

Learning to Recognize Promoter Sequences in *E. coli* by Modeling Uncertainty in the Training Data

Steven W. Norton
Department of Computer Science
Hill Center for the Mathematical Sciences
Rutgers University, Busch Campus
New Brunswick, NJ 08903
norton@cs.rutgers.edu

Abstract

Automatic recognition of promoter sequences is an important open problem in molecular biology. Unfortunately, the usual machine learning version of this problem is critically flawed. In particular, the dataset available from the Irvine repository was drawn from a compilation of promoter sequences that were pre-processed to conform to the biologists' related notion of the *consensus sequence*, a first-order approximation with a number of shortcomings that are well-known in molecular biology. Although concept descriptions learned from the Irvine data may represent the consensus sequence, they do not represent promoters. More generally, imperfections in preprocessed data and statistical variations in the locations of biologically meaningful features within the raw data invalidate standard attribute-based approaches. I suggest a dataset, a concept-description language, and a model of uncertainty in the promoter data that are all biologically justified, then address the learning problem with incremental probabilistic evidence combination. This knowledge-based approach yields a more accurate and more credible solution than other more conventional machine learning systems.

Introduction

Understanding cellular biology at the level of gene expression would enable tremendous advances in pharmaceuticals, gene therapy, and more. Part of understanding gene expression involves understanding the complex regulatory signals present in DNA. A *promoter* is a signal that identifies specific segments of DNA that are transcribed into RNA, a necessary precursor to the production of protein (Watson *et al.* 1987). RNA polymerase is the enzyme that produces RNA on the DNA template (Losick & Chamberlin 1976). Before it produces RNA, the polymerase must recognize and bind to a promoter sequence. Characterizing the three-dimensional structure of the polymerase would help in understanding the promoter/polymerase interaction, but the size and complexity of the polymerase have made the approach impractical. Much of

Thanks to Haym Hirsh, Ringo Ling, Mick Noordewier, Mark Schwabacher, and Ke-Thia Yao for careful readings of drafts and endless technical discussions. This work was partially supported by NSF grant IRI-9209795.

the research effort has concentrated instead on understanding the structure of the promoter sequence itself.

Double-stranded DNA is made up of nucleotides, each containing a sugar, a phosphate group, and a base. DNA sequences are represented as strings of characters taken from a four character alphabet (A, G, C, or T) representing the bases that distinguish one nucleotide from another. Biologists believe that raw sequence information governs most polymerase/promoter interactions, and that the interactions are essentially localized to a handful of bases.

In 1975, Pribnow published a seminal paper describing a pattern of bases occurring imperfectly in a region just upstream of the transcriptional start sites of six of the promoters he examined (Pribnow 1975). He also suggested the existence of an important region 35 bases upstream. These regions have come to be known as the Pribnow box and the recognition region. Figure 1 is a highly stylized illustration of the DNA, the polymerase, and the various elements of the promoter.

Further research seemed to support the presence of these regions, *e.g.* (Siebenlist, Simpson, & Gilbert 1980), and their biological significance (Youderian, Bouvier, & Susskind 1982). The Pribnow box and the recognition region are thought to be the contact points between the polymerase and the promoter. (The actual contacts can be determined in the laboratory by *base conservation studies*.) Together the two recurrent patterns are now known as the *consensus sequence*. In the bacteria *E. coli*, the consensus sequence consists of two specific sequences of six

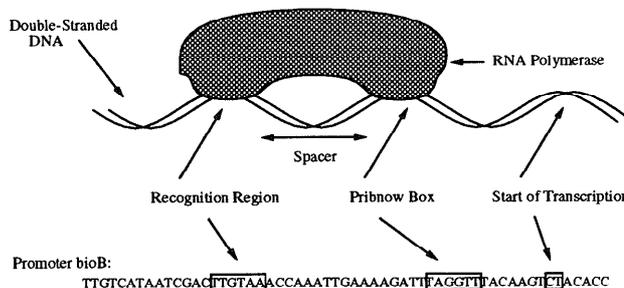


Figure 1. Abstract Promoter Structure

bases, TTGACA and TATAAT, separated by a gap of exactly 17 bases. No *E. coli* promoter has precisely this structure, and most have many differences. Still, the very idea of a canonical sequence capturing the essence of promoter structure and function was so influential that biologists produced compilations of promoter sequences aligned specifically to enhance correspondence to the consensus sequence (Hawley & McClure 1983; Harley & Reynolds 1987).

