Dictionary Models for Haplotype Analysis

Kenneth Lange
Kristin Ayers
Chiara Sabatti

UCLA Departments of Biomathematics,
Human Genetics, and Statistics

January, 2006

Dictionary Models for DNA Sequence Data

1. Proposed by Bussemaker et al (2000) PNAS 97:11096-10100

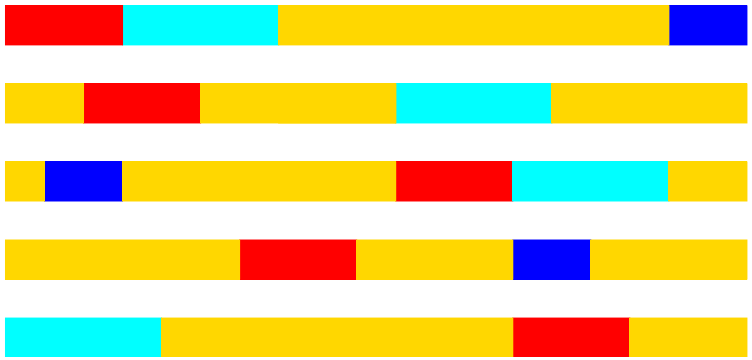
2. A DNA sequence is constructed by concatenating words chosen randomly from a predefined dictionary.

3. The sequence includes no spaces or punctuation marks.

4. The DNA dictionary consists of words and their probabilities.

5. The 4 bases A, T, C, and G are viewed as words of length 1. These trivial words supply background variation.

Sequences



Dictionary

Our Extensions to the Dictionary Model

1. We extended the model to permit alternative spellings and Bayesian priors.

2. Showed how to correctly compute likelihoods under the model.

3. Derived algorithms for maximum likelihood and maximum a posteriori estimation of parameters.

4. Provided an algorithm for computing posterior probabilities of motif occurrence.

5. Implemented the algorithms in code and applied them to *E. Coli* sequence and microarray data.

Successes of the Motif Model

1. In collaboration with Lars Rohlin and James Liao of UCLA, we were able to find many known binding sites in *E. Coli* and predict new ones.

2. We were able to model spelling variations in a systematic way.

3. Our worst results occur when motifs overlap.

4. Nonetheless, we do better than Robinson et al. (1998), who rely on similarity scores.

5. The next slide gives some statistics for known motifs using a posterior probablity cutoff of 0.5 for motif calling.

Binding Domain	recovered sites	missed sites	imputed sites
araC	6	0	6
arcA	8	5	28
argR	15	2	24
cpxR	11	1	29
creB	8	0	9
crp	36	13	131
cspA	4	0	4
cytR	2	3	7
dnaA	7	1	41
fadR	7	0	8
fis	8	7	36
fliA	12	0	14
fnr	12	0	14
fruR	12	0	18
fur	8	1	18
galR	7	0	10
gcvA	4	0	4
glpR	7	6	20
hipB	2	2	2
lexA	19	0	24
malT	4	6	6

Binding Domain	recovered sites	missed sites	imputed sites
metJ	6	3	8
metR	5	3	10
nagC	6	0	9
narL	7	3	9
narP	8	0	4
ntrC	4	1	4
ompR	5	4	28
oxyR	4	0	4
phoB	10	2	12
purR	21	1	25
rpoH2	6	1	6
rpoH3	8	0	8
rpoN	6	1	11
rpoS17	5	10	9
rpoS18	4	3	8
soxS	11	6	22
torR	3	1	5
trpR	4	0	4
tus	5	0	5
tyrR	13	4	19
Total	340	90	663

Application to Microarray Data

1. We used the motif model to explain the gene expression data of Courcelle (2002), who exposed *E. Coli* to UV light. This is known to affect the genes regulated by LexA.

