Prediction of Genomic Functional Elements

Steven J.M. Jones

Genome Sciences Centre, British Columbia Cancer Research Center, Vancouver, British Columbia, V5Z 1L3, Canada; email: sjones@bogsc.ca

Key Words
gene prediction, comparative genomics, genome annotation, gene regulation, intron splice sites, microRNA

Abstract
As the number of sequenced genomes increases, the ability to deduce genome function becomes increasingly salient. For many genome sequences, the only annotation that will be available for the foreseeable future will be based on computational predictions and comparisons with functional elements in related species. Here we discuss computational approaches for automated genome-wide annotation of functional elements in mammalian genomes. These include methods for ab initio and comparative gene-structure predictions. Gene features such as intron splice sites, 3′ untranslated regions, promoters, and cis-regulatory elements are discussed, as is a novel method for predicting DNaseI hypersensitive sites. Recent methodologies for predicting noncoding RNA genes, including microRNA genes and their targets, are also reviewed.
INTRODUCTION

Computational methods for the prediction of genomic features rarely supplant molecular biology–based experiments, but instead provide a powerful starting point for further studies. However, the rate at which genomes, including mammalian genomes, are being sequenced currently far outpaces the capacity for any systematic biochemical analysis. For many genomes the only annotation available will be derived from ab initio algorithmic predictions or through computational methods that can transfer knowledge from related organisms. Concentrating on methods for metazoan and particularly mammalian genomes, this review provides an overview of the approaches possible for functional element prediction on a genome-wide scale, including the more recent developments and an assessment of the state-of-the-art methods in their ability to provide reliable results.

GENE PREDICTION

From the beginning of genomic sequencing an established goal has been the accurate ab initio prediction of genes, that is, the identification of gene structure using only information inherent in the DNA sequence. Although the task of identifying the open reading frames of genes is somewhat simplified in prokaryotes and lower eukaryotes such as *Saccharomyces cerevisiae*, the problem remains largely unsolved in metazoan genomes where coding elements are substantially punctuated with introns.

Determining methods for ab initio gene detection has been an active area of research and a number of gene prediction algorithms have been developed. Despite this effort in the creation of computational tools, the gold standard in gene structure determination is still through biochemical confirmation and relatively large efforts have been set up to expedite the experimental determination of gene structures in numerous vertebrates including mouse and human (28, 50, 70, 108, 150). Nonetheless, computational modeling of gene structure represents a key way in which we can understand the underlying biological process and identify the salient genomic signals and features that are employed. Such ability will be important in the next phase of biology, where synthetic biology approaches will be employed to design and produce novel functional gene constructs. The object of this section is not to provide an exhaustive discussion of the original gene prediction algorithms, as these have been addressed in reviews elsewhere (21, 97, 168), but to introduce the underlying methods utilized and recent directions within the field.

The most heavily used gene prediction approaches have been Genefinder (P. Green, unpublished), FGENESH (135), GeneID (111) and Grail I and II (165), GeneMark.hmm (94), Genie (124), MZEF (167), and Morgan (136), among others. Typically, gene prediction employs two phases. First, a signal sensor detects gene features such as ribosome attachment sites, intron donor and acceptor splice sites, initiation codons, codon biases, and open reading frames. Second, the programs will attempt to optimally combine these features to form the final gene prediction. For example, Genefinder uses log-likelihood scores from sequence conservation matrices to detect gene features and a dynamic programming approach to join these to create the gene prediction. Geneparser, Genie, and Grail implement a neural network approach trained on known examples to identify gene features. Although both Grail and Geneparser use dynamic programming, Genie was the first to implement a generalized hidden Markov model to generate gene structures, an approach now heavily exploited in the gene prediction field for both feature detection and gene structure generation. In 2004, Eddy (40) presented a review of the utility of hidden Markov models in DNA sequence analysis.

In 1997 GENSCAN was published (26) and showed a significant improvement in gene prediction over the existing methods, using...
a generalized hidden Markov model to generate the gene structures. GENSCAN incorporated a number of methods for feature detection, including a first-order weight-array model (WAM) (169) for predicting splice acceptors, a second-order WAM to identify the splice branch point, a novel maximal dependence decomposition (MDD) method for the splice donor site, and a fifth-order hidden Markov model for other features. GENSCAN also incorporated exon-specific length distributions, allowing exon scores to correlate with a biologically observed distribution of exon lengths (55), as opposed to the geometric distribution of scores that decays with exon size usually generated by a hidden Markov model approach. GENSCAN also was able to simultaneously predict multiple genes on both DNA strands, thus avoiding a common deficiency in predicting overlapping genes. Since the publication of GENSCAN, it remains arguable whether the overall accuracy of pure ab initio methods has been substantially improved upon.

Pragmatically, the utility of an ab initio algorithm is questionable if the goal is to determine the best possible gene prediction at a particular locus and experimental data exists to help define the gene structure. Genewise (16) and Procrustes (49) were both developed to predict gene structures through the alignment of known protein sequences to genomic sequence. GenomeScan (166) is a retrofit of GENSCAN that can also incorporate sequence similarity based on BLASTX or BLASTP search results (3). The utility of such approaches naturally depends on how closely related the proteins used are to the target genome. For many genomes, especially mammalian genomes where a number have now been sequenced and where several confirmed gene structures exist (28, 50), this represents the easiest and most accurate approach for gene prediction. As the position of intron-exon boundaries is generally conserved, this represents further information available to aid gene prediction. Intron-exon boundary information has been implemented by the Projector algorithm (100), which rewards correlation of exon boundaries between the known gene and the prediction. This is more effective than the Genewise algorithm when the similarity of the protein to the prediction is less than 80%. At higher degrees of similarity the protein alignment alone is sufficient for accurate gene prediction.