Machine learning experiments in recognizing promoter sequences typically rely on the promoter recognition database from the UCI Repository of Machine Learning Databases and Domain Theories (Cost & Salzberg 1993; Langley, Iba, & Thompson 1992; Towell & Shavlik 1992). Its 53 promoter sequences were selected from the compilation of Hawley and McClure (1983), and are left-aligned on the putative recognition region. (Sequence data from the Irvine dataset for the *bioB* promoter is shown at the bottom of Figure 1. The annotations are taken from the original compilation.) The 53 non-promoter sequences were taken from a longer sequence of DNA known not to exhibit promoter activity (Towell, Shavlik, & Nordewier 1990).

Hawley and McClure's compilation was based on smaller, earlier compilations and on the consensus sequence for *E. coli*. They aligned the promoters by hand to enhance correspondence with the consensus sequence. There is no mention of a computer program or an algorithm. More recent approaches perform the alignment automatically (Harley & Reynolds 1987), but are imperfect none the less. In fact, no published compilation is the result of optimal alignment to the consensus sequence, because the complexity of optimal multiple sequence alignment is exponential in the number of sequences to be aligned (Waterman 1989). Heuristic alignment fails when the predicted consensus alignment differs from regions of actual base conservation, the *biological foundation* of the consensus sequence. For example, in the compilation of (Harley & Reynolds 1987), the predicted alignments of fully 70 of the 263 promoters have notable deviations from one or more aspects of the laboratory data.

The point is that while the consensus sequence has proven to be a useful concept (Youderian, Bouvier, & Susskind 1982), it is still only a first-order approximation to an as yet unknown promoter concept. Consensus-sequence alignment may well associate consensus regions with consensus regions. But because the consensus sequence is an imperfect predictor of the contact regions, consensus-sequence alignment does not necessarily associate contact regions with contact regions. This means that while the data in the Irvine dataset could be used to learn about the consensus sequence, it should not be used to learn about promoter sequences.¹ A better choice for the alignment would

¹ Researchers working on other problems involving multiply-aligned data, such as to learning to recognize ribosomal binding sites, should beware of this pitfall too.

come from the DNA itself. For learning to recognize promoter sequences, the only natural alignment is the start of transcription, the site where the polymerase begins to produce the RNA product. It is not a theoretical construct, but a real biological entity present after every promoter, identifiable in the laboratory, and recorded in the original compilation.

It may seem that with such an alignment the promoter-recognition problem is ready to be solved, but even data with a biologically-justified alignment is insufficient. There are two specific reasons why this learning problem is harder than most: 1) There are often multiple transcriptional start sites, and during transcription the relevant one is chosen nondeterministically (Hawley & McClure 1983). 2) The length of the gaps between the start site, the Pribnow box, and the recognition region vary from promoter to promoter (von Hippel *et al.* 1984; Youderian, Bouvier, & Susskind 1982). What this means is that it is *impossible* to represent promoters by single contiguous sequences of DNA and simultaneously align them so that *each* attribute has a unique and consistent biological significance. For example, if the dataset is constructed so that the Pribnow boxes are in alignment, the recognition regions will necessarily be out of alignment. Since the attributes in the misaligned areas have no consistent biological significance from example to example, it is inappropriate to learn recognition rules directly from such data.

I have taken an alternative approach that does not depend on consensus alignment. Instead, it is based on laboratory research described in the open literature of molecular biology and on recent work in machine learning. By using biologically-based evidence, the promoter recognition problem in *E. coli* is addressed using incremental probabilistic evidence combination [Norton and Hirsh, 1992, 1993]. In particular, biological research on promoter structure and function justify the dataset, the concept-description language, and the characterization of the uncertainty present in the training data. Consequently the learned classifier is more credible and more accurate than those produced by CN2, C4.5, and the k-nearest-neighbor classifier.

