BMC Bioinformatics

Research article

The p53HMM algorithm: using profile hidden markov models to detect p53-responsive genes

Todd Riley*[1,2], Xin Yu[4], Eduardo Sontag[2,3] and Arnold Levine[1,4]

Address: [1]The Institute for Advanced Study, Princeton, New Jersey, USA, [2]The BioMaPS Institute at Rutgers University, Piscataway, New Jersey, USA, [3]The Mathematics Department, Rutgers University, Piscataway, New Jersey, USA and [4]The Cancer Institute of New Jersey, New Brunswick, New Jersey, USA

Email: Todd Riley* - tr2261@columbia.edu; Xin Yu - yuxi@umdnj.edu; Eduardo Sontag - sontag@math.rutgers.edu; Arnold Levine - alevine@ias.edu

* Corresponding author

Published: 20 April 2009

BMC Bioinformatics 2009, **10**:111 doi:10.1186/1471-2105-10-111

Received: 25 January 2008
Accepted: 20 April 2009

This article is available from: http://www.biomedcentral.com/1471-2105/10/111

Abstract

Background: A computational method (called p53HMM) is presented that utilizes Profile Hidden Markov Models (PHMMs) to estimate the relative binding affinities of putative p53 response elements (REs), both p53 single-sites and cluster-sites. These models incorporate a novel "Corresponded Baum-Welch" training algorithm that provides increased predictive power by exploiting the redundancy of information found in the repeated, palindromic p53-binding motif. The predictive accuracy of these new models are compared against other predictive models, including position specific score matrices (PSSMs, or weight matrices). We also present a new dynamic acceptance threshold, dependent upon a putative binding site's distance from the Transcription Start Site (TSS) and its estimated binding affinity. This new criteria for classifying putative p53-binding sites increases predictive accuracy by reducing the false positive rate.

Results: Training a Profile Hidden Markov Model with corresponding positions matching a combined-palindromic p53-binding motif creates the best p53-RE predictive model. The p53HMM algorithm is available on-line: http://tools.csb.ias.edu

Conclusion: Using Profile Hidden Markov Models with training methods that exploit the redundant information of the homotetramer p53 binding site provides better predictive models than weight matrices (PSSMs). These methods may also boost performance when applied to other transcription factor binding sites.

Background

The p53 protein plays a crucial role in cancer suppression in the human body. In response to cancer-inducing, DNA-damaging stress conditions, the tetrameric p53 proteins can activate different pathways that lead to DNA repair, cell cycle arrest, inhibition of angiogenesis, and apoptosis [1]. A highly degenerative, palindromic consensus DNA binding site, consisting of a half-site RRRCWWGYYY, followed by a variable length spacer, then followed (almost always) by a second half-site RRRCWWGYYY sequence, has been discovered for the protein, where R is a purine, Y a pyrimidine, W is either A or T (adenine or thymine) and G is guanine and C is cytosine (see Figure 1) [2,3]. By labeling each quarter-site RRRCW as → and its reverse-complement WGYYY as ←, the first discovered p53 consensus sequence can be graphically represented by → ← *spacer* →

Clone	5′ Region	1st Half-site R R R C W W G Y Y Y	*Spacer*	2nd Half-site R R R C W W G Y Y Y	3′ Region
s57	CGACCTGTCA caccg	G G G C C T G T C A		C A G C A T GaC C T	acctgtcacaccggg
N22	atttt CACCATGCTT	C T G C A T G T C T		A G G C A A G T C A	ccttctc CACTGGCC
11A2	ccccatcctccatcc	A A A C AaT G C C C		A G A C T T G T C T	ct CCGCCTGAAT ga
W211	tttgtcctaccatcc	A G G C A T G C C T		- - - - T T G C C T	CACTCGTTA tttcct
W7B2	tatct GTGCAGCTG t	G G G C A T G T T T	t	A G G C A A G C T T	cct GTGCTAGTTC cc
3H	AACTAGATC cttttc	A G A C A T G T T A		T A A C A A G T C A	GTACAAGTTT atttt
8A	gctggt GCACAAGAG	T G A C A T G T C C		C G A C G T G T T T	tgtc
532	CATCATGCCA cctgc	A G G C A T G T T C	tggat	G G G C - T G T C T	t GTGCTTTGTTG ttt
64A2	c AAACCAGGGT gtct	T G A C T T G C C T	atcctgggaggt	T G A C A T G T T C	ctccccttcccctc
W7A1	gccaaacataaccac	C A G C - T G C C A		A G G C A T G C A G	tacc ACGCTCAGCCC
s61	c	C A A C T T G T C T	attctgtgttgat	G G A C A T G T T C	ccgtttttggctatt
11B3	actgttgatgatgaa	A G A C A A G C C T	a	G G G C A G G T C C	tgggggggtgggg
N42	gcagtgtggtggagg	A A A C A A G C C C	a	G G A T G T G C C C	a GGGCAGGCTG ggac
s201	tgttc ATACCTGTCC	A C A C T T G T C T		A T A C C T G C C T	ACACCTGTCT tgttt
s1583	ctttaattcagttgt	A A A C A T GaC T T	gttcattata	T G A C A T G T T C	aattacaattcgatt
s592I	ctcagttctcagctg	G G A C T T G C C C		T G G C C A G C C C	tgg GGTCACTGCTG c
s592II	tgcctcagcacctcc	A G G T TcT G C C -		G G G C T T G T T C	ctttcctttcagcat
2NB	gcctttgttgtgccc	T G A C T T G C C C		A G A C A T G T T T	gggaa TGTCTTGTGC
9H	gtattctcttttcct	A A G C A T G C C T		T G A C T T G T T C	tttcatctcctctga
CBE10d	tgaaagcaggtagat	T G C C T T G C C T		G G A C T T G C C T	GGCCTTGCCT tttct

Figure 1

Original Data from El-Deiry et al., Used To Define The p53 Consensus Binding Site. The original DNA fragments collected from a genome-wide, p53-antibody immunoprecipitation, that were used to define the head-to-head (HH) p53 Consensus Binding Site, are graphically presented [3]. The yellow columns corresponding to the 1st and 2nd half-sites were used to define the consensus p53 motif. The p53 binding site is highly degenerative. Within the yellow columns, notice that 7 of the 20 DNA target sites (35%) had at least one nucleotide insertion (green), deletion (red), or both (magenta) relative to the discovered 10 bp-spacer-10 bp consensus. Since insertions and deletions throw off the reading frame of a weight matrix, any PSSM approach will inherently mis-score at least 35% of these 20 sites. Alignments of the 160 experimentally validated p53 binding sites also reveal that any PSSM approach would inherently mis-score at least 30% of them as well. Another observation is that additional p53 half-sites are immediately adjacent (in yellow) to the ones used to define the consensus in 15 of the 20 target sites (75%). Since the genome-wide immunoprecipitation study was designed to pull down the highest affinity sites, the fact that 75% of the target sites are actually p53 cluster-sites is the first indication that cluster-sites of 3 or more half-sites confer higher binding affinity [22].

←. This configuration of the four quarter-sites is often referred to as the head-to-head (HH) orientation, and represents the vast majority of experimentally-validated p53 binding sites to date.

The degeneracy of the p53-RE

In the influential paper "Definition of a Consensus binding Site for p53", by El-Deiry et al., 7 of the 20 DNA target sites (35%) used to form the head-to-head (HH) p53 consensus sequence had at least one nucleotide insertion or deletion relative to the discovered 20 bp consensus after proper alignment (see Figure 1) [3]. Alignments of the roughly 160 experimentally-validated p53 binding sites to date also show that approximately 30% of presently known sites have at least one nucleotide insertion or deletion relative to the consensus matrix [4]. Discovery of p53 binding sites with such degeneracy cannot be reliably made with a PSSM approach, since prevalent insertions and deletions in the consensus sequence misalign the PSSM reading frame, and lead to improper scoring. Therefore, PSSM binding site discovery algorithms inherently mis-score at least 30% of the known p53 binding sites.

PHMMs can model nucleotide insertions and deletions

Profile Hidden Markov Models provide a coherent theory for probabilistic modeling of degenerate binding sites where random nucleotide insertions into and deletions from the motif are tolerated at certain positions [5,6]. Natural selection suggests that critical nucleotides are conserved over evolutionary time, while non-critical nucleotides (including tolerated insertions in the motif) are not conserved. The match-state emissions of the PHMM serve to model the critical positions in the motif with their observed nucleotide frequencies. The additional hidden deletion and insertion states at each position enable the model to train for (relatively rare) observed deletions and insertions at different positions in the motif (see Figure 2). Although the probability of any particular insertion or deletion of a nucleotide at a certain position in a functional motif may be rare, the accumu-

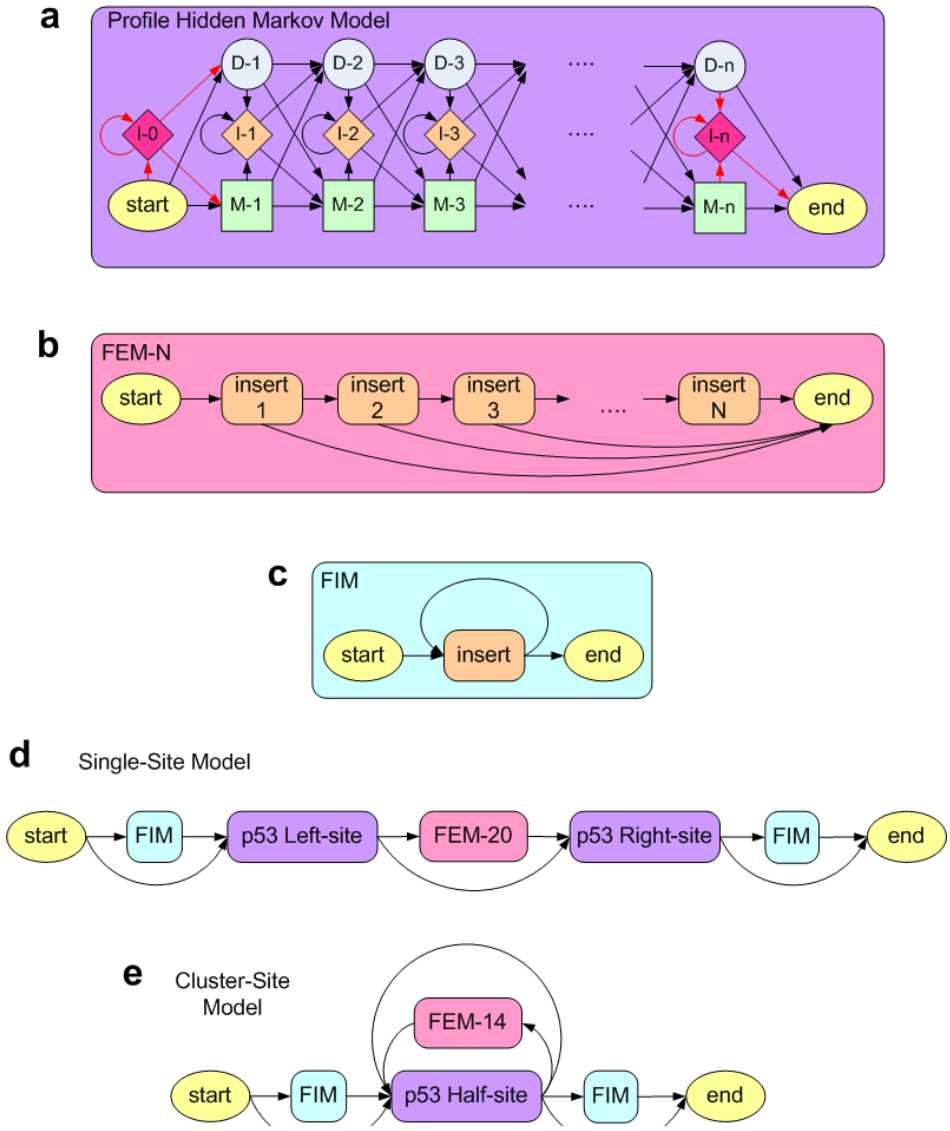

Figure 2

The Topologies of p53 Single-site and Cluster-site Models. **(a)** A Profile Hidden Markov Model (PHMM) contains three hidden states for each position in a sequence motif of length *n*: a match state (green squares), an insertion state (orange diamonds), and a delete state (gray circles). The arrows represent allowed transitions between states and have associated probabilities. The match and insertion states also have associated nucleotide emission probabilities. The first and last insertion states (I-0 and I-n) and associated transitions (in red) are shown for completeness. However, they are not present in the p53 models since they are replaced by FIM and FEM models. **(b)** The topology of the Finite Emission Module (FEM) of length *N* allows the ability to model any distribution of spacer-lengths between 1 and N. For the p53 models, the model and background probabilities within the FEM modules are identically uniform so that there is no-cost for spacer-lengths between 1 and *N*, and are referred to as "no-cost FEMs". **(c)** The topology of the Free Insertion Module (FIM) allows for the ability to model an exponentially decaying distribution of spacer-lengths. However, by setting the model and background probabilities to identically uniform, the FIM can model any sequence of infinite length with no associated cost to the overall score (hence the word "Free"). **(d)** The main components of the p53 single-site model are the left and right half-site PHMMs, which potentially contain corresponding positions between them. These two half-site models are separated by a no-cost FEM model that limits the length of any intervening spacer sequence to 20 bp. The half-site models are also wrapped by two FIMs that allow the Viterbi algorithm to find the best matching motifs anywhere in the candidate sequences. **(e)** The topology of the p53 cluster-site model consists of a single PHMM that models a general half-site, and two back-transitions that allow for modeling an infinite number of half-sites within the cluster-site. The back-transition through the no-cost FEM-14 model limits the spacer-sequence between the half-sites to lengths $\leq$ 14 bp.