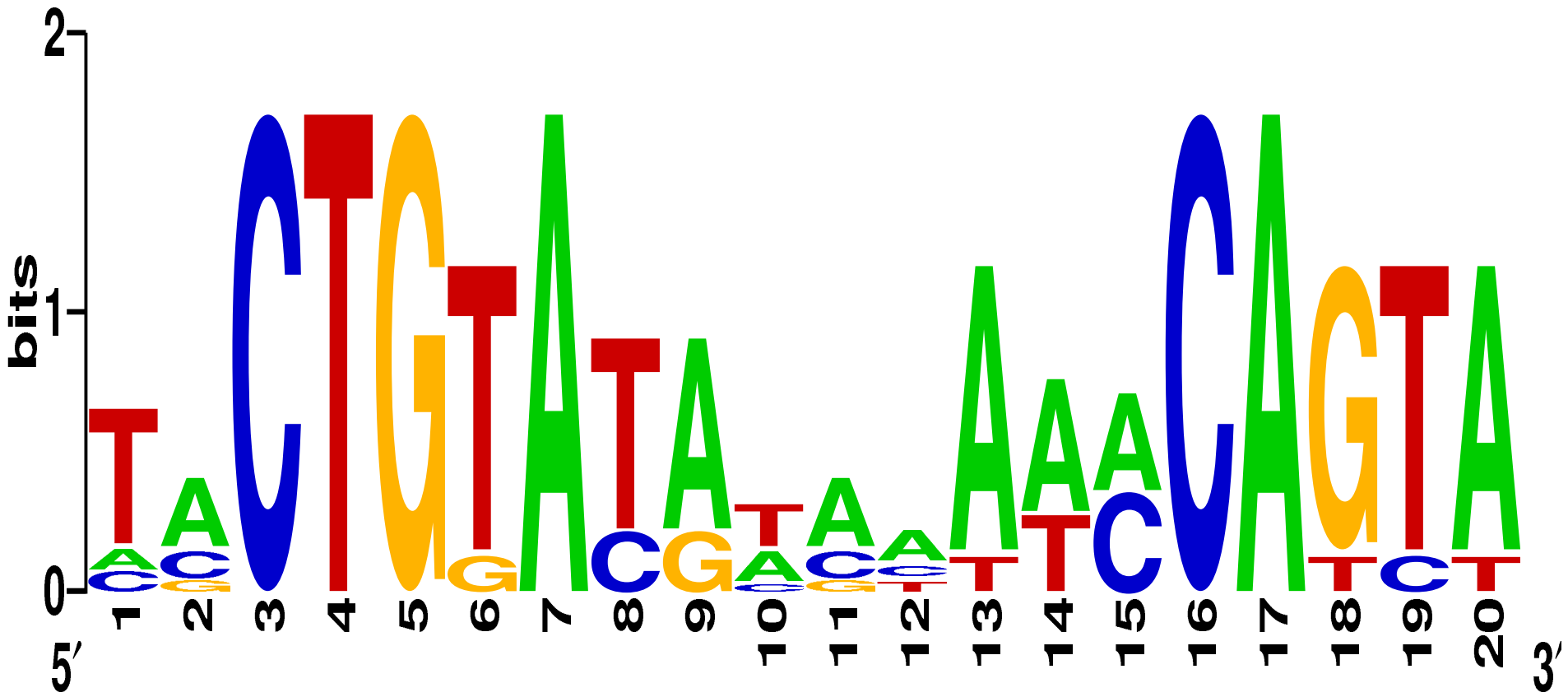
2. We regressed the log change in gene expression against the expected number of binding sites in the upstream region of each gene, for all regulatory proteins in our dictionary.

3. The explanatory variable that gives the most significant result is LexA.

4. Similar work appears in (Bussemaker et al, Nat Gen 2001; Conlon et al, PNAS 2003; Keles et al, Bioinformatics 2002).

Some Predictors for the UV Experiment

	Estimate	Std. Err	t value	Pr(>|t|)	
Intercept	-2.021e-02	7.223e-03	-2.798	0.00518	**
malT	2.201e-01	3.320e-01	0.663	0.50737	
:					
cspA	5.521e-02	1.176e-01	0.469	0.63889	
lexA	3.841e-01	4.572e-02	8.401	< 2e-16	***
fnr	3.284e-02	7.832e-02	0.419	0.67501	
:					
hipB	4.159e-01	2.494e-01	1.668	0.09552	.
fis	1.428e-01	3.312e-02	4.313	1.67e-05	***
oxyR	4.259e-02	2.878e-01	0.148	0.88236	
:					


bits
2
1
0
5′
3′
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
weblogo.berkeley.edu

