Selecting relevant features for gene structure prediction

Yvan Saeys[1], Sven Degroeve[1], Dirk Aeyels[2], Pierre Rouzé[1,3] and Yves Van de Peer[1]

[1] Bioinformatics group, Department of Plant Systems Biology, Ghent University,
Flanders Interuniversity Institute of Biotechnology (VIB), Technologiepark 927, Ghent, 9052, Belgium
[2] SYSTeMS Research Group, Ghent University, Technologiepark - Zwijnaarde 9, Ghent,
9052, Belgium
[3] Laboratoire associé de l'INRA (France)

1 Introduction

Gene structure prediction is an important, yet complex task in bioinformatics. A reliable gene prediction forms the basis of many applications in functional, structural and comparative genomics.

Apart from using machine learning techniques merely for classification or prediction, biologists are also interested in the principles behind complex processes, such as splicing. Such a discovery of domain-specific knowledge is a main challenge for scientists working in this field, because many issues involved in gene transcription are not yet well understood.

A well-known method to gain more insight into data is the application of feature selection techniques. By eliminating irrelevant or redundant features, a subset of relevant features can be discovered, often improving both classification performance as domain understanding.

The paper is structured as follows. We start by explaining the biological background, needed to understand gene prediction. Then we introduce the main topics of the paper: gene structure prediction and feature subset selection. The next section then discusses the specific problems that arise when combining gene structure prediction with feature selection techniques. We end with some concluding remarks and future perspectives.

2 Methods

2.1 Biological background

At different stages during the life of a cell, different proteins are synthesized, interacting with each other and performing a specific function. When a certain protein is needed, the gene that codes for that protein will be expressed and the appropriate protein will be synthesized.

The mechanisms behind gene expression depend on the complexity of the organism being studied. In this paper, we focus on *eukaryotes*, i.e. higher organisms.

To know when and where to be expressed, each gene is preceded by a promoter region. The promoter region can be seen as a gene's identity card. When the appropriate proteins bind to this region, the start sign is given to produce the protein. This process is illustrated in Figure 1.

When the start of the gene (promoter region) is recognised, a copy of the rest of the gene is made. This step is called the *transcription*, and produces a single stranded copy of the gene: the pre-messenger RNA (pre-mRNA). In eukaryotes, this transcript contains both non-coding (introns) and coding regions (exons), the introns have to be spliced out, producing the mature mRNA[1]. The mRNA will then be transported from the nucleus to the cell's cytoplasm, where it is translated into a protein. To this end, a special marker on the mRNA is looked for: the start codon. From this position on, the mRNA is translated until another marker (stop codon) is encountered.

The prediction of the complete structure of a gene can be described as a two-step process. In a first step, the various structural elements are predicted: promoter elements, start- and stop-codon, and boundaries between the introns and exons (these are the so called splice sites). In a second step, all of these predictions have to be combined into an overall, consistent gene structure.

[1] It has to be noted that in RNA, the nucleotide T is replaced by another nucleotide: U. However, for reasons of simplicity we will use the DNA alphabet (A,T,C,G) throughout the paper, even if describing RNA

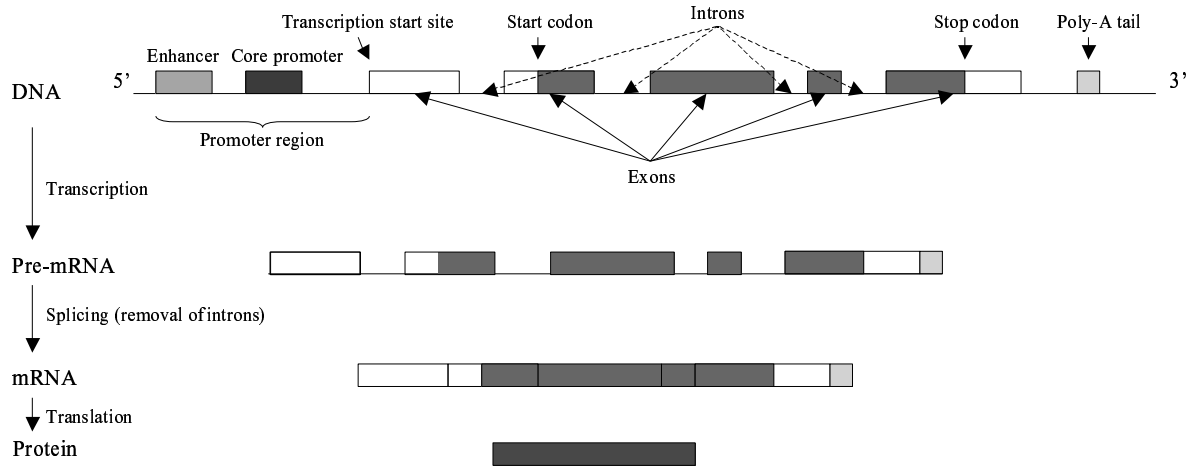

Figure 1: Transcription and translation.