lated probability over all the positions in the motif that an insertion or deletion event may occur can be significant. The training set of observed insertions and deletions serves to fine-tune the model to be properly sensitive to tolerated deviations from the most prevalent consensus motif. The main strength of the PHMM is this *trained flexibility* to properly model variable length motifs. The major drawback is that more data is required to train the extra parameters not found in weight matrices (PSSMs).

Using PHMMs to estimate binding affinities

Like weight matrices (PSSMs), Profile Hidden Markov Models can be used to estimate the relative binding affinity of a protein for a particular binding site sequence [7]. Under ideal conditions, the log-odds scores $G^s(x)$ that a Profile Hidden Markov Model (trained on training set S) calculates for any candidate site x is directly proportional to the free energy $-\Delta G(x)$ of the TF-protein binding to that candidate site [see Additional file 1 for details] [7-9]. The log-odds scores are given by:

$$G^s(x) = \ln\left(\frac{P_{hmm}(x)}{P_{background}(x)}\right) \quad \text{(Site Log-odds Score)}$$
$$= \sum_{j=1}^{\text{length}(x)} G_j^s(b)$$

where we define:

$$G_j^s(b) = \ln\left(\frac{P_{hmm}(j,b)}{P_{background}(j,b)}\right) \quad \text{(Nucleotide Log-odds Score)}$$
j = position in the sequence x, $j \in \{1 \ldots \text{length}(x)\}$
b = observed nucleotide base, $b \in \{A, C, G, T\}$
$P_{hmm}(j,b)$ = probability of base b at position j in the PHMM model
$P_{background}(j,b)$ = probability of base b at position j in the null (background) model
(1)

With these definitions, and assuming independence of positions, we have:

$P_{hmm}(x)$ = probability of candidate site x in the PHMM model
$P_{background}(x)$ = probability of candidate site x in the null (background) model

The dynamic programming *forward* and *backward* algorithms are used to calculate the probabilities $P_{hmm}(x)$ and $P_{hmm}(j, b)$. These two probabilities are calculated by summing up the probability of observing the sequence x, and the base b at position j, for all the paths through the linear PHMM, respectively. The dynamic programming *Viterbi* algorithm is used to find the best alignment of the candidate site x to the binding-site motif modeled by the PHMM. The best (optimal) alignment of the sequence x is obtained by finding the path through the PHMM that gives the highest log-odds score for the sequence [8]. In the case of transcription factor binding sites, the log-odds score of this optimal path (also called the *Viterbi score*) is

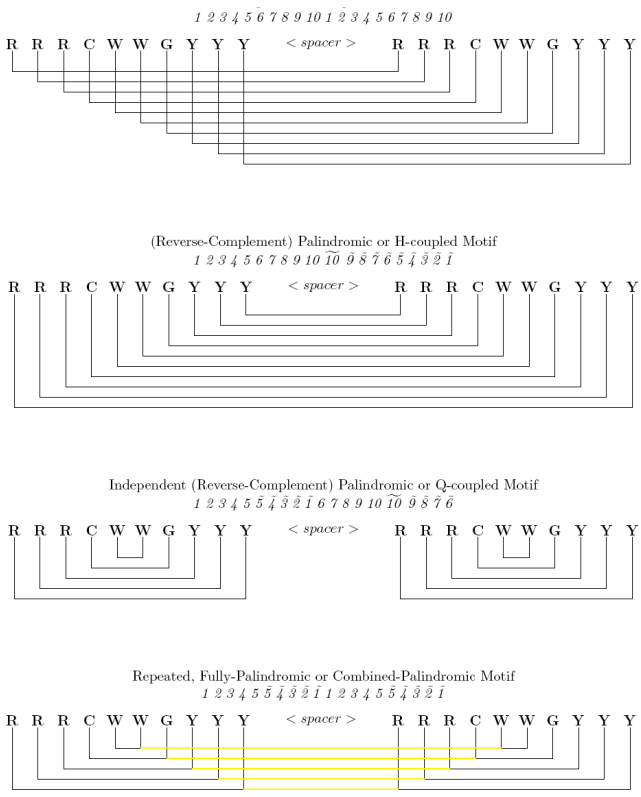

Figure 3

The Four p53 Correspondence Motifs. The four correspondence motifs for the repeated, palindromic p53 RE are graphically represented. In the top three motifs, each line corresponds 2 synonymous positions. In the bottom motif, the previously independent half-sites are made corresponding (tied) by the yellow connecting lines so that now 4 synonymous positions are corresponded. The completely un-tied motif (not shown) has no correspondence, and thus no connecting lines, between any of the positions in the motif. (R = A or G, W = A or T, and Y = C or T. Position $\bar{a}$ has the complement nucleotide emission distribution of a.)

commonly used to provide adequate approximations to the probabilities $P_{hmm}(x)$ and $P_{hmm}(j, b)$ [see Additional file 1 for details]. When using the *Viterbi score* for the probability $P_{hmm}(x)$ we are assuming that there is generally only one major set of binding interactions between specific nucleotides and amino acids for a given protein-DNA complex, and that all other possible binding locations in the response element can be ignored.

Training a PHMM with validated binding sites

Before a PHMM can be used to estimate the relative binding affinity for any putative binding site, the PHMM must be trained to properly model a functional binding site of interest. When training a PHMM for a particular motif, the goal is to choose the parameters of the model in order to maximize the likelihood of the sequences in the training

set, without over-fitting. Again, under ideal conditions the log-odds score (*log-likelihood ratio*) $G^s(x)$ to be maximized for the collection of binding sites in the training set is proportional to the estimated binding free energy $-\Delta G(x)$ of these binding sites. When the state paths for the training sequences are not known, no known closed form solution exists for the parameter estimations [8]. The *Baum-Welch* algorithm is the most commonly used iterative Expectation Maximization (EM) method to train the parameters of the model. The Baum-Welch algorithm always climbs the gradient (to increase the combined scores of the training set) and uses the optimized dynamic programming *forward* and *backward* algorithms [8].

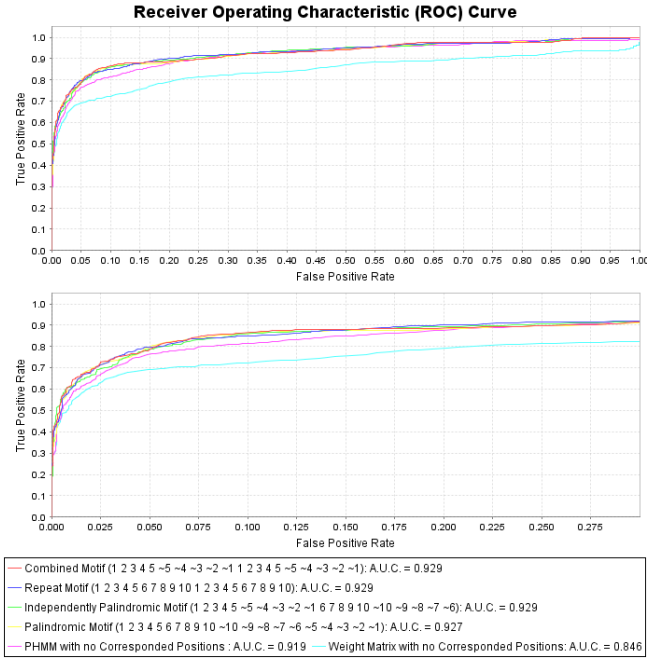

Figure 4
Cross Validation with Receiver Operating Characteristic (ROC) curves reveals increased predictive power over weight matrices. 1000 iterations of 10-fold random-split cross validation reveal that the most predictive models utilize the correspondence structures. The combined-palindromic model is the best model since it contains roughly half as many parameters as the other three correspondence models. The positive set contains 160 experimentally validated p53 binding sites, and the negative set contains 40 bp random samples from the mononucleotide content of the training set. The true positive and false positive rates are calculated and plotted for all possible threshold values for each model. The predictive measure for comparing the curves is the AUC (Area Under the Curve). In all the PHMM models the insert-state emissions are fixed to the A, G, C, T nucleotide distribution of the training set. The best classifier uses the combined-palindromic training motif. (Position $\tilde{a}$ has the complement nucleotide emission distribution of a).

Results and discussion

A novel training method that boosts predictive power

To increase the predictive power of our p53-motif PHMMs, we attempt to exploit the *a priori* knowledge that when proteins bind as homodimers or homotetramers, their corresponding binding sites typically have a *palindromic*, *repeat*, and/or *reverse complement* structure (see Figure 3). This prior knowledge can be used to correspond (fully or partially tie) the parameters between positions in order to exploit the inherent redundancy in the information of the motif. Within a set of corresponding positions, the updating of emission and transition probabilities can borrow strength from each other by sharing information. In addition, the degree of sharing of information for any set of corresponding positions can be optimized during training. The process of corresponding parameters can greatly reduce the parameter search-space during the training of the model, and provide the ability to train for rare occurrence insertion and deletion events. This general technique has been effectively used when HMMs have been applied to speech and handwriting recognition problems, and has been referred to as *parameter tying* [10]. We introduce an extension to this method that allows for the setting or training for an optimal level of partial or full parameter tying. In the domain of protein-DNA binding sites, even if a palindromic, repeat, or reverse complement structure of a binding site is not known *a priori*, all the known structural motifs can be tested, and the structure can be *discovered* (inferred) from the ROC curve that maximizes predictive accuracy. For example, of the six structural models tested for the p53 binding motif, the combined-palindromic motif that completely corresponds the four quarter-sites is the *discovered* motif, since it is the best classifier (see Figure 4).

The Corresponded Baum-Welch algorithm

In order to include the prior knowledge of the structural motif (or in an attempt to discover it), a novel "Corresponded Baum-Welch" algorithm is proposed to enforce or learn the optimal correspondence between expectations of parameters for corresponding positions after each iteration of the Baum-Welch algorithm (see Methods). For example, assume that we have prior knowledge that a transcription factor protein binds to the DNA in homodimer form, where each monomer interacts with 5 DNA base pairs. Then a corresponding palindromic motif for the nucleotide positions would be: $1\ 2\ 3\ 4\ 5\ 5\ 4\ 3\ 2\ 1$, while a reverse-complement palindromic motif would be: $1\ 2\ 3\ 4\ 5\ \tilde{5}\tilde{4}\tilde{3}\tilde{2}\tilde{1}$ (where $\tilde{a}$ has the complement nucleotide emission distribution of a). All the emission distributions for each of the five sets of synonymous positions would be made corresponding, as well as all the transition probabilities between synonymous positions. In this example, if

all the parameters between synonymous positions were fully corresponding (tied), then the parameter search space would be roughly cut in half. The level of correspondence between the parameters for synonymous positions can be given *a priori*, or trained for if the training set is sufficiently large. One optimal level of correspondence, *c*, can be calculated for the whole motif (for all the corresponding positions), or a separate one can be found for each set of corresponding positions. (See Methods for details.)