Dictionary Models for Haplotypes

1. There is limited haplotype diversity over short chromosome segments in humans.

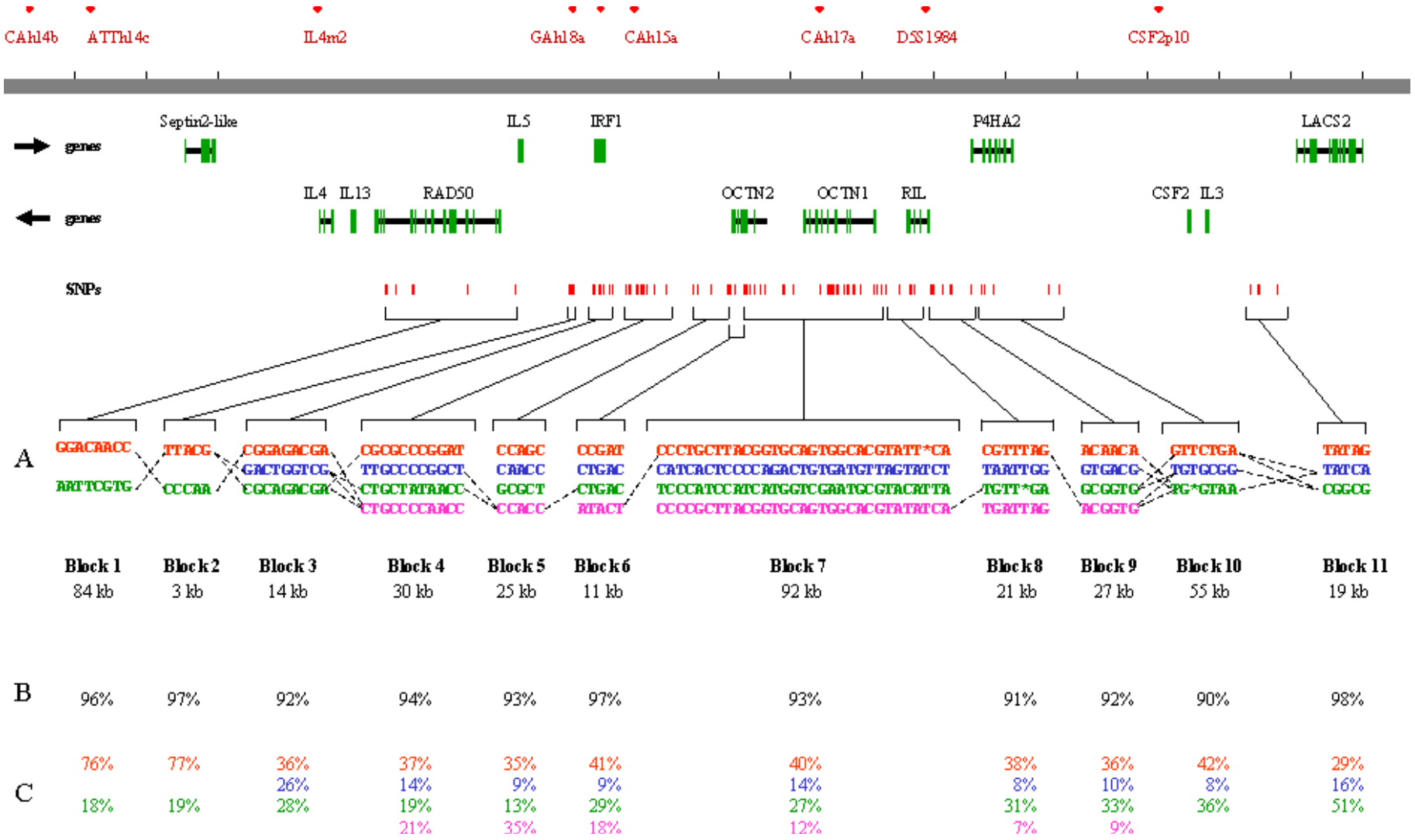
2. Extreme linkage disequilibrium (LD) prevails in many cases.

3. Just a few tagging markers are often sufficient to identify a haplotype.

4. Daly et al (Science, 2001) argue that haplotypes break along block boundaries.

5. In fact, haplotypes often straddle block boundaries.

6. We need phenomenological models as well as mechanistic models to interpret results. Dictionary models fall in the former category.


CAh14b
ATTh14c
IL4m2
GAh18a
CAh15a
CAh17a
DS51984
CSF2p10
genes
Septin2-like
IL5
IRF1
P4HA2
LACS2
genes
IL4
IL13
RAD50
OCTN2
OCTN1
RIL
CSF2
IL3
SNPs
A
B
C
Block 1
84 kb
Block 2
3 kb
Block 3
14 kb
Block 4
30 kb
Block 5
25 kb
Block 6
11 kb
Block 7
92 kb
Block 8
21 kb
Block 9
27 kb
Block 10
55 kb
Block 11
19 kb
96% 97% 92% 94% 93% 97% 93% 91% 92% 90% 98%
76% 77% 36% 37% 35% 41% 40% 38% 36% 42% 29%
26% 14% 9% 9% 14% 8% 10% 8% 16%
18% 19% 28% 19% 13% 29% 27% 31% 33% 36% 51%
21% 35% 18% 12% 7% 9%

Haplotype Blocks

1. Some combinations of marker alleles almost always occur together. This kind of variation is best described by a limited collection of haplotypes.

2. Block boundaries may correspond to recombination hotspots.

3. Even without recourse to recombination hotspots, genetic drift and population bottlenecks can partially explain haplotype uniformity.

4. Consecutive blocks are a parsimonious way of describing the genome even if they do not reliably reflect human evolutionary history.