Incremental Probabilistic Evidence Combination

In a noisy and uncertain domain, knowledge of the probabilistic processes affecting available data can help solve a learning problem. Incremental probabilistic evidence combination has been used successfully to learn conjunctions from noisy synthetic data (Norton & Hirsh 1992) and to learn DNF expressions from real and synthetic data (Norton & Hirsh 1993). The high-level idea behind the approach is to guess what the true data are, based on the observations and the probabilistic background knowledge, then return a concept description consistent with that data. Guesses supporting no concept descriptions are ruled out as inconsistent. Other guesses are ruled out as too unlikely,

leaving only the plausible guesses. Using the principle of maximum *a posteriori* probability, a concept description consistent with the best of the plausible guesses is returned as the result of learning.

Consider, for example, learning a conjunctive concept description from binary data subject to a uniform label noise process with a 10% noise rate. Suppose that three attributes can take on values 0, 1, or * (which matches either 0 or 1). What should be learned from these five observations: $\{(010, +) (011, +) (101, +) (110, +) (111, -)\}$? Since noise events are unlikely, the most probable single guess is that noise did not effect the true data, and that the observed data are the same as the true data. But since no term correctly classifies this data, one or more noise events *must* have occurred. Five other guesses suppose single noise events. Of those only one is consistent, namely that $(111, +)$ was changed to $(111, -)$. Furthermore, several consistent guesses involve two or more noise events, but each is less probable. A concept description consistent with the most probable consistent guess $\{(010, +) (011, +) (101, +) (110, +) (111, +)\}$ should be considered.

The remainder of this section describes in more detail the evidence-combination framework that implements the above reasoning process and will be instantiated with knowledge of promoter-specific probabilistic processes in order to build the final application program.

In the framework of incremental probabilistic evidence combination, knowledge takes the form of a probability distribution describing noise processes and/or other uncertainties working on the data. The uniform label-noise process of the preceding example changes class labels from + to - or from - to + with probability $\eta = 10\%$. It can be described by a probability distribution $P(o|s)$ where o is the label of the observation and s the uncorrupted class label. When η is low, noise events are unlikely, and the observed labels usually correspond to the uncorrupted labels. The four elements of the noise model for a uniform label noise process are $P(+|+) = P(-|-) = 1 - \eta$ and $P(+|-) = P(-|+) = \eta$.

If the true and correct training data (S) are known, the best concept description (H_i) is the one with maximum *a posteriori* probability $P(H_i|S)$. But of course the true and correct training data are unavailable, having been corrupted somehow. The best thing to do is select a concept description that seems most probable given the observations. Let O be the sequence of observations, S_j a particular series of guesses about the nature of the true but unavailable data, $VS(S_j)$ the set of concept descriptions strictly consistent with those guesses, and $P(O|S)$ the product of probabilities from the noise model. Norton and Hirsh (1992) show that the posterior probabilities are proportional to sums of noise probabilities. The first expression given below is for the posterior probability $P(H_i|O)$. It shows that every sequence of guesses consistent with a hypothesis

gives it a measure of probabilistic support. Expressions for the sets of consistent concept descriptions and the various posterior probabilities of the observation sequence are given as well. s_{jk} and o_k denote the k -th element of sequences S_j and O respectively.

$$P(H_i|O) \propto \sum_{H_i \in VS(S_j)} P(O|S_j)$$

$$VS(S_j) = \bigcap_k VS(s_{jk})$$

$$P(O|S_j) = \prod_k P(o_k|s_{jk})$$

Given these formulae, it is natural to view the computation of posterior probabilities as evidence combination. Evidence and current belief are represented by sets of tuples, each tuple consisting of a set of concept descriptions and a probability. Initial belief is represented by the singleton set $\{VS(\emptyset), 1.0\}$, where $VS(\emptyset)$ denotes the set of all concept descriptions. Each observation o suggests several evidence tuples, $\{VS(s_1), P(o|s_1)\}, \dots, \{VS(s_m), P(o|s_m)\}$, where m is the number of supposed examples (guesses) that could account for the observation. Each probability is essentially a weight associated with a guess, and hence with a corresponding set of consistent concept descriptions. The more probable it is that the guess is correct, the more probable it is that the correct hypothesis is in the corresponding set of consistent concept descriptions.