Interestingly, for whole-genome annotation such as that employed within the Ensembl database, ab initio gene prediction is not considered necessary and protein-coding gene models are generated mainly through the alignment of existing protein and cDNA sequences (61). The assumption in this approach is that for each genome studied sufficient information exists for the species or closely related species to essentially define the entire proteome, and using prediction algorithms would simply not be worth the inaccuracies they would introduce. In this case, the Exonerate alignment algorithm (146) is used to conduct this alignment task in a rapid manner using a dynamic programming implementation incorporating heuristics, thus enabling the approximation of an exhaustive alignment. The speed up achieved by Exonerate allows such an alignment strategy to be implemented on a genome-wide scale with relatively insignificant computational demands.

ASSessment of gene prediction algorithms

With an abundance of gene prediction algorithms, it has become important to accurately assess their relative performance. This is not as simple as one might think. Most gene prediction algorithms were trained or developed using a set of known gene structures, the identities of which may or may not be obvious to the appraiser. Including such genes in the appraisal process will almost certainly incur an overestimate in the assessment of the ability of the algorithm to correctly predict genes. Likewise, the sets of training genes used will typically be only partially overlapping and depend somewhat on when the algorithm was...
developed. Remaking feature tables and test sets, and then retraining algorithms is a time-consuming endeavor and is typically done infrequently, if at all, by gene prediction software authors. Therefore, another area of ambiguity is how these algorithms would perform given a more updated training set or how they would compare given the same set of training genes. For instance, new methods proposed to have greater performance, either for gene prediction or for a specific gene feature, may merely benefit from an improved and more accurate training set as opposed to superior algorithmic approaches. To address some of these issues, the SNAP gene finder (76), which like GENSCAN uses a generalized hidden Markov model, has been designed for easy configuration and training for different organisms. The GAZE gene prediction framework (60) uses a dynamic programming approach but allows the easy incorporation of new signal sensors so that their contribution within a gene prediction context can easily be gauged and utilized. Gene-finding algorithms also tend to weight differently the specificity and sensitivity of the output. Is it better to provide high-quality gene predictions, while missing many genes, or to provide higher gene coverage at the cost of a greater number of false positive predictions? Due to authors’ differing emphasis on sensitivity and specificity, direct comparisons of software performance are not always easy.

A number of appraisals have been carried out on the ability of algorithms to predict genes in mammalian sequence (27, 129, 147). These indicate that ab initio methods can correctly predict more than 90% of the coding bases and 70–75% of exons. Because eukaryotic genes are formed from multiple exons [for example, human genes have, on average, five exons per gene with a standard deviation of 4.3 (37)], it is unsurprising that with the cumulative effect of errors, less than 50% of complete gene predictions correspond precisely to known structures. Note that these assessments only refer to the coding sections of genes, with untranslated regions (UTRs) typically being ignored in such appraisals. However, rigorous assessment of the behavior of gene predictors led to the observation that they often had similar predictive values yet identified only partially overlapping sets of gene features. This has led to the concept of combining the output of gene predictors for improved performance (27, 101, 130, 139). For example, Rogic et al. (130) derived an approach to combine the outputs of GENSCAN (26) and HMMgene (79) that improved coding nucleotide and exon detection by 21% and 32.5%, respectively.

Another confounding issue in gene prediction is that it is not necessarily obvious when an algorithm is making a mistake. For example, novel splice variants are constantly being discovered (120, 152). Therefore, it is not clear how many apparently incorrect predictions are simply reporting valid yet currently unknown alternative gene structures. Although the goal is to replicate the ability of the cell to accurately recognize gene structures and construct functional transcripts, the true ability of a cell to achieve this has yet to be elucidated. While our molecular biology dogma may lead us to believe that the cell can perform this with great accuracy, this may be far from true. The abundance of alternate transcripts may simply reflect a certain plasticity in splicing, which although possibly relevant in an evolutionary sense, may not always produce functional gene products. We may be interpreting such events as providing functional variation adding to the complexity of the proteome, but many of these may simply be the result of erroneous splicing. Also, other mechanisms may cause the accuracy of the splicing machinery to be overestimated. Nonsense mediated decay and premature stop codon recognition may serve to remove the more erroneously spliced transcripts (98). A possible challenge to gene prediction algorithms and their authors is that it remains possible that for a number of genes the transcript most likely to be created is one that is subsequently removed by the cell using such machinery.
USING COMPARATIVE GENOMICS FOR GENE PREDICTION

Recently, the focus of development in gene prediction has been to take advantage of the wealth of genomic sequence data now being generated from other organisms. Twinscan (77), SGP2 (110), SLAM (2), and Doublescan (99) were developed as comparative gene prediction algorithms exploiting sequence conservation between mouse and human to identify gene structures. Such approaches can be considered a variant of ab initio gene prediction as the genes are predicted using only two related sequences and in the absence of experimental information such as cDNA and homologous protein sequences. These approaches are highly dependent on the ability to align DNA sequences on a genomic scale and a number of applications have been designed to approach this challenge (17, 19, 20, 23). Because multiple mammalian genomes are now available this approach has also graduated from pairwise alignments to the simultaneous use of multiple genomes for gene prediction. Exoniphy was the first to utilize multiple alignments to predict exons in this way (142), using a phylogenetic hidden Markov model or “phylo-HMM.” This strategy, guided by the phylogenetic tree, models molecular evolution to predict the state of each base, characterizing DNA sequences as conserved coding, conserved noncoding, or noncoding. Although implemented using a geometric length distribution for predicting exons, such approaches will become increasingly relevant as more mammalian genomes are sampled (42, 96). N-SCAN is another comparative gene predictor using multiple alignments (53). It uses a phylogenetic Bayesian network as opposed to a phylo-HMM, which has the advantage of allowing the problem to be much more feasible in terms of the time required for computation. However, all approaches using comparative information for protein-coding gene prediction are potentially confounded by conserved noncoding sequences, which continue to occupy a larger proportion of the genome than previously anticipated (2, 34, 46). For highly related genomes, annotation can become trivial. Effective gene annotation in the chimpanzee genome (32) was simply carried out by aligning the genome to the human sequence using BLASTZ (138) and BLAT (72) and transferring the human gene coordinates to the chimpanzee sequence.