5. Many algorithms for identifying blocks of markers have been suggested, implemented, and refined.

Reasons for Modeling Haplotype Blocks

1. To understand the forces behind the formation of blocks.

2. To predict the extent of LD in different regions of the genome.

3. To compare different populations and understand their demographic histories.

4. To take advantage of limited haplotype diversity in gene mapping and in the selection of tagging SNPs.

5. To inform genotype calling and permit construction of haplotypes in the absence of family data.

Dictionary Models for Haplotypes

1. Haplotyping is similar to parsing DNA sequence data.

2. The alphabets from which letters are drawn vary from marker to marker.

3. Sharp block boundaries are unnecessary.

4. The observed data usually consist of multimarker genotypes; the implied phase ambiguities complicate haplotyping.

5. Phase can be deduced from relatives or by typing single isolated chromosomes.

Haplotype Segment Dictionary

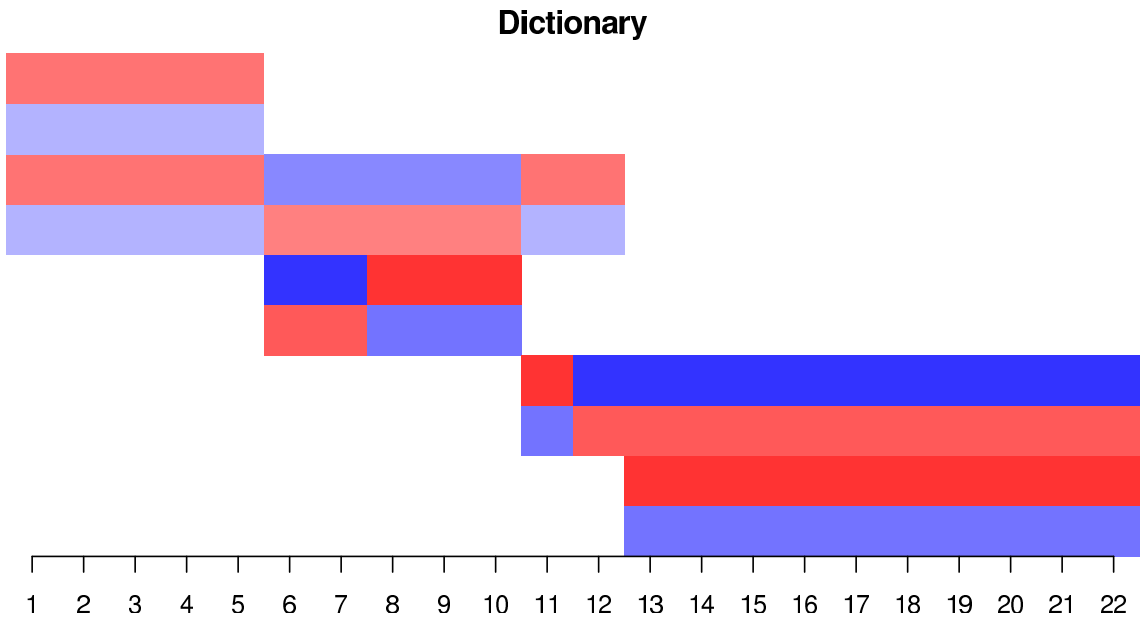
1. Each haplotype sequence is constructed by random concatenation of haplotypes segments.

2. Genetically, a haplotype segment represents an ancestral combination of alleles that are almost always co-transmitted.

3. In contrast to a word in a DNA motif dictionary, a haplotype segment always spans the same marker segment.

4. The haplotype segments spanning a given marker segment constitute a haplotype block. The marker segments corresponding to different blocks can overlap.

5. Haplotypes may have misspellings due to mutation and genotyping errors. If a marker within a conserved haplotype segment is polymorphic, then it is probably best to replace the segment by two different segments.


Dictionary
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22

Notation

1. A marker segment consists of a set of consecutive indices $[i:j] = (i, i+1, \ldots, j-1, j)$.

2. The sequence of alleles of a haplotype h on the marker segment $[i:j]$ is denoted $h_{[i:j]}$.

3. A haplotype block $\mathcal{B}_{[i:j]}$ is assigned probability $q_{[i:j]}$, with the implied constraint $\sum_{j \geq i} q_{[i:j]} = 1$.

4. A haplotype segment $s \in \mathcal{B}_{[i:j]}$ is assigned conditional probability r_s, with the implied constraint $\sum_{s \in \mathcal{B}_{[i:j]}} r_s = 1$.

5. A partition π divides m markers into consecutive marker segments π_1 through $\pi_{|\pi|}$. Segment π_1 begins with marker 1, and segment $\pi_{|\pi|}$ ends with marker m. If segment π_i ends with marker j, then segment π_{i+1} starts with marker $j+1$.

Forward Likelihood Algorithm for a Haploytpe

1. Let E_i be the event that a random haplotype H of length m is constructed with a haplotype segment ending at marker i.

2. The forward algorithm computes $f_i = \Pr(H_{[1:i]} = h_{[1:i]}, E_i)$.

3. Starting with $f_0 = 1$, we update these joint probabilities by

$$f_i = \sum_{k=1}^{\min\{d,i\}} f_{i-k} q_{[i-k+1:i]} r_{h_{[i-k+1:i]}},$$

   where d is the maximum length of a haplotype segment.

4. f_m is then the probability that $H = h$.

Backward Likelihood Algorithm for a Haploytpe

1. Let E_i be the event that a random haplotype H of length m is constructed with a haplotype segment ending at marker i.

2. We compute $b_i = \Pr(H_{[i:m]} = h_{[i:m]} \mid E_{i-1})$ in the backward algorithm.

3. Starting with $b_{m+1} = 1$, we update these conditional probabilities by

$$b_i = \sum_{k=1}^{\min\{d, m-i+1\}} q_{[i:i+k-1]} r_{h_{[i:i+k-1]}} b_{i+k},$$

   where d is the maximum length of a haplotype segment.

4. b_1 is then the probability that $H = h$.

Incorporation of Genotyping Error

1. The forward and backward algorithms just given do not incorporate genotyping error.

2. To handle such errors, we introduce a penetrance function $\phi(h_k \mid s_k)$, which is the probability of the observed allele h_k given the true allele s_k at marker k.

3. Under the uniform error model, $\phi(h_k \mid s_k)$ equals $1 - \epsilon_k$ (a match) or ϵ_k (a mismatch).

4. In the forward and backward algorithms, the factor $r_{h_{[i:j]}}$ is replaced by

$$p(h_{[i:j]}) = \sum_{s \in \mathcal{B}_{[i:j]}} r_s \prod_{k=i}^{j} \phi(h_k \mid s_k)$$

assuming independent genotyping errors.

