Chapter 3 HMMs for whole *Plasmodium Falciparum* Chromosomes

3.1 Introduction

Observation of chromosomes in a variety of organisms appears to show that they are composed of a number of distinct blocks. For example, there are the banding patterns observed in condensed eukaryotic chromosomes (Rooney 2001). With the primary sequence of these chromosomes becoming available it is now possible to investigate what relationship if any there is between these patterns and the sequence.

By using unsupervised learning techniques it is possible to look for natural patterns in the sequence without being biased by prior expectations. We can then compare these natural patterns with the annotated biological function to look for correlations. If the chromosomes are constructed from blocks that have one of a small number of sequence composition biases, it should be possible to estimate both the number of and the compositional bias for each distinct bias and use these to partition the chromosome into regions. In the general case, the chromosome could be modelled as being made up of regions of DNA which each have a reasonably constant sequence composition but which noticeably vary in composition from their neighbours.

The compositional bias parameters and the likely order in which blocks follow one another can be estimated using Hidden Markov Models (HMMs) (see Section 1.3.2). In contrast to the complex HMM methods commonly employed for modelling biological sequences, such as gene finders, our models do not need to be concerned with the fine structure of the DNA, concentrating instead on large-scale chromosomal structures.

It has been observed that gross sequence content correlates particularly strongly with function in the Malarial parasite *Plasmodium Falciparum*. The chromosomes as a whole have a very high ratio of AT to GC. However, since coding for amino acids require all nucleotides to be used, exons tend to have a slightly lower ratio (Escalante, Lal et al. 1998). Annotators use sequence composition plots as a tool to aid annotation. Since this is particularly useful in the *P. Falciparum* annotation process (K. Rutherford, personal communication) this genome was selected as a target for investigation using these approaches.

P. Falciparum has a genome estimated to be about 30[30] megabases (mb) in length, divided into 14 chromosomes (Pollack, Katzen et al. 1982). The genome exhibits a strongly biased AT/GC ratio with an overall (A + T) content estimated at 82 %. Recently, the complete sequence of chromosomes two and three have been sequenced and published (Bowman, Lawson et al. 1999; Gardner, Tettelin et al. 1999)[31]. The telomeric regions of chromosomes two and three are similar in structure, containing a shared pattern of terminal telomeric repeats followed by the repeats R-CG7 and rep20, a member of the *var* gene family, the R-FA3 repeat and finally a *riffin* gene. This arrangement of repeats and genes appears to be functional, promoting shuffling of the sub-telomeric regions between multiple chromosomes (Figueiredo, Freitas-Junior et al. 2002).

---

[30] The total size of the genome has since been found to be closer to 23 mb in length Gardner, M. J., N. Hall, et al. (2002). "Genome sequence of the human malaria parasite Plasmodium falciparum." Nature **419**(6906): 498-511..

[31] Since this time, the entire genome of *P. Falciparum* has been sequenced Ibid..

By eye, it is possible to identify regions of chromosome 3 with extreme base composition. Some of these are clearly correlated with biologically important features. In addition to the GC enrichment associated with exons, there is a region with extreme AT content (95-100%) that is believed to be the centromere (Hall, Pain et al. 2002). It is interesting to speculate as to how many types of sequence composition exist within Malaria chromosomes, or even whether the chromosome can meaningfully be grouped into regions that have one of a small number of compositional biases, or are in fact part of a continuum.

The BioJava HMM APIs are very flexible and allow many different architectures and parameter sets to be evaluated rapidly. The representation of the underlying alphabets in BioJava enables us to reuse architectures for different representations without recoding the core recursions. This makes them an ideal tool for investigating this type of open-ended question.

3.2 Simple HMM Architectures

3.2.1 Methods

A simple HMM was constructed using the BioJava HMM APIs (Section 2.6) with two states each with independent emission distributions. This was expected to segregate the chromosome into regions of high and low AT/GC ratio. A second model was generated with four independent states expected to segregate the chromosome into regions of relatively very low, low, high and very high AT/GC ratio. In both cases, these models were fully connected (transitions existed between all states). Transitions and emissions parameters were initiated to random values, but with the constraint that the transition from any state to itself was initialised to a value approximately 1000 times more likely than the transition to any other state. All model

scores are presented in units of log probability due to the extreme dynamic range of these probabilities. These models were trained using Baum-Welch with sampling (as described in Section 1.3.2).

3.2.2 Results

The two-state model reached a stable set of parameters within a very few cycles. The log likelihood remained almost constant from cycle 40 to completion at cycle 1214 (–1246786 at cycle 40, with mean –1246786 between cycles 40 and 1214). The Viterbi state paths from the model at cycles 40 and 1000 are 98.6 % identical. The model with four states showed similar convergence behaviour (data not shown).

The emission probabilities of the model with two states were complementary rather than being segregated into high and low GC. Over multiple training sessions with different initial parameters, the model with four states learned two distinct sets of model parameters.

Both four state models contained a pair of states that were similar to the states in the two state model. This pair of states aligned to the major part of the chromosome. The other two states of the four state model trained differently.

In the first set of model parameters, the two additional states aligned to the chromosome ends (telomeric regions) and were complementary to each other, i.e. for each telomere one state aligned to one strand and the other to the reverse complement of it.

In the second set of model parameters, the two additional states instead modelled a strong first order relationship in the telomeres. Specifically, one state modelled ‘A’ rich regions and the other modelled ‘T’ rich regions. Frequently these ‘regions’ were

only single nucleotides in length. Transition probabilities favoured them moving from one to another. The other pair of states modelled the internal regions as before.

3.3 HMM Architectures with Complementary Emission Distributions

The above results demonstrate that the chromosome must be considered in terms of being a double-stranded DNA molecule rather than as a single-stranded sequence. In particular, if there is a block with a characteristic sequence composition on one strand, this, by definition, implies a block with the complementary distribution on the other strand. This pair of states should be modelling a single set of parameters. To achieve this we developed a `Distribution` that implements a complementary view onto another `Distribution`. A pair of states can then be added to the model, one with the forward strand distribution and one with its complement. We call these complementary states pair-states. During training, all counts associated with the complementary distribution are first un-complemented and then forwarded as counts to the forward-strand distribution. This guarantees that the total number of parameters is minimized and that all available evidence for emissions is used during training.

3.3.1 Methods

Models were constructed with two, three, four or five pair-states (4, 6, 8 and 10 total states respectively). During training, all emission probabilities and all transition probabilities were initially set to random values, with transitions from each state to itself initially being approximately 1000 times more likely than any other transitions. Each model was then trained using Baum-Welch with sampling, as described above, as well as by Baum-Welch, using the sequence of Malaria chromosome 3. Training was stopped after 100 cycles due to a combination of computational constraints and the observation that models appear to converge before cycle 100. The different

models were then aligned to both chromosome 3 and chromosome 2 without additional training.

Models with more than 5 pair-states were not trained as both the memory and computational requirements for the estimation of training parameters becomes prohibitive. The space required is approximately equal to length_of_sequence x number_of_states x size_of(double), which for large sequences with many states quickly reaches the limits of a machine with hundreds of megabytes.