IDENTIFICATION OF SEQUENCES UNDER POSITIVE SELECTION

It is tempting to speculate that all functional features in the genome could be identified through evolutionary conservation. The identification of bases under positive selection pressure remains one of the most unbiased approaches to infer function, and is possible as a number of mammalian genomes have already been sequenced and numerous others are in progress or are being planned (96). The ultimate goal would be to have the ability to determine whether any given base in the genome is subjected to positive selection pressure. Both Eddy (41) and Cooper et al. (33) provided an informative discussion and modeled the number of genomes required to achieve such predictive power. From these models, if positive selection pressure is defined as a substitution rate that is 20% of that of a base under no selective pressure, it is estimated that to detect 50% of such bases with a false positive rate of better than 0.006 (i.e., $e^{-5}$), approximately 25 genomes with similar phylogenetic branch distances as between mouse and human (approximately 0.31 substitutions per neutral site) would provide the cumulative branch distance required of 7.75 neutral substitutions per base pair. To detect 99% of such sites with a false positive rate of better than 1 in $10^{-4}$ would require around 120 similarly distanced genomes. A more stringent approach modeled by Cooper et al. indicated that even if a base were essentially
invariant among mammals, approximately 17 genomes would be required to reach a false positive rate of 0.006. However, both of these models assume the ready availability of equidistant, independently evolved mammalian lineages. It is unlikely that 25 mammalian genomes that provide the cumulative branch distance required could be selected, let alone 120 genomes. Margulies et al. (96), using empirical distances and normalizing to a more generous estimate for mouse/human divergence, determined that 22 well chosen mammals (excluding monotremes and marsupials) will only provide a cumulative neutral substitution rate of 3.805 per site. Because further genome sampling results in rapidly diminishing returns in terms of adding to the cumulative branch length, significantly more than 25 genomes would likely be needed to provide the required 7.75 neutral substitutions per site modeled above. It is not clear that sampling all of the approximately 5000 extant mammalian species could give the cumulative branch length of 37.2 neutral substitutions per site, equivalent to 120 independent mouse/human distances.

Fortunately, single base pairs are unlikely to provide function independently, and so Eddy suggests that a better model would be to consider features 8 and 50 nucleotides in length, modeling the sizes of regulatory elements and small coding exons. In such cases, 3 to 16 mouse-human distance equivalent genomes would be required to detect an 8-base pair feature with false positive rates of between 0.006 and 0.0001, whereas a 50-base pair feature would only require 1 to 3 similarly distanced genomes for detection with the same range of false positive rates.

One caveat is that even with an abundance of sequenced genomes, a comparative analysis still requires that each of the sequences be correctly aligned. Typically, alignments become more error-prone for genomes with less than 70% pairwise identity (41, 118), leading to a further underestimate of the actual number of genomes required. One response to this difficulty in generating accurate sequence alignments is to use “phylogenetic shadowing,” whereby a number of very closely related species are sampled, leading to a multiple sequence alignment that is far more likely to be correct. In Boffelli et al.’s (18) analysis, primate sequences were used to study conservation around several human genes. However, the use of such closely related species comes at a predictable statistical cost, with sequence from up to 17 species required to begin to identify exons and putative regulatory features.

IDENTIFICATION OF INDIVIDUAL GENE FEATURES

Intron Splicing Signals

A number of studies have been directed at determining more effective computational signal detectors for specific gene features. The most studied have been the intron donor and acceptor splice sites. Clearly, more accurate models of splice sites would be the largest determinant in more effective gene prediction, and would represent key advances. The example of the dystrophin gene (103), which spans 2.3 Mb, indicates that the splicing machinery can retain specificity over great lengths of genomic sequence. Noncanonical splice sites (GC-AG) are not handled by gene prediction programs, yet seem to be perfectly acceptable to the cellular splicing machinery (140).

For splice site prediction a number of algorithms have been created including Spliceview (131) and Splicepredictor (74). Typically, splice sites have been modeled using the fairly simple approach of a positional weight matrix without the perceived need to include insertions and deletions. Lacking a requirement to have to deal with insertions and deletions around these sites has also meant that they have been easy to explore with artificial neural network approaches, as was done by Brunak et al. (24), NetPlantGene (56), NetGene2 (154), and NNSplice (123), allowing covariances and interbase dependencies around the splice sites to be detected. Other
machine-learning approaches have also been used, such as support vector machines, utilized in the Splicemachine algorithm (170), genetic algorithms (133), and Bayesian networks (29); the latter study also does the rare job of demonstrating how the prediction method actually performs as a signal sensor within a gene prediction algorithm.

Within GENSCAN, Burge & Karlin (26) utilized MDD, a decision-tree-like approach, to effectively model the significant interdependencies that were observed between both adjacent and nonadjacent positions at the splice donor site. The authors inferred that these interdependencies arise from the requirement of the donor site to interact with the U1 snRNP ribonucleoprotein complex during splicing. The authors determined that the MDD approach is only necessary for the splice donor, and for the splice acceptor, a simpler first-order WAM model, which only considers dependencies at adjacent bases, was sufficient. GeneSplicer (117) takes this approach further, adding a second-order hidden Markov model to discriminate differences in coding and noncoding sequences, which would be expected to occur at a splice junction.