Multimarker Genotypes rather than Haplotypes

1. Let S_g denote the set of observed haplotype pairs (h^m, h^p) consistent with the observed multimarker genotype g. Here h^m is a maternal haplotype and h^p a paternal haplotype.

2. The likelihood of g is

$$\Pr(G = g) = \sum_{(h^m,h^p)\in S_g} \Pr(H^m = h^m) \Pr(H^p = h^p).$$

3. The set S_g contains 2^n elements if S_g has n heterozygous genotypes.

4. When parents are typed, some heterozygous genotypes can be resolved. With codominant alleles, no genotyping error, and both parents typed, a marker has a probability of at most $\frac{1}{8}$ of presenting a phase ambiguity. If only one parent is typed, this increases to at most $\frac{1}{4}$.

EM Algorithms for Parameter Estimation

1. Each entry of the parameter vector $\theta = (q, r, \epsilon)$ is a success probability for a hidden multinomial trial.

2. If N equals the random number of trials for component i and N_i the number of successes over these trials, then the EM update for θ_i is

$$\theta_i^{n+1} = \frac{\mathsf{E}(N_i \mid \text{obs}, \theta^n)}{\mathsf{E}(N \mid \text{obs}, \theta^n)}.$$

3. The outcome vectors f_i and b_i of the forward and backward algorithms furnish the raw material for computing these conditional expectations. Sandwich formulas apply.

4. Maximum a posteriori estimation can be implemented by trivial modification of the EM updates if we impose independent Dirichlet priors on the hidden multinomial trials.

Dictionary Construction

1. The inverse problem of constructing a dictionary from observed haplotype data is much harder than parameter estimation.

2. The minimum description length (MDL) approach strikes a balance between completeness and parsimony.

3. The MDL approach chooses the dictionary that minimizes the sum of the negative loglikelihood and penalty terms reflecting the complexity of the dictionary.

4. MDL penalizes (a) the number of parameters, (b) the size of each haplotype block, and (c) the complexity of the haplotype segments in each haplotype block.

Heuristics of Growing and Pruning

1. We start with a fairly large dictionary assembled from short to medium-length haplotype segments that are over represented in the data. Any method of defining rigid block boundaries will give such a dictionary.

2. In a growing phase, we check whether the concatenation of two adjacent haplotype segments is over represented in the data. If so, the concatenated segment is added.

3. In a pruning stage, we rank blocks by their apparent usage probabilities in the data. Seldom used blocks are eliminated by a bisection strategy based on MDL.

4. Pruning can also drop redundant haplotype segments such as those used in adding concatenated segments.

5. Growing and pruning are alternated until no further progress is made.

Toy Example of Dictionary Construction

1. Using 400 fully-phased chromosomes generated randomly without typing error, we were able to perfectly reconstruct the toy dictionary shown earlier involving 22 SNPs and 10 non-trivial haplotype segments. Part (a) of the next figure shows the numerical details of the original dictionary. Haplotype segments of length 1 are omitted for clarity.

2. When we introduce a 5% error rate and include 5% missing data in generating the 400 chromosomes, we get the reconstruction show in part (b) of the next figure.

3. Now reconstruction is imperfect because of confusing block overlap at SNPs 11 and 12.

4. Nonetheless, the reconstruction preserves major details.

BLOCK	BLOCK PROB	START	FINISH	SEGMENT PROB	HAPLOTYPE SEGMENT
1	.5000	1	5	.6000	11111
1	.5000	1	5	.4000	22222
2	.5000	1	12	.6000	111112222211
2	.5000	1	12	.4000	222221111122
3	1.0000	6	10	.6000	11222
3	1.0000	6	10	.4000	22111
4	1.0000	11	22	.6000	122222222222
4	1.0000	11	22	.4000	211111111111
5	1.0000	13	22	.6000	1111111111
5	1.0000	13	22	.4000	2222222222

(a)

BLOCK	BLOCK PROB	START	FINISH	SEGMENT PROB	HAPLOTYPE SEGMENT
1	.3653	1	5	.5280	11111
1	.3653	1	5	.4720	22222
2	.6306	1	12	.0425	111112211112
2	.6306	1	12	.6485	111112222211
2	.6306	1	12	.3089	222221111122
3	.6039	6	12	.5704	1122212
3	.6039	6	12	.4296	2211112
4	.9690	11	12	1.0000	12
5	.9990	13	22	.5349	1111111111
5	.9990	13	22	.4651	2222222222

(b)