3.3.2 Results

Training using Baulm-Welch with sampling exhibited quicker convergence properties than Baulm-Welch, and was also computationally less expensive due to the decreased number of counts which needed to be summed. Multiple training runs with sampling produced models with more similar parameters and alignment scores than with Baulm-Welch training (data not shown).

We then considered a representative from the replicates of the two, three and four pair-state models. The Viterbi paths at 20 and 100 cycles for the two, three and four state-pair models differed by 0.23 %, 0.45 % and 1.41 % respectively. This indicates that by cycle 100, the models were not changing significantly in their predictions. The model with five pair-states did not use one pair of states at all, indicating that this family of models could only distinguish four types of gross genomic content, and is therefore not discussed further.

In all cases, some transition probabilities in the trained models have moved greatly from their original values, and the most used states have emission probabilities that lie close to the ratios found in the chromosome (Figure 3-1).

The Viterbi paths for all three pair-state models at cycle 100 against chromosome 3 (Figure 3-2) show use of paired states at the beginning and end of the chromosome, and a similar banding pattern of states within the chromosome. In all three alignments, a single state emits the first 276 bases of the chromosome. The corresponding complementary state then emits the final 186 (±2) bases of the chromosome. This corresponds to the regions of sequenced telomere. In addition, in all three models, a single pair of complementary states emits the majority of the body of the chromosome. The models with more states show additional features, such as the appearance of a band near the ends of the chromosome that resembles telomeric sequence, and blocks of sequence corresponding to repeat elements. Not all states were used by the more complex models. For example, the three pair-state model learned two telomeric states, two internal states and a final state that matches a region within the genes PFC005w and PCFC1120c. This state did not use the complement of this final state anywhere.

The four pair-state model has corresponding states for each of these regions and two complementary states that match a region between the telomeres and the *var* genes. These overlap significantly with the repeat elements rep20, rep11 and R-CG7, and show striking similarities with the state-paths for chromosome 3. Again, the telomeres have been correctly identified, and the exons on each strand seem to segregate with the two main states. In addition, the regions near the telomeres are predicted to have a very similar structure, including a telomeric-like section within the *var* genes, and the use of states that overlap the repeat elements.

In the four state-pair model, the coding region state pair has one state associated with each strand. 86% of all bases in exons on the positive strands are matched by the

first state, and only 14% by second. 94% of all bases in exons on the negative strand are matched by the first state and 5% by the second. Overall, these states predict the strand correctly 90% of the time on a per-nucleotide basis. Most errors are made on the boundaries between genes on opposite strands.

The Viterbi state path for chromosome 2 (Figure 3-3) shows striking similarities with those obtained for chromosome 3 (Figure 3-2). This is evidence that the two chromosomes share a common architecture. Labellings of randomised sequences do not show these similarities in patterns (data not shown). Therefore, we believe that the models have indeed learned some general properties of malarial chromosomes.

It is possible that the consistent structures predicted at the beginning and end of each chromosome is an artefact of transition probabilities associated with entering and exiting the model. To test this, artificial chromosome sequences were constructed. The first half of the chromosome was appended to the second half so that the central regions of the sequence were now at the ends, and the ends of the sequence were now in the centre. State-paths were predicted using the same models as before. The regions at the ends of the artificial sequences were labelled with the states associated with the body of the chromosome, and the regions corresponding to the telomeres now located in the centre of the sequence were labelled with the telomere-associated state-pair. This indicates that the models are making predictions on the basis of the sequences, and not any edge-effect artefacts.

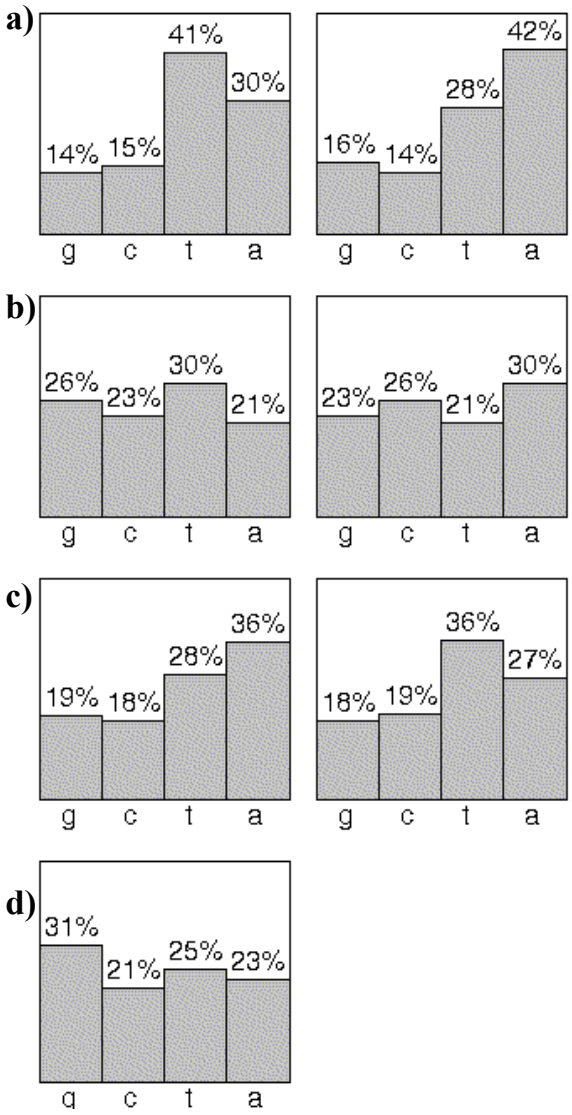

Figure 3-1 Emission probabilities for the four pair-state model

Each row shows a state-pair with complementary emission probabilities. They match **a)** chromosome body; left and right associated with (-) and (+) strand exons respectively **b)** telomere-like sequence; left and right associated with telomeres at the right and left of the chromosome respectively **c)** near-telomere repeat associated regions **d)** (G + C) rich region in the *var* genes (only one of this state-pair is used).