Even with this body of work addressing computational prediction of splice sites, there has been little appraisal to determine how many of these methods perform as a signal sensor embedded within a gene-finding algorithm. The presence of other gene features and the necessity to maintain an open frame will dictate which splice site predictions can be selected, making assessment of splice site prediction algorithms outside of this context less valuable.

Patterson et al. (113) investigated how modeling the secondary structure of the mRNA around splice sites can be used to aid splice site prediction. In this study Mfold (65) was used to fold 100 base pairs centered on each putative splice site and three specific structural metrics were determined: (a) the free energy of the optimal folding, (b) the helical propensity, where real splice sites showed a decrease in “Max Helix” score approximately 10 base pairs before the splice junction, and (c) considering whether the bases at each position in a fold structure were paired and either stacked or unstacked with the preceding base pair. Using a decision tree to combine the outputs of the structural assessments and a first-order weight-array model of the sequence consensus, the authors showed that small but significant improvements in accuracy can be achieved using structural information. However, it seems inevitable that an explicit modeling of the interaction between the U1 RNP and the transcript would be required to utilize structural information successfully to predict intron splice sites.

### 3′ Ends of Genes

The 3′ terminus of a gene defines the end of the encoded protein and regions of the transcript containing elements that direct polyadenylation. Features in 3′ UTR can influence the transcript decay rate, translational rate, and transport of the transcript (64, 87, 162). Eukaryotic mRNA transcripts typically possess the canonical polyadenylation signal (AAUAAA) that precedes the endolytic cleavage site at which the polyadenylation occurs. A GU-rich tract is typically located 20–40 nucleotides downstream of the cleavage site. Computational approaches to detect the polyadenylation signal have utilized word frequencies around the polyadenylation signal (134) and also position weight matrices of the sequence around the sequence signal (84, 151). More recently, by exploiting the annotated 3′ ends of genes from the entire *Caenorhabditis elegans* genome, Hajarnavis and colleagues (54) created a generalized hidden Markov model describing this feature. This provides an interesting example where detailed modeling has not only been able to provide a tool for feature prediction, but has also provided further insight into the fine structure and sequence biases of the nematode 3′ UTR (Figure 1). The resulting model could
accurately predict 3’ ends of genes even in the presence of multiple AAUAAA sites and multiple cleavage sites.

Xie et al. (164) performed a systematic computational search for regulatory motifs within the 3’ UTR region of human genes, which has also been successful in further identifying functional elements. In this study the alignment of human UTRs with orthologous UTR sequences from mouse, rat, and dog allowed the identification of 106 novel motifs. These motifs possessed the unusual properties of having strong strand specificity, being mostly eight base pairs in length, and preferentially terminating with the nucleotide adenine. These properties led the authors to propose that approximately half of these discovered motifs were targets for microRNA genes (12).

**Promoters**

The prediction of RNA polymerase II promoters, which transcribe protein-coding genes, represents a key determinant in defining gene structure and for subsequent activities such as regulatory element detection and the creation of engineered constructs for gene therapy and protein production. Such promoters often contain a TATA box approximately 25 base pairs upstream of the transcriptional start site. Around the transcriptional start site is the initiation site (Inr), which possesses a weak sequence consensus. Promoters can often possess only one of the TATA box or the Inr and still remain functional. Approximately one third of eukaryotic promoters also contain a CCAAT box (25). CpG islands, which are associated with
approximately 60% of RNA polymerase II transcribed genes (161), are a powerful signal in detecting promoter regions, and are strongly associated with promoters.

A number of algorithms have been developed to detect promoter elements by incorporating the statistical properties of their DNA sequences (8–10, 35, 38, 75, 122). Other methods rely on promoter prediction through identification of transcriptional start sites from experimentally derived mRNA transcript sequences (28, 90). Such approaches are greatly enhanced by chromosomal tiling paths placed on hybridization chips that can map transcribed sequences at high resolution (31).

An assessment of the performance of these algorithms was specifically designed to address how promoter prediction algorithms would fare when used in genome-scale studies (11). Although no method provided a positive predictive value greater than 65%, the authors concluded that DragonGSF (8, 9) and Eponine (38) provided a useful combination of positive predictive value and coverage. Other algorithms, CpGProD (119) and McPromoter (105, 107), displayed higher positive predictive values (70% and 78%, respectively), albeit at a much higher cost in sensitivity. The assessment also revealed the overall failure of algorithms to detect promoters lacking CpG islands, further indicating that this feature represents a strong component of the signal. In the absence of CpG islands some of the promoter algorithms failed to perform better than would be expected by random chance.

To identify promoters DragonGSF uses overlapping pentamer matrix models of promoters, exon, and introns to score sequences as well as quantify G+C content and CpG frequencies. An artificial neural network is subsequently used to classify whether the features are promoter-like. Eponine takes a hybrid machine-learning approach, using a relevance vector machine and a Bayesian support vector machine to select from a number of possible sensors, each of which is a weight matrix with a particular sequence position distribution relative to the promoter. The weight matrices selected by Eponine (Figure 2) demonstrate the paucity of the signal around the transcriptional start site, where, apart from the TATA box, the signal detected is comprised of differing patterns of CG richness.