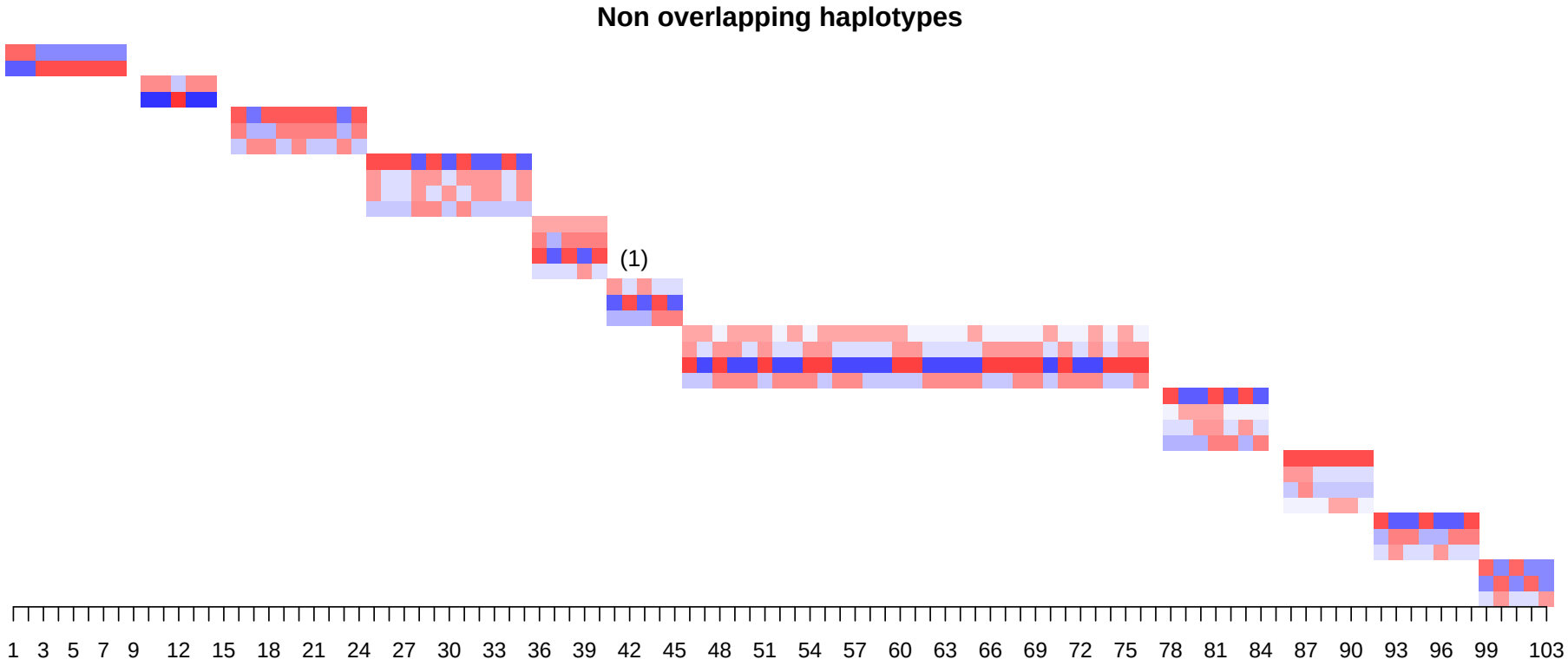
Dictionary Construction of Daly et al. Data

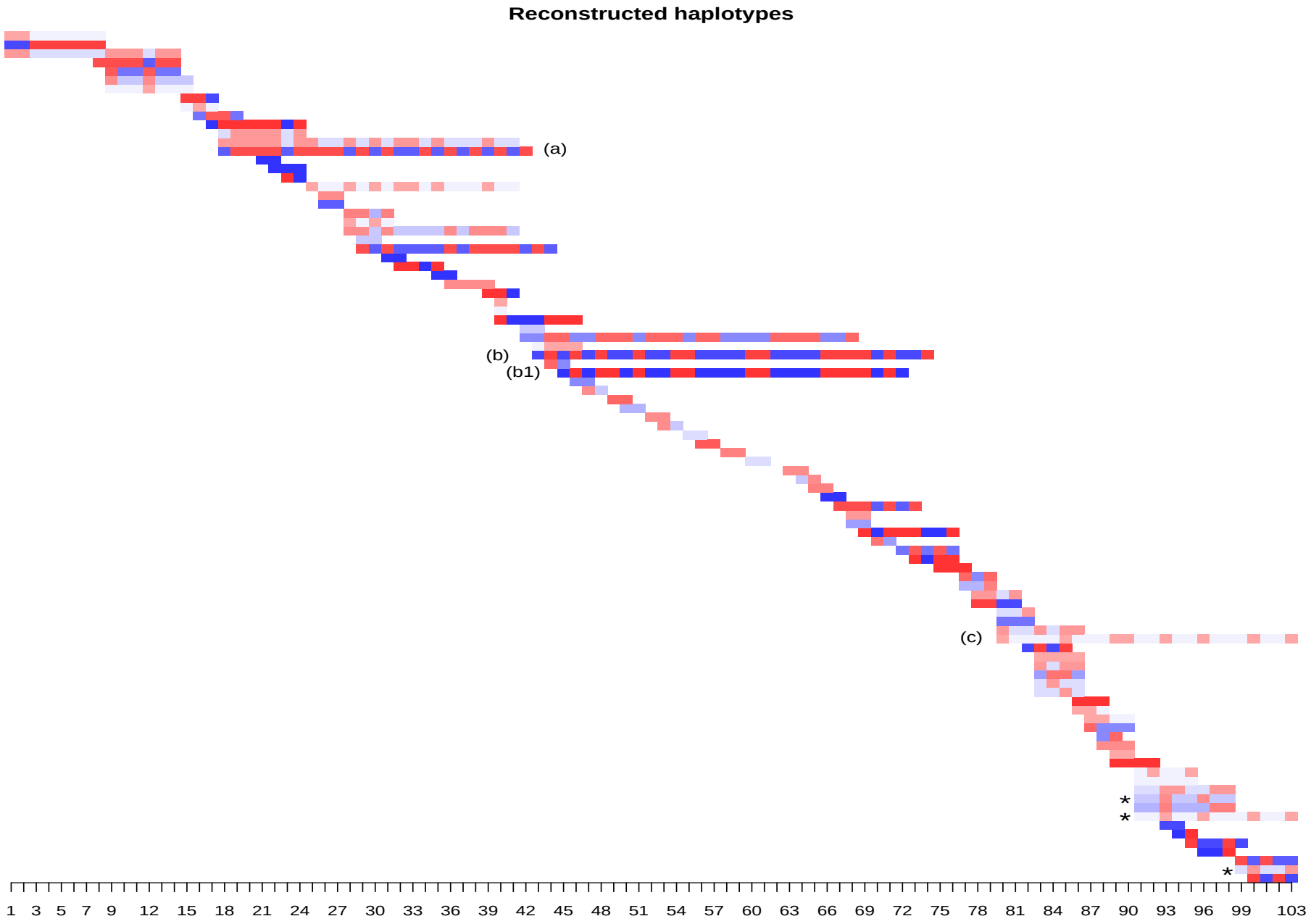
1. The Daly et al. data set involves 103 SNPs. Based on 129 fully-phased multimarker genotypes, we reconstructed two dictionaries, one with sharp block boundaries and one with overlapping boundaries.

