Chapter 6

FUTURE WORK

Genome-based research and development have opened a new era in biology, biomedicine and biotechnology. Significance of the genomic information cannot be fully understood without uncovering every component that a genome contains. This process of identification or annotation in the post-genomic era requires use of computational methods that can handle huge volume of sequence data. Researchers all over the world are attempting to develop programs that scan through the genome and identify the genomic components. Despite the all-round development of programs and algorithms for the analysis of different components of the eukaryotic genomes, the requirement for accurate and fully automated in silico annotation of the genomes remains unfulfilled. In the present work an attempt is made to develop methods to predict to of the most important components of the eukaryotic genomes, namely—protein coding genes and the repetitive sequences in genomic DNA.

The work on identification of protein coding genes in eukaryotic genomes described in this thesis makes use of similarity information that is obtained on searching against protein and intron sequences. Like most existing gene finders, the present method also has some limitations. One of the perennial problems of computational prediction is the assumption by programs that biological processes are being independent, simple events. However, the information-flow during transcription and translation events in a cell neither is static nor can be determined by a simple, single method as the existing gene finding programs suggests. There are several variants to how the information is passed from DNA to RNA to Protein. Some of the possible events include alternative splicing, use of nonconsensus splice sites, exon skipping (where a possible second transcript from a gene does not include one or more exons that are included in the first transcript from the same gene), and nested genes (a gene that is present inside intron of another protein-
coding gene). Similarity evidence from protein database could possibly provide valuable information regarding alternative splice sites and use of non-consensus splice sites in a gene.

Similarly, the prediction of repeats using Fourier transformation technique has its demerits. One of the major limitations of the Fourier methods is that it cannot be used alone to differentiate the boundaries of the repeat units or even regions. This requires the use of a method that locates the individual units of a repetitive DNA sequence. Locating the repeat units individually increases the time required for completion in the present method. This presents one of the primary challenges in any repeat prediction algorithm. Another important limitation that repeat-finding programs have to deal with regularly is the problem of occurrence of mutations. Mutations are reported to be relatively at a higher rate in the repeat regions than in the regions of conservation like the protein-coding regions. These mutations can be substitutions, insertions or deletions of some bases in one or more copies of a motif. SRF algorithm does not consider the occurrence of indels in motifs. Another problem that is evident from the algorithm is that SRF cannot locate or identify repeat units that lie inverted along the sequence or are complementary to the parent seed pattern. The method is also unable to discriminate the repeat units that matches only partially to the seed pattern and are the result of segmental deletion or insertion events.

Most of these problems discussed above are among the foremost challenges in the field of gene prediction and repeat prediction for quite some time. Efforts are presently underway worldwide to solve some of these problems. Experience from the present work in genome annotation has given some insight into the probable solutions for the above problems. Models on the some of the probable method to solve some of the problems mentioned above are discussed here.
6.1 Models for prediction of alternative splice sites and alternate translation start sites (TSSs)