2.2 Gene prediction

Current methods of gene prediction mostly focus onto the problem of identifying the coding and non-coding parts of the sequence, often leaving the detection of the promoter region aside. Two approaches can be distinguished: a) the prediction of genes by aligning them with genes already known (extrinsic prediction), and b) the use of classification models based on features extracted from the sequence (intrinsic prediction). Here we will only discuss intrinsic gene prediction, as these methods allow us to perform feature selection.

At present, a wide variety of classification models for gene prediction is used. A large number of methods make use of Markov models of different kinds and flavours (Burge and Karlin, 1997; Lukashin and Borodovsky, 1998; Salzberg et al., 1999; Schiex et al., 2001). Other methods include Linear or Quadratic Discriminant Analysis (Solovyev et al., 1995; Zhang, 1997), Neural Networks (Hebsgaard et al., 1996), and more recently also Support Vector Machines (Sonnenburg, 2002; Degroeve et al., 2002). A more detailed overview of gene prediction methods can be found in Mathé et al. (2002) and Zhang (2002).

2.3 Feature subset selection

Feature subset selection (FSS) is a technique to reduce data dimensionality, and is often used as a pre-processing task for classification (Kohavi and Pfleger, 1994). The reduction of data dimensionality has a number of advantages. Irrelevant or redundant features often behave like noise in the data, confusing the classification model and degrading its performance. The removal of such features results in a restricted subset of features with equal or better classification performance than the full feature set. Another benefit is the need to store less features that describe the data, and a faster classification. Additionally, reducing the number of features helps the human expert focus on a subset of relevant features, providing the ability to get a better insight in the processes described by the data. The different FSS techniques can be divided into three major classes: filter approaches, wrapper approaches and embedded approaches (Blum and Langley, 1997).

2.3.1 Filter methods

Filter approaches compute a feature relevance score, using only the intrinsic properties of the data. Afterwards these scores are sorted and low scoring features are removed. The computation of these scores is mostly calculated using measures from probability or information theory and provides a mechanism that is independent of the classification method to be used afterwards (Doak, 1992; Ben-Bassat, 1982). A common criticism to the traditional filter techniques is the fact that these models do not take into account the dependencies that might exist between features (Kohavi and John, 1997). On the other hand, such a simplifying view significantly speeds up the selection process, making the filter approach the preferred option when dealing with a large number of features.

To overcome the problem of ignoring feature dependencies, more advanced filter methods have been developed, taking into account feature-feature correlations. Examples of these include the method of Koller and Sahami, based on Markov blankets (Koller and Sahami, 1996) and the correlation-based feature selection method (CFS, Hall and Smith (1999)). However, taking into account feature correlations heavily increases the computational complexity of these algorithms, rendering them at least quadratic in the number of features.

2.3.2 Wrapper methods

In the wrapper approach various subsets of features are generated and evaluated. The evaluation of a specific subset of features is obtained by training a classification model, and using either cross-validation on the training set, or a separate training and holdout set. As a consequence, the wrapper method is tailored to a specific classification model.

To search through the space of all feature subsets, a search algorithm is "wrapped" around the classification model. The search can be either greedy, using e.g. sequential forward or backward methods (Kohavi and John, 1997), or heuristic. Stochastic models like Genetic Algorithms (GAs) or Estimation of Distribution Algorithms (EDAs) have been applied successfully as wrapper methods to FSS (Vafaie and De Jong, 1993; Kudo and Sklansky, 2000; Larrañaga and Lozano, 2001).

2.3.3 Embedded methods

In the embedded approach, the feature selection mechanism is built into the classification model, making directly use of the parameters of the induction model to include or reject features. Examples of embedded methods are the pruning of decision trees (Blum and Langley, 1997) and iterative feature elimination using the weight vector of a linear Support Vector Machine (Brank et al., 2002).