2. In the next figure, the two SNP alleles are colored red and blue. The color intensity of a haplotype segment is proportional to the probability of its appearance on a random chromosome.

3. The MDL criterion overwhelmingly favors the overlapping dictionary.

4. This dictionary suggests a phylogeny for some of the haplotype segments. For example, the haplotype segment labeled (c) can be interpreted as ancestral to the segments indicated with an asterisk (*).


Non overlapping haplotypes
(1)
1 3 5 7 9 12 15 18 21 24 27 30 33 36 39 42 45 48 51 54 57 60 63 66 69 72 75 78 81 84 87 90 93 96 99 103


Reconstructed haplotypes
(a)
(b)
(b1)
(c)
*
*
*
1 3 5 7 9 12 15 18 21 24 27 30 33 36 39 42 45 48 51 54 57 60 63 66 69 72 75 78 81 84 87 90 93 96 99 103

Application to Haplotyping and Genotype Calling

1. The dictionary model can be used to haplotype isolated individuals.

2. Haplotyping works best when block boundaries are not sharp. With sharp boundaries, the two haplotype segments within each block may be well determined, but it is impossible to decide which is maternal and which is paternal. The overall phase uncertainty increases geometrically with the number of blocks. Boundary straddling alleviates this problem.

3. Because of the number of possible haplotype pairs, it is preferable to explore haplotype-pair space by MCMC methods.

4. In an MCMC run, the state is a pair of true haplotypes, not a complete decomposition of two true haplotypes into haplotype segments. This makes for a smaller state space.

Typing Error at the Genotype Level

1. A multimarker genotype g is generated by a maternal haplotype h^m plus a paternal haplotype h^p.

2. Genotyping error can be incorporated at the level of the conditional probability $\Pr\left(G = g \mid \frac{H^m}{H^p} = \frac{h^m}{h^p}\right)$.

3. Under a product multinomial model,

$$\begin{aligned}\Pr\left(G = g \mid \frac{H^m}{H^p} = \frac{h^m}{h^p}\right) &= \prod_{k=1}^{m} \Pr\left(G_k = g_k^1/g_k^2 \mid \frac{H_k^m}{H_k^p} = \frac{h_k^m}{h_k^p}\right) \\ &= \prod_{k=1}^{m} [\phi(g_k^1 \mid h_k^m)\phi(g_k^2 \mid h_k^p) \\ &\quad + 1_{\{g_k^1 \neq g_k^2\}}\phi(g_k^2 \mid h_k^m)\phi(g_k^1 \mid h_k^p)],\end{aligned}$$

   where $\phi(g_k^j \mid h_k^i)$ is the probability of calling allele g_k^j given the true allele h_k^i. Under the uniform error model, $\phi(g_k^j \mid h_k^i)$ equals $1 - \epsilon_k$ (a match) or ϵ_k (a mismatch).

MCMC Steps

1. In executing an MCMC run, we alternative two kinds of steps.

2. In a Gibbs step, we choose a random marker k and replace an ordered genotype $\frac{h_k^m}{h_k^p}$ at this marker by a random ordered genotype in proportion to its posterior probability.

3. In a Metropolis swap, we choose a random marker k and swap the terminal portions of the haplotype pair $\frac{h^m}{h^p}$ from this marker onward. This kind of move is accepted with the usual Metropolis probability.

4. The two kinds of moves provide for local and global rearrangements of the existing state.

5. Both moves are quick to execute because it is easy to compute the three probabilities $\Pr(H^m = h^m)$, $\Pr(H^p = h^p)$, and $\Pr\left(G = g \mid \frac{H^m}{H^p} = \frac{h^m}{h^p}\right)$.

