Telomerase Gene Prediction Using Support Vector Machines

David Luper [1] **and Spandana Makeneni** [2]
[1]Computer Science Department, UGA, Athens, GA, USA
[2]Institute of Bioinformatics, Complex Carbohydrate Research Center, UGA, Athens, GA, USA

Abstract - *Telomerase genes have been said to be of great importance in various aspects of biology. Currently their composition and purpose is a topic of much research. Finding and validating telomerase genes in different species is of great importance and is also a difficult task that consumes many resources. In this research a method for isolating potential telomerase gene regions within a genome is discussed. A Support Vector Machine will be used to differentiate regions of DNA containing telomerase genes from those that do not. The Support Vector Machine will be trained on identified telomerase genes from related species, and then it will be used to classify sequences encompassing an entire chromosome of a different species as either potential telomerase gene regions or non-telomerase regions. Ultimately, a fast algorithm is presented that can act as an initial filter to remove large portions of a genome, allowing more time intensive routines to better target optimal regions of a genome.*

Keywords: Data Mining, Computational Biology, Machine Learning

1 Introduction

Telomerase (Fig. 1), also called telomere terminal transferase, is an enzyme made of protein and RNA subunits that dictates the synthesis of telomere terminal repeats. This mechanism is required for the maintenance of chromosome termini, as the structure and integrity of telomeres are essential for genome stability. Telomere deregulation can lead to cell death, cell senescence, or abnormal cell proliferation. It has been identified that telomerase plays very important roles in aging and cancer. Telomerase activity is detected during development and has a very low, almost undetectable, activity in somatic (body) cells. These somatic cells age as a result of telomerase inactivity. So, if telomerase is activated in a cell, the cell will continue to grow and divide leading to exciting possibilities. In the past several years of research, it has been found that cancer cells are immortal and divide uncontrollably. Such immortal cancer cells have 10-20 times more active telomerase than in normal body cells. Reducing this activity could eventually lead to the death of those cells.