Evidence (tuples derived from the current example) and current belief (tuples summarizing sequences of guesses based on previous examples) are combined by taking cross products, multiplying pairwise probabilities and intersecting corresponding sets of concept descriptions. When the current belief is inconsistent with the new evidence, the intersection for the resulting tuple becomes empty. When this happens, the inconsistent sequence is discarded. My implementation of this approach controls its combinatorics in two more ways. It imposes a strict upper bound on the number of stored sequences and a limit on the difference between the probability of the most likely and the least likely sequences of guesses. In the end, when all the observations are processed, the most specific concept description from the most probable set of concept descriptions is returned as the result of learning.

Learning individual conjunctions this way is straightforward. The IPEC-DNF learner is an iterative application of the conjunction learner, with a modification to accommodate *representational noise*. IPEC-DNA, the program used in my promoter-recognition experiments, has the same iterative control structure. Refer to (Norton & Hirsh 1993) for more details.

Background Knowledge for Promoter Recognition

Application of the framework just described requires a concept-description language, a method for enumerat-

ing guesses about the true but unknown data, and a method for assigning a probability to each guess. Fortunately the biology literature is considerable, containing many helpful results. This section presents biological requirements for the concept-description language and characterizations of the uncertainties present in promoter sequences. It also shows how the background knowledge is used to construct evidence tuples for the IPEC-DNA learning program.

The molecule that transcribes DNA into RNA is RNA polymerase. Laboratory data suggest that it loosely binds to the DNA then moves along the molecule until it finds a promoter. Abortive initiation studies on the promoter/RNA polymerase complex interrupt the formation of RNA in the earliest stages. They indicate that the recognition region, Pribnow box, and the spacer are clearly important for characterizing promoter function (Borowiec & Gralla 1987). IPEC-DNA reasons about these entities by encoding them in its concept-description language. For instance, **STTGAC (17 18) TATAAT** matches any sequence starting with **C** or **G** (**S** is a shorthand from the biology literature), followed immediately by **TTGAC**, a gap of any 17 or 18 bases, and finally by **TATAAT**.

Recent studies in molecular biology argue that a single consensus-like sequence is inadequate. One argument suggests that a single sequence could not distinguish between promoters biologically optimized in different ways (McClure 1985). This criticism suggests that a disjunctive concept description language is necessary. IPEC-DNA learns disjunctions of the basic promoter descriptions described above, using the iterative control structure of IPEC-DNF (Norton & Hirsh 1993).

What makes the promoter-recognition problem especially difficult is uncertainty inherent in the training data. In particular, it is unclear where the actual recognition region and the Pribnow box lie within each promoter training datum. Uncertainties result from multiple transcriptional start sites, variable separation between the start site and the Pribnow box, and variable separation between the Pribnow box and the recognition region. Discrete probability distributions over these values allow IPEC-DNA to enumerate possible configurations of the contact regions and assign them probabilities. Here are four of the candidates the program considers for *bioB*:

TGTAAA (17 17)	AGGTTT	0.160
TTGTAA (17 17)	TAGGTT	0.118
TTGTAA (18 18)	AGGTTT	0.080
CTTGTA (18 18)	TAGGTT	0.059

Each candidate was assigned the probability on the right, by combining three *independent* models of uncertainty into a single model of domain uncertainty. Each of these models is justified by the molecular-biology literature, as explained below.

Mutational studies examine the effects of individual base insertions, deletions, or replacements within a promoter region. They show that the preferred spacer

length is 17 bases (Youderian, Bouvier, & Susskind 1982). This is consistent with consensus-sequence analysis that indicates spacers of 17 ± 1 base pairs represent 92% of promoters (Harley & Reynolds 1987). In helical DNA, each base contributes about 35 degrees of twist (Dickerson 1983). It follows that the length of the spacer influences the preferred orientation of the Pribnow box relative to the recognition region by altering helical twist. Other research (Borowiec & Gralla 1987) suggests that twisting of the DNA has a quadratic effect on the rate of closed complex formation, one of several steps in the initiation of transcription. Since rates are proportional to probabilities, the form of the probability distribution over the spacer length should be roughly quadratic. In the experiments reported here, IPEC-DNA uses a spacer-length distribution that assigns a 50% probability to the 17 base spacer, and 25% probabilities to the 16 base and 18 base spacers.