Comparing the different p53 corresponding (structural) motifs

Since the 20 bp-tetrameric p53 binding site has a repeated and nested palindromic structure, different correspondence motifs can be constructed to train the PHMM models, and cross validation can be used to compare their predictive properties. The motifs that are compared are: the repeat or T-coupled motif (*1 2 3 4 5 6 7 8 9 10 1 2 3 4 5 6 7 8 9 10*), the (reverse-complement) palindromic or H-coupled motif (*1 2 3 4 5 6 7 8 9 10* $\widetilde{10}$ $\tilde{9}\tilde{8}\tilde{7}\tilde{6}\tilde{5}\tilde{4}\tilde{3}\tilde{2}\tilde{1}$), the independently (reverse-complement) palindromic or Q-coupled motif (*1 2 3 4 5* $\tilde{5}\tilde{4}\tilde{3}\tilde{2}\tilde{1}$*6 7 8 9 10* $\widetilde{10}$ $\tilde{9}\tilde{8}\tilde{7}\tilde{6}$, the repeated, fully-palindromic or combined-palindromic motif (*1 2 3 4 5* $\tilde{5}\tilde{4}\tilde{3}\tilde{2}\tilde{1}$*1 2 3 4 5* $\tilde{5}\tilde{4}\tilde{3}\tilde{2}\tilde{1}$), and the completely un-tied motif with no correspondence between any positions (see Figure 3) [11]. We perform 1000 iterations of ten-fold random-split cross validation on each model to gain statistics on their predictive accuracy. The positive set contains 160 experimentally validated p53 binding sites from [4], and the negative set contains 40 bp random samples from the mononucleotide content of the training set. Then we utilize Receiver Operating Characteristic (ROC) curves in order to compare the predictive power of the classifiers in an unbiased, threshold-independent (non-parametric) manner. This is achieved by calculating the true positive and false positive rates for all possible threshold values for each model. The summary statistic for comparing the ROC curves is the AUC (Area Under the Curve). AUC values lie somewhere between 1.0 and 0.5 (where an AUC of 1.0 would correspond to a perfect classifier, and an AUC of 0.5 would correspond to a classifier that is no better than random coin flipping.)

Training Insert-State Emissions

A major consideration when training Profile Hidden Markov Models (PHMMs) is which parameters to train for at each position, and which parameters to fix at each position to the over-all average. The more non-fixed parameters that must be trained for at each position in the motif, the more data that is needed to properly train the model. Ideally, a sufficiently large training set is available to be able to train for all the parameters in the PHMM at each position. Unfortunately, in the case of transcription factor binding sites, this is rarely the case. Typically, when using PHMMs to model DNA binding sites, both the insert probabilities and insert state nucleotide emissions probabilities are set to the binding site averages, since there are rarely enough examples of these rare occurrence events at a particular position to train those parameters for that position alone [12]. By corresponding (fully or partially tying) positions and in effect increasing the training data

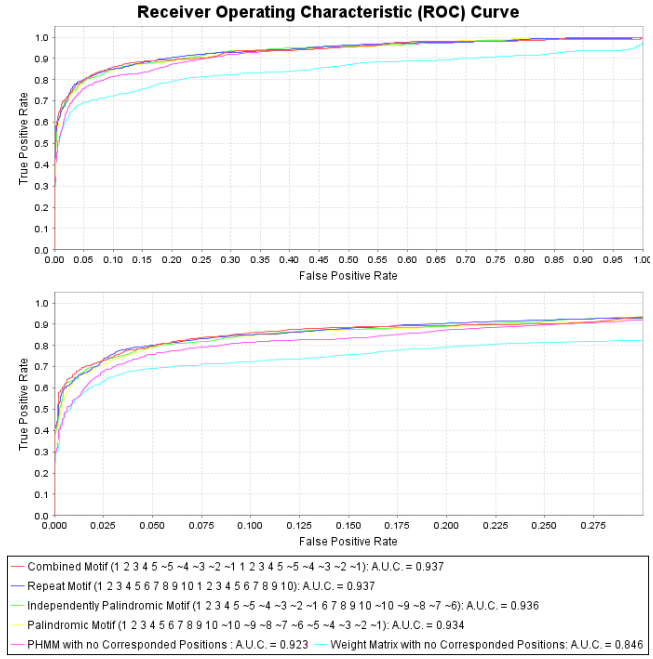

Figure 5
Cross Validation with Receiver Operating Characteristic (ROC) curves reveals increased predictive power when training insert-state emissions. All the PHMM models in this comparison train the insert-state emission distributions based on positional insertions occurring in the training set. Again, 1000 iterations of 10-fold random-split cross validation reveal that the most predictive models utilize the correspondence structures. The positive set contains 160 experimentally validated p53 binding sites, and the negative set contains 40 bp random samples from the mononucleotide content of the training set. The true positive and false positive rates are calculated and plotted for all possible threshold values for each model. The predictive measure for comparing the curves is the AUC (Area Under the Curve). The AUC values improve for all the PHMM models compared to Figure 4, but not for the weight-matrix model (which does not use the insert states). The best classifier (with the combined-palindromic training motif) was used for the p53HMM algorithm. (Position $\tilde{a}$ has the complement nucleotide emission distribution of *a*).

for each position, it may be possible to train the insertion-state emissions distributions for these corresponding positions. This could possibly boost predictive power of the models, if the p53 protein is selective as to which nucleotides can be inserted into the motif at certain positions without compromising the binding affinity of the site. A common example of such selective sequence insertions can be found in functional protein families, whereby hydrophobic or hydrophilic amino acid insertions may be tolerated at certain positions, provided that the insertions are present either in the core or at the surface of the protein, respectively, after folding. Notice that fixing the insertion-state emission distributions at every position to the amino-acid average for the whole sequence would be very inappropriate in this example.

The final results

The combined-palindromic motif (*1 2 3 4 5* $\tilde{5}\tilde{4}\tilde{3}\tilde{2}\tilde{1}$ *1 2 3 4 5* $\tilde{5}\tilde{4}\tilde{3}\tilde{2}\tilde{1}$) performs on par with or better than all other structural motifs, although it contains comparably half the degrees of freedom (see Figures 4 and 5). In addition, all four of the structural motifs perform on par with each other. These results suggest that there exist correlations between the positions in the repeat, independently palindromic, and palindromic motifs, and that the combined-palindromic motif leverages the correlations found in all of them. Furthermore, it can be seen that training the insert-state emissions per corresponding position also boosts the predictive power of all the models (see Figures 4 and 5). Analysis of the AUC measurements reveals some interesting features. Adding insert-state emission training to the base PHMM (with no motif-corresponded positions) has an AUC improvement of .923 - .919 = .004, but with motif training has one of .937 - .929 = .008. Adding motif training (motif-corresponded positions) to the PHMM when not insert-state emission training has an AUC improvement of .929 - .919 = .010, but with insert-state emission training has one of .937 - .923 = .014. Therefore the improvements are not additive. There is "positive synergy" when performing both motif training and insert-state emission training together that further boosts the predictive accuracy of the model. This observation confirms our hypothesis that training insert-state emissions can significantly boost the accuracy of the model after corresponding positions in the PHMM according to a binding-site motif.

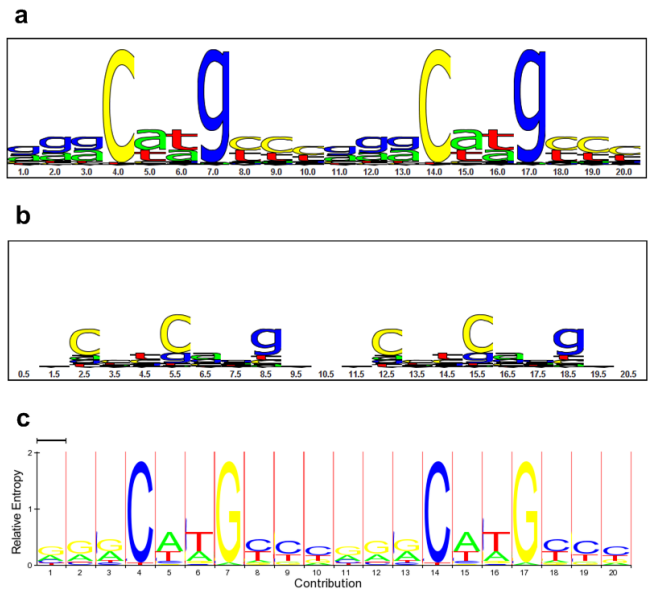

Figure 6

The p53HMM Match and Insert Emissions. **(a)** The match-state sequence logo for the combined-palindromic p53 motif: *1 2 3 4 5* $\tilde{5}\tilde{4}\tilde{3}\tilde{2}\tilde{1}$ *1 2 3 4 5* $\tilde{5}\tilde{4}\tilde{3}\tilde{2}\tilde{1}$. (Motif position $\tilde{a}$ has the complement nucleotide-emission distribution of *a*.) The height of each letter is made proportional to its frequency at each position, and the letters are sorted in descending frequency order. The height of the entire stack at each position is then adjusted to signify the information content (in bits) of that position [25]. The match-state nucleotide positions 4, 7, 14, and 17 (motif positions *4*, 7, $\tilde{7}$, and $\tilde{4}$ respectively) are the most conserved and are the main points of contact with the p53 protein. **(b)** The insert-state sequence logo for the same combined-palindromic p53-model. These nucleotide insertions occur in-between the nucleotide positions shown in part **a**. The specificity motif of the insert-state emissions is different from that of the match-state emissions. **(c)** The HMM logo that combines parts **a** and **b** and state transition information into one graph. The wide, white-background stacks correspond to the match states in part **a**, while the narrow, red-background stacks correspond to the insert states in part **b**. (A weakness of this HMM logo is that the insert-state stacks are so narrow that it is difficult to accurately see the stack specificity depicted in part **b**.) The y-axis is the same for all three graphs. However, the width of a stack in the HMM Logo is proportional to the expected contribution of that match or insert state to an emitted sequence of the model [26].

In addition, the more correspondence placed between the synonymous positions during each training iteration, the better the resulting classifier at that point in the training (results not shown). For this training set, all the combined-palindromic models with fixed correspondence factors between c = 0.4 and c = 1.0 eventually converged to the same predictive model, although lower correspondence factors required more iterations to do so. All the models converged on correspondence factors between c = .98 and c = .999 when training for optimum correspondence. Therefore the best predictive model completely corresponds (ties) the four quarter-sites in a combined-palindromic structure during each iteration of the training. Our published p53HMM algorithm is this best predictive model: trained on the dataset of 160 functional p53 REs, fully corresponding the data per position based on the combined-palindromic structural motif, and training the insert-state emissions (see Figure 6).

Validation of the p53HMM algorithm

The new p53HMM algorithm was used to screen for putative p53 binding sites in the endosomal compartment genes, which led to the discovery of a functional p53 site and a new p53-regulated gene, CHMP4C [13]. The putative p53RE sequence AAACAAGCCC agtagcagcagctgctcc GAGCTTGCCC was predicted in the promoter region (-497 to -460 bp) of the CHMP4C gene. The data from the chromatin immunoprecipitation and the luciferase reporter assays showed that p53 protein can bind to this sequence and induce CHMP4C gene expression. Additionally, analysis by p53HMM found an alternative putative p53 binding site in the LIF gene that corresponds to a 6 bp upstream shift of the downstream half-site relative to the recently published putative site in intron 1 [14]. The p53HMM algorithm predicted the site GGACATGTCG-GGACA-GCTC, which matches the consensus RRRCWW-GYYYRRRCWWGYYY perfectly except for the low-conserved position 10 and the gap ("-", deletion) at position 16. A PSSM approach predicted the shifted site GGA-CATGTCGggacagCTCCCAGCTC, which is the best "gapless" p53 site in the region conferring p53 regulation, but it still matches the consensus very poorly with five mismatches (the putative spacer sequence is in lowercase) [14]. A few genes in the dataset of 160 functional p53 binding sites have a deletion relative to the consensus exactly between the well-conserved C and G as seen above, including the genes: EGFR, TYRP1, EEF1A1, HSP90AB1, and BAI1. This discovery of an alternative p53 binding site that better matches known functional sites, by modeling for observed insertions and deletions, highlights some of the advantages of the new p53HMM algorithm.

