensates for the lack of information obtained by a weak pooling design with the knowledge of the overlapping structure of the BACs

- In this way, the number of pools required is reduced ⇒ less expensive/time-consuming

Hybridization data

h(b,p)=1 (pool *p* hybridizes to BAC *b*)

- b must contain at least one of the probes/genes represented by *p*
- positive information

h(b,p)=0 (pool *p* does not hybridize to BAC *b*)

- *b* cannot contain any of the probes/genes represented by *p*
- negative information

Deconvolution problem

- Given *h(b,p)* for all pairs *(b,p)* the *deconvolution problem* is to establish a one-to-many assignment between the probes *p* and the clones *b* in such a way that it satisfies the value of *h*

1. Basic deconvolution: uses only on information obtained from group testing
2. Improved deconvolution: also uses the physical map

Input to the basic deconvolution

Hybridization table

h	p_1	p_2	p_3	p_4
b_1	1	0	0	0
b_2	1	1	0	0
b_3	0	1	1	0
b_4	0	0	1	1
b_5	0	0	0	1

p_i is a pool
b_j is a BAC
u_k is a probe/gene

Input to the basic deconvolution

Hybridization table

h	p_1	p_2	p_3	p_4
b_1	1			
b_2	1	1		
b_3		1	1	
b_4			1	1
b_5				1

Pool content table

	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9
p_1	1	1	1						
p_2			1	1	1				
p_3					1	1	1		
p_4							1	1	1

p_i is a pool
b_j is a BAC
u_k is a probe/gene

Positive information

	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9
b_1,p_1	1	1	1						
b_2,p_1	1	1	1						
b_2,p_2			1	1	1				
b_3,p_2			1	1	1				
b_3,p_3					1	1	1		
b_4,p_3					1	1	1		
b_4,p_4							1	1	1
b_5,p_4							1	1	1

p_i is a pool
b_j is a BAC
u_k is a probe/gene

Negative information

	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9
b_1			0	0	0	0	0	0	0
b_2					0	0	0	0	0
b_3	0	0	0				0	0	0
b_4	0	0	0	0	0				
b_5	0	0	0	0	0	0	0		

p_i is a pool
b_j is a BAC
u_k is a probe/gene

Combining positive & negative

	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9
b_1,p_1	1	1	1						
b_2,p_1	1	1	1						
b_2,p_2			1	1	1				
b_3,p_2			1	1	1				
b_3,p_3					1	1	1		
b_4,p_3					1	1	1		
b_4,p_4							1	1	1
b_5,p_4							1	1	1

p_i is a pool
b_j is a BAC
u_k is a probe/gene

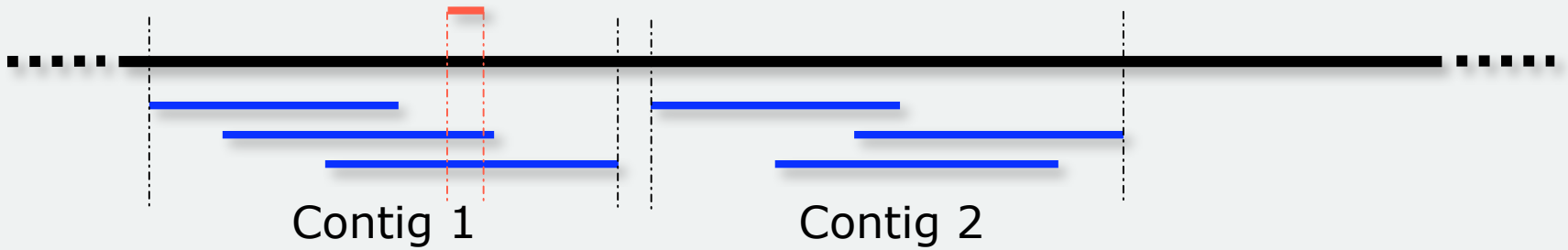
Combining positive & negative

	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9
b_1,p_1	1	1							
b_2,p_1	1	1	1						
b_2,p_2			1	1					
b_3,p_2				1	1				
b_3,p_3					1	1			
b_4,p_3						1	1		
b_4,p_4							1	1	1
b_5,p_4								1	1

- Each row represents a *constraint* to be satisfied
- If a row contains only one “1”, then the relationship between the BAC and probe is resolved exactly

p_i is a pool
b_j is a BAC
u_k is a probe/gene

Physical map-assisted deconvolution



- Basic deconvolution is not sufficient
- BACs are assembled into contigs by FPC (a *contig* is a set of BAC clones)
- We assume the probes are unique ⇒ each probe can belong to exactly one contig