3 Combining gene prediction with feature selection

The motivation for combining gene structure prediction with FSS techniques consists of two parts. The first part concerns the purely computational aspects: a reduced set of features, allowing a better and faster classification of the data. The second part focuses more on the gain of insight into the biological processes related to transcription and translation. During the rest of this section we describe the three major difficulties that arise when applying FSS techniques to the different subtasks: 1) the problem of constructing features that describe DNA/RNA sequences, 2) the existence of many correlations between the features, and 3) the great diversity of the various subtasks involved in gene prediction.

3.1 Feature construction

Biology is not an exact science. Although already a number of important mechanisms regarding transcription and translation have been discovered, many things still remain unknown. As a result, there is no clear-cut answer to the question which features to include.

A problem related to the uncertainty about the underlying biological processes is how to map *biological* features to *experimental* features. By biological features, we here mean the true, but largely unknown features of molecules and proteins responsible for the underlying process. Based on biological knowledge and statistical analyses of data, we can then construct approximations of the biological features, and use them afterwards to train classification models, hence the term experimental features.

The traditional way to deal with this uncertainty is to include a large number of *potentially useful* features, hoping that (most of) the features that are relevant to the classification task are included. However, increasing the size of the feature set has a number of drawbacks.

A first problem is the so called *peaking phenomenon*: as more features are included, more training data will be needed in order to allow reliable parameter estimations (Raudys and Jain, 1991). Moreover, as it is uncertain whether the additional features are useful, many irrelevant or redundant features will be introduced. These features behave like noise for the classification model, often degrading its classification performance. A last, yet evident, shortcoming of adding lots of features is the fact that the time needed to train the classification model will be longer, if at all the model is capable of handling that many features. As a consequence, the need for FSS techniques arises, providing a way to eliminate the redundant or irrelevant features

that have been introduced.

To exemplify the amount of features that can be extracted from a genomic sequence we shortly describe one of the subtasks in gene prediction: the prediction of acceptor sites. An acceptor site is defined as the splice site occurring at the transition from intron to exon. Acceptors are characterized by a conserved "AG" dinucleotide in the intron part.

However, as this dinucleotide occurs very frequently throughout the genome, only a very small percentage of them will be true acceptor sites. As a consequence, the local context around the putative acceptor site has to be taken into account when deciding if a site is either a true or a pseudo acceptor site (see Figure 2).

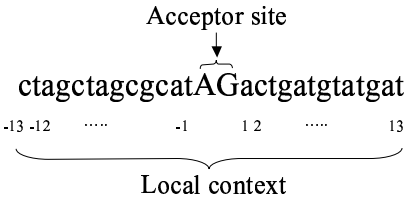

Figure 2: Acceptor prediction: local context around the acceptor site.

To allow the classification algorithm to learn the difference between true and pseudo acceptors, features are extracted from the local context. The most simple type of features are the occurrences of nucleotides at fixed positions in the pre-mRNA string, e.g. a c at position -13, a t at position -1, etc.

Another type of features are the *position-independent* features. These features can capture the importance of certain motifs that occur in the sequence. This can be done by extracting words of length k from the sequence. Commonly used features include the trimers ($k = 3$) and hexamers ($k = 6$). It is clear that the longer words drastically increase the size of the feature set, e.g. extracting only the position-independent hexamers from a local context already results in 8192 features (4096 hexamers on either side of the splice site).

To handle such large feature sets, traditional filter methods seem the most appropriate solution, as the computational cost of using a wrapper method will be too high. Still, when lots of features are present, the use of some wrapper methods might still be feasible when features are sparsely encoded (Saeys et al., 2003).

3.2 Many correlations between features exist

Nucleic acid sequences are more than just sequences of A, T, C and G. In the cell, these nucleic acids are organized into a three-dimensional structure, and from biology it is known that the structural properties of DNA/RNA are also very important in identifying its function. For the subtasks involved in gene prediction, structural properties need to be considered in order to understand the correlations or dependencies between features. Two important types of correlations between features are binding information and RNA secondary structure.

Binding information is related to the fact that certain protein-complexes bind to a specific location on the DNA/RNA. In the example of splice sites, a protein-complex called the *spliceosome* binds to the pre-mRNA, and then splices out the intron. The place where the protein-complex then binds is usually a conserved subsequence, where some of the nucleotide positions might show a stronger conservation than others. Correlations then exist between each of the features capturing the information at the positions of the conserved subsequence.