The physical properties of the DNA have also been used to identify promoters. The

![Figure 2](http://www.sanger.ac.uk/Users/td2/eponine/)

(Figure 2)

Representation of the positional distribution and composition of the weight matrices relative to the transcriptional start used by the Eponine promoter predictor. The figure is adapted from T. Down.
sequence-dependent curvature of the DNA observed around transcriptional start sites has been utilized within the McPromoter algorithm, which scores the bendability of the sequence (106). Others have determined that promoter regions have particular mechanical properties containing both intrinsically rigid and flexible regions (47, 69). These may promote or enable the binding of the transcriptional initiation complexes and represent another distinctive feature that could aid in computational prediction of transcriptional start sites.

Comparative methods for detecting promoter and transcriptional start sites have also been developed. The first, PromH, developed by Solovyev & Shahmuradov (148), was inspired by Wasserman et al.’s work (159) in detecting regulatory elements through the comparison of orthologous sequences from mouse and human. This approach was built on a previous promoter finder, TSSW (45), but also rewarded features that appeared conserved based on a pairwise alignment of two orthologous sequences. The N-SCAN comparative gene predictor (53) incorporated a model of the 5′ UTR in a gene-finding approach that can identify the transcriptional start site and is based on the technology used to create the comparative gene prediction algorithm TWINSCAN (77). By including information from multiple genomic alignments comprised of human, mouse, rat, and chicken sequences, the authors demonstrated that this algorithm was the most accurate prediction tool for identifying transcriptional start sites (22).

Due to the weakness of the signal, successful ab initio prediction of promoters may require further consolidation of other features that are currently difficult to model. Transcription depends on the chromatin status, methylation patterns, other cis-regulatory elements both proximal and distal, and the plethora of transcriptional factors actually present in the nucleus, all of which need to be considered. A high apparent false positive rate may also reflect the underlying possible biological reality that genomic sequence can inherently promote promiscuous gene expression. High-resolution transcript-mapping studies indicate numerous transcribed regions or “transfrags” with no obvious gene product (31). Although initial inferences are often that such transcripts are functional, many will undoubtedly reflect a basal level of background expression. As each cell type possesses a different pattern of open chromatin and methylation while containing a specific combination of transcription factors, many of these spurious transcripts will also appear tissue specific. Such background transcription may be energetically wasteful but evolutionarily fruitful by enabling the selection of alternative promoters and subfunctionalization after duplication events.

**RNA Gene Prediction**

There is considerable interest in the identification of genes that produce functional RNA moieties as opposed to encoding a protein. Historically, studies have been confined to looking at the free energy of binding of potential stem-loop structures, and programs such as Mfold (65) and RNAfold (59) have been created to identify such features. These approaches are useful in determining the structure of a sequence already known to be an RNA gene. However, these methods are prone to false positive prediction, and no real means exist to appraise the putative functionality of a structure. Hence, even random sequence will often provide the user, on visual inspection, with a plausible structure. Therefore, such methods cannot be considered for the scanning of large tracts of genomic sequences for RNA gene discovery. In 2000, Rivas & Eddy (127) showed that approaches relying on prediction from a single sequence lack any real statistical power to detect RNA genes, ushering the field into comparative methods of RNA gene prediction. One notable exception to this is tRNA gene prediction by tRNAscan-SE (92). This can identify 99.5% of tRNA genes with an expected false
positive rate of less than one per 15 billion nucleotides of genomic sequence searched. Although tRNA gene prediction represents an extremely defined problem, with the well characterized structure containing a signature clover leaf motif and a restricted range of sequence lengths, this represents a clear example where a computational approach has essentially solved a particular aspect of gene detection.

Comparative methods of RNA gene detection typically look for not only sequence conservation but also variation where compensatory changes occur elsewhere in the alignment to retain a putative RNA stem structure. QRNA was the first RNA gene finder to utilize comparative information; taking a fixed pairwise alignment as input, it attempts to classify sequence as RNA coding, protein coding, or noncoding (128). More recent applications, such as comRNA (66) and CARNAC (116, 156), use the predicted secondary structure to help guide the actual alignment. A further refinement of the comparative approach has been the development of RNA-Decoder to identify conserved RNA structures residing within protein-coding sequences (115).

MicroRNAs and Their Targets

MicroRNAs (miRNAs) are small noncoding RNA products approximately 22 base pairs in length that were originally described in C. elegans (112). They provide a regulatory role in that they function to bind to target mRNAs and either repress transcription (109) or induce cleavage of the target mRNA (63). The proposed roles of miRNA genes have become increasingly important and although it was previously thought that approximately 0.5–1% of all mammalian genes would be miRNAs (88, 89), more than 300 human miRNA genes have already been described in the miRNA database, miRBase (51). It has been estimated that the total number may be at least 800 (14). miRNAs derive from larger RNA precursors, which form an extended stem-loop structure from which the miRNA is usually processed out of one of the hairpin arms (82, 125). Although miRNAs are typically determined from size-fractionated RNA preparations, there have been some attempts to computationally predict the presence of miRNA genes. miRNASEEKER was developed to look for miRNA genes in the Drosophila melanogaster genome (81). In this heuristic approach trained on the ability to detect the 24 known Drosophila miRNAs, the D. melanogaster genome was aligned with the D. pseudoobscura genome using the AVID alignment tool (19). Next, conserved sections of 100 or more base pairs of sequence identity were subjected to RNA folding prediction using the Mfold algorithm (65). The top 25% scoring folds were also folded in D. pseudoobscura and the candidates were ranked based on the average score from both species. Finally, Boolean filters were put in place that required a block of at least 22 nucleotides of perfectly conserved sequence (i.e., the miRNA) between the two species approximately 10 base pairs from the terminal loop. If both arms of the hairpin structure satisfied such a test then the sequence was predicted as an miRNA gene. Genes for which a conserved region between D. melanogaster and D. pseudoobscura was detected in only one of the hairpin arms were retained depending on the pattern of conservation within the hairpin arms and the absence of conservation with the stem loop. In 2004, Rhoades & Bartel (67) used a similar approach in Arabidopsis thaliana in which the rice genome was used to evaluate gene conservation. This study determined that the more stringent requirement of the predicted miRNA gene products to possess a near-perfect complimentarity to targets within the Arabidopsis and rice transcriptomes could be utilized successfully. This suggests that the miRNA molecules in plants possess greater complimentarity to their targets than is observed in animal genomes and may therefore allow both miRNA genes and their targets to be more easily determined. Bentwich et al. (14) described an approach that has
been applied to the human genome. In this case a number of parameters are considered for each predicted miRNA including hairpin length, loop length, a stability score and free energy calculation, paired nucleotides in 22 nucleotide windows, bulge size, repeat characteristics, and GC content. Scores from each of these features are combined to form an integrated optimal score and a threshold that is able to best distinguish known miRNA genes from background is chosen. However, searching a mammalian sized genome with such an approach does not predict miRNA at a specificity where wet-lab validation is not essential. Of the approximately 5300 above-threshold predictions, 89 were subsequently confirmed by expression and sequence analysis. However, the prediction deliberately excluded sequence conservation as a prediction metric, allowing 53 miRNA genes to be found that appear not to be conserved outside of the primate lineage.