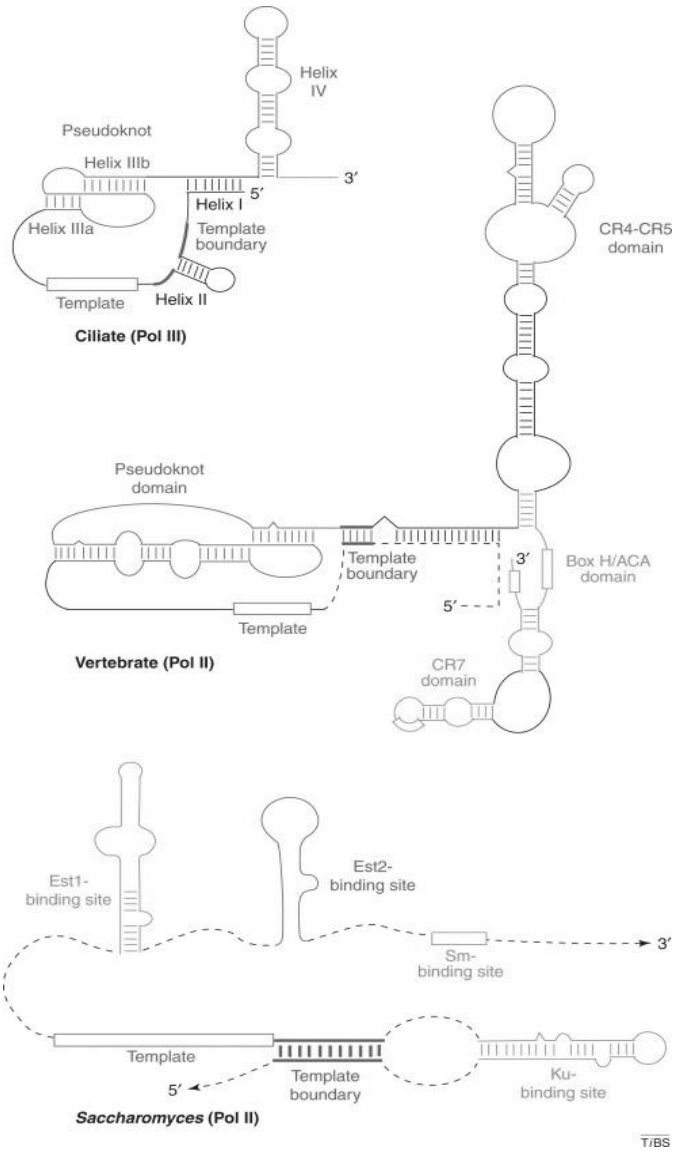

Figure 1. Detailed telomerase RNA secondary structure for humans and yeast

This could be a great therapy especially in the early stages of cancer. In the later stages, inhibition or absence of telomerase may result in cell crisis in cancer cells and tumor regression in cancer patients. Research on telomerase continues to be a very exciting field with potential for discovering many more facts about what might help fight cancer and the aging process.

In this research both yeast telomerase genes and telomerase genes of vertebrates are used to teach a supervised machine learning algorithm what telomerase genes look like. Once the algorithm builds a model to represent these genes, it can look through entire genomes to narrow down the search for new telomerase genes in species they have not been identified. The type of supervised machine learning algorithm used for this research is a Support Vector Machine (SVM). Support Vector Machines have been used for instance classification in complex biological domains with great effectiveness [1] [3]. SVMs have been shown to obtain better results over a wide variety of problems in comparison with other algorithms used in supervised machine learning. This is because they generalize better due to the nature of how they learn.

This paper will show how an SVM can be used to narrow the search for new telomerase genes. It will be laid out in the following manner. First, supervised machine learning and support vector machines will be briefly discussed. After this, the methodology section will outline the steps used in this research to isolate regions of chromosomes labeled as having potential to house a telomerase gene. Next, the experiments from this work will be presented along with results. Finally, future enhancements to the methodology will be discussed before concluding remarks.

2 Machine Learning

Support Vector Machines are a type of supervised machine learning algorithm. Supervised machine learning algorithms are used to approximate non-linear functions for instance classification. These algorithms build models from a group of data instances called training data, and use these models to classify new instances where the class is not known. Each data instance in the training data consists of n features, from an n dimensional feature space S, and a label that tells the algorithm which class the data instance belongs to. These instances describe locations for each class in S, and they are treated as a representation of a non-linear function f(i) where i is a an input vector of features. Once the training data is assembled a model is constructed. While constructing the model a portion of the training data is placed into another data set called the validation data. The validation data is withheld from the learner while training it and used to test how effective the model generalizes to instances outside of the training data. There are different schemes for segmenting and utilizing the validation data, this research uses a method called n fold cross validation. N fold cross validation divides the training data into n data sets and builds n – 1 models where one of the n data sets is used as the validation data. The n – 1 models are combined to produce a single model that can be used for classification. This technique helps the model generalize better when there are relatively few instances in the training data. Once the final model is built, it can be tested with a separate group of disjoint data instances called production data. These instances are labeled as belonging to a particular class, but this label is withheld from the algorithm during classification to see how accurately the model approximates the targeted non-linear function on data it has never seen.

SVMs have been shown to obtain better results over a wide variety of problems in comparison with other algorithms used in supervised machine learning. This is because they generalize better, due to the nature of how they learn. SVMs learn concepts by separating data distributions into classes of data and treating them as two generalized sets of vectors in a feature space. The SVM will find a separating hyperplane between these two datasets (or concepts) which is the maximum distance from either of the two (Fig. 2). Other machine learning algorithms can find hyperplanes that separate datasets but the power of the SVM comes from the fact that the hyper plane found by the SVM is the one with the greatest distance between either of the two classes. The SVM finds support vectors, which are data instances from either class that are the closest to the opposite class. Once these support vectors are found, geometric operations are applied to find the hyperplane that is equally distant from both sets of support vectors. Finding support vectors and computing the maximum marginal hyperplane is a standard quadratic programming problem [4]. This explanation assumes a linearly separable feature space because that is the easiest way to explain the concept. SVMs can be generalized to support nonlinear features spaces as well as more than two classes of data, but these topics are beyond the scope of this paper.