Special considerations for the p53HMM algorithm

Although the spacer within a p53 RE has been shown to greatly affect the binding affinity for p53 protein, the ability to properly quantify this effect for all possible spacers of lengths 0–21 base pairs has been elusive. Therefore like previous algorithms, we have chosen to initially ignore the spacers of the training set and putative REs [15]. We are able to ignore arbitrary-length spacers by inserting a no-cost *Free Insertion Module* (FIM) between the two half-sites of the single-site PHMM [16,17]. Similarly, we can

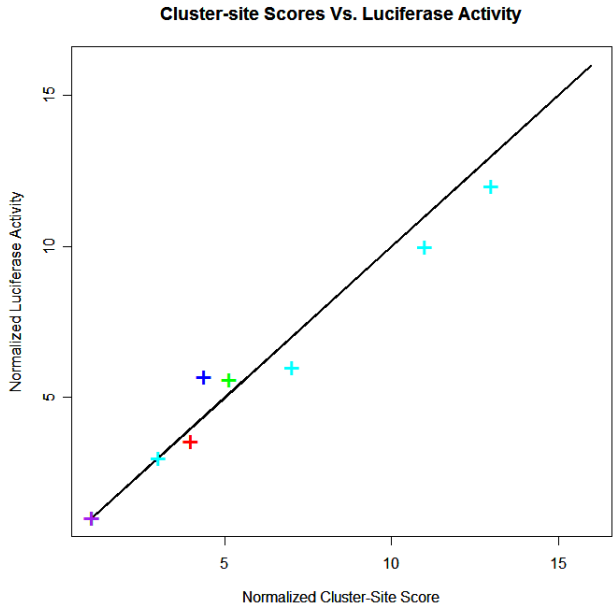

Figure 7
Comparison of Cluster-site scores and Luciferase Activity. This graph compares the estimated relative binding affinity given by the cluster-site score to the luciferase activity from four experiments for four different p53 cluster-sites. The four cluster-sites regulate the genes DDB2 (blue), CKM (red), IGFBP3 (green), and TP53I3 (cyan). In all four experiments the luciferase activity of truncated mutants of the respective p53 cluster-site were compared to the luciferase activity of the full cluster-site. In the case of the TP53I3 cluster-site, four different mutants of varying lengths were measured for luciferase activity. All cluster-site scores and activity measurements are normalized by the full-site (two half-sites) measurement. The cluster-site scores are attained by summing the estimated binding affinity of all viable full-sites in the cluster-site that have an affinity above a lower bound and spacer-lengths below an upper bound. The full-site affinity lower bound and spacer-length upper bound were chosen to best match the experimental data. The best fit was attained by enforcing that spacer-lengths not exceed 14 bp and affinity scores exceed 27.5.

Table 1: Normalized Experimental Affinity of Cluster-sites

	Number of Half-sites													
	2	**3**	**4**	**5**	**5.5**	**6**	**7**	**7.5**	**8**	**8.5**	**9**	**10**	**11**	**12**
Cluster Site	**Relative Binding Affinity**													
DDB2	1		5											
TP53I3		3			6			10		12			16	
Theoretical Affinity Approximations														
# of Full-sites with spacers ≤ 14 bp	*1*	*3*	*5*	*7*	*8*	*9*	*11*	*12*	*13*	*14*	*15*	*17*	*19*	*21*
# of Full-sites with spacers ≤ 24 bp	1	3	6	9	10.5	12	15	16.5	18	19.5	21	24	27	30
# of Full-sites with any size spacer	1	3	6	10		15	21		28		36	45	55	66

This table contains the normalized experimental affinities of different cluster-sites dependent upon the number of half-sites contained in the RE. These affinity measurements were obtained by mutating or truncating p53 cluster-sites in the genes DDB2, and TP53i3 [20,21]. These two p53 cluster-sites are chosen because they match the assumption of the theoretical models that no spacer sequences are present between the half-sites. All affinities are normalized by the 2 half-site (full-site) affinity respective of the RE. The theoretical models assume that all the half-sites in each cluster-site are identical, which is not the case for either of the two cluster-sites. Experimental results support a linear affinity growth model based upon the number of full-sites with spacers no longer than 14 bp (in italics).

ignore spacers with lengths between 1 and *N* base pairs by inserting a no-cost *Finite Emission Module* (FEM-N) between the two half-sites (see Figure 2). A prior p53 RE search algorithm (p53MH) was based upon a PSSM approach and a novel filtering matrix [15]. Unfortunately, the tables were not symmetric and the filtering table overfit the available data at the time. The combined result was that the p53MH method completely rejects 58 of the 160 experimentally validated sites to date (receiving a score of 0 out of 100, where 100 represented the maximum relative binding affinity). Additionally, some sites received very high scores approaching 100, while the reverse-complement received a score of 0, and vice-versa. Due to these observations, we have purposely designed the p53HMM algorithm to be symmetric, so as to give identical scores for putative sites and their reverse complements. Secondly, we chose to abandon the filtering matrix to avoid over-fitting the available data. A feature that we preserved from p53MH is the normalizing of scores by the highest possible affinity for the motif (×100), so that the highest possible normalized score is 100.

Modeling dependencies between positions

PSSMs assume that all nucleotide positions within the motif contribute independently to the binding affinity of the binding site, which has been shown experimentally to not always be the case [7]. Recent research has focused on modeling dependencies between positions in protein-DNA binding sites [18,19]. Typically *Tree Bayesian Networks* and *Mixtures* of trees have been used to attempt to model these dependencies between positions, which have been shown through cross validation to increase the predictive power of these models [18]. Our PHMM models do not attempt to model dependencies between the positions, however they can be extended to do so by using higher-order Profile Hidden Markov Models. Unfortunately, the ability to train for positional dependencies, and boost predictive power, is dependent upon the sampling size of the training set and requires larger training sets to train the extra parameters.

A novel p53 cluster-site algorithm

Binding affinity measurements have been obtained for certain p53 cluster-sites of different lengths by mutating or truncating known p53 cluster-sites in the genes: DDB2, TP53i3, CKM, IGFBP3, and RGC (see Table 1 and Figure 7) [20-23]. Based on the relative binding affinities of these p53 cluster-sites, we propose a new p53 cluster-site algorithm that utilizes the trained PHMM to calculate and sum up the relative estimated binding-affinities, above a certain threshold, of all viable full-sites in the cluster with a spacer of ≤ 14 bp or less (see Methods). This model predicts a linear increase in p53 binding affinity dependent upon the number of half-sites in the cluster-site and the length of spacers between them. For example, for p53 cluster-sites with 2, 3, 4, 5, or 6 adjacent p53 half-sites, the number of possible full-sites with spacer-lengths = 14 bp would be 1, 3, 5, 7, and 9, respectively. Let N be the number of half-sites in the cluster-site, then the number of full-sites (to calculate binding affinities for and sum up) is

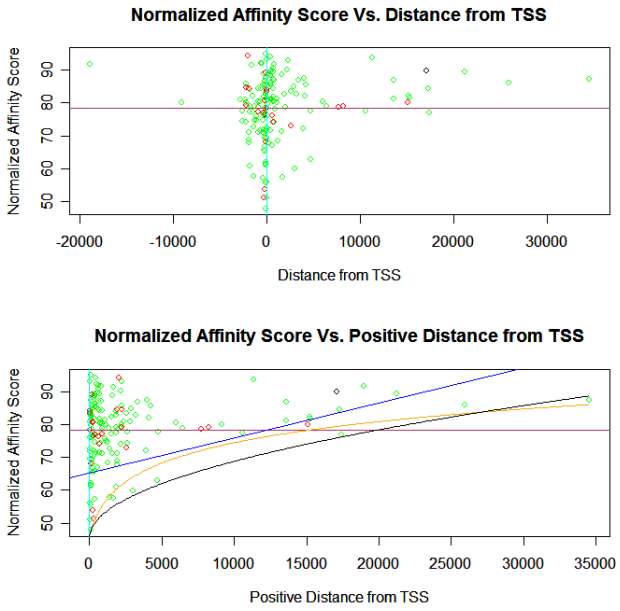

Figure 8
Normalized affinity scores versus distances from the TSS. (Upper) This plot presents the normalized affinity scores returned from the p53 single-site model versus the distance from the Transcription Start Site (TSS) for 158 experimentally validated p53-binding sites. Low affinity sites exist in a tight band around the TSS (cyan vertical line). p53 activation-sites are plotted in green, repression-sites in red, and both activation and repression in black. All sites ≥ 11 Kb from the TSS have relative affinity scores above the average of ≈ 78 (purple horizontal line). **(Lower)** This plot presents the estimated normalized affinity scores versus the positive distance (absolute value) from the TSS. Three dynamic acceptance thresholds are shown for scoring for putative p53 binding sites. The blue linear threshold corresponds to the formula .00107·Δ*X* + 65.16 and has a false negative rate of 18 out of 158 validated p53 sites (11.4%). The orange logarithmic threshold corresponds to the formula 9.6854·log(Δ*X* + 593.31) - 15.308 and has a false negative rate of 5 out of 158 validated p53 sites (3.2%). Finally, the black square-root threshold corresponds to the formula .23186·sqrt(Δ*X* + 7.5231) + 45.6 and has a false negative rate of 1 out of 158 validated p53 sites (0.63%). (Δ*X* = distance from TSS)

given by the expression 2*N* - 3 (*N* ≥ 2). Although there exist functional sites with spacers ≥ 15 bp, experiments suggest that their contribution to the overall binding affinity within a cluster-site is negligible.

These p53 cluster-site scores are attained through a two step process. The first step uses the cluster-site model which contains a generalized p53 half-site PHMM and a back-transition that limits any spacer between two half-sites to no more than 14 bp (see part e of Figure 2). The dynamic programming *Viterbi* algorithm is used to find the highest scoring p53 half-sites in the sequence (that are separated by no more then 14 bp). The second step then parses the state-path generated from step 1 and generates viable p53 full-sites with any spacers removed, while conserving the property that the half-sites in the cluster-site were not separated by more than 14 bp. Now we use the more flexible p53 single-site model to score these viable full-sites using the *Viterbi* algorithm (see part d of Figure 2). We maintain a running sum of the log-odds scores of the candidate full-sites that are above a certain threshold. The log-odds score threshold and spacer-length limit (14 bp) are chosen so as to best fit the experimental data (see Figure 7).

Additionally, this p53 cluster-site model follows statistical mechanics, in that the overall binding affinity for the complete RE is proportional to the probability of any p53 protein binding to any of the allowed motifs found in the cluster-site. (See Methods for more details.)

Dynamic acceptance thresholds as a function of the distance from the TSS

An interesting finding from the analysis of our dataset of 160 functional p53 binding sites is that the low relative affinity scores from our model are significantly correlated with short distances from the Transcription Start Site (TSS). We find that low affinity sites exist only in a tight band around the TSS (see part a of Figure 8). Therefore a dynamic binding-affinity acceptance threshold, dependent upon the putative site's distance from the TSS, can greatly reduce the false positive rate of our classifier. With a dynamic acceptance threshold, putative sites will require higher calculated binding affinities as their distance from the TSS increases in order to be accepted as potentially functional. For example, consider the linear dynamic acceptance threshold .00107 · Δ*X* + 65.16 shown in Figure 8, with the additional restriction that the putative sites must be within 5,000 bp upstream and 1,000 bp downstream of the gene. Let the static acceptance threshold be all normalized scores above 70 with the same restriction that the putative sites must be within 5,000 bp upstream and 1,000 bp downstream of the gene. Even though the restricted dynamic threshold has a false negative rate of 22 out of 158 validated p53 sites (13.9%), and the restricted static threshold 32 out of 158 (20.3%), the restricted static threshold generates over 3.2 times as many positive hits when scoring all 39,288 isoforms of known genes in the human genome (hg18). Thus, the dynamic acceptance threshold has a lower known false negative rate and a considerably lower false positive rate. Different dynamic acceptance thresholds can be chosen to match desired levels of the known false negative rate and the genome hit rate (see part b of Figure 8). An important consideration when choosing an acceptance threshold is that a decrease

in the threshold will in general produce an exponential increase in the number of positive hits.