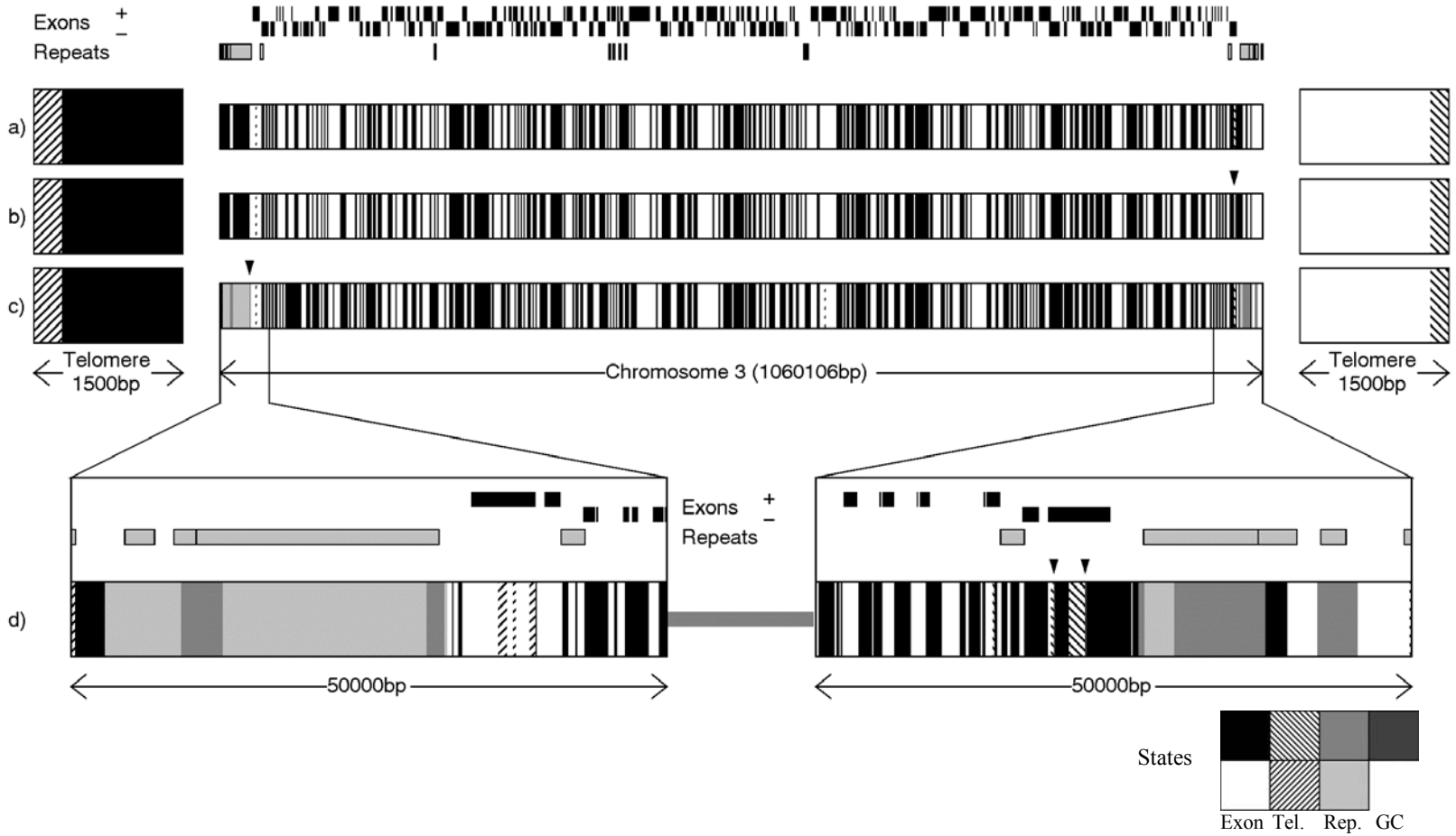

Figure 3-2 Diagram of the *P. Falciparum* chromosome 3 and the state paths through three models (legend continued on next page)

Sections **a**, **b** and **c** represent the state paths of the two, three and four state-pair models respectively across the entire chromosome with insets to the left and right showing the extreme telomeric region. The relative positions of exons and repeat elements are indicated above these diagrams. Section **d** shows an enlarged view of the state paths for the first and last 50,000 bp of the chromosome, with the corresponding exons and repeat elements above. Within each diagram, a different shading pattern is used for each state, as indicated by the key (Exon - exon-related, Tel. - telomeric-like, Rep. - repeat-associated, GC - high (G + C) content). Arrows above the diagrams indicate the positions of narrow regions that may not be easily visible.

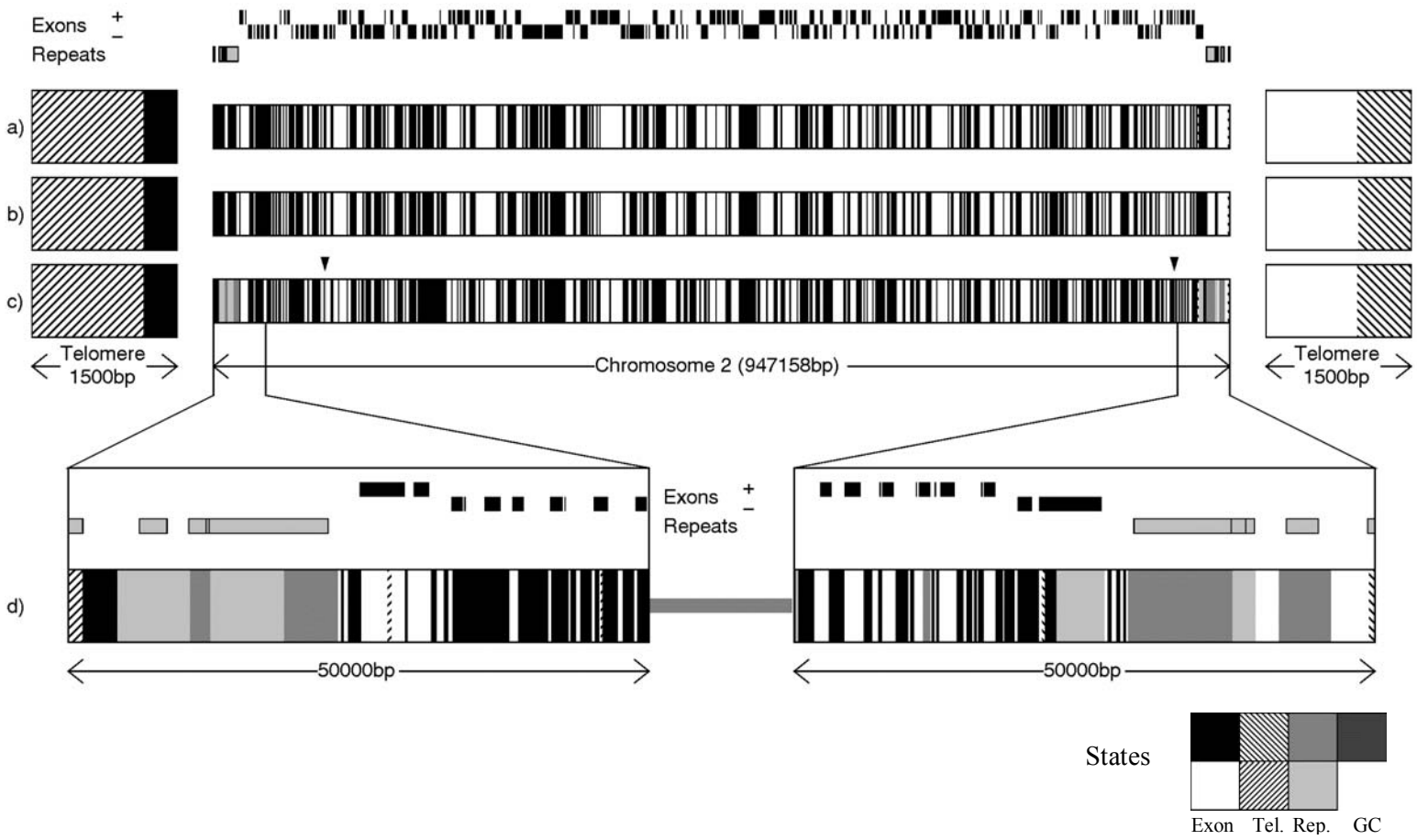

Figure 3-3 Diagram of the *P. Falciparum* chromosome 2 and the state paths through three models (legend continued on next page)