3 Methodology

Given feature sets representing data instances, SVMs learn concepts and identify instances as belonging to specific classes of data. This project uses SVMs for locating regions within chromosomes that have potential to contain telomerase genes. The classes of data instances in this project are + and – where + is a segment of chromosome that potentially holds a telomerase gene and – is any other region of chromosome. The training data used to construct the classification model for the SVM is used to tell the SVM what telomerase regions look like (+), and what they do not (-). Since each species has only one telomerase gene region, telomerase gene regions from related species are used to in the training data as + instances. The - instances in the training data were randomly sampled non-telomerase regions from the same group of species. Five times as many – instances were included in the training data as available + instances. The production data

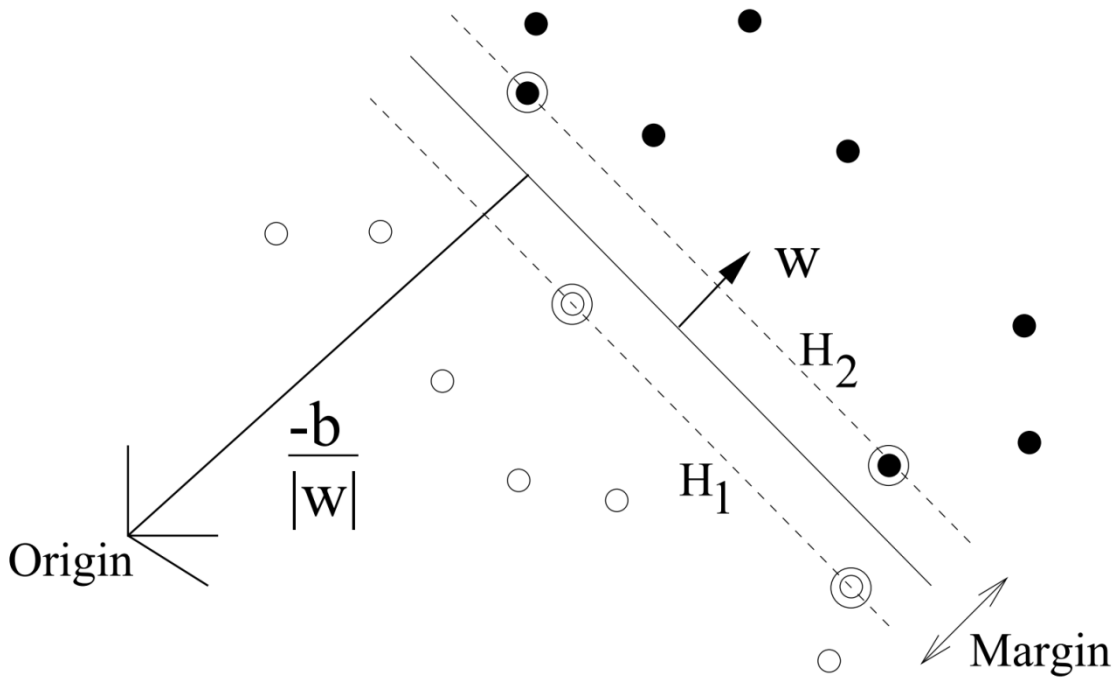

Figure 2. Illustration of what a maximum marginal hyperplane looks like between two set of data instances in a feature space. (Image taken from Christopher J.C. Burges [4])

used to test the correctness of the SVM was an entire chromosome from species X, where X was related to the species in the training data, and the telomerase gene region for X had already been positively identified. This production data allows the correctness of the SVM to be effectively measured by whether the SVM successfully classifies the telomerase region from species X and how much of the rest of the chromosome from species X gets correctly labeled non-telomerase. The data for this research was gathered from two sources. The telomerase gene regions were obtained from http://telomerase.asu.edu/sequences.html, and the chromosome in which those regions reside were taken from ftp://ftp.ncbi.nih.gov/.

During the construction of data sets for this research it was imperative that the telomerase gene region from the species being used in the production data not be included in the training data. The inclusion of this telomerase region would skew the results for the experiment as the SVM would be trained specifically on one of the instances it is also being evaluated for correctness on. This kind of scenario leads to over fitting on the training data and as a result an SVM generalizes poorly.