Conclusion

Profile Hidden Markov Models (PHMMs) can boost predictive power over weight matrices (PSSMs) when the binding motif is highly degenerative and tolerates insertions and/or deletions at various positions. The increase in predictive power for the p53-binding motif can be seen in Figures 4 and 5. When the RE has a known repeated and/or palindromic motif, this prior knowledge can be used to correspond parameters in the model to exploit the redundancy in the information in the motif. We propose a novel "Corresponded Baum-Welch" training algorithm that significantly boosts the predictive power of the p53-RE model, as seen in Figures 4 and 5. When the motif is not known, all possible motifs for the given size can be sampled and cross-validation techniques leveraged to infer the correct motif that maximizes predictive power. For example, Figure 5 reveals that the maximally predictive p53-binding motif corresponds the four quarter-sites in a combined-palindromic structure.

Our algorithms demonstrate the best predictive capability to date in classifying putative p53 binding sites. One algorithm uses a novel "Corresponded Baum-Welch" training method that exploits the repeated, palindromic structure of the p53 motif to train for allowed insertions and deletions relative to the consensus. The second algorithm properly models the relative increase in binding affinity for p53 cluster-sites (REs with ≥ 3 adjacent half-sites) by using a two step process that scores all viable full-sites in the cluster-site while restricting the spacer-length to 14 bp. This new cluster-site algorithm best matches the experimental data (see Figure 7).

Functional low-affinity p53-sites only exist near the TSS. Therefore the binding affinity threshold for accepting a putative site should be dependent on the putative site's distance from the TSS. By this method, putative sites with relatively low calculated binding affinities that are near the TSS may be accepted, while those sites with equal scores but more distant from the TSS will be rejected. A dynamic threshold, as a function of the distance from the TSS, can greatly reduce the false positive rate when searching for putative p53-sites in genes.

Methods

The Corresponded Baum-Welch algorithm

In order to exploit the redundancy of information in a homodimer or homotetramer binding motif, we wish to share information between corresponding positions. The level of sharing of information for any set of corresponding positions is given by a correspondence factor c such that $0 \le c \le 1$. At the end of each round of the iterative Baum-Welch algorithm we calculate the average values of each of the newly updated emission probabilities $e'_k(b)$ and transition probabilities a'_{kl} for all k and l in the set of corresponding positions, represented as $\overline{e'(b)}$ and $\overline{a'}$ respectively. Each of these average values represents the expected probability if the corresponding positions are fully tied ($c = 1$), and are referred to as the "corresponding average". Then we update the new emission and transition probabilities within the set of corresponding positions, using the current correspondence factor and corresponding average, according to:

$$\begin{aligned} a''_{kl} &= a'_{kl} + c[\overline{a'} - a'_{kl}] \\ e''_k(b) &= e'_k(b) + c\left[\overline{e'(b)} - e'_k(b)\right] \end{aligned} \tag{2}$$

If we wish to train for the optimum correspondence factor, then we calculate a new c' for each emission and transition probability at each position in the set of corresponding positions:

$$\begin{aligned} c'_{kl} &= \frac{c\cdot\overline{a'}}{a'_{kl}+c[\overline{a'}-a'_{kl}]} = \frac{c\cdot\overline{a'}}{a''_{kl}} \\ c'_k(b) &= \frac{c\cdot\overline{e'(b)}}{e'_k(b)+c\left[\overline{e'(b)}-e'_k(b)\right]} = \frac{c\cdot\overline{e'(b)}}{e''_k(b)} \end{aligned} \tag{3}$$

Now, we can calculate a new correspondence factor c' by averaging over sets of the c'_{kl} and $c'_k(b)$ values. The one optimum correspondence factor for the whole motif or separate correspondence factors for sets of corresponding positions are obtained by averaging over different sets:

$$c' = \overline{c'_k(b)} \quad \begin{array}{l}\text{(over all bases } b \text{ and all emissions and transitions } k\text{)} \\ \text{or} \\ \text{(over all bases } b \text{ and corresponding emissions and transitions } k\text{)}\end{array} \tag{4}$$

The Corresponded Baum-Welch algorithm will converge at (local) optimum emission and transition probabilities and correspondence factors that maximize the likelihood of observing the training set with possible pseudo-counts. Please see the Additional file 1 for further details.

The p53 cluster-site algorithm

The p53 cluster-site algorithm is a two step process designed to sum the estimated relative binding affinities of all viable full-sites within a cluster-site. The first step uses the cluster-site model that contains a generalized p53

half-site PHMM and a back-transition through a no-cost FEM-14 module (see part e of Figure 2). The no-cost Finite Emission Module (FEM) of length 14 can match any sequence of length ≤ 14 bp with no contribution to the over-all score. We score the entire putative cluster-site using the p53 cluster-site model and the *Viterbi* algorithm to find the best-supported path through the cluster-site. This path provides the strongest affinity half-sites that are not separated by more than 14 bp. If we use the notation "14" for any spacer sequence of length 0 to 14 and *H* for a half-site sequence, then we can represent the cluster-site sequence path as:

$H_1[14]H_2[14]H_3[14]....[14]H_N$ (where N = number of half-sites in the path)

Step 2 now parses the cluster-site sequence path and generates a list of all viable full-sites, which are concatenations of any two half-sites such that they are not separated by more than 14 bp:

$$\text{Set of viable full-sites} = \{H_1H_2, H_1H_3, H_2H_3, ...\}$$

Now we use the more flexible (and more accurate) single-site model with the Viterbi algorithm to estimate the relative binding affinity of all the viable full-sites in the cluster-site. The cluster-site affinity score is the sum of all viable full-site scores that exceed a certain threshold. If *F* denotes a viable full-site then:

$$\text{cluster-site affinity score} = \sum_{F}^{\{H_1H_2, H_1H_3, H_2H_3, ...\}} \text{contribution}(F)$$

$$\text{contribution}(F) = \begin{cases} Viterbi(F) & \text{if} \quad Viterbi(F) \geq \text{threshold}; \\ 0 & \text{if} \quad Viterbi(F) < \text{threshold}. \end{cases} \quad (5)$$

The spacer-length upper bound and the affinity-score lower bound were fit to best match the experimental results. In the case for p53-binding sites, the best fit is a spacer-length of no more than 14 bp and a log-odds score of at least 27.5 (see Figure 7).

The p53HMM implementation

The p53HMM algorithm is implemented in Java and is available on-line at http://tools.csb.ias.edu. The implementation makes extensive use of the BioJava Toolkit [24].

Authors' contributions

TR participated in the design of the algorithms, wrote the code, performed the computational analysis, and drafted the manuscript. XY performed all experiments. ES participated in the design of the algorithms and helped to draft the manuscript. AL conceived of the study and helped to draft the manuscript. All authors read, edited, and approved the final version of the paper.

Additional material

Additional File 1

***Supplementary Material**. The Supplementary Material contains more theory behind modeling TF-binding sites with PHMMs, details of the Corresponded Baum-Welch algorithm, and a proof that the PHMM log-odds score of a TF-binding site estimates its relative binding affinity given certain assumptions.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-10-111-S1.pdf]

Acknowledgements

We thank Michael Krasnitz, Amar Drawid, Anirvan Sengupta, Sean Eddy, Jiri Vanicek, and Raúl Rabadán for helpful discussions.

References

1. Levine AJ: **p53, the cellular gatekeeper for growth and division.** *Cell* 1997, **88(3):**323-331.
2. Funk WD, Pak DT, Karas RH, Wright WE, Shay JW: **A transcriptionally active DNA-binding site for human p53 protein complexes.** *Mol Cell Biol* 1992, **12(6):**2866-2871.
3. el Deiry WS, Kern SE, Pietenpol JA, Kinzler KW, Vogelstein B: **Definition of a consensus binding site for p53.** *Nat Genet* 1992, **1:**45-49.
4. Riley T, Sontag E, Chen P, Levine A: **Transcriptional control of human p53-regulated genes.** *Nat Rev Mol Cell Biol* 2008, **9(5):**402-412.
5. Krogh A, Brown M, Mian IS, Sjölander K, Haussler D: **Hidden Markov models in computational biology. Applications to protein modeling.** *J Mol Biol* 1994, **235(5):**1501-1531.
6. Eddy SR: **Profile hidden Markov models.** *Bioinformatics* 1998, **14(9):**755-763.
7. Stormo G, Fields D: **Specificity, free energy and information content in protein-DNA interactions.** *Trends in Biochemical Sciences* 1998, **23(5):**109-113.
8. Durbin R, Eddy S, Krogh A, Mitchison G: *Biological sequence analysis* 1st edition. Cambridge University Press; 1998.
9. Djordjevic M, Sengupta AM, Shraiman BI: **A Biophysical Approach to Transcription Factor Binding Site Discovery.** *Genome Res* 2003, **13(11):**2381-2390.
10. Lee J, Kim J, Kim J: **Data-driven design of hmm topology for online handwriting recognition.** 2000 [http://citeseer.ist.psu.edu/lee00datadriven.html].
11. Ma B, Pan Y, Zheng J, Levine AJ, Nussinov R: **Sequence analysis of p53 response-elements suggests multiple binding modes of the p53 tetramer to DNA targets.** *Nucleic Acids Res* 2007, **35(9):**2986-3001.
12. Marinescu VD, Kohane IS, Riva A: **MAPPER: a search engine for the computational identification of putative transcription factor binding sites in multiple genomes.** *BMC Bioinformatics* 2005, **6:**79.
13. Yu X, Riley T, Levine AJ: **The regulation of the endosomal compartment by p53 the tumor suppressor gene.** *FEBS Journal* 2009, **276(8):**2201-2212.
14. Hu W, Feng Z, Teresky AK, Levine AJ: **p53 regulates maternal reproduction through LIF.** *Nature* 2007, **450(7170):**721-724.
15. Hoh J, Jin S, Parrado T, Edington J, Levine AJ, Ott J: **The p53MH algorithm and its application in detecting p53-responsive genes.** *Proc Natl Acad Sci USA* 2002, **99(13):**8467-8472.
16. Hughey R, Krogh A: **Hidden Markov models for sequence analysis: extension and analysis of the basic method.** *Comput Appl Biosci* 1996, **12(2):**95-107.
17. Barrett C, Hughey R, Karplus K: **Scoring hidden Markov models.** *Comput Appl Biosci* 1997, **13(2):**191-199.
18. Barash Y, Elidan G, Friedman N, Kaplan T: **Modeling dependencies in protein-DNA binding sites.** In *RECOMB '03: Proceedings of the*

seventh annual international conference on Research in computational molecular biology New York, NY, USA: ACM Press; 2003:28-37.
19. Zhou Q, Liu JS: **Modeling within-motif dependence for transcription factor binding site predictions.** *Bioinformatics* 2004, **20(6):**909-916.
20. Tan T, Chu G: **p53 Binds and activates the xeroderma pigmentosum DDB2 gene in humans but not mice.** *Mol Cell Biol* 2002, **22(10):**3247-3254.
21. Contente A, Dittmer A, Koch MC, Roth J, Dobbelstein M: **A polymorphic microsatellite that mediates induction of PIG3 by p53.** *Nat Genet* 2002, **30(3):**315-320.
22. Bourdon JC, Deguin-Chambon V, Lelong JC, Dessen P, May P, Debuire B, May E: **Further characterisation of the p53 responsive element-identification of new candidate genes for transactivation by p53.** *Oncogene* 1997, **14:**85-94.
23. Kern SE, Pietenpol JA, Thiagalingam S, Seymour A, Kinzler KW, Vogelstein B: **Oncogenic forms of p53 inhibit p53-regulated gene expression.** *Science* 1992, **256(5058):**827-830.
24. Holland RCG, Down TA, Pocock M, Prlić A, Huen D, James K, Foisy S, Dräger A, Yates A, Heuer M, Schreiber MJ: **BioJava: an open-source framework for bioinformatics.** *Bioinformatics* 2008, **24(18):**2096-2097.
25. Schneider TD, Stephens RM: **Sequence logos: a new way to display consensus sequences.** *Nucleic Acids Res* 1990, **18(20):**6097-6100.
26. Schuster-Böckler B, Schultz J, Rahmann S: **HMM Logos for visualization of protein families.** *BMC Bioinformatics* 2004, **5:**7.