Because miRNA genes work through complimentary base pairing, a logical progression has been the development of computational approaches for predicting their targets and thus establishing novel forms of regulatory networks. The partial complimentarity often observed between animal miRNAs and their targets means that more sophisticated approaches than sequence scanning need to be implemented. TargetScan (86) takes into account the free energy of binding of the predicted RNA duplex that would be formed and also the evolutionary conservation at the binding site. Therefore, to use this approach, the orthologous miRNA genes need to be identified in a number of other organisms and their transcriptomes, specifically 3’ UTRs, need to be well characterized. In reality, miRNA products are identical within mammals, thus mitigating the need to identify orthologous miRNA genes in most cases (80). This study also infers that, at least for animal miRNAs, base pairing in positions two to eight of the 22nd sequence appears to be the most critical for target recognition, referring to this seven-nucleotide region as the miRNA seed.

Kiriakidou et al. (73) used a similar approach to determine miRNA targets utilizing a free energy of binding and evolutionary conservation of the target site; they also experimentally developed a number of rules reflecting the size and position of bulges and loops outside of the miRNA seed region that can be tolerated in the miRNA target duplex while retaining functionality. Stark et al. (149) used a similar methodology of utilizing sequence complimentarity and conservation, but also exploited the fact that previously validated miRNA target genes often have multiple miRNA-binding sites in a 3’UTR and indicated that without using this attribute many bona fide target sites would remain statistically insignificant, even in a *Drosophila* 3’UTR database. Enright et al. (43) also used sequence complimentarity, free energy, and conservation to look for miRNAs in *Drosophila*, in this case allowing G-U wobble pairings and using strict conservation requirements in the putative target region with both *D. pseudoobscura* and *A. gambiae*.

**Gene Regulatory Elements**

The current understanding is that much of the specificity of gene expression can depend on how proteins bind to *cis*-regulatory DNA sequences and facilitate or repress the assembly of the transcriptional machinery at the promoter. Historically, *cis*-regulatory elements have been determined through labor-intensive molecular biology reporter assays (102) and DNaseI footprinting (48), whereas the potential binding sites of known transcription factors have been determined through in vitro DNA selection assay (SELEX) (157) experiments and gel-retardation assays. Recently, the ChIP-chip methodologies were developed in which DNA bound to transcription factors was isolated through cross-linking and immuno-precipitation of the transcription factor and hybridized to microarrayed genomic sequences (126); this latter method has the potential, in a high-throughput way, to determine potential binding sites of any
transcription factor. For mammals in particular, the results of such experiments remain noisy and incomplete due to the lack of generally available whole genomic arrays at the required resolution; further, experiments have to be repeated to include specific biological states (e.g., tissues/cell types and various conditions/stimuli). At present only a small number of transcription factors have been used in mammalian ChIP-chip-type experiments. A bioinformatic approach to identifying cis-regulatory elements on a genome-wide scale remains a powerful and complementary contribution to understanding gene regulatory control. The yield of bioinformatic approaches has been increasing steadily, not just due to the presence of the near-complete human genomic sequence, but also to the increasing numbers of vertebrate sequences that are available, including those of fish, mouse, rat, dog, cow, and a range of primates as well as a rapidly improving quality in computational genome annotations.

One approach in cis-regulatory element detection relies on determining evolutionary conserved motif sequences, and such so-called phylogenetic footprinting approaches have become increasingly powerful in determining the presence of potential regulatory elements (reviewed in Reference 160). The fecundity of such approaches will continue to improve as the number of sequenced genomes increases. For example, such approaches will benefit from the 11 low-coverage mammalian genomes to be sequenced by the Broad Institute (96) and the deep phylogenetic sampling of a few genomic regions provided by the ENCODE project (42). By investigating the sequence alignments between multiple mammals, Margulies and colleagues (95) were able to use a phylogenetically weighted scoring approach to identify “multi-species conserved sequences,” many of which were inferred to be regulatory elements. Because searching with degenerate motif models of transcription factor–binding sites typically generates a large number of false positives (44), searching only conserved regions with known motif models from databases such as TRANSFAC (163) and Jaspar (137) leads to substantial improvements in prediction accuracy (85). However, such approaches rely on correct sequence alignments, which is a challenge with current methods (118). Unlike protein-coding elements, which are typically conserved in a colinear manner, regulatory elements can change their relative order or undergo inversion without any obvious functional consequences. Studies have also shown that there is significant turnover of cis-regulatory elements, where elements can be lost and reappear within lineages at different positions (36, 62, 93), indicating that even with a correct alignment such functional motifs would not be detectable.