Information flow during transcription and translations events in a cell is not static or a simple, single method as the predictions from gene finding programs suggests. There are several variants to how the information is passed from DNA to RNA to Protein. One of the possible events is the alternative splicing event with or without the use of non-consensus splice sites. Similarity evidences from protein database could possibly provide valuable information regarding alternative splice sites and use of non-consensus splice sites in a gene. One of the logical methods to solve these problems would be to keep the complete 'global transcript' information of all protein hits from a database similarity search. From such information, it is possible to derive overlapping hits to probable exon regions. The differences in the length and signal content of overlapping exon regions from different transcripts would easily provide information regarding complex events such as use of alternate translation start site (TSSs), and alternate splicing sites. Use of non-consensus splice sites is also easily identifiable with the similarity-based approach against protein database. Figure 6.1a shows the diagrammatic representation of a probable solution to the problem of alternate start sites. In this example, a particular query \( Q \) nucleotide sequence has two different protein hits, \( P_1 \) and \( P_2 \) from the protein sequence database with very high E-value for each. \( Q \) has one initial exon, one terminal exon and two internal exons. Of these four exons, the latter three matches exactly in coordinates between \( P_1 \) and \( P_2 \), whereas the first or the initial exon uses different start codons. In the present algorithm, only the longest exon with lowest E-value is selected for modeling the probable gene structure. However, strong database hit for the second smaller transcript suggest that alternate start codons could be used for the translation of proteins. The model suggested here argues to also keep the second possible transcript variant for the initial exon and present two different gene structure models to the user. Since it is genuinely possible to obtain only partial protein hits, it would be logical to include only those transcripts that completely align with the query genomic sequence. The given argument will also hold true in case of alternate splice sites and would allow annotators to identify the non-consensus splice sites very easily (Figure 6.1b).
Figure 6.1a: Diagrammatic representations of the two probable gene structure models of a multi-exon gene that can be derived when two protein hits have alternate start codons in first exon. FE – First exon; IE – Internal exon; LE – Last exon; S – Start codon; T – Termination codon; Q – Query sequence; P1 – First protein hit; P2 – Second protein hit.
Figure 6.1b: Alternate splice sites can be verified using multiple coding region hits from protein sequences. FE1 and FE2 – First exons; IE1-IE3 – Internal exons; LE1 and LE2 – Last exons; S – Start codon; T – Termination codon; Q – Query sequence; P1 – First protein hit; P2 – Second protein hit; a – acceptor sites; d – donor sites. Red lines indicate the path of probable alternative gene structure models. Alternate splice sites are observed in hits to the first internal exon and last exon. Eight gene structure models that can possibly be derived in the diagrammatic representation above include the {FE1 – IE1 – IE2 – LE1}; {FE1 – IE3 – IE2 – LE1}; {FE1 – IE1 – IE2 – LE2}; {FE1 – IE1 – IE2 – LE2}; {FE2 – IE1 – IE2 – LE2}; {FE2 – IE3 – IE2 – LE2}; {FE2 – IE1 – IE2 – LE1}; and {FE2 – IE1 – IE2 – LE2}. 
6.2 Models for prediction of alternate eukaryotic gene transcripts in case of 'exon skipping' and 'nested genes'.

Some of atypical cases of translation involve events like exon skipping and nested genes and are difficult to predict using regular gene finding programs. Exon skipping describes biological processes where the second transcript of a gene does not include one or more exons that are included in the first transcript of the same gene. Nested genes are biological processes where a protein-coding gene is present inside intron of another protein-coding gene.

Nested genes could possibly be identified using a negative evidence (intron) approach. In this method, any such gene that will be completely masked by similarity to intron sequences can be considered as a potential nested gene. The hypothesis for this model is that because the protein-coding regions are conserved across related sequences, it could be used to extract any smaller nested gene that would show similarity to a single intron sequence from another gene. Figure 6.2a diagrammatically depicts the model where a potential single-exon gene lies inside the intron sequence of another gene. In this example, the intron sequences from the database that show similarity to query genomic sequence in one model completely overlaps the single exon gene from the other model of the same query genomic sequence.

Similar strategy can also be used to identify events like exon skipping (Fig 6.2b). In this case, multiple hits may be observed from the similarity search against sequence databases. Any exon or exons that are present in some hits but are completely absent in other hits could be a potential candidate for exon skipping. Complete masking of an exon by an intron sequence would confirm the probability of the occurrence of such an exon-skipping event. The process of exon skipping can also be verified on similarity search against complete cDNA sequences.
Figure 6.2a: Diagrammatic representation of a nested gene inside an intron of another gene. Q – Query genomic sequence; S – Start codon; T – Termination codon; CDS – Annotated protein coding region from query sequence; PROTEIN – BLASTX hits on similarity search against protein database; INTRON – BLASTN hits on similarity search against intron database that overlaps BLASTX hits from protein database; thin line – intron hits from BLASTN against intron database.
Figure 6.2b: Diagrammatic representation of 'exon skipping' event in a eukaryotic gene. Q - Query genomic sequence; S - Start codon; T - Termination codon; CDS - Annotated protein coding region from query sequence; PROTEIN - BLASTX hits on similarity search against protein database; INTRON - BLASTN hits on similarity search against intron database. The gene model 1 represents the first transcript where three exons are located from complete protein transcripts whereas in model 2, four exons are identified. The first internal exon of gene model 2 is completely masked by the intron from model 1. Multiple lines in each model represent each hit to database sequences using BLAST program.
6.3 Importance of genome-to-genome comparisons to aid in gene prediction