To obtain data instances for submission to the SVM, features are computed from segments of DNA. For the + examples used in the training data, the segments of DNA used were the telomerase genes from each of the species involved in defining the + examples. For the - examples used in the training data, and the examples used in the production data the chromosomes were segmented into regions of length m using a sliding window over the chromosome. This sliding window was started at index n = 0 and between segments n was incremented by x. For this project x was set to 75 and m was set to the average length of the telomerase regions used as + instances in the training data. After the necessary chromosomes were segmented and assigned to the training and production data sets features for the segments could be calculated.

There are nine features used in this research to classify instances. These features were either taken from or inspired by Guo et al [5] and Schattner [6]. Schattner's work was of particular use to this research. In Schattner's paper the base composition of sequences are used to determine RNA gene regions. Schattner only uses statistical analysis of these regions to infer their class, but these features work very well for machine learning. The features used by Schattner are (G+C)%, (G-C)%, (A-T)%, and RO(AB). The features used in this research are the following:

Percentage A:

The percentage nucleotides in the DNA sequence that were A.

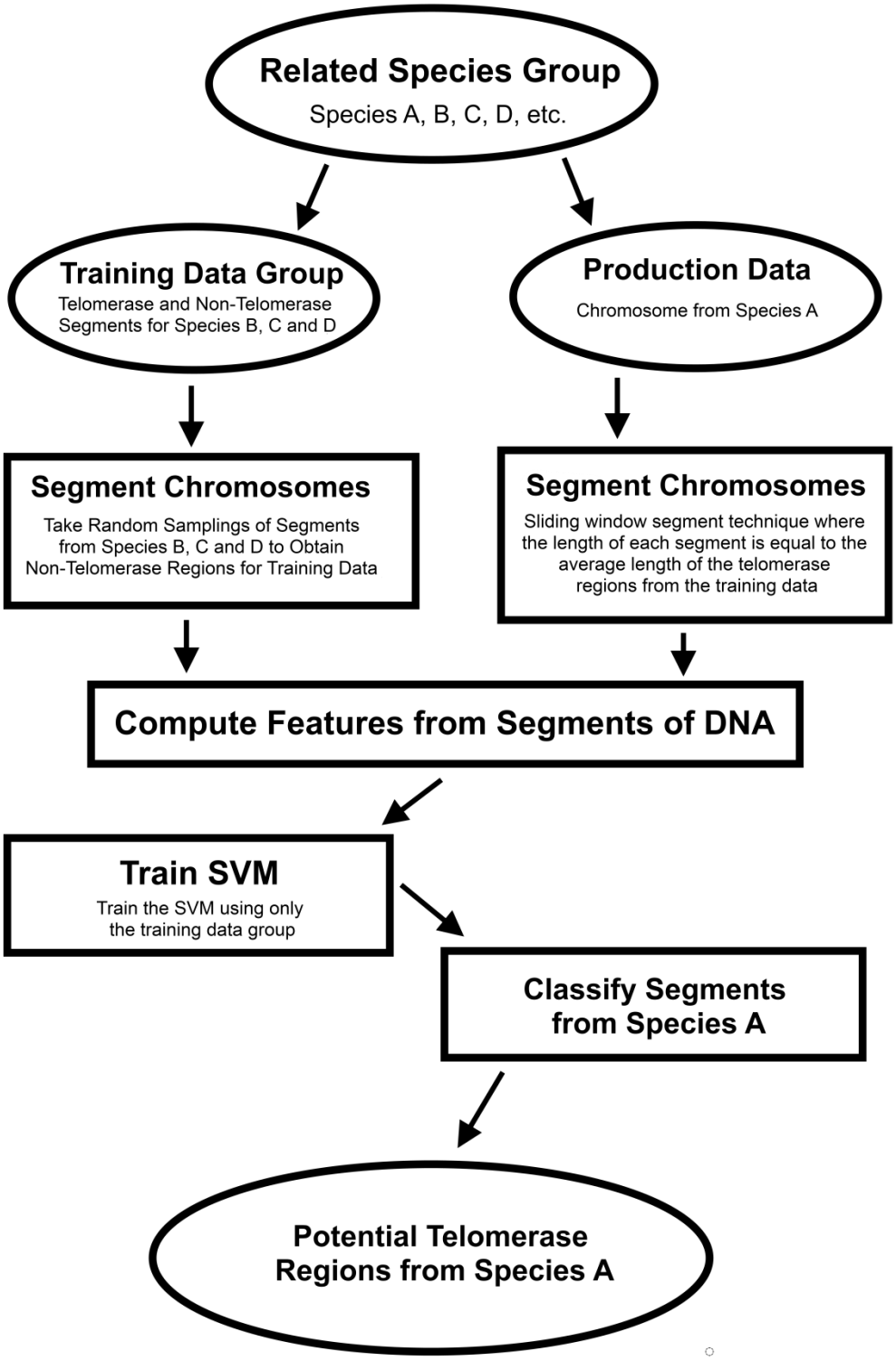

Figure 3. Methodology flow diagram to illustrate the process of obtaining potential telomerase gene regions.

Percentage T:

The percentage nucleotides in the DNA sequence that were T.

Percentage G:

The percentage nucleotides in the DNA sequence that were G.

Percentage C:

The percentage nucleotides in the DNA sequence that were C.

Percentage (X + Y):

The percentage of the nucleotides in the DNA sequence that were either X or Y summed, with this feature each possible combination of nucleotides were computed.

Percentage (X – Y):

The percentage of the nucleotides in the DNA sequence that were X subtracted from the percentage of nucleotides in the sequence that were Y, with this feature each possible combination of nucleotides were computed.

Percentage (X / Y):

The percentage of the nucleotides in the DNA sequence that were X divided by the percentage of nucleotides in the sequence that were Y, with this feature each possible combination of nucleotides were computed.

RO(XY):

The frequency count of XY (FREQ_XY) multiplied by the length of the sequence then divided by the percentage X times the percentage Y.

ex. (length * FREQ_XY) / (Percentage X * Percentage Y)

Standard Deviation:

The standard deviation of the percentages of A, T, G, and C.

Once the features are computed for each of the DNA segments the SVM can be trained. Due to the small size of the training data, cross fold validation was used to help prevent over fitting. After the SVM was trained the production data was classified, and then overlapping segments of + classifications were merged together. This results in regions of DNA, of various lengths, that potentially house the telomerase gene. The number of nucleotides in the calculated regions can be used against the total number of nucleotides in the entire chromosome to compute the percentage of the chromosome classified + or -.