To expose the template strand once the polymerase has bound to the promoter, 17 ± 1 bases are unwound from the middle of the Pribnow box to six or eight bases past the start of transcription (Gamper & Hearst 1982). Allowing three bases in the Pribnow box leaves between five and nine bases between the start of transcription and the downstream end of the Pribnow box. As to the probability distribution over this gap, we only know that 64% of uniquely identified transcriptional start sites are six or seven bases downstream of the Pribnow box (von Hippel *et al.* 1984). Orientation is likely to be key again, suggesting a quadratic form for this distribution. IPEC-DNA models this uncertainty by assigning 32% probability to gaps of six or seven bases, 15% probability for gaps of five or eight bases, and 6% probability for a gap of only four bases.

All that remains is the uncertainty concerning multiple start sites. The various compilations indicate each start site, but do not indicate the preferred one (if any). In the absence of stronger information, transcriptional start site uncertainty was modeled as a uniform probability distribution over the candidate sites. This policy is adopted in IPEC-DNA.

A promoter with three adjacent start sites generates 24 evidence tuples with probabilities between 0.3% and 11.75%. Each possible start site is considered in turn. Given a start site, each possible value from four to eight bases is used to locate the putative Pribnow box. Then 16, 17, and 18 base spacers are used to locate the recognition region. If a particular combination of these values is indeed correct, the others are necessarily incorrect. Evidence tuples consist of a probability and a set of concept descriptions. A given tuple generates that set by treating exactly one of these combinations as a positive example while treating the remainder as negative examples. The corresponding concept descriptions are consistent with at least that one positive example and inconsistent with at least the other negative examples.

Non-promoters are handled differently, because they contain no special regions. Knowing that the polymerase does not bind *anywhere* in these fragments

(Towell, Shavlik, & Noordewier 1990), I generated 50 negative examples from each non-promoter at random, and combined them into a single evidence tuple. Specifically, spacers were chosen at random according to the distribution given previously. A segment 12 bases wider (to accommodate the ersatz Pribnow box and recognition region) was randomly selected from the original non-promoter and used to construct a negative example with the same form as the four examples shown earlier. Since none of these 50 components binds to RNA polymerase, each evidence tuple so constructed has unit probability.

Experimental Results

Learning to recognize promoters required that I construct a dataset with a biologically-justified alignment, left or right aligned at the start of transcription. By examining the Irvine dataset and identifying corresponding elements in the original compilation (Hawley & McClure 1983) I decided that aligning the sequence by the rightmost transcriptional start site most preserved the relative locations of the recognition regions and the Pribnow boxes.² Trimming just enough bases from the left and right of each promoter so that they are a uniform length leaves 51 bases. I trimmed non-promoters to the same length by removing bases from the left side. Six promoters were eliminated because no transcriptional start was given, leaving a total of 100 examples. I will refer to this dataset as the biologically-aligned dataset.

I began by performing leave-one-out cross-validated trials on the biologically aligned dataset using IPEC-DNF (Norton & Hirsh 1993), CN2 (Clark & Niblett 1989), C4.5 (Quinlan 1993), and a k-nearest-neighbor classifier. Each of these conventional learners uses the 51 individual bases as features, even though this approach is invalid as indicated in the Introduction. IPEC-DNF computes DNF expressions. CN2 produces rules or an ordered decision list. The k-nearest-neighbor classifier was run with $K = 1$, $K = 3$, and $K = 5$. Increasing K increased the false-positive rate and decreased the false-negative rate without changing the overall error rate, so $K = 1$ is reported here. C4.5 learns decision trees. Tree pruning was found to be helpful and is used here. I performed the same experiment using the IPEC-DNA evidence-combination program described in previous sections. The lowest error rate, 19%, is attributed to IPEC-DNA. The results are summarized in Table 1 under the "CV Rate" heading.