Another approach to determine regulatory motifs is to identify shared or similar motifs between coexpressed genes, with the presumption that such genes would be under the same genetic control. Other genomic resources now exist to identify potentially coregulated genes such as the Gene Ontology database (5), biochemical pathway information (68), and protein-protein interaction databases (for example, the BIND database) (6). The oPOSSUM software demonstrates an attractive approach for identifying overrepresented motifs shared between such sequences by using phylogenetic footprinting to identify evolutionary conserved regions that are subsequently searched for similarity to known motif models (58). Such approaches are suited for the study of previously characterized motifs. For the de novo discovery of functional motifs, a number of algorithms have been developed (143), and four general approaches have become popular. First is a Gibbs sampling method, which provides a computationally optimized approach to determine similar motifs between multiple sequences (83); second are expectation maximization algorithms such as MEME (7); third are greedy approaches such as Consensus (57) that find sequence motifs that share the greatest information content; and fourth are approaches that involve exhaustive enumeration of motif sequences such as...
Recently, the NestedMICA algorithm (39), which represents a novel approach to motif detection was developed. This approach optimizes a probabilistic mixture model of motif-like and background-like sequences. This algorithm uses the Nested Sampling Inference strategy that allows determination of the optimal solution without the need for a heuristic method to choose a reasonable starting point or “seed” to build the motif. Nested Sampling is a way of inferring sequence similarity by exploring in parallel an ensemble of states that is less likely to be trapped in local minima. This approach was determined to outperform algorithms such as MEME, and the approach taken by NestedMICA will be particularly suited to motifs for which there are few highly conserved bases and the absence of a conserved core.

There have been numerous refinements to these approaches which take advantage of some of the features typically displayed by regulatory motifs, such as the presence of motifs as spaced dyads (158) and the presence of sequence palindromes (91, 153). Because such methods search for the presence of motifs at any point within the input sequences, the typical degeneracy and short length of motifs mean that large sequence tracts will contribute too much noise and cannot be usefully searched. Generally, only sequences of 1–2 kb in length can be searched with such approaches. Another important consideration for regulatory element detection is the expectation of motif occurrence within the input sequence sets. For example, initial implementations of Gibbs sampling strategies assumed that all sequences contained the same motif and that the size of the expected motif was known, requiring a level of knowledge about the motif to be discovered that might mitigate the purpose of the exercise. Most approaches now incorporate motif occurrence models, allowing sequences to contain either zero or multiple occurrences of the same motif. Motif detection algorithms are also now able to capitalize on the availability of orthologous sequences from other species, and algorithms such as PhyloGibbs (141) and PhyME (144) can account for the lack of evolutionary divergence between closely related species and effectively reward the conservation of motifs over longer evolutionary distances.

Tompa and colleagues (155) recently assessed the algorithms used for motif discovery and assessed the algorithms on their ability to identify known motifs from the TRANSFAC database (163). To create the test set of motifs, TRANSFAC motifs were embedded not only within their native promoter sequences, but also in randomly selected promoter sequences and artificially generated DNA sequences. The exhaustive enumerator Weeder outperformed other algorithms, with much of the success attributed to its ability to not report motifs for many sequence sets in contrast to more verbose algorithms that will always report a best hit.

The Toucan software (1) also provides an interesting hybrid of approaches, as it identifies conserved regions between sequences through sequence alignment and then allows the de novo discovery of motifs in these regions using the Gibbs sampling implementation within MotifSampler (153). The use of sequence alignments means that much larger tracts of DNA can be studied than with a typical motif detection algorithm alone. However, the search space limitations remain such that de novo motif detection is limited to relatively well-defined search regions such as promoter regions (121) and 3′ UTRs (164).

Even when only a fraction of the DNA is exposed as open chromatin for binding by transcription factors, the small size of regulatory motifs and the variation tolerated by transcription factors to their known sites suggests that they will bind in a promiscuous way across the genome. The question thus remains of how transcriptional control is maintained by the cell. ChIP-chip experiments using the transcription factors Sp1, cMyc, and
p53 suggested 12,000, 25,000, and 1600 binding sites within the genome for each of these proteins, respectively (30). This data was collected using fairly stringent thresholds for hybridization and in only one cell line for each transcription factor, indicating that many weaker and cell-type-specific binding sites have yet to be determined. This unexpectedly large number of sites suggests that many of these binding events have little or no consequence on their own, but that the combinatorial binding of transcription factors in close physical proximity ultimately determines the specificity of gene expression. The cooperativity of transcription factors, where functional motifs typically occur in sparsely distributed clusters or modules, provides the potential for computational approaches to search for such over-represented patterns of motifs. Kreiman (78) described a technique that can identify conserved motif modules within sets of putatively coregulated genes. Zhu et al. (172) also used this property to discover novel motifs, first identifying regions possessing similarity to a known motif and then searching for enriched motifs in their immediate vicinity, e.g., 50–100 base pairs on either side. Motifs were found to most likely co-occur with themselves, correlating with previous biological data (4) that such repetition is important in emphasizing the regulatory signal. The tendency for regulatory motifs to act cooperatively has also been exploited for de novo motif discovery, where less well conserved motifs can be reported if they co-occur among the input sequences (171).

MODELING OPEN CHROMATIN

Within the nucleus, DNA forms chromatin fibers through association with histone proteins, creating the densely packaged nucleosome. Transcriptional control depends not only on the transcription factors present but also on their ability to co-operatively displace the histone proteins and bind to the DNA. Such exposed regions to which regulatory factors bind are also sensitive to cutting by the endonuclease DNaseI (52). DNaseI hypersensitivity has been used as a powerful empirical indicator of a regulatory role for a given DNA sequence. This approach has been particularly useful in determining regulatory regions that are distal to the transcriptional start site, where other methods such as reporter assays are impractical.