4 Experiment and Results

The experiments for this research were run in two different groups (vertebrates and fungi). A flow diagram outlining the experimental procedure can be seen in Fig. 3. The *training data* for each group consisted of + instances of telomerase genes from as many related species as could be found. For each group three experiments were run. The experiments consisted of removing species *X* from the *training data* for use as the *production data*. After training the SVM, results were obtained from classification of the entire chromosome containing the telomerase gene region from species *X*. The results were defined by the recall and precision of the classification of gene regions in the chromosome. The recall was whether the SVM classified the telomerase region in species *X* correctly, and the precision was how much of the rest of the chromosome was classified correctly as non-telomerase. For this experiment species *X* had to meet two constraints. First, its telomerase gene region must be known, and second, the rest of the chromosome in which the telomerase region resided must have been sequenced. For vertebrates the three species experimented on were Mus musculus, Rattus norvegicus and Equus caballus, and for fungi the three species were Schizosaccharomyces pombe, Saccharomyces cerevisiae and Kluyveromyces lactis. The results are shown in table 1. The results show the number of potential telomerase regions detected and the percentage of the chromosome those regions accounted for. The percentage of the chromosome the potential telomerase regions account for minus one depicts the amount of the chromosome that is excluded from being a potential telomerase region. This shows how far the SVM narrowed the search for the telomerase gene. In each of the experiments run, the SVM classified the actual telomerase gene correctly. This puts the recall at 100%. Since there is only one telomerase gene within a genome for any species, the percentage of the genome classified as potential telomerase regions can be seen as the false positive rate, within a very small statistical margin of error. The results show significant information gain. However, the results on the vertebrates are significantly better than the results on the fungi. Possible explanations for this could be that the groups of vertebrates used in the *training data* were more closely related. This could make their telomerase genes more alike and provide a better representation for the SVM. Another more likely explanation could be that the results on the vertebrates were better because the SVM had more data to learn from with the vertebrates. The number of known telomerase gene regions in the *training data* for the vertebrates was 22, but only 13 telomerase gene region examples were available for the fungi. A final explanation for the better results on the vertebrates could be that in the vertebrates the telomerase genes were simply more distinct from the rest of the chromosome than they were in the fungi.

