B-factor profile prediction for RNA flexibility using Support Vector Machines

Ivantha Guruge[1], Ghazaleh Taherzadeh[1], Jian Zhan[1], Yaoqi Zhou[1], and Yuedong Yang[1,2]

[1]School of Information and Communication Technology and Institue for Glycomics, Griffith University, Parklands Drive, Southport, Queensland 4215, Australia. [2]School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510275, China.

Correspondence to: Yaoqi Zhou (E-mail: yaoqi.zhou@griffith.edu.au) and Yuedong Yang (Email: yangyd25@mail.sysu.edu.cn).

ABSTRACT

Determining the flexibility of structured biomolecules is important for understanding their biological functions. One quantitative measurement of flexibility is the atomic Debye-Waller factor or temperature B-factor. Most existing studies are limited to temperature B-factors of proteins and their prediction. Only one method attempted to predict temperature B-factors of ribosomal RNA. Here we developed and compared machine-learning techniques in prediction of temperature B-factors of RNAs. The best model based on Support Vector Machines yields Pearson's correction coefficient at 0.51 for five-fold cross validation and 0.50 for the independent test. Analysis of the performance indicates that the model has the best performance on rRNAs, tRNAs and protein-bound RNAs, for long chains in particular. The server is available at http://sparks-lab.org/server/RNAflex.

Introduction

Three-dimensional structures of proteins and RNA determined by X-ray crystallography are the average positions of atoms. Thermal fluctuation around the average positions can be measured by temperature B-factor, or Debye-Waller factor[1]. Atoms with high B-factor values are in general more flexible. Structural flexibility and dynamic motions are essential for protein catalysis and allostery[2] and for secondary structure formation and the folding of RNA catalysts as well as protein–RNA recognition[3]. Temperature B-factors have been used in many applications including analysis of protein active sites[4], protein disorder regions[5] and protein thermal stability[6].

Importance of temperature B-factors has led to development of methods for sequence and structure-based analysis and predictions. If a protein structure is known, molecular dynamics simulations have been used to correlate root-mean-squared fluctuations and temperature B-factors[7]. However, MD simulations are too time consuming to perform large-scale studies. As a result, various simple models using normal-mode analysis[8], graph theory[9], mean-field theory[10,11] have been developed. Normal-mode analysis, for example, achieved correlation coefficients between experimental B-factor and computational calculations at about 0.6[12-14]. Interestingly, similar correlation can be achieved by using a weighted contact number[15].

For sequence-based approaches, various machine models such as Support Vector Machines (SVM)[16], Random Forest (RF)[17], Support Vector Machines[18,19], and neural network[20] were employed. These sequence-

based methods with real-value prediction achieved a correlation coefficient between predicted and actual B-factors at about 0.5.

Unlike multiple methods developed for protein temperature B-factors, there is only one study that predicted ribosomal RNA B-factor profiles based on their sequence and structure information[21]. The correlation coefficient between predicted and actual temperature-B-factor is 0.39 for the best sequence-based method and 0.48 for the best structure-based method. Only a small dataset was used (13 crystal structures of ribosomal 50S subunits) without performing necessary normalization for consistence of structural flexibility across different structures that are crystallized at different conditions and environments[24]. More importantly, there is no web-based server available for the academic community.

The objective of this work is to establish a method for predicting temperature B-factor based on RNA sequence information only. Developing a sequence-based method for B-factor prediction is important because the vast majority of genome were coded for RNAs, rather than proteins and more and more of these non-coding RNAs are found functional[22]. Compared to proteins, RNAs are more difficult to fold into unique three-dimensional structures (i.e., inherently more flexible) because the differences between four bases are small (all hydrophilic with similar sizes). In fact, only a few hundreds of non-redundant RNA structures were determined in high resolution. Majority of RNAs without known structures make it imperative to develop sequence-based methods.

In this study, we have built a non-redundant dataset of 142 RNA structures that are randomly separated into training (108 RNAs) and test (34 RNAs) sets. We examined single-sequence and sequence-profile-based features and several machine-learning techniques (Support Vector Machines (SVM), Artificial Neural Networks (ANN) and Random Forest (RF) for their ability to predict temperature B-factors. The best model was based on SVM and produced a Pearson's correlation coefficient (PCC) of 0.51 for the five-fold cross validation and 0.50 for the independent test set.

Methods