Sections **a**, **b** and **c** represent the state paths of the two, three and four state-pair models respectively across the entire chromosome with insets to the left and right showing the extreme telomeric region. The relative positions of exons and repeat elements are indicated above these diagrams. Section **d** shows an enlarged view of the state paths for the first and last 50,000 bp of the chromosome, with the corresponding exons and repeat elements above. Within each diagram, a different shading pattern is used for each state, as indicated by the key (Exon - exon-related, Tel. - telomeric-like, Rep. - repeat-associated, GC - high (G + C) content). Arrows above the diagrams indicate the positions of narrow features that may not be visible at this scale.

3.4 First Order HMMs with Time-Reversible Transition Probabilities

The model with four independent distributions in some cases learned a pair of states that crudely represented a 1st order distribution (encapsulating dinucleotides). To explore this further, models were constructed that contained emission `Distribution` objects encapsulating a 1st order Markov process.

The 0th order model used the third pair-state to identify a high-G region. However, the transition probabilities learned for this model only allowed one of the pair of states to be used. This was due to the lack of association between the transition probabilities. The re-architecting process required to introduce higher order emission probabilities also provided an opportunity to constrain the transition probabilities such that the resulting HMM is truly time-reversible (if the sequence being analyzed is played back in reverse with the appropriate complementation, it would induce a state labelling that is the reverse-complement of the forward state labelling). The time-reversed transition probabilities should in theory remove the kind of artefacts observed in the 0th order model's third pair-state.

3.4.1 Methods

The chromosome was viewed through an `NthOrderSymbolList` instance to translate it into all overlapping pairs of symbols, and the HMM emission alphabet was set to DNAxDNA.

The 1st order emission distributions presented additional challenges to ensure that they were correct estimates in both the forward and reverse directions. The probability of observing a given dinucleotide is defined as being the probability of observing the second nucleotide conditioned upon the first. That is, there is a 0th order probability distribution over each second nucleotide that is chosen according to the identity of the

first nucleotide. If the distribution associated with the complementary state is calculated by simply reverse-complementing the dinucleotide and finding its probability in the original distribution, this will not be a true probability distribution (the sum over all probabilities given all dinucleotides starting with a given nucleotide will not be guaranteed to be 1). This is because in this case we are effectively conditioning upon the second nucleotide rather than the first. This causes the models to be non-probabilistic and the training algorithms to fail.

We address this by reverse-complement the table of observations and then re-normalize to give the complementary 1st order probabilities. During training, a standard `DistributionTrainer` is registered with the forward-strand probability distribution. The reverse-complement distribution registers a `DistributionTrainer` that forwards all counts on to this after reverse-complementing the dinucleotide. Probabilities for the forward distribution are estimated as normal and those for the reverse distribution are estimated by normalizing the reverse-complemented counts.

This scheme ensures that all available information is pooled (both evidence for forward and reverse strand are aggregated) and that the result is a strand-reversible probabilistic Markov process. As a concrete example, given the short sequence 'AATGCGT' we can estimate both a forward 1st order distribution, that would produce this and a reverse-strand 1st order distribution, that would produce 'ACGCATT' with an equal probability using the counts in Table 3-1 and a suitable normalization (such as pseudocounts). It is clear from this example that the probability of observing a dinucleotide is not equivalent to observing its reverse-complement (for example, AG is half in the forward strand, but CT is not observed at all in the reverse strand).

Table 3-1 Forward-strand and reverse-strand counts

	A	**G**	**C**	**T**	**Sum**
A	1			1	2
G			1	1	2
C		1			1
T		1			1

	A	**G**	**C**	**T**	**Sum**
A			1		1
G			1		1
C	1	1			2
T	1			1	2

The transition distributions present a more complex problem. The first naïve approach was to constrain the transition probabilities of a reverse-strand state to be the transition probability from the forward-strand state to the complement of the destination. This does yield a probabilistic model. However, it is not fully time-reversible. This is because if we consider both strands, the model effectively treats entry to a forward-strand state as being equivalent to exiting a reverse-strand state.

This was again addressed by estimating the transition probabilities from tables of counts. However, we run into a problem that prevents us using the same `DistributionTrainer` solution as for the 1st order probabilities. Table 3-2 enumerates every possible transition from state 'a' to state 'b' given that neither, one, or both may be complemented (indicated as a' or b' respectively).

Table 3-2 State-transitions and their reverse-complements

Forward	Reverse Complement
a-b	b'-a'
a-b'	b-a'
a'-b	b'-a
a'-b'	b-a

For three of the four cases, either the forward or reverse-complement forms start with a forward-strand state. These can all use the reverse-complement forward-state counts to calculate the backward-state probabilities. However, in one case (a'-b:b'-a), there is no count associated purely with the forward-strand process. In the naive probability model described above, this case is considered interchangeable with (a-b':b-a'). However, the two are clearly distinct transitions. This issue did not arise for the emission probabilities as we only considered the cases of a-b, or b'-a', which are a well-behaved subset of all the interactions in Table 3-2 (in particular a-b:b'-a').

The problematic transitions do not arise for more restrictive model architectures for which forward and reverse model regions are separated by an a-directional region. During training, a table of counts for all pair-wise combinations of states was kept, and while collecting observations, the count was split into two parts, which were then forwarded to each count cell representing the two possible time-reversed transitions. Then, during training, the distribution was estimated by normalizing the aggregates of each of the two time-reversed transitions.

3.4.2 Results

Models with 2, 3, 4 or 5 pairs of states were trained on chromosome 3 of *P. Falciparum* using Baum-Welch training until the forwards probability did not vary by more than $e^{0.01}$ between two cycles (changes of > 0.01 relative to scores in the range of tens of thousands). The models took 116, 103, 77 and 90 cycles respectively to converge. Models with more states were again not trained due to the memory and computational constraints.

The transition probabilities for all models are dominated by state-to-self transitions. Emission probabilities for all models (Figure 3-4) show a progression in complexity

with the number of state-pairs available. The additional distributions seem to model additional sub-types of sequence, and in every case, each of the distributions in the simpler models are represented in the more complex models. This indicates that the additional available complexity is being used to model distinct populations of sequences. This is in contrast to the 0^{th} order model, which could model no more than four compositional biases. Presumably, the 1^{st} order probability distributions are capturing some more biologically relevant information. It is also interesting in that each model was trained entirely independently with different starting parameters but learned very similar final parameters. This is good evidence that the models are learning some legitimate signals embedded within the chromosomal sequence rather than using the extra parameters in an arbitrary manner to memorise the training sequence.