Species	# of Regions Classified +	Percentage of Chromosome Classified +
Schizosaccharomyces pombe	680	0.25377804949519833
Saccharomyces cerevisiae	165	0.44209262916606207
Kluyveromyces lactis	196	0.30870646879168767
Mus musculus	960	0.011700037718866478
Rattus norvegicus	1536	0.010321241324534453
Equus caballus	777	0.024617685567403513

Table 1

5 Future Research

Future research should be invested in at least two areas for this work. First, the SVM used in this project utilized default settings in the WEKA machine learning software package (i.e. complexity parameter and a linear kernel). Different settings for these parameters such as an RBF kernel, or different numeric values for the complexity parameter, should be explored to see if the results for the experiment could be improved. Second, new features should be explored to see if they can better detect potential telomerase regions. One such feature could reflect base pairings within the sequence of DNA. Telomerase genes should have a unique and learnable base pairing signature that sets them apart from the rest of the chromosome (i.e. the way a telomerase region folds to create its secondary structure should be distinctive). Another feature that should be looked into would be to isolate the most commonly repeated *l-mer* in the + examples from the training data and provide the number of times the particular subsequence (either exactly or with some accommodation for mutation allowed) appears in the instance. A third feature would be to create a multiple alignment from the + instances in the training data to obtain a median string (or consensus string) used for computing global and local alignment scores for each instance. In telomerase genes from related species there should exist conserved regions, and thus telomerase genes could have a unique scoring signature against this median string.

6 Conclusion

The work presented in this paper provides substantial results showing an SVM can definitively narrow the search for telomerase genes within a genome. A methodology has been outlined that segments a chromosome into DNA sequences that are treated as data instances in a machine learning application. Features are computed from these DNA sequences and the feature vectors are classified by an SVM as either potential telomerase gene regions (+) of not (-). The results from the experiment show significant information gain, however they have potential to be improved through the exploration of new features and parameter refining in the SVM.

7 References

[1] Terrence S. Furey, Nello Cristianini, Nigel Duffy, David W. Bednarski, Michel Schummer, David Haussler, 2000, Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data, Vol. 10 (2000), 906 – 914

[2] Simon Tong, Edward Chang, Support Vector Machine Active Learning for Image Retrieval, 2001, ACM International Conference Proceedings, Vol. 1, 107 – 118

[3] Isabelle Guyon, Jason Weston, Stephen Barnhill, Vladimir Vapnik, 2002, Gene Selection for Cancer Classification Using Support Vector Machines, Machine Learning, 1 – 39

[4] C.J.C. Burges, A Tutorial on Support Vector Machines for Pattern Recognition, Data Mining and Knowledge Discovery, vol. 2, no. 2. pp. 121-167, 1998.

[5] Feng-Biao Guo, Hong-Yu Ou, Chun-Ting Zhang, 2003, ZCURVE: A New System for Recoginizing Protein-Coding Genes in Bacterial and Archaeal Genomes, Nucleic Acids Research, 2003, Vol. 31, No.6, 1780 – 1789

[6] Peter Schattner, 2002, Searching for RNA Genes Using Base Composition Statistics, Nucleic Acids Research, 2002, Vol. 30, No. 9, 2076 – 2082

[7] Tom M. Mitchell, [1997], Machine Learning, International Edition, MIT Press and The McGraw-Hill Companies, Inc.

[8] Kim N.W. Clinical implications of telomerase in cancer (1997) European Journal of Cancer Part A, 33 (5), pp. 781-786.

[9] Toren Finkel, Jan Vijg & Jerry W. Shay. Time, tumours and telomeres. Meeting on Cancer and Aging

[10] Jiunn-Liang Chen, Carol W. Greider, Telomerase RNA structure and function: implications for dyskeratosis congenita, Trends in Biochemical Sciences, Volume 29, Issue 4, April 2004, Pages 183-192, ISSN 0968-0004, DOI:10.1016/j.tibs.2004.02.003.