Extracting the complete functional information encoded in a genome—including the genie, regulatory and structural elements—is a central challenge in biological research. *De novo* identification of the complete set of protein-coding sequences remains imperfect, even in well-studied organisms with compact genomes. Comparative genome analysis of related species should provide a powerful and general approach for identifying functional elements without previous knowledge of function. This is based on the hypothesis that coding DNA sequences are more conserved than non-coding sequences (intronic and intergenic). Because evolution relentlessly tinkers with genome sequence and tests the results by natural selection, such elements should stand out by the virtue of having a greater degree of conservation across related species. The approach has the advantage that one can increase its power by increasing the number of species studied. All the comparative genome methods have, theoretically, the advantage of not being species specific. In practice, their performance will depend on the evolutionary distance between the compared sequences. A greater evolutionary distance allows some algorithms to more accurately discriminate between coding and non-coding sequence conservation. One of the main drawbacks is that these methods will be computationally intensive. The accuracy of the strategy will also be affected by the presence of repeated sequences and therefore these methods have necessarily to be coupled or integrated with other programs to eliminate these repeated sequences (SINES, LINES, etc), which are very frequent in higher organisms like humans. Some of the programs that try to retrieve information on conservation or synteny between organisms from genomic DNA to genomic DNA alignments include, MUMmer (Delcher et al. 1999), WABA (Kent et al. 2000), PipMaker (Morgenstern, 2000) and DIALIGN (Batzoglou et al. 2000). Some recent major works that use the genomic alignment strategy for prediction of protein-coding genes include the work on yeast species (Kellis et al. 2003) and more on Drosophila (Sinha et al. 2004).
6.4 Model for an alignment based repeat identification method.

Finding repeat patterns of a motif in any DNA sequence is a complex problem, particularly where there are lots of mutations, including insertions and deletions (or indels). In such a case, an exact method will not be able to accurately find all the occurrence of a particular pattern in a DNA sequence. Though allowing for substitutions to be present in the pattern will increase accuracy, these methods will still be unable to find patterns with indels. We propose to develop an alignment-based method to locate all the occurrences of a pattern that has been modified due to insertions and deletions. The algorithm would work in following steps

1) For a subject sequence of Length N, and query pattern of length D, alignment is performed between query sequence and sub-patterns M of length D + (%mutation allowed in bp with respect to query sequence), derived from subject sequence starting from position 1 to position (N-D).

2) For every alignment performed, percent similarity is computed between the query sequence and corresponding aligned sub-portion of sub-pattern M.

3) Percent Score: For every match +1 is added to Percent score (P) and total percent score of each pattern is divided by length of pattern D to normalize the value for each pattern.

\[
P = \frac{\text{total exact match}}{\text{length of pattern (D)}};
\]

In event of Insertion and deletions being permitted, another scoring is also calculated called the Total Score.

4) Total Score: For every match +1 is added and for every mismatch a penalty of -0.5 is subtracted from the score (S) and for each insertions and deletions -0.25 is subtracted from score (S).

\[
S = (\text{total exact match}) \times 1 - (\text{total substitutions})/2 + (\text{total del+ins})/4
\]
5) For any group of consecutive sub-patterns that has less than 100% match (percent score < 1), first occurring sub-pattern that has the highest score is selected.

6) For alignment, the program ClustalW could be utilized. It is a global alignment program that aligns the query patterns to sub-patterns and provides an alignment that could be utilized to reasonably represent any insertions and deletions in terms of gaps in query sequence and sub-pattern, respectively.

**Drawbacks:** The algorithm is computationally intensive since alignment methods like ClustalW would be used for aligning sequences. The time required for completing analysis would be directly proportional to the length of subject pattern, number of mutations allowed, and number of motifs present in the sequence.