The entire chromosome was classified into the following biological feature types; exon, intron, repeat and other, using the annotation associated with the malarial chromosome. This classification was then projected onto the state labelling from the 5 pair-state model (shown graphically in Figure 3-5 and Figure 3-6 for chromosomes 3 and 2 respectively). From this, a count of the number of times a particular state and feature are co-located was calculated (Figure 3-7). These counts show dramatic trends for certain features and states to be associated with one another. There are clearly two state-pairs (3± and 5±) associated with exons. States 2± are also associated to some degree with the 'other' category while States 4± accounts for the majority of repeats. Figure 3-8 and Figure 3-9 are normalized views of these counts for the 5 pair-state model representing the conditional probabilities of observing a particular state given that the feature is known, and observing a given feature given that a state is known.

From Figure 3-8 it is clear that if a region is an exon, the states 3± and 5± together account for nearly the entire feature (> 93 % for + strand, > 94 % for –strand). Introns do not have a single state associated, but show a predisposition towards the pair 2± (2- preferred over 2+ for forward strand introns and 2+ preferred over 2- for backward strand introns). Indeed, the predispositions in forward and reverse strand introns appear to show a distinctly strand-dependant pattern despite there not being a single indicator state. This indicates that the introns contain important strand-dependant information. Repeats are associated with the states 4±. The other category most closely resembles the average of the intron distributions. It is interesting that the repeat distribution seems not to include a large proportion of states 1±, despite these being found in introns and 'other'.

The most striking feature of Figure 3-9 is that almost all states are associated primarily with only one feature type. The second observation is that no state is predictive of introns. States 1± and 2± are associated with 'other'. States 3± and 5± are associated with exons. States 4± are associated with repeat regions.

We can see from Figure 3-7 how as state pairs were added, the correlation between states and features altered. In the 2 pair-state model, exons are only labelled by states 2±, but these states are also frequently found labelling 'other', and accounts for almost all repeats. In the 3 pair-state model, the exons and the repeats are modelled by their own states (3±), while 2± remain the major 'other' states. In the 4 pair-state model, states 4± now take on the role of specifically modelling the repeats. Finally, in the 5 pair-state model, states 5± model a sub-set of exons. Clearly, as more states are added to the models, they are making finer distinctions over how to model the chromosome.

Figure 3-5 and Figure 3-6 are graphical representations of the state-paths of the five pair-state model to chromosomes 3 and 2 of Malaria respectively. Again, from these figures, the co-localisation of some feature types with biological features are clear to see, particularly at the extreme ends of the chromosomes. These results also demonstrate how as the complexity of the models increase, finer distinctions in the assignments are identified.

Figure 3-7 displays the frequency with which different states align to each of the different biological feature classes. In Figure 3-8, this data has been normalized to give the observed probability of any given state given a particular type of feature. This gives an indication of how strongly a given state labelling of a region of chromosome indicates a particular biological function for that region. In Figure 3-9, this same data has been normalized to give the observed probability of any feature type given a particular state. This indicates how predictive each state is of the different feature classes.

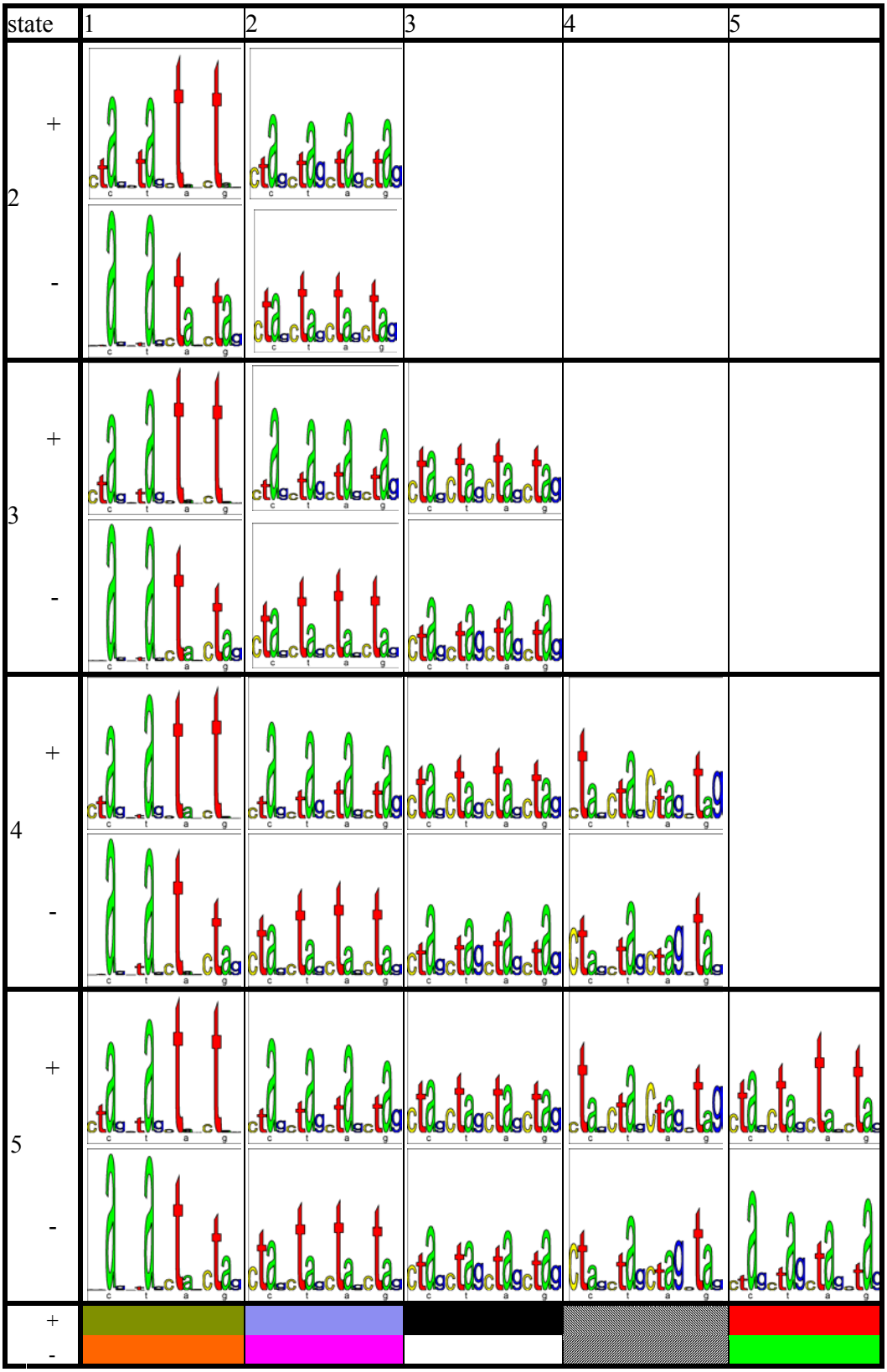

Figure 3-4 Emission Spectrums for all Pair-State Models (legend continued on next page)

Sections for the 2-5 pair-state models contain two rows, the first contains graphs of the 1st order emission probabilities and the second contains graphs of the reverse-strand emission probabilities. The emission probabilities for each model are arranged so that those that appear similar are in the same column (1-5). The final area displays a key that associates the states with colours in the whole-chromosome diagrams Figure 3-5 and Figure 3-6.