The p53HMM algorithm: using profile hidden markov models to detect p53-responsive genes

Todd Riley*[1,2] , Xin Yu[4] , Eduardo Sontag[2,3] and Arnold Levine[1,4]

[1]The Institute for Advanced Study, Princeton, New Jersey
[2]The BioMaPS Institute at Rutgers University, Piscataway, New Jersey
[3]The Mathematics Department, Rutgers University, Piscataway, New Jersey
[4]The Cancer Institute of New Jersey, New Brunswick, New Jersey

Email: TR: tr2261@columbia.edu; XY: yuxi@umdnj.edu; ES: sontag@math.rutgers.edu; AL: alevine@ias.edu;

*Corresponding author

Supplementary Material

The Theory of Modeling TF-Binding Sites with Profile Hidden Markov Models

Given a set S of experimentally validated binding sites s for a TF-protein (and a few assumptions) it is possible to use the set S to estimate the relative binding free energy $-\Delta G(x)$ of any putative site x (without having to perform direct experimental measurements of binding constants). This bioinformatic approach using PHMMs (and PSSMs) is an attractive alternative, if a sufficient set S of experimentally validated binding sites is available.

The Assumptions:

1. The positions of a binding site contribute independently and additively to the binding free-energy
2. Background DNA sequences are generally random samples from some k-mer distribution

Neither of these assumptions are always true [1]. The first assumption can be relaxed by calculating di-nucleotide, tri-nucleotide,....,n^{th}-nucleotide frequencies from the training set S, but at some point an additivity assumption must be applied. Also, genomes are generally not random, but can be closely approximated by a 3^{rd} or 4^{th} Order Markov Model [2]. For simplicity in the examples here, we will assume that the background DNA can be modeled by a simple 0^{th} Order Markov Model (i.e. by mononucleotide content alone). This assumption greatly simplifies the calculation of the partition function [1].

From the additivity assumption we have that for any putative site x:

$$-\Delta G(x) = \sum_{i=1}^{length(x)} -\Delta G_j(b)$$

where we define...

$$-\Delta G_j(b) = \text{the independent contribution of base } b \text{ observed at position } j \text{ to the over-all binding free energy} \quad (1)$$

The Profile Hidden Markov Model (PHMM) provides a completely probabilistic model for observing a sequence x within the modeled motif. The PHMM achieves this by incorporating the probabilities of different nucleotide insertions, deletions, and motif matches at each position in the motif [3]. In this application, the PHMM model is used to calculate the probability $P_{hmm}(x)$ of observing the putative site x in a real transcription factor binding site that is modeled by the PHMM. The probability $P_{hmm}(x)$ is used to find the *site log-odds score* of a putative site x. The *site log-odds score* $G^s(x)$ calculated by a PHMM trained by S is given by:

$$\begin{aligned} G^s(x) &= \log_e\left(\frac{P_{hmm}(x)}{P_{background}(x)}\right) \qquad \text{(Site Log-odds Score)} \\ &= \sum_{j=1}^{\text{length}(x)} G_j^s(b) \end{aligned}$$

where we define:

$$\begin{aligned} G_j^s(b) &= \log_e\left(\frac{P_{hmm}(j,b)}{P_{background}(j,b)}\right) \qquad \text{(Nucleotide Log-odds Score)} \\ j &= \text{position in the sequence } x,\ j \in \{1 \ldots \text{ length}(x)\} \\ b &= \text{observed nucleotide base, } b \in \{A, C, G, T\} \\ P_{hmm}(j,b) &= \text{probability of base } b \text{ at position } j \text{ in the PHMM model} \\ P_{background}(j,b) &= \text{probability of base } b \text{ at position } j \text{ in the null (background) model} \end{aligned} \quad (2)$$

With these definitions, and assuming independence of positions, we have:

$$\begin{aligned} P_{hmm}(x) &= \text{probability of candidate site } x \text{ in the PHMM model} \\ P_{background}(x) &= \text{probability of candidate site } x \text{ in the null (background) model} \end{aligned}$$

The *Site Log-odds Score* $G^s(x)$ can be considered proportional to the relative binding free energy $-\Delta G(x)$ when the Fermi-Dirac Equation for the equilibrium probability of a protein-bound binding site can be

approximated by the Maxwell-Boltzmann Equation [4]. Another assumption is that the training set S consists of a proper sampling of functional binding sites that were collected under similar experimental conditions (like temperature T). However, this is likely not the case. A last assumption is that we are able to perfectly train the PHMM from our training set S, so that we can accurately predict the probability $P_{hmm}(x)$ for all possible putative sites x. However, properly training a PHMM from a limited training set S is a challenging problem. But with our idealizations and assumptions, the *Nucleotide Log-odds Score* $G_j^s(b)$ (calculated by our perfectly trained PHMM) is directly proportional to the binding free energy contribution of each observed base b at each position j in the sequence x.

Thus, under ideal conditions the log-odds scores that a trained Profile Hidden Markov Model calculates for any candidate site x is directly proportional to the free energy of binding to that candidate site. (Typically, proper scaling of G^s if not performed to make $G^s(x) \approx -\Delta G(x)$. Instead, G^s is only proportional to $-\Delta G(x)$.) [5] If the Profile Hidden Markov Model has no insertion or deletion states, then the PHMM is essentially a PSSM (weight matrix), and the probability $P_{hmm}(j, b)$ is equivalent to the $(b, j)^{th}$ entry in the (probability) weight matrix.

Three dynamic programming algorithms are used to calculate the probability $P_{hmm}(x)$ of observing the putative site x in the model. The *forward* and *backward* algorithms calculate $P_{hmm}(x)$ by summing up the probability of observing x for all possible paths π through the model:

$$forward(x) \quad = \quad backward(x) \quad = \quad P_{hmm}(x) \quad = \quad \sum_{\pi}^{all\ paths} P(x, \pi) \tag{3}$$

The *Viterbi* algorithm calculates both the optimal alignment of the putative site x which produces the path $\pi^*(x)$ with the highest log-odds score, and the probability $P_{hmm}^{\pi^*}(x)$ of observing that optimal path in the model. These two results of the Viterbi algorithm are commonly referred to as the *Viterbi path* and the *Viterbi score*, respectively:

$$\begin{array}{lcll} Viterbi\ path(x) & = & \pi^*(x) & = & \underset{\pi}{\text{argmax}}\,[P(x, \pi)] \\ Viterbi\ score(x) & = & P_{hmm}^{\pi^*}(x) & = & P_{hmm}(x, \pi^*(x)) \end{array}$$

In the case of modeling transcription factor binding sites, it is commonly assumed that the log-odds score of the optimal path that best aligns the putative site x to the model is the only significant contributor to the over-all log-odds score. When this is indeed true, the *Viterbi score* can be used as a good

approximation to $P_{hmm}(x)$:

$$Viterbi\ score(x) \quad = \quad P_{hmm}(x, \pi^*(x)) \quad \approx \quad \sum_{\pi}^{all\ paths} P(x, \pi) \quad = \quad P_{hmm}(x) \quad = \quad forward(x) \quad (4)$$

However, we see that this assumption is not true when modeling p53 cluster sites, where experiments suggest that the p53 protein can bind to overlapping combinations of adjacent half-sites. In this scenario, the true probability $P_{hmm}(x)$ provided by the *forward* and *backward* algorithms is needed to properly model experimental results.

All three dynamic programming algorithms are highly efficient, and when applied to PHMMs run in $O(NM)$ time and $O(NM)$ space for a PHMM with M states and a sequence of length N [6]. For further details about the *forward*, *backward*, and *Viterbi* algorithms please see [5].

The Corresponded Baum-Welch Algorithm

The standard Baum-Welch EM algorithm is used to estimate the expected transition and emission probabilities from the training set. The Baum-Welch algorithm is an optimized, iterative EM method that always climbs the gradient and uses the dynamic programming *forward* and *backward* algorithms [5].

Let:

s	$=$ *binding site*	The nucleotide sequence of a binding site
s_i	$=$ *nucleotide*	The i^{th} nucleotide in the binding site s
S	$=$ *training set*	The training set of binding sites s_j
π	$=$ *path*	The state sequence of a binding site s
π_i	$=$ *state*	The i^{th} state in the path π
ps_{kl}	$=$ *pseudocount*	Prior bias of probability of transition from k to l
$ps_k(b)$	$=$ *pseudocount*	Prior bias of probability of emitting symbol b in state k
ψ	$=$ $\{ps_{kl}, ps_k(b)\}, \forall k, l, b$	The set of all pseudocounts in the model
a_{kl}	$=$ $P(\pi_i = l \mid \pi_{i-1} = k)$	The probability of transition from state k to state l
$e_k(b)$	$=$ $P(s_i = b \mid pi_i = k)$	The probability of emitting symbol b in state k
θ	$=$ $\{a_{kl}, e_k(b)\}, \forall k, l, b$	The set of all parameters in the model
$a_{kl}^{background}$	$=$ $P_{background}(\pi_i = l \mid \pi_{i-1} = k)$	The probability of transition from state k to state l in the null (background) model
$e_k^{background}(b)$	$=$ $P_{background}(s_i = b \mid pi_i = k)$	The probability of emitting symbol b in state k in the null (background) model
A_{kl}	$=$ *expected* a_{kl} *counts*	Number of transitions from k to l in the training set
$E_k(b)$	$=$ *expected* $e_k(b)$ *counts*	Number of emissions of b from state k in the training set
$f_k(i)$	$=$ $P(s_1 \ldots s_i, \pi_i = k)$	The probability of the sequence up to and including s_i, requiring that $\pi_i = k$
$f_k(i+1)$	$=$ $e_k(s_{i+1}) \cdot \sum_j^{states}(f_j(i) \cdot a_{jk})$	Recursive formula for $f_k(i+1)$ going forward
$b_k(i)$	$=$ $P(s_i \ldots s_L, \pi_i = k)$	The probability of the sequence from s_i to the end, requiring that $\pi_i = k$, L = length of the sequence s
$b_k(i-1)$	$=$ $e_k(s_{i-1}) \cdot \sum_j^{states}(b_j(i) \cdot a_{jk})$	Recursive formula for $b_k(i-1)$ going backward

The goal is to choose the parameters θ of the model in order to maximize the log-likelihood of the sequences s in the training set S, without over-fitting. To avoid over-fitting, the goal is to find the Posterior Mean Estimator (PME), a Bayesian approach that uses the pseudo-counts ψ as a prior from a Dirichlet family of distributions and all the paths π for all sequences s in the training set S [5]:

$$\theta^{PME} = \underset{\theta}{\text{argmax}} \left[\sum_{s \in S} \log P(s|\theta, \psi) \right] = \underset{\theta}{\text{argmax}} \left[\sum_{s \in S} \sum_{\pi} \log P(s, \pi|\theta, \psi) \right]$$

The Baum-Welch algorithm climbs the gradient during each iteration and is guaranteed to converge within some epsilon to a local maximum, which may or may not be the PME [5]. Theoretically, the Corresponded Baum-Welch algorithm has the advantage of using prior motif knowledge to greatly reduce the parameter space and to potentially "flatten" the space. Both of these improvements can increase the probability of the algorithm converging to the PME.

In each iteration, the Baum-Welch algorithm calculates the expected number of times each transition and emission is used by the training set sequences (calculates A_{kl} and $E_k(b)$), given the current model parameters (a_{kl} and $e_k(b)$). Then the model parameters are updated to the new posterior mean estimators

a'_{kl} and $e'_k(b)$, calculated from the new expectation counts (A_{kl} and $E_k(b)$).

Notice that the probability that a_{kl} is used at position i of binding site sequence s with current model parameters θ is given by:

$$P(\pi_i = k, \pi_{i+1} = l | s, \theta) = \frac{f_k(i) \cdot a_{kl} \cdot e_l(s_{i+1}) \cdot b_l(i+1)}{P(s)}$$

By summing over all training sequences and positions, we can derive A_{kl} and $E_k(b)$, the expected number of times that a_{kl} and $e_k(b)$ are used by the training set, given the current model parameters θ:

$$\begin{aligned} N &= \text{number of training sequences} \\ L &= \text{length of the sequence } s^j \\ W(s^j) &= \text{sequence weight of } s^j \\ A_{kl} &= \sum_{s^j \in S}^{N} \frac{W(s^j)}{P(s^j)} \sum_{i=1}^{L} f_k^j(i) \cdot a_{kl} \cdot e_l(s_{i+1}^j) \cdot b_l^j(i+1) \\ E_k(b) &= \sum_{s^j \in S}^{N} \frac{W(s^j)}{P(s^j)} \sum_{i | s_i^j = b}^{L} f_k^j(i) \cdot b_k^j(i) \end{aligned} \tag{5}$$

The sequence weight $W(s^j)$ is used to vary the importance of different sequences in the training set S and to vary their influence in training the model. A weight $W(s^j) > 1$ increases the expected counts in sequence s^j, and a weight $W(s^j) < 1$ decreases them. Sequence weights are used when we do not fully trust that the training set S provides a proper distribution of valid binding sites, and we attempt to remedy that deficiency by weighting the known sequences. Most sequence weighting methods attempt to penalize the expected counts of similar sequences and to enhance the expected counts of distant sequences [5].