The IPEC-DNA solution is the four term DNF given below. The nucleotide codes (*e.g.* D stands for A or G or T.) are standard (Cornish-Bowden 1985). The spacer (17 17) is exactly 17 bases. (16 18) matches 16, 17, or 18 base gaps. (17 18) matches 17 or 18 base gaps.

²This choice was meant to be most favorable to the conventional learners. Performing the same series of experiments using left-aligned data gives substantially similar results.

Table 1. Error Rates Comparison

Learning System	CV Rate	FP Rate
IPEC-DNF	43%	11%
CN2 (rules)	32%	21%
CN2 (ordered)	31%	36%
1-Nearest-Neighbor	30%	33%
C4.5 (pruned)	23%	34%
IPEC-DNA	19%	1.5%

NDDNHN (17 17) TANHDW
 or NWDNNN (17 17) VNWAUV
 or KHBVMD (16 18) HMTRNT
 or KYKHHN (17 18) RTDVVV

On-line genetic databases are growing rapidly. GenBank currently contains about 130 million nucleotide bases from all sources (Benson, Lipman, & Ostell 1993). *E. coli* itself contains about five million nucleotides. Much of this data has been automatically sequenced, and its biological significance is unknown. Learned classifiers could shed some light on this data, and would be used by molecular biologists to suggest laboratory experiments if they were accurate enough. The key factor is the false-positive rate. Because regulatory signals such as promoters occur so infrequently, false positives translate directly into wasted laboratory time. The "FP Rate" column in Table 1 shows the false-positive rates for these classifiers. The scores were computed by counting the number of locations that they recognize as promoters in a 1500 base DNA sequence known not to bind to RNA Polymerase (Towell, Shavlik, & Noordewier 1990). The IPEC-DNA classifier is the clear winner in this respect, with a 1.5% false-positive rate.

To characterize the contributions of the different pieces of background knowledge, I performed a series of experiments in which uniform probability distributions were substituted for the biologically-justified distributions. Replacing both the spacer distribution and the distribution of the separation between the start of transcription and the Pribnow box with uniform distributions should indicate the contribution of the concept-description language. Replacing either distribution alone should help quantify the contribution of the other. In each experiment the resulting error rates were greater than 50%, indicating that each piece of background knowledge is necessary for the solution.

I performed the same series of experiments using the more up-to-date and extensive promoter database given in (Lisser & Margalit 1993). I aligned the data on the rightmost transcriptional start site, and trimmed each instance to 65 bases (-50 to +15). Four promoters had to be removed because the compilation listed too few upstream bases (*argCBH-P2*, *speC-P1*, *speC-P2*, and *speC-P3*). The remainder of my dataset consists of an equal number of non-promoters (296) generated at random from the 1500-base non-binding sequence

Table 2. Error Rates Comparison, Large Dataset

Learning System	CV Rate	FP Rate
C4.5 (unpruned)	37%	33%
IPEC-DNF	34%	26%
CN2 (ordered)	32%	29%
1-Nearest-Neighbor	29%	32%
CN2 (rules)	27%	50%
IPEC-DNA	2.5%	0.2%

mentioned earlier. The dramatic results of learning are presented in Table 2. The 12-term DNF learned by IPEC-DNA is far superior to the other classifiers, at a statistical significance level better than 10^{-5} . 10-fold cross-validated error rates appear under the heading "CV Rate". Once again, the false-positive rate was estimated by applying the learned classifiers to each position of the non-binding DNA strand, and appears under the heading "FP Rate".

Related Work

To establish a basis of comparison for IPEC-DNA, I experimented with C4.5, CN2, IPEC-DNF, and the k-nearest-neighbor classifier. Decision tree, decision list, nearest-neighbor, and DNF learners are among the most popular of the general-purpose machine-learning methods available today. They are efficient and widely applicable, but knowledge poor. Aside from the error-rate comparisons already given, when applied to the biologically-aligned promoter data the lack of knowledge manifests itself in unfocussed concepts that depend importantly on bases that do not play a role in promoter function. The branches of the C4.5 decision tree and the CN2 rules are insufficiently specific to describe promoters or particular promoter behaviors (O'Neill 1989). The false-positive rates given in Tables 1 and 2 bear this out. On the other hand, the multitude of bases referenced by IPEC-DNF's classifier, chiefly outside the contact regions, hurt more than they help. Classifiers that reference so many specific bases outside the contact regions lose credibility. In contrast, IPEC-DNA's classifier only references bases in the putative contact regions.