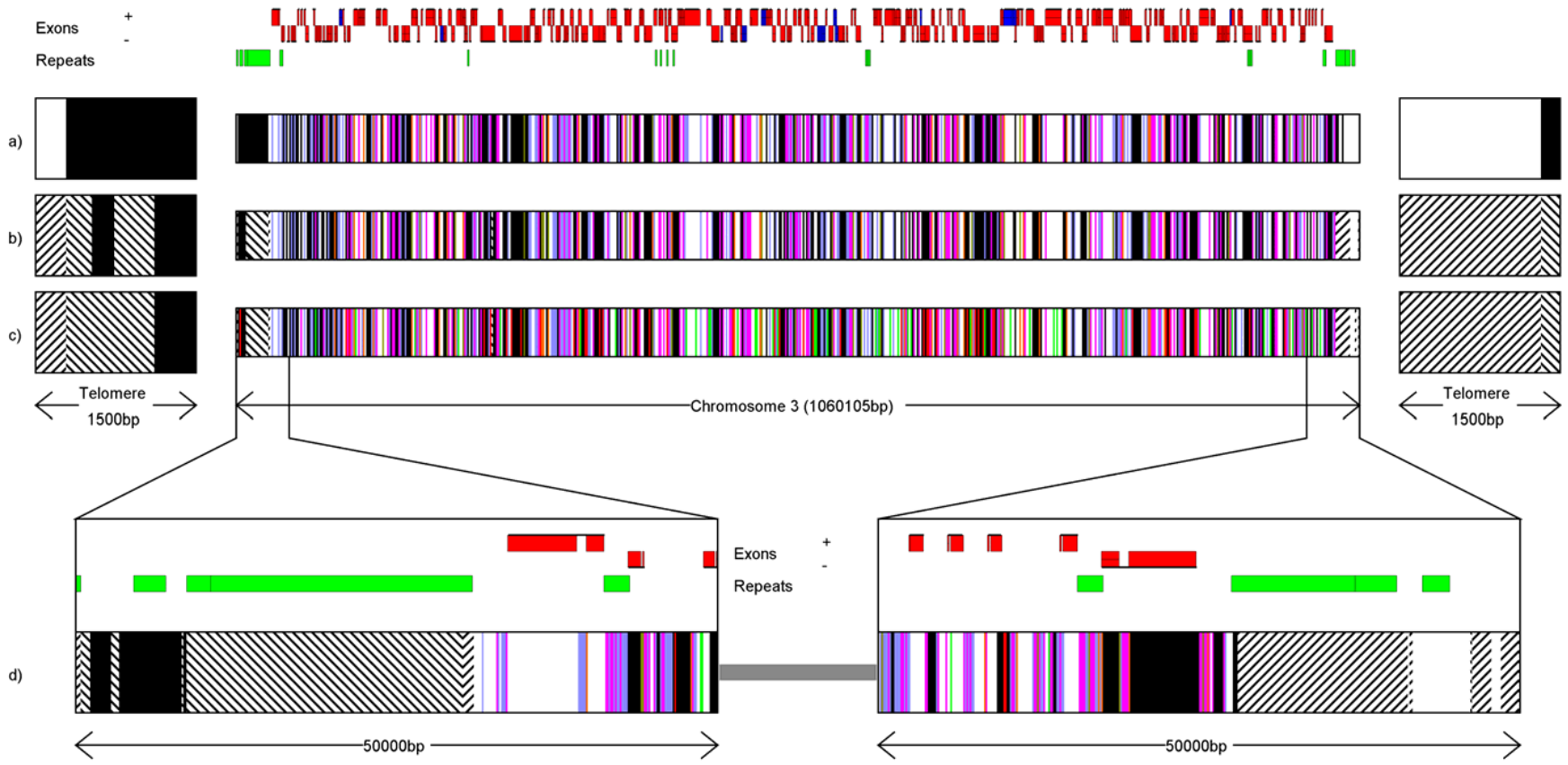

Figure 3-5 Diagram of the alignments of the 3,4 and 5 state-pair models to Malaria chromosome 3

(legend continues on next page)

Lines **a**, **b** and **c** display the alignments of the 3,4 and 5 state-pair models respectively. The colours are as in the key in Figure 3-4. The red exons belong to ‘normal’ genes. The blue exons belong to ‘bob’ genes (Bowman, Lawson et al. 1999).