Optimization problem

- We formulate the following optimization problem

MAXIMUM CONSTRAINT SATISFYING PROBE-CONTIG ASSIGNMENT (MCSPCA)
Instance: A set of probes $\mathbb{O}$, a set of contigs $\mathbb{C}$ and a list of constraints Ω', where each item of Ω' has the form $(\mathbf{c}, \mathbb{O}_{b,\mathbf{p}})$, $\mathbf{c} \in \mathbb{C}$ and $\mathbb{O}_{b,\mathbf{p}} \subseteq \mathbb{O}$.
Objective: Assign each probe to at most one contig in $\mathbb{C}$ such that the number of satisfied constraints in Ω' is maximized. A constraint $(\mathbf{c}, \mathbb{O}_{b,\mathbf{p}})$ is satisfied if one or more of the probes from $\mathbb{O}_{b,\mathbf{p}}$ is assigned to $\mathbf{c}$.

- The problem is NP-complete (proof in the paper, reduction from 3SAT)

Integer Linear Programming

- The optimization problem can be solved via integer linear programming (ILP)

$$
\begin{aligned}
&Maximize \sum_{q \in \Omega'} Y_q \\
&Subject\ to \sum_{\mathbf{c} \in \mathbb{C}} X_{o,\mathbf{c}} \leq 1 && \forall o \in \mathbb{O} \\
&Y_{q=(\mathbf{c},S)} \leq \sum_{o \in S} X_{o,\mathbf{c}} && \forall q \in \Omega' \\
&X_{o,\mathbf{c}} \in \{0,1\} && \forall o \in \mathbb{O}, \mathbf{c} \in \mathbb{C} \\
&Y_q \in \{0,1\} && \forall q \in \Omega'
\end{aligned}
$$

LP and randomized rounding

- The ILP is relaxed to the corresponding LP, then the LP is solved exactly (via the GLPK package)
- Optimal solution to the LP is mapped to a valid solution to the ILP via randomized rounding
- We prove that our method achieves approximation ratio $(1-e^{-1})$

Experimental results on rice genome

- Whole genome sequence for rice is available
- BAC library and fingerprinting data are available from AGI
- BAC-end sequences are also available from Genbank
- Physical map was built using FPC
- Coordinates of the BAC on the genome were determined by BLASTing BAC-end sequences against the genome

Experimental results on rice genome

- Rice unigenes are available from NCBI
- Unique probes for the unigenes were designed by the Oligospawn software
- Experiments focused on chromosome I
- Probe pools were designed following the shifted transversal design (STD)
- Dataset: 2,002 probes and 2,629 BACs

Experimental results

1-decodable pooling design

pooling	#pools	# true assigns	basic recall	MCSPCA recall	MCSPCA precision
$P = 13, L = 3$	39	14742	0.0103	0.199	0.2647
$P = 47, L = 2$	94	14742	0.0173	0.4005	0.5236

Experimental results

2-decodable pooling design

pooling	#pools	# true assigns	basic recall	MCSPCA recall	MCSPCA precision
$P = 13, L = 5$	65	14742	0.2726	0.618	0.7668
$P = 47, L = 3$	141	14742	0.763	0.9069	0.9446

Experimental results

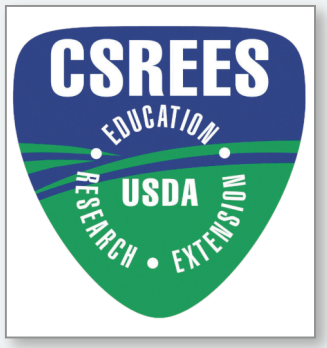
pooling	#pools	# true assigns	basic recall	MCSPCA recall	MCSPCA precision
$P = 13, L = 3$	39	14742	0.0103	0.199	0.2647
$P = 13, L = 5$	65	14742	0.2726	0.618	0.7668
$P = 47, L = 2$	94	14742	0.0173	0.4005	0.5236
$P = 47, L = 3$	141	14742	0.763	0.9069	0.9446

Findings

- We proposed a new method to solve the BAC-gene deconvolution problem based on integer linear programming
- Experimental results show that our method is accurate and effective

Thank you

- Funding



- Serdar Bozdag (UC Riverside) for providing the rice data (fingerprinting and hybridization)