Additionally, the process by which the training set S was ascertained may be biased toward a certain subset of sites independent of their sequences (*ascertainment bias*). In the derivation for our approximation for $-\Delta G(x)$ in the next section, we relied on the assumption that the probability $P_{extract}(x)$ of extracting a TF-bound binding site was independent of the sequence in or around x. This may not always be the case. For example, if we know that a certain antibody preferentially binds to adjacent binding sites compared to ones with no neighbors, then after precipitation our training set S would be biased toward adjacent binding sites that appear in tight clusters in the DNA. We could attempt to compensate for this inherent bias by penalizing those sequences found adjacent to each other in the genome and promoting the

ones with no neighbors. Different sequence weighting schemes can be found in [7–13].

From these new expected counts, we can now calculate new maximum likelihood estimators for each position:

$$\begin{aligned} a'_{kl} &= \frac{A_{kl}}{\sum_{m}^{states} A_{km}} \\ e'_k(b) &= \frac{E_k(b)}{\sum_{n}^{\{A,C,G,T\}} E_k(n)} \end{aligned} \tag{6}$$

However, if we believe the training set S to be incomplete and intend to avoid over-fitting the data, we add pseudocounts as priors to our expected counts. Here, pseudocounts are distributed in proportion to the null (background) model. The pseudocount weight w represents how many counts from the null (background) model we want to include in the expected counts of our model. From the expected counts, we calculate the new posterior mean estimators using pseudocounts for each position:

$$\begin{aligned} w &= \text{pseudocount weight} \\ ps_{kl} &= w \cdot a_{kl}^{background} \\ ps_k(b) &= w \cdot e_k^{background}(b) \\ a'_{kl} &= \frac{ps_{kl} + A_{kl}}{w + \sum_{m}^{states} A_{km}} \\ e'_k(b) &= \frac{ps_k(b) + E_k(b)}{w + \sum_{n}^{\{A,C,G,T\}} E_k(n)} \end{aligned} \tag{7}$$

Now we use the prior knowledge (or make a guess) of the repeat and/or palindromic motif and correspond (partially or fully tie) the new posterior mean estimators based upon corresponding positions. This prior knowledge can be used to reduce the parameter space and increase the statistical accuracy of the model. The degree of sharing of information between corresponding positions is controlled by a correspondence factor c, which can be fixed or trained to an optimum value. One can estimate a correspondence factor

based on the initial conditions by the following:

$$\begin{aligned}
dist &= \text{a probability distribution in the set of corresponding distributions} \\
var &= \text{a variable in the probability distributions} \\
N &= \text{number of corresponding distributions} \\
\overline{P(var)} &= \text{average probability of a variable over all corresponding distributions} \\
c_0 &= \text{initial correspondence factor} \\
&= 1 \quad - \quad \frac{1}{N-1} \sum_{dist} \sum_{var} \left| \overline{P(var)} - P(var) \right|
\end{aligned} \tag{8}$$

We calculate the corresponding posterior mean estimator (PME) after calculating the average emission and transition probabilities for all the corresponding positions:

$$\begin{aligned}
c &= \text{correspondence factor} \\
\overline{a'} &= Avg(a'_{kl}) \quad \text{(over all transitions from } k \text{ to } l \text{ in the set of corresponding positions)} \\
\overline{e'(b)} &= Avg(e'_k(b)) \quad \text{(over all emissions in the set of corresponding positions)} \\
a''_{kl} &= a'_{kl} + c\left[\overline{a'} - a'_{kl}\right] \\
e''_k(b) &= e'_k(b) + c\left[\overline{e'(b)} - e'_k(b)\right]
\end{aligned} \tag{9}$$

If we wish to train for the optimum correspondence factor, then we calculate a new c' for each emission and transition probability at each position in the set of corresponding positions:

$$\begin{aligned}
c'_{kl} &= \frac{c \cdot \overline{a'}}{a'_{kl} + c\left[\overline{a'} - a'_{kl}\right]} = \frac{c \cdot \overline{a'}}{a''_{kl}} \\
c'_k(b) &= \frac{c \cdot \overline{e'(b)}}{e'_k(b) + c\left[\overline{e'(b)} - e'_k(b)\right]} = \frac{c \cdot \overline{e'(b)}}{e''_k(b)}
\end{aligned} \tag{10}$$

Now, we can calculate a new correspondence factor c' by averaging over sets of the c'_{kl} and $c'_k(b)$ values. The one optimum correspondence factor for the whole motif or separate correspondence factors for sets of corresponding positions are obtained by averaging over different sets:

$$\begin{aligned}
c' &= \overline{c'_k(b)} \quad \text{(over all bases } b \text{ and all emissions and transitions } k\text{)} \\
&\qquad\qquad \text{or} \\
&\qquad\qquad \text{(over all bases } b \text{ and corresponding emissions and transitions } k\text{)}
\end{aligned} \tag{11}$$

We can now update the parameters of the model to the new posterior mean estimators that have been made corresponding (fully or partially tied) by our prior knowledge (or guess) of the motif:

$$\begin{aligned} a_{kl} &\Longrightarrow a''_{kl} \\ e_k(b) &\Longrightarrow e''_k(b) \\ c &\Longrightarrow c' \end{aligned} \tag{12}$$

This process is then iterated to obtain new A_{kl} and $E_k(b)$ values from the new model parameters. At each iteration the log likelihood of the training set increases to a local maximum. Since convergence is in a continuous-valued space, the maximum is never actually reached. Typically, the iterations are stopped when the change in the total log likelihood is sufficiently small or after some fixed number of iterations, whichever comes first [5].

Derivation of finding optimum correspondence. The method of finding the locally optimum degree of correspondence (sharing of information) between corresponding positions starts by introducing the new parameter c for each set of corresponding positions. If we interpret the correspondence factor c as the probability $P(\text{identical})$ that the positions are completely synonymous, then we can interpret that every emission and transition probability $P(x)$ for each corresponding position in the model can now be replaced by a new probability $P'(x)$:

$$\begin{aligned} P'(x) &= P(\text{identical}) \cdot \overline{P(x)} \quad + \quad (1 - P(\text{identical})) \cdot P(x) \\ &= P(x) + c\left[\overline{P(x)} - P(x)\right] \end{aligned} \tag{13}$$

where $\overline{P(x)}$ is the average of the corresponding emission and transition probabilities. Now we can calculate new correspondence factors c' for each corresponding emission and transition probability in the set of corresponding positions:

$$\begin{aligned} c' &= \frac{P(\text{identical}) \cdot \overline{P(x)}}{P(\text{identical}) \cdot \overline{P(x)} \quad + \quad (1 - P(\text{identical})) \cdot P(x)} \\ &= \frac{c \cdot \overline{P(x)}}{P(x) + c\left[\overline{P(x)} - P(x)\right]} \\ &= \frac{c \cdot \overline{P(x)}}{P'(x)} \end{aligned} \tag{14}$$

Now we can calculate a new correspondence factor c'' for the set of corresponding parameters by averaging over the new c' for all the corresponding emission and transition probabilities:

$$c'' \;=\; \overline{c'} \quad \text{(over all } c' \text{ in the set of corresponding positions)} \tag{15}$$

Example. Assume that we have prior knowledge (or we guess that) the binding motif of a 10-bp binding site is singly palindromic: *1 2 3 4 5 5 4 3 2 1*. Then the positions that have been made corresponding are: *1* and *10*, *2* and *9*, *3* and *8*, *4* and *7*, *5* and *6*. (There are five sets of corresponding positions in this example.) First, each of the 10 distributions of the posterior mean emission probabilities for each of the 10 positions in the motif are now corresponding and sharing data with its partner position. Then the posterior mean transition distributions between positions are similarly made corresponding (for example *1-2* and *2-1*). Separate correspondence calculations are performed for each of the sets of corresponding positions. A correspondence factor of $c = 1$ would fully correspond (tie) the parameters between synonymous positions to the average over all corresponding parameters. (In this case, the parameter space would roughly be cut in half, and the training data per parameter would roughly double.) A correspondence factor of $c = 0$ would not change the initial distributions of emission and transition probabilities at a position at all, thus creating no correspondence between the positions. The correspondence factor c can be regarded as our *known* prior belief in the level of correspondence between synonymous positions in a palindromic, repeat, and/or reverse-complement binding-site motif. Alternatively, the correspondence factor c can be regarded as the *unknown* probability of correspondence between synonymous positions that needs to be determined. In the latter case, the Corresponded Baum-Welch algorithm will converge on the (locally) optimum c that maximizes the total log likelihood of the training set.

The Proof that the Log-odds Score $G^s(x)$ is proportional to $-\Delta G(x)$

It has been shown experimentally that in general, transcription factor proteins have a weak affinity for background DNA (any non-consensus sequence) and a strong affinity for consensus sites. Within the nucleus (or general cell in prokaryotes) the DNA concentration is high enough that an activated TF-protein is bound somewhere on the DNA essentially all the time (to a 1^{st} approximation) [14]. Therefore, the binding specificity (the ability of the TF protein to distinguish a functional site from background DNA) must be adequately high for proper regulation to occur [14]. The goal is to quantify the free energy of binding to a candidate site x through statistical mechanics, thermodynamics and Information Theory. We

start with the mass action kinetics of a TF-protein binding to a site:

$$\begin{aligned} p &= \text{transcription factor protein} \\ x &= \text{a candidate DNA binding site} \\ px &= \text{Bound Protein-Binding Site Complex} \\ k^+ &= \text{forward equilibrium binding constant} \\ k^- &= \text{backward equilibrium binding constant} \\ p + x &\underset{k^-}{\overset{k^+}{\rightleftharpoons}} px \\ K_{eq}^x &= \frac{k^+}{k^-} \quad = \text{equilibrium association constant for site } x \end{aligned} \tag{16}$$

We normalize K_{eq}^x in order to obtain the specific association constant K_s^x that quantifies specificity:

1. K_{eq}^{avg} = Average K_{eq} for all sites x
2. $K_s^x = \frac{K_{eq}^x}{K_{eq}^{avg}}, \quad (avg(K_s^x) = 1)$
3. Specificity of Valid Site: $K_s^{\text{valid site}} \approx 10^6$
4. Specificity of Background: $K_s^{\text{background}} < 1$

In experiments performed in E. Coli cells, with about 5×10^6bp of DNA, a single TF-protein and a single binding site with a specificity of 10^6 will be bound together only about 20% of the time. During the other 80% of the time, the protein will be transiently bound to other random places along the genome. However, with 20 copies of the protein the binding site will be occupied about 99% of the time [15].