One way to address the problem of uncertainty in training data is to invent a set of higher-level features that abstract the uncertainty away. This is precisely what is done in (Hirsh & Noordewier 1994). By discarding the raw data in favor of the higher-level features, they avoid the criticisms set out in the Introduction. These are general features taken from the molecular-biology literature that they feel will be useful for a variety of related problems. A key difference between that approach and the one presented here is the level of detail of the background knowledge. Here the motivation is to provide a knowledge-based solution to a single learning problem rather than to a family of learning problems. Hirsh and Noordewier have 'coarsened' the background knowledge to achieve

a measure of generality across sequence learning tasks. For instance, there are 12 features describing sharp bends in the DNA. These are used singly in (Hirsh & Noordewier 1994), even though it is "the periodic occurrences of hexamers with identical, large twist angles on the left-hand side of the axis of symmetry" that seemed "strikingly non-random" to the original researchers (Nussinov & Lennon 1984). For a general sequence learner, abstracting from *periodic occurrences* of these features to *one or more occurrences* of these features is fine, provided over-generalization is not a problem. But for IPEC-DNA, a promoter-specific learner, augmenting the feature set would only be appropriate after tightening up the biological significance of the new features. They report an 8.7% error rate for C4.5 rules and a 10.2% error rate for the neural network when the raw data are discarded. These values can be compared to IPEC-DNA's 2.5% error rate because their dataset is very similar to the large one described here.

Towell *et al* (1990) also take a knowledge-based approach to the promoter problem. In particular, a set of rules describing consensus-like sequences and certain conformational properties is used to construct a back-propagation neural network. But as discussed in the Introduction, the original alignment changes the nature of the problem, so that the network recognizes the consensus sequence rather than the promoter sequence. Though the background knowledge could be applied to the biologically-aligned data, additional uncertainty due to the variable separation between start of transcription and the Pribnow box, and between the Pribnow box and the recognition region would cause the network to emphasize the wrong bases. If a more complete dataset was used (Lisser & Margalit 1993), one with increased variability in the separation between the putative contact regions, limitations of the background knowledge might be highlighted that were not apparent in the original study.

Closing Remarks

Learning systems depend critically on the assumption that each attribute has the same meaning, across multiple examples, an assumption not satisfied by consensus-aligned promoter data. In particular this alignment does not always align the biologically-active sites where promoter and polymerase bind. At best the consensus-sequence alignment introduces an inappropriate bias and changes the problem from learning to recognize promoter sequences to learning to recognize the consensus sequence. More generally, alignment is a potential problem for any learner using raw sequence data, whether it is DNA, RNA, or protein.

Machine learning research has produced a number of excellent general-purpose techniques that often perform well, but are necessarily knowledge-poor. IPEC-DNA outperforms these conventional learners because it is able to exploit biologically-justified background knowledge that others cannot. This work supports a claim that knowledge-based learners with problem-

specific background knowledge can be expected to produce more accurate, credible concept descriptions.

Using the biology literature I justified a dataset, a concept-description language, and a model of uncertainty in promoter data. The knowledge-based approach using incremental probabilistic evidence combination yields a more accurate solution than more conventional machine learning systems. Equally important, the knowledge-based solution is more credible since it only references bases biologically implicated in promoter structure and function.