Binding information introduces correlations between features that model nearby sequence characteristics. However, this need not necessarily be the case to explain correlations between features. The reason for this is that nucleotides that are far apart in the sequence, might be closely together in the three-dimensional structure.

Biologists often refer to this three-dimensional structure as the *tertiary structure* while the RNA sequence itself is termed the *primary structure*. The *secondary structure* describes the sequence at a level in between these two, using structural elements like loops and stems. This is illustrated in Figure 3 where a structure with three loops and two stems is shown. From this figure it can be easily seen how dependencies between distant parts of the sequence may exist.

The existence of many feature dependencies has some important consequences for feature

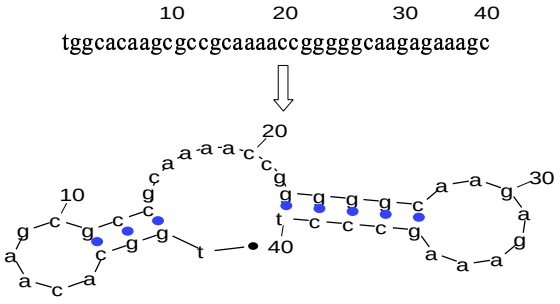

Figure 3: Going from sequence to secondary structure.

selection, and suggests the use of FSS techniques that take into account these dependencies: the more advanced filter methods (Koller-Sahami method and CFS) or the wrapper methods. However, the computational complexity of these methods does not allow them to scale to domains with a large number of features. An interesting application therefore would be the use of a correlation-based filter method, recently introduced by Yu and Liu (2003), which has a complexity that is less than quadratic in the number of features. Other solutions might be the application of hybrid filter/wrapper approaches (Das, 2001; Xing et al., 2001).

3.3 Diversity of gene prediction subtasks

As explained above, gene structure prediction consists of a number of subtasks, each task being the prediction of a certain structural element that makes up the gene. Each subtask models a certain underlying, biological process, and these processes are very different from each other, turning the subtasks into very diverse problems.

As a consequence, each of the subtasks will have its own type of features, and features that might be useful for one problem, may turn out useless for another one. Therefore, each of these subproblems has to be tackled individually, extracting potentially useful features and subsequently applying FSS techniques, to tune the feature set for each problem separately.

This observation can be extended further to sequences of different organisms: as biological processes (e.g. splicing) differ from one organism to another, other features might be relevant for a similar subtask when comparing different organisms.

4 Related work

Contrarily to other bioinformatics problems such as the selection of genes from microarray data (Guyon et al., 2000; Grate et al., 2002; Inza et al., 2003), FSS techniques have hardly been applied to gene structure prediction. A first attempt to define a set of relevant features can be found in Fickett and Tung (1992) where the authors provide a statistical analysis of several measures that can discriminate between coding and non-coding regions. The use of model-based FSS techniques for splice site prediction was pioneered in Degroeve et al. (2002) using an embedded approach within a Support Vector Machine (SVM), and in Saeys et al. (2003), where an EDA-based wrapper method was developed that is able to handle large feature sets when features are sparsely encoded. For the problem of translation initiation site prediction, Zeng et al. (2002) combined the CFS method with various classification methods, reporting improved classification performance.

In contrast to the other subtasks in gene prediction, the prediction of promoter elements is still in its infancy (Rombauts et al., 2003), and remains a largely unexplored area. In Pedersen et al. (1999) an overview of the biological issues involved in promoter prediction is discussed, suggesting the use of several structural profiles, yet the combination of such structural features, and the application of FSS techniques still needs to be explored.

As already mentioned before, feature construction/extraction is closely related to feature selection. Due to the increasing popularity of SVMs, special kernel functions have been developed to tackle specific biological problems, e.g. for predicting translation initiation sites (Zien et al., 2000) or splice sites (Sonnenburg, 2002). On the other hand, embedded FSS techniques have already been developed for SVM (Hermes and Buhmann, 2000; Sindhwani et al., 2001). An interesting topic of research would be to combine both approaches and perform both feature extraction and selection inside the SVM.

5 Concluding remarks and future work

In this paper we provided an overview of the issues involved when applying feature selection to the biological domain of gene structure pre-

diction. We showed that there is indeed a need for appropriate FSS techniques, taking into account the specific problems inherent to gene prediction: problems regarding feature construction, many correlations between features and the heterogeneity between the different subtasks involved in gene structure prediction.