The specific association constant K_s^x is related to the binding free energy $-\Delta G(x)$ by the following:

$$-\Delta G(x) = -k_\beta \cdot T \cdot \ln(K_s^x)$$

and

$$-K_s^x = \frac{k^+}{k^- \cdot K_{eq}^{avg}} = e^{-\Delta G(x)/k_\beta T} \tag{17}$$

Now lets estimate the probability that a putative binding site x is bound by a TF-protein in a well-mixed solution at equilibrium. Let $P(x \text{ bound})$ be the probability that the binding site x is bound by a

TF-protein. Then we have:

$$\begin{aligned} P(x \text{ bound}) &= \frac{\text{binding rate}}{\text{binding rate} + \text{unbinding rate}} \\ &= \frac{[p] \cdot k^+}{[p] \cdot k^+ + k^-} \\ &= \frac{[p] \cdot K_{eq}^{avg} \cdot e^{-\Delta G(x)/k_\beta T}}{[p] \cdot K_{eq}^{avg} \cdot e^{-\Delta G(x)/k_\beta T} + 1} \end{aligned} \tag{18}$$

which can be re-written into the form known as the Fermi-Dirac Equation, where $\mu = k_\beta T \ln(K_{eq}^{avg} \cdot [p])$ is the *chemical potential* dependent on the protein concentration $[p]$:

$$P(x \text{ bound}) = \frac{1}{e^{(\Delta G(x) - \mu)/k_\beta T} + 1} \qquad (Fermi - Dirac)$$

In the low concentration limit the Fermi-Dirac Equation for the probability $P(x \text{ bound})$ can be approximated by the Maxwell-Boltzmann Equation:

$$\begin{aligned} P(x \text{ bound}) &\approx \frac{1}{e^{(\Delta G(x) - \mu)/k_\beta T}} \quad \text{when } \Delta G(x) \gg \mu \\ &\approx e^{\mu/k_\beta T} \cdot e^{-\Delta G(x)/k_\beta T} \qquad (Maxwell - Boltzmann) \\ &\approx z e^{-\Delta G(x)/k_\beta T} \qquad (z = e^{\mu/k_\beta T} = \text{fugacity}) \end{aligned} \tag{19}$$

Now we are ready to analyze a sampling set S of known transcription factor binding sites for a given TF-protein. A version of this proof exists for weight matrices (PSSMs) in [4,16]. Here we provide a general proof that it is applicable for any fully probabilistic model that calculates $P_{background}(x)$ and $P_{setS}(x)$.

Assume that we attain the set S from a single experiment so that all the sites are collected under identical conditions. Assume that we have a very large number of DNA sequences of roughly similar length from a given genome mixed in solution with a certain concentration of TF-proteins. At equilibrium some of the DNA sequences with bound TF-protein are extracted (precipitated) and sequenced to create our sampling set S.

The probability of observing exactly the set S is given by:

$$\begin{aligned} P(\text{observing the set } S) &= \prod_{x \in S} (P_{exist}(x) \cdot P_{bound}(x) \cdot P_{extract}(x)) \cdot \prod_{x \notin S} (1 - P_{exist}(x) \cdot P_{bound}(x) \cdot P_{extract}(x)) \\ &\approx \prod_{x \in S} (P_{exist}(x) \cdot P_{bound}(x) \cdot P_{extract}(x)) \cdot e^{\sum_{x \notin S} (P_{exist}(x) \cdot P_{bound}(x) \cdot P_{extract}(x))} \end{aligned} \tag{20}$$

The likelihood function $\mathscr{L}$ for the P(observing the set S) can now be approximated:

$$\begin{aligned}
\mathscr{L} &= \ln\left[P(\text{observing the set } S)\right] \\
&\approx \ln\left[\prod_{x\in S}\left(P_{exist}(x)\cdot P_{bound}(x)\cdot P_{extract}(x)\right)\cdot e^{\sum_{x\notin S}\left(P_{exist}(x)\cdot P_{bound}(x)\cdot P_{extract}(x)\right)}\right] \\
&\approx \sum_{x\in S}\ln\left(P_{exist}(x)\cdot P_{bound}(x)\cdot P_{extract}(x)\right) - \sum_{x\notin S}\left(P_{exist}(x)\cdot P_{bound}(x)\cdot P_{extract}(x)\right) \qquad (21)
\end{aligned}$$

Now plug-in the Maxwell-Boltzmann approximation $ze^{-\Delta G(x)/k_\beta T}$ for $P(x \text{ bound})$, and for simplicity assume that $P_{extract}(x) = P_{extract}$ is identical for all x:

$$\begin{aligned}
\mathscr{L} &\approx \sum_{x\in S}\ln\left(P_{exist}(x)\cdot ze^{-\Delta G(x)/k_\beta T}\cdot P_{extract}\right) - \sum_{x\notin S}\left(P_{exist}(x)\cdot ze^{-\Delta G(x)/k_\beta T}\cdot P_{extract}\right) \\
&\approx N_s\cdot\ln(z\cdot P_{extract}) + \sum_{x\in S}\left(\ln(P_{exist}(x))\cdot\frac{-\Delta G(x)}{k_\beta T}\right) - z\cdot P_{extract}\sum_{x\notin S}\left(P_{exist}(x)\cdot e^{-\Delta G(x)/k_\beta T}\right) \qquad (22)
\end{aligned}$$

Where N_s is the size of the sampling set S. We are now ready to maximize the likelihood function $\mathscr{L}$ by taking the partial derivatives with respect to $zP_{extract}$ and $\Delta G_i(b)$ and setting them equal to 0. We have From the additivity assumption that for any putative site x:

$$-\Delta G(x) = \sum_{i=1}^{length(x)} -\Delta G_i(b)$$

where we define...

$$\begin{aligned}
-\Delta G_i(b) &= \text{the independent contribution of base } b \text{ observed at position } i \\
-\Delta G_i(x,b) &= -\Delta G_i(b)\cdot x(i,b) \\
x(i,b) &= 1 \text{ if } x_i = b, \text{ and } 0 \text{ if } x_i \neq b \qquad (23)
\end{aligned}$$

After taking the partial derivatives we have:

$$\begin{aligned}
\frac{\partial\mathscr{L}}{\partial(zP_{extract})} &= \frac{N_s}{z\cdot P_{extract}} - \sum_{x\notin S}\left(P_{exist}(x)\cdot e^{-\Delta G(x)/k_\beta T}\right) = 0 \\
\frac{\partial\mathscr{L}}{\partial(\Delta G_i(b))} &= \frac{\sum_{x\in S}x(i,b)}{k_\beta T} - \left[\frac{z\cdot P_{extract}}{k_\beta T}\cdot P_{exists}(i,b)\cdot e^{-\Delta G_i(b)/k_\beta T}\cdot\prod_{j\neq i}\sum_{b'}P_{exists}(j,b')\cdot e^{-\Delta G_j(b')/k_\beta T}\right] = 0 \qquad (24)
\end{aligned}$$

We can combine the results from the partial derivatives to obtain:

$$\frac{1}{N_s}\sum_{x\in S}x(i,b) = \frac{P_{exists}(i,b)\cdot e^{-\Delta G_i(b)/k_\beta T}\cdot\prod_{j\neq i}\sum_{b'}P_{exists}(j,b')\cdot e^{-\Delta G_j(b')/k_\beta T}}{\sum_{x\notin S}\left(P_{exist}(x)\cdot e^{-\Delta G(x)/k_\beta T}\right)} \qquad (25)$$

If we make the observation that:

$$\sum_{x \notin S} \left(P_{exist}(x) \cdot e^{-\Delta G(x)/k_\beta T} \right) = \sum_{b'} P_{exists}(i,b') \cdot e^{-\Delta G_i(b')/k_\beta T} \cdot \prod_{j \neq i} \sum_{b'} P_{exists}(j,b') \cdot e^{-\Delta G_j(b')/k_\beta T}$$

then we have that:

$$\begin{aligned}
\frac{1}{N_s} \sum_{x \in S} x(i,b) &= \frac{P_{exists}(i,b) \cdot e^{-\Delta G_i(b)/k_\beta T} \cdot \prod_{j \neq i} \sum_{b'} P_{exists}(j,b') \cdot e^{-\Delta G_j(b')/k_\beta T}}{\sum_{b'} P_{exists}(i,b') \cdot e^{-\Delta G_i(b')/k_\beta T} \cdot \prod_{j \neq i} \sum_{b'} P_{exists}(j,b') \cdot e^{-\Delta G_j(b')/k_\beta T}} \\
&= \frac{P_{exists}(i,b) \cdot e^{-\Delta G_i(b)/k_\beta T}}{\sum_{b'} P_{exists}(i,b') \cdot e^{-\Delta G_i(b')/k_\beta T}} \\
&= \frac{P_{exists}(i,b) \cdot e^{-\Delta G_i(b)/k_\beta T}}{C} \\
\frac{\frac{1}{N_s} \sum_{x \in S} x(i,b)}{P_{exists}(i,b)} \cdot C &= e^{-\Delta G_i(b)/k_\beta T} \\
\ln \left[\frac{\frac{1}{N_s} \sum_{x \in S} x(i,b)}{P_{exists}(i,b)} \right] + \ln C &= -\frac{\Delta G_i(b)}{k_\beta T} \\
\ln \left[\frac{\frac{1}{N_s} \sum_{x \in S} x(i,b)}{P_{exists}(i,b)} \right] &\approx\propto -\Delta G_i(b)
\end{aligned} \tag{26}$$

Now we make the following observations:

$$\begin{aligned}
\frac{1}{N_s} \sum_{x \in S} x(i,b) &= \text{probability of observing base b at position i in our set } S \\
&= P_{setS}(x_i(b)) \\
P_{exists}(i,b) &= P_{background}(i,b)
\end{aligned} \tag{27}$$

So now we have:

$$\begin{aligned}
G_i^s(b) &= \ln \left[\frac{P_{setS}(x_i(b))}{P_{background}(i,b)} \right] &&\approx\propto -\Delta G_i(b) \\
G^s(x) &= \ln \left[\frac{P_{setS}(x)}{P_{background}(x)} \right] &&\approx\propto -\Delta G(x) \qquad \text{(by the additivity assumption)}
\end{aligned}$$

□

References

1. Stormo GD: **DNA binding sites: representation and discovery**. *Bioinformatics* 2000, **16**:16–23, [http://bioinformatics.oxfordjournals.org/cgi/content/abstract/16/1/16].
2. Thijs G, Lescot M, Marchal K, Rombauts S, De Moor B, Rouze P, Moreau Y: **A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling**. *Bioinformatics* December 2001, **17**:1113–1122(10), [http://www.ingentaconnect.com/content/oup/cabios/2001/00000017/00000012/art01113].

3. Krogh A, Brown M, Mian IS, Sjölander K, Haussler D: **Hidden Markov models in computational biology. Applications to protein modeling.** *J Mol Biol* 1994, **235**(5):1501–1531, [http://dx.doi.org/10.1006/jmbi.1994.1104].
4. Djordjevic M, Sengupta AM, Shraiman BI: **A Biophysical Approach to Transcription Factor Binding Site Discovery**. *Genome Res.* 2003, **13**(11):2381–2390, [http://www.genome.org/cgi/content/abstract/13/11/2381].
5. Durbin R, Eddy S, Krogh A, Mitchison G: *Biological sequence analysis*. Cambridge University Press, 1^{st} edition 1998.
6. Eddy SR: **Profile hidden Markov models.** *Bioinformatics* 1998, **14**(9):755–763.
7. Thompson JD, Higgins DG, Gibson TJ: **Improved sensitivity of profile searches through the use of sequence weights and gap excision.** *Comput Appl Biosci* 1994, **10**:19–29.
8. Gerstein M, Sonnhammer EL, Chothia C: **Volume changes in protein evolution.** *J Mol Biol* 1994, **236**(4):1067–1078.
9. Altschul SF, Carroll RJ, Lipman DJ: **Weights for data related by a tree.** *J Mol Biol* 1989, **207**(4):647–653.
10. Sibbald PR, Argos P: **Weighting aligned protein or nucleic acid sequences to correct for unequal representation.** *J Mol Biol* 1990, **216**(4):813–818.
11. Eddy SR, Mitchison G, Durbin R: **Maximum discrimination hidden Markov models of sequence consensus.** *J Comput Biol* 1995, **2**:9–23.
12. Henikoff S, Henikoff JG: **Position-based sequence weights.** *J Mol Biol* 1994, **243**(4):574–578.
13. Krogh A, Mitchison G: **Maximum entropy weighting of aligned sequences of proteins or DNA.** *Proc Int Conf Intell Syst Mol Biol* 1995, **3**:215–221.
14. Stormo G, Fields D: **Specificity, free energy and information content in protein-DNA interactions**. *Trends in Biochemical Sciences* 1 March 1998, **23**:109–113(5), [http://www.ingentaconnect.com/content/els/09680004/1998/00000023/00000003/art01187].
15. Fields D, He Yy, Al-Uzri A, Stormo G: **Quantitative Specificity of the Mnt Repressor**. *Journal of Molecular Biology* August 1997, **271**:178–194(17), [http://www.ingentaconnect.com/content/ap/mb/1997/00000271/00000002/art01171].
16. Heumann JM, Lapedes AS, Stormo GD: **Neural networks for determining protein specificity and multiple alignment of binding sites.** *Proc Int Conf Intell Syst Mol Biol* 1994, **2**:188–194.