References

- Benson, D.; Lipman, D. J.; and Ostell, J. 1993. GenBank. *Nucleic Acids Research* 21(13):2963-2965.
- Borowicz, J. A., and Gralla, J. D. 1987. All three elements of the *lac p^s* promoter mediate its transcriptional response to DNA supercoiling. *Journal of Molecular Biology* 195:89-97.
- Clark, P., and Niblett, T. 1989. The CN2 induction algorithm. *Machine Learning* 3:261-284.
- Cornish-Bowden, A. 1985. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Research* 13(9):3021-3030.
- Cost, S., and Salzberg, S. 1993. A weighted nearest neighbor algorithm for learning with symbolic features. *Machine Learning* 10(1):57-78.
- Dickerson, R. E. 1983. Base sequence and helix structure variation in B and A DNA. *Journal of Molecular Biology* 166:419-441.
- Gamper, H. B., and Hearst, J. E. 1982. A topological model for transcription based on unwinding angle analysis of *E. coli* RNA polymerase binary, initiation and ternary complexes. *Cell* 29:81-90.
- Harley, C. B., and Reynolds, R. P. 1987. Analysis of *E. coli* promoter sequences. *Nucleic Acids Research* 15(5):2343-2361.
- Hawley, D. K., and McClure, W. R. 1983. Compilation and analysis of *Escherichia coli* promoter DNA sequences. *Nucleic Acids Research* 11(8):2237-2255.
- Hirsh, H., and Noordewier, M. 1994. Using background knowledge to improve inductive learning of DNA sequences. In *The Tenth Conference on Artificial Intelligence for Applications*.
- Langley, P.; Iba, W.; and Thompson, K. 1992. An analysis of Bayesian classifiers. In *AAAI92: Proceedings of the Tenth National Conference on Artificial Intelligence*, 223-228. AAAI Press.
- Lisser, S., and Margalit, H. 1993. Compilation of *E. coli* mRNA promoter sequences. *Nucleic Acids Research* 21(7):1507-1516.
- Losick, R., and Chamberlin, M. J., eds. 1976. *RNA Polymerase*. Cold Spring Harbor Laboratory.
- McClure, W. R. 1985. Mechanism and control of transcription initiation in prokaryotes. *Annual Review of Biochemistry* 54:171-204.
- Norton, S. W., and Hirsh, H. 1992. Classifier learning from noisy data as probabilistic evidence combination. In *AAAI92: Proceedings of the Tenth National Conference on Artificial Intelligence*, 141-146. AAAI Press / MIT Press.
- Norton, S. W., and Hirsh, H. 1993. Learning DNF via probabilistic evidence combination. In *Proceedings of the International Conference on Machine Learning*, 220-227. Morgan Kaufmann Publishers.
- Nussinov, R., and Lennon, G. G. 1984. Periodic structurally similar oligomers are found on one side of the axes of symmetry in the *lac*, *trp*, and *gal* operators. *Journal of Biomolecular Structure and Dynamics* 2(2):387-395.
- O'Neill, M. C. 1989. *Escherichia coli* promoters: I. Consensus as it relates to spacing class, specificity, repeat substructure, and three-dimensional organization. *Journal of Biological Chemistry* 264:5522-5530.
- Pribnow, D. 1975. Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter. *Proc. Nat. Acad. Sci.* 72(3):784-788.
- Quinlan, J. R. 1993. *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann Publishers.
- Siebenlist, U.; Simpson, R. B.; and Gilbert, W. 1980. *E. coli* RNA polymerase interacts homologously with two different promoters. *Cell* 20:269-281.
- Towell, G. G., and Shavlik, J. W. 1992. Using symbolic learning to improve knowledge-based neural networks. In *AAAI92: Proceedings of the Tenth National Conference on Artificial Intelligence*, 177-182. AAAI Press.
- Towell, G. G.; Shavlik, J. W.; and Noordewier, M. O. 1990. Refinement of approximate domain theories by knowledge-based neural networks. In *AAAI90: Proceedings of the Eighth National Conference on Artificial Intelligence*, 861-866. Morgan Kaufmann Publishers.
- von Hippel, P. H.; Bear, D. G.; Morgan, W. D.; and McSwiggen, J. A. 1984. Protein-nucleic acid interactions in transcription: A molecular analysis. *Annual Review of Biochemistry* 53:389-446.
- Waterman, M. S. 1989. *Mathematical Methods for DNA Sequences*. CRC Press, Inc.
- Watson, J. D.; Hopkins, N. H.; Roberts, J. W.; Steitz, J. A.; and Weiner, A. M. 1987. *Molecular Biology of the Gene*. Benjamin/Cummings Publishing Company, Inc.
- Youderian, P.; Bouvier, S.; and Susskind, M. M. 1982. Sequence determinants of promoter activity. *Cell* 30:843-853.