Obviously, a computational approach to identifying DNaseI hypersensitive regions would be a useful indicator of regulatory potential in a genome, yet no strong generic signals or consensus sequences have been associated with such regions. Therefore, it is of particular interest that Noble and colleagues
(104) devised a computational methodology to detect DNaseI hypersensitive regions. Fueled by a large data set of experimentally confirmed hypersensitive sites from high-throughput assays (132), the authors trained a support vector machine as a classifier. Using sequence words between two and six base pairs as input DNA, the classifier showed an impressive positive predictive value of at least 73.9%, the assessment being confounded by the fact that apparent false positive predictions may represent real sites in other cell types. This work also determined that the human genome contains more than 26,500 sites that are DNaseI sensitive and therefore likely to contain regulatory features.

**CONCLUSION**

There is now a vast array of algorithms and computational methods that can aid in the layering of functional information onto genomic sequence, thus providing a mechanism to rapidly annotate entire genome sequences. These include ab initio methods as well as approaches that rely on comparative sequence analysis and the consolidation of experimental data. It is also appealing to speculate that future computational approaches alone will play an increasingly important role in the identification of new functional classes of sequence, as exemplified by the elucidation of ultraconserved elements (MCEs) in the human genome (13).

**SUMMARY POINTS**

1. The more successful gene prediction programs have incorporated hidden Markov models as their machine-learning approach. Methods of gene prediction that incorporate information from orthologous sequences or known proteins are more successful than approaches that utilize a single DNA sequence.

2. Various machine-learning approaches have been applied to the recognition of intron splice sites. Further improvement in the ability to accurately determine such sites has been made through consideration of the three-dimensional structure of the RNA message.

3. Transcriptional promoters represent a subtle feature for computational prediction, the strongest components of the signal being the CpG island and the TATA box.

4. Computational techniques now exist to predict both microRNAs and also the targets of microRNAs. Free energy of binding and orthologous sequences has been used to improve the fecundity of such methods.

5. Numerous approaches exist to detect regulatory elements, and the prediction of exposed regions of DNA where transcription factors can bind using a support-vector machine approach represents an innovative advance.

**ACKNOWLEDGMENTS**

I would like to thank Mikhail Bilenky, Thomas Down, Paul Flicek, Obi Griffith, Ian Korf, Irmtraud Meyer, Stephen Montgomery, Erin Pleasance, and Gordon Robertson for their helpful discussions, assistance, and comments during the creation of this review. Steven Jones is a scholar of the Michael Smith Foundation for Health Research.
LITERATURE CITED


Contents

A 60-Year Tale of Spots, Maps, and Genes  
Victor A. McKusick ................................................................. 1

Transcriptional Regulatory Elements in the Human Genome  
Glenn A. Maston, Sara K. Evans, and Michael R. Green .................. 29

Predicting the Effects of Amino Acid Substitutions on Protein Function  
Pauline C. Ng and Steven Henikoff ........................................ 61

Genome-Wide Analysis of Protein-DNA Interactions  
Tae Hoon Kim and Bing Ren ................................................... 81

Protein Misfolding and Human Disease  
Niels Gregersen, Peter Bros, Sören Vang, and Jane H. Christensen .......... 103

The Ciliopathies: An Emerging Class of Human Genetic Disorders  
Jose L. Badano, Norimasa Mitsuma, Phil L. Beales, and Nicholas Katsanis .......... 125

The Evolutionary Dynamics of Human Endogenous Retroviral Families  
Norbert Bannert and Reinhard Kurth ....................................... 149

Genetic Disorders of Adipose Tissue Development, Differentiation, and Death  
Anil K. Agarwal and Abbimanyu Garg ....................................... 175

Preimplantation Genetic Diagnosis: An Overview of Socio-Ethical and Legal Considerations  
Bartha M. Knoppers, Sylvee Bordet, and Rosario M. Lasi .................. 201

Pharmacogenetics and Pharmacogenomics: Development, Science, and Translation  
Richard M. Weinshilboum and Liewei Wang ................................ 223

Mouse Chromosome Engineering for Modeling Human Disease  
Louise van der Weyden and Allan Bradley .................................. 247
The Killer Immunoglobulin-Like Receptor Gene Cluster: Tuning the Genome for Defense
Arman A. Bashirova, Maureen P. Martin, Daniel W. McVicar, and Mary Carrington ................................................................. 277

Structural and Functional Dynamics of Human Centromeric Chromatin
Mary G. Schueler and Beth A. Sullivan ............................................. 301

Prediction of Genomic Functional Elements
Steven J.M. Jones ........................................................................... 315

Of Flies and Man: Drosophila as a Model for Human Complex Traits
Trudy F.C. Mackay and Robert R.H. Anholt .................................. 339

The Laminopathies: The Functional Architecture of the Nucleus and Its Contribution to Disease
Brian Burke and Colin L. Stewart .................................................. 369

Structural Variation of the Human Genome
Andrew J. Sharp, Ze Cheng, and Evan E. Eichler ......................... 407

Resources for Genetic Variation Studies
David Serre and Thomas J. Hudson .............................................. 443

Indexes

Subject Index .................................................................................. 459
Cumulative Index of Contributing Authors, Volumes 1–7 .................. 477
Cumulative Index of Chapter Titles, Volumes 1–7 .............................. 480

Errata

An online log of corrections to Annual Review of Genomics and Human Genetics chapters may be found at http://genom.annualreviews.org/