Running the Chain

1. After a burnin, we run the chain for 10^6 steps and sample every 100 steps.

2. We record the ordered and unordered genotypes at each marker. The proportion of time each unordered genotype occurs provides the posterior probability of that genotype. These posterior probabilities help in spotting poor genotype calls, particularly when SNPs redundantly define haplotype segments.

3. We record the proportion of time that each unordered pair of haplotypes occur. This is the raw material for haplotyping.

4. If the ordered genotype for a marker is taken as given, say on the basis of evidence from parents, then this marker is not visited or revised in either the Gibbs resampling or Metropolis swaps.

LOCUS #	TRUE HAPLO-TYPE PAIR		OBSERVED GENOTYPE
1	2	1	1 - 2
2	2	1	2 - 1
3	2	1	2 - 1
4	2	1	1 - 2
5	2	1	1 - 2
6	1	2	2 - 1
7	1	2	2 - 1
8	1	2	1 - 2
9	1	2	1 - 2
10	1	2	2 - 1
11	2	1	2 - 1
12	2	1	2 - 1
13	1	2	2 - 1
14	1	2	1 - 2
15	1	2	1 - 2
16	1	2	2 - 1
17	1	2	2 - 1
18	1	2	2 - 1
19	1	2	0 - 0
20	1	2	1 - 2
21	1	2	2 - 2
22	1	2	2 - 1

Pair 1 (.1883)		Pair 2 (.1571)		Pair 3 (.1791)		Pair 4 (.1564)		Posterior Probabilities
1	2	1	2	1	2	1	2	
1	2	1	2	1	2	1	2	
1	2	1	2	1	2	1	2	
1	2	1	2	1	2	1	2	
1	2	1	2	1	2	1	2	
2	1	2	1	2	1	2	1	
2	1	2	1	2	1	2	1	
2	1	2	1	2	1	2	1	
2	1	2	1	2	1	2	1	
2	1	2	1	2	1	2	1	
1	2	1	2	1	2	1	2	
1	2	1	2	1	2	1	2	
1	2	1	2	2	1	2	1	
1	2	1	2	2	1	2	1	
1	2	1	2	2	1	2	1	
1	2	1	2	2	1	2	1	
1	2	1	2	2	1	2	1	
1	2	1	2	2	1	2	1	
1	2	1	2	2	1	2	1	
1	2	1	2	2	1	2	1	
1	2	2	2	2	1	2	2	
1	2	1	2	2	1	2	1	

Expected Copies Employed for a Haplotype Segment

1. Partial haplotype information is better than none.

2. We compute the expected number of copies of each nontrivial haplotype segment s in an observed multimarker genotype g. This done by taking the MCMC time average of the sum $\mathsf{E}(N_s^m \mid H^m = h^m) + \mathsf{E}(N_s^p \mid H^p = h^p)$, where N_s^m and N_s^p are random indicators of whether s occurs in the maternal and paternal haplotypes h^m and h^p.

3. At a recorded epoch of the chain, either h^m is inconsistent with $s \in \mathcal{B}_{[i:j]}$ or

$$\mathsf{E}(N_s^m \mid H^m = h^m) \;=\; \frac{f_{i-1} q_{[i:j]} r_s b_{j+1}}{\mathsf{Pr}(H^m = h^m)}.$$

   is readily computed as a byproduct of the forward and backward recurrences.

START	FINISH	EXPECTED	HAPLOTYPE SEGMENT
1	12	0.8620	111112222211
1	12	0.8704	222221111122
13	13	0.5290	1
13	22	0.4685	1111111111
13	22	0.9012	2222222222
14	14	0.5289	1
15	15	0.5350	1
16	16	0.5350	1
17	17	0.5429	1
18	18	0.5228	1
19	19	0.5414	1
20	20	0.5203	1
21	21	0.5907	2
22	22	0.5328	1

PAIR	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
1-1	.0189	.0013	.0010	.0029	.0075	.0059	.0029	.0096	.0032	.0069	.0148	.0004	.0001	.0010	.0071	.0035	.0118	.0036	.0520	.0004	.0000	.0040
1-2	.9761	.9912	.9910	.9890	.9827	.9762	.9792	.9904	.9938	.9886	.9802	.9879	.9973	.9954	.9893	.9965	.9878	.9841	.9059	.9880	.5081	.9933
2-2	.0050	.0075	.0080	.0081	.0098	.0179	.0179	.0000	.0030	.0045	.0050	.0117	.0026	.0036	.0036	.0000	.0004	.0123	.0421	.0116	.4919	.0027

References

1. Ayers KL, Sabatti C, Lange K (2005) Reconstructing ancestral haplotypes with a dictionary model. *J Comp Biol*

2. Bussemaker HJ, Li H, Siggia ED (2000), Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. *PNAS* 97:10096–10100

3. Conlon E, Liu X, Lieb J, Liu J (2003) Integrating regulatory motif discovery and genome-wide expression analysis. *PNAS* 100:3339–3344

4. Sabatti C, Lange K (2002) Genomewide motif identification using a dictionary model. *Proceedings IEEE* 90:1803–1810

5. Sabatti C, Rohlin L, Lange K, Liao JC (2005) Vocabulon: a dictionary model approach for reconstruction and localization of transcription factor binding sites. *Bioinformatics* 21:922–931