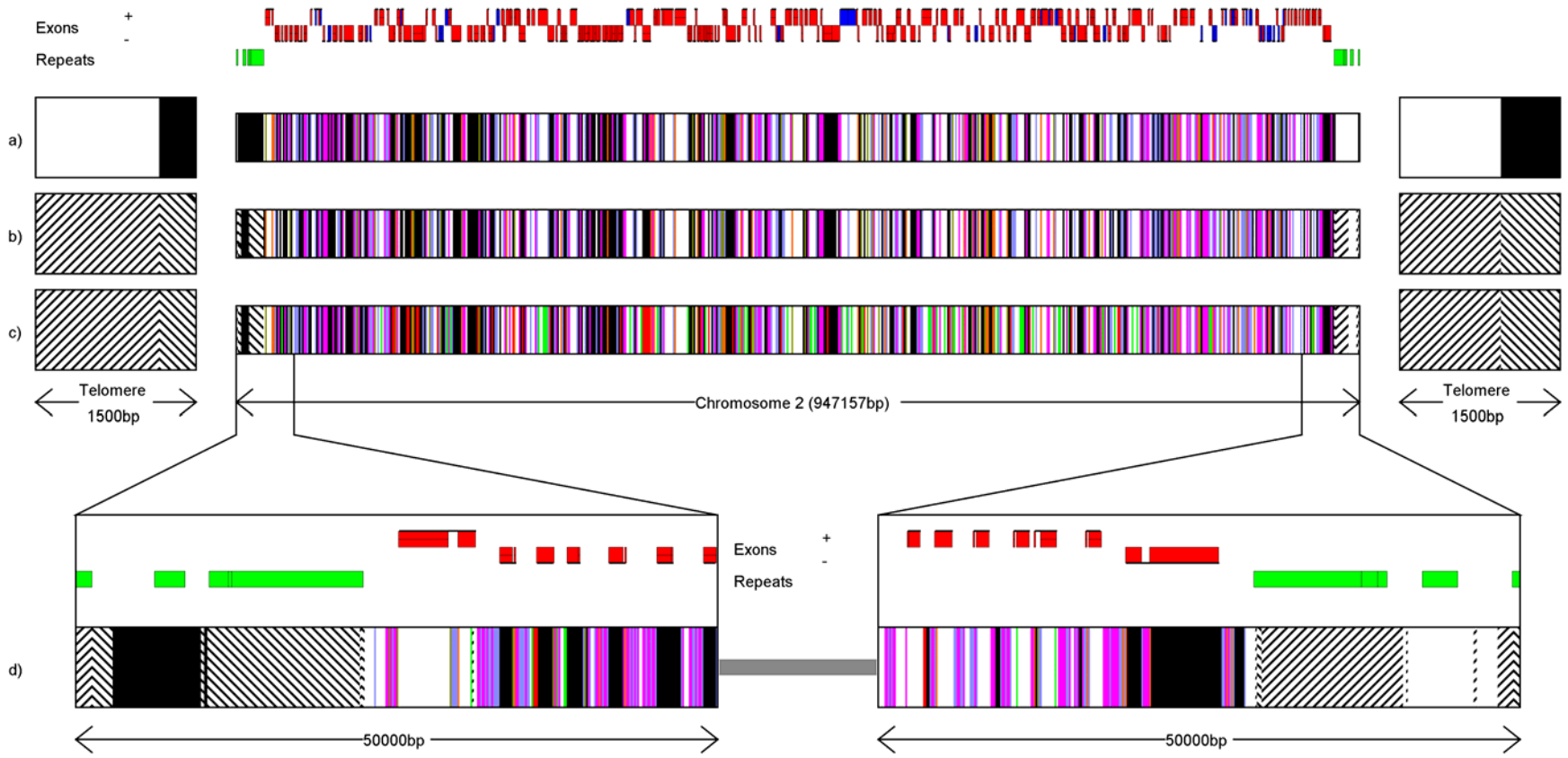

Figure 3-6 Diagram of the alignments of the 3,4 and 5 state-pair models to Malaria chromosome 2

(legend continues on next page)

Lines **a**, **b** and **c** display the alignments of the 3, 4 and 5 state-pair models respectively. The colours are as in the key in Figure 3-4. The red exons belong to ‘normal’ genes. The blue exons belong to ‘bob’ genes (Bowman, Lawson et al. 1999).

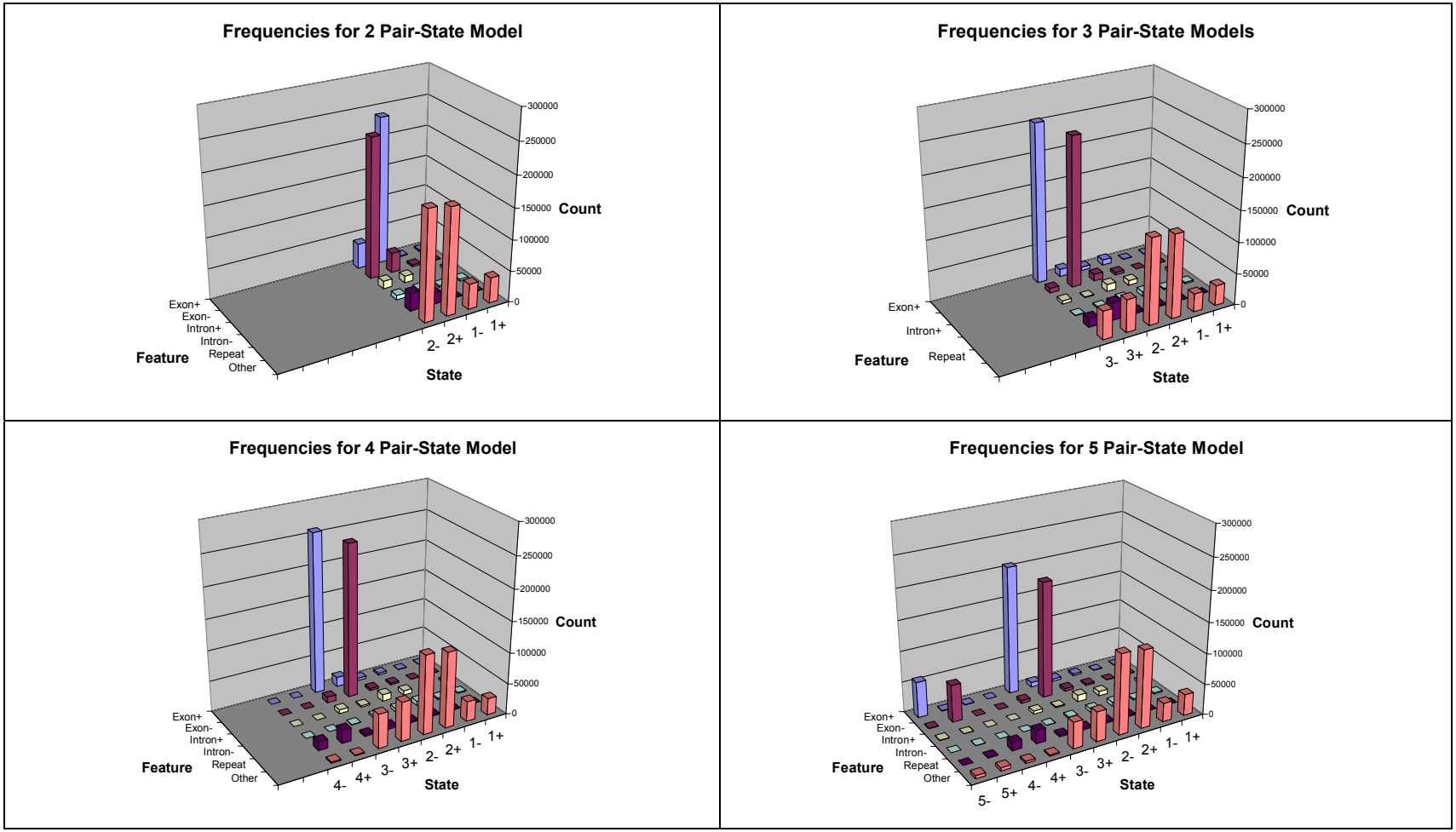

Figure 3-7 Counts for Biological Feature and States for the 2-5 Pair-State Models

(legend continues on next page)

The state labels are consistent with the labels in Figure 3-4. Each model in turn was used to label chromosome 3. The state labelling was then compared to the location of known genes and repeats.

Error! Not a valid link.

Figure 3-8 Normalized Counts of States for Biological Features

This bar chart displays the data in Figure 4-8 grouped by biological feature. For each feature, the probability of observing a given state is displayed as a bar. Exons are primarily accounted for by states 3±, with states 5 aligning to only approximately one fifth of exon sequence. No other biological feature class is predicted so clearly by any state.

Error! Not a valid link.

Figure 3-9 Normalized counts of Biological Features for States

This bar chart displays the data in Figure 4-8 grouped by state. For each state, the probability of observing a given feature is displayed as a bar. States 1+, 1-, 2+ and 2- are specific for the Other category. States 3+, 3-, 5+ and 5- are specific to exons States 4+ and 4- are specific to repeats.