Benefits of applying FSS techniques do not only include an increase in classification performance, they also provide useful insights into the underlying domain. This is especially important when classification models are used that are less transparent to the user (e.g. Neural Networks or Support Vector Machines).

From our research, some novel ideas regarding the application of FSS techniques to gene prediction emerge. A first idea regards the dependencies between features. So far, we have only considered dependencies between features within a specific subtask of gene prediction. At the higher level of the complete gene structure however, it may well be that there exist some dependencies between features of different subproblems. Such higher order correlations would add a new layer of complexity to the feature selection process.

A second idea concerns the way redundant features are handled. Traditionally, these features are eliminated by the FSS method, just like irrelevant features are removed from the feature set. However, for the biological expert, redundant features might be of interest when trying to extract domain knowledge. The rationale behind this is the fact that the underlying biological mechanisms should be robust, inherently implying some form of redundancy.

So far, these ideas have never been explored, as they make the problem of FSS for gene prediction even more complex. On the other hand, they provide an interesting and challenging task for scientists involved in machine learning.

Future work will focus on the integration of new features and feature selection techniques to improve classification, and get more insight into the different subtasks of gene prediction.

References

Ben-Bassat, M. 1982. *Pattern recognition and reduction of dimensionality.* In *Handbook of Statistics, P.P. Krishnaiah and L.N. Kanal (eds), pp. 773-791, North Holland.*

Blum, A.I. and Langley, P. 1997. *Selection of relevant features and examples in machine learning. Artificial Intelligence, 97:245-271.*

Brank, J., Grobelnik, M., Milic-Frayling, N., Mladenic, D. 2002. *Feature selection using linear support vector machines. Microsoft Research Technical Report MSR-TR-2002-63, 12 June 2002.*

Burge, C. and Karlin, S. 1997. *Prediction of complete gene structures in human genomic DNA. J. Mol. Biol., 268:78-94.*

Das, S. 2001. *Filters, wrappers and a boosting hybrid for feature selection.* In *Proc. 18th Intl. Conf. on Machine Learning.*

Degroeve, S., De Baets, B., Van de Peer, Y. and Rouzé, P. 2002. *Feature Subset Selection for Splice Site Prediction. Bioinformatics, 18-2:75-83.*

Doak, J. 1992. *An evaluation of feature selection methods and their application to computer security. Technical Report CSE-92-18, Univ. of California at Davis.*

Fickett, J.W. and Tung, C.S. 1992. *Assessment of protein coding measures. Nucleic Acids Res., 20(24):6441-50.*

Grate, L.R., Bhattacharyya, C., Jordan, M.I. and Mian, I.S. 2002. *Simultaneous relevant feature identification and classification in high-dimensional spaces.* In *Proceedings of the Workshop on Algorithms in Bioinformatics, WABI 2002.*

Guyon, I., Weston, J., Barnhill, S. and Vapnik, V. 2000. *Gene selection for cancer classification using support vector machines. Machine Learning, 46:389-422.*

Hall, M.A. and Smith, L.A. 1999. *Feature selection for machine learning: comparing a correlation-based filter approach to the wrapper.* In *Proceedings of the Florida Artificial Intelligence Symposium (FLAIRS-99).*

Hermes, L. and Buhmann, J.M. 2000. *Feature Selection for Support Vector Machines.* In *Proceedings of the International Conference on Pattern Recognition (ICPR'00)-Volume 2 September 03 - 08, 2000 Barcelona, Spain.*

Inza, I., Larrañaga, P., Blanco, R. and Cerrolaza, A.J. 2003. *Filter versus wrapper gene selection approaches in DNA microarray domains. Artificial Intelligence in Medicine, accepted.*

Hebsgaard, S.M., Korning, P.G., Tolstrup, N., Engelbrecht, J., Rouzé, P. and Brunak, S. 1996. *Splice site prediction in Arabidopsis thaliana pre mRNA by combining local and global sequence information. Nucleic Acids Res., 24:3439-3452.*

Kohavi, R. and Pfleger, K. 1994. *Irrelevant features and the subset selection problem. Proc. 11th Intl. Conf. on Machine Learning, pp. 121-129.*