3.5 Discussion

In order to investigate the gross structure of malarial chromosomes, we explored a number of different HMM architectures using the BioJava HMM APIs. The initial model contained just two states. It was trained to classify every base within the Malaria chromosome as being emitted by one or other of these states. The transition probabilities were set to initial values that favoured a single state emitting a long region of the chromosome. The belief was that this model would segregate the chromosome into high and low AT/GC content. After training, the emission spectrums of the two states were very close to being complementary. One interpretation of this is that the underlying biological process learned was strand-dependant, so that in effect the model reflected a single probability distribution, but learned it once for each strand.

This observation lead to the construction pair-state HMMs with pairs of states that emit nucleotides according to complementary distributions. The two pair-state model revealed that the telomeric regions were distinct from the internal chromosomal sequence, and that the body of the chromosome aligned to a single pair of states which flipped between one another. This pair of states appears to correspond to the positions at which Malaria utilizes one strand or the other for coding exons, predicting the strand with an accuracy of 90 %. This is surprisingly accurate, given the extreme simplicity of the model.

The more complicated models additionally predict a feature resembling telomeric sequence followed by a small region of sequence that is distinguished from the rest of the chromosome by its very high (G + C) content (52 %). This interesting pattern is visible only in the genes PFC005w and PCFC1120c, which are putative members of

the *var* family (Bowman, Lawson et al. 1999), involved in evading the host immune system. There is some evidence that *var* genes are subject to epigenetic control and undergo frequent intragenic recombination (Corcoran, Thompson et al. 1988), so the telomeric-like fragments within these genes may be the remnants of chromosomal rearrangement events resulting in the shuffling of these sub-telomeric regions.

The models trained on chromosome 3 were aligned without further training to *P. Falciparum* chromosome 2, to check whether the models had learned features specific chromosome 3, or more general features of Malarial chromosomes. Without further training, the models correctly recognise the telomeres, predict the exon directions and also identified the telomere-associated repeats in chromosome 2. In addition, the *var* genes on chromosome 2 appear to contain a band of telomeric sequence in the corresponding locations to the *var* genes located on chromosome 3. In chromosome 2, the band of high (G + C) content appears not to be present.

The blocks of telomeric base composition within, and beyond the *var* genes, may be a relic of recent recombination events between these genes and other telomeric *var* loci. Other *var* loci also appear to share this feature (data not shown), although these types of model may not be appropriate for analysing short sequences. The high (G + C) region found in the chromosome 3 alignments may be specific to that chromosome as none of the other *var* genes analysed shares this feature. It was not possible to train simple pair-state models that used more than four pairs of states, which is evidence that the models were not over-fitting the training data, and were characterizing real information about the chromosomes.

In addition to the observation that both strands of the chromosome must be considered, the original model indicated that some of the processes observed were not

easily modelled by 0^{th} order probabilities. This inspired the creation of the fully time-reversible 1^{st} order models. These models were able to learn more subtle signals that were associated with or were indicators of exons (+ and – strand), introns (again + and – strand), repeat elements and 'other' (assumed to be intergenic sequence). These models were capable of consistently learning the same signals given different initial training parameters and different numbers of paired states. Additionally, they were able to learn additional and more complex signals as the number of parameters was increased. The most interesting feature of the 5 pair-state model is the sub-division of exons into those with high and low adenine content (states 3± and 5± respectively). This does not coincide with any obvious properties of the genes.

None of these models learned a state associated with the putative centromere, which has been predicted to lie in a region which is almost entirely (A + T) in composition (Bowman, Lawson et al. 1999). However, the centromere is a comparatively small structure that may not be distinctively different from the already extreme A/T bias of the chromosome in general. The 0^{th} order model did model a very small region of high G+C content, but this had sequence-composition characteristics that are radically different to those associated with the other states. It is possible that a 1^{st} order models with more states would have recognized the centromere.

All of these models were trained using unsupervised learning techniques, and had no supplied data to indicate the location or type of biological features. However, all of these models have learned signals that are co-located with biologically significant structures. Given the relative simplicity of the models, this is clearly a potentially powerful method.

3.6 Future Directions

At the time this work was done the model size was limited due to the physical memory required to store training parameters for large sequences. With newer machines with greater physical memory it has now become practical to extend this work to consider more states and larger sequences, such as human chromosomes. It would also be interesting to look at orders greater than one. However, to train models with large numbers of transitions and high-order emission states would most likely require a more sophisticated regularization framework and possibly a more complex representation of the HMM than simple pseudo-counts or these probability parameterized finites state machines can afford.

The memory requirements for the dynamic-programming matrices used during training scales linearly with the length of the training sequences, and also with the number of states in the model. On the computer hard ware used in this study, this becomes prohibitive for sequences that exceed more than a megabase in length, and for models with more than ten states.

One solution would be to calculate one matrix completely, and then calculate each column of the other in turn using the space-saving implementation of the recursion, adding counts associated with each completed row of the matrix as we go. However, one of the matrices must still be held in memory, so this still scales in proportion to the length of the sequence, allowing us only to double the training sequence length, or the number of states.

Another solution would be to calculate the space-saving version of one recursion, and as each matrix row is completed, calculate the other recursion back to that point. Although the memory required for this is trivial, the computation will scale by the

square of the sequence length. This is likely to become prohibitive even quicker than the memory constraints of the above approaches.

A combination of the two methods can be developed that has a computational and space cost proportional to the length of the sequence. Firstly, a chunk size is chosen. Then, the forwards recursion is calculated using the space-saving version of the recursions. The first matrix row encountered is then stored in a list. Each time a number of rows have been calculated that is a multiple of the chunk size, this is also stored in the list. This is done until the complete recursion has been calculated. The complete sub-matrix running from any stored row to the next (or the end of the sequence) can now be calculated as needed using the normal forward recursion, initialized on the stored row. The space-saving implementation of the backwards matrix can then be used to provide the backwards scores for each region, starting with the last and working towards the first, and counts can be added to the model trainer as normal.

This method requires the forwards matrix to be calculated twice, and also will need enough memory to store the forwards matrix rows for each of the chunks. However, this is significantly lower than the cost of storing the complete matrix. If the largest sequence that can be used for training with the current method is one megabase, then the chunk size can be set to once per half megabase. This would allow half a million chunks to be processed before using half of the available memory (the other half being required for calculating the sub-matrices). There are no sequences that we are aware of that are likely to exceed the order of a million, million nucleotides. Therefore, we propose that this method will allow single-head HMMs to be trained on

any practically available sequences without exceeding readily available memory resources. We plan to implement this training method in BioJava in the near future.

The investigations described in this chapter have demonstrated that the BioJava HMM APIs are highly adaptable to different model architectures, with potentially complex relationships between the values of parameters. The same implementation code was successfully used here for models with different numbers of states and for different emission alphabets. The results are numerically stable and fully probabilistic. These APIs have been used by others for modeling biological signals, for example, see (Hasan 2003). We hope that as the APIs mature, they will become used even more widely for different modeling tasks.