Kohavi, R. and John, G. 1997. *Wrappers for feature subset selection. Artificial Intelligence Journal, 97:273-324.*

Koller, D. and Sahami, M. 1996. *Toward optimal feature selection.* In *Proceedings Thirteenth International Conference on Machine Learning (1996):pp. 284-292.*

Kudo, M. and Sklansky, J. 2002. *Comparison of algorithms that select features for pattern classifiers. Pattern Recognition, 33:25–41.*

Larrañaga, P. and Lozano, J.A. 2001. *Estimation of Distribution Algorithms. A New Tool for Evolutionary Computation, Kluwer Academic Publishers.*

Lukashin, A.V. and Borodovsky, M. 1998. *GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res., 26:1107-1115.*

Mathé, C., Sagot, M.F., Schiex, T. and Rouzé, P. 2002. *Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Res., 30(19):4103-17.*

Pedersen, A.G., Baldi, P., Chauvin, Y. and Brunak, S. 1999. *The biology of eukaryotic promoter prediction–a review. Comput Chem. 1999 Jun 15;23(3-4):191-207.*

Raudys, S.J. and Jain, A.K. 1991. *Small sample size effects in statistical pattern recognition: recommendations for practitioners. IEEE Trans. Pattern Anal. Machine Intell., 13-3: 252-264.*

Rombauts, S., Florquin, K., Lescot, M., Marchal, K., Rouzé, P. and Van de Peer, Y. 2003. *Computational approaches to identify promoters and cis-regulatory elements in plant genomes. Plant Physiol. 132:1162-1176.*

Saeys, Y., Degroeve, S., Aeyels, D., Van de Peer, Y. and Rouzé, P. 2003. *Fast feature selection using a simple Estimation of Distribution Algorithm: a case study on splice site prediction. Bioinformatics 19-2: 179-188.*

Salzberg, S.L., Pertea, M., Delcher, A.L., Gardner, M.J. and Tettelin, H. 1999. *Interpolated Markov models for eukaryotic gene finding. Genomics, 59:24-31.*

Schiex, T., Moisan, A. and Rouzé, P. 2003. *EuGène: an eukaryotic gene finder that combines several sources of evidence.* In *Gascuel, O. and Sagot, M.-F. (eds), Lecture Notes in Computer Science, Vol. 2006, First International Conference on Biology, Informatics, and Mathematics, JOBIM 2000, Springer-Verlag, Germany, pp. 111-125.*

Sindhwani, V., Bhattacharyya, P. and Rakshit, S. 2001. *Information theoretic feature crediting in multiclass support vector machines.* In *SIAM International Conference on Data Mining, Chicago, USA, April, 2001, Apr 2001.*

Solovyev, V.V., Salamov, A.A. and Lawrence, C.B. 1995. *Identification of human gene structure using linear discriminant functions and dynamic programming.* In *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology (eds. C. Rawling et al.) AAAI Press, Menlo Park, CA, pp. 367-375..*

Sonnenburg, S. 2002. *New Methods for Splice Site recognition. Diploma thesis, Humboldt-Universität zu Berlin, 2002.*

Vafaie, H. and De Jong, K. 1993. *Robust feature selection algorithms.* In *Proceedings of the Fifth International Conference on Tools with Artificial Intelligence, pages 356–363.*

Xing, E.P., Jordan, M.I. and Karp, R.M. 2001. *Feature Selection for High-Dimensional Genomic Microarray Data.* In *Proc. 18th International Conf. on Machine Learning, pp. 601–608.*

Yu. L. and Liu, H. 2003. *Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution.* In *Proceedings of The Twentieth International Conference on Machine Leaning (ICML-2003), pages 856-863, Washington, D.C..*

Zeng, F., Yap, R.H.C, and Wong, L. 2002. *Using feature generation and feature selection for accurate prediction of translation initiation sites. Genome Informatics, 13:192-200.*

Zhang, M.Q. 1997. *Identification of protein coding regions in the human genome by quadratic discriminant analysis. Proc. Natl. Acad. Sci. 94: 565-568.*

Zhang, M.Q. 2002. *Computational prediction of eukaryotic protein-coding genes. Nat. Rev. Genet.,3:698-709.*

Zien, A., Raatsch, G., Mika, S., Schoelkopf, B., Lengauer, T. and Mueller, K.R. 2000. *Engineering support vector machine kernels that recognize translation initiation sites. Bioinformatics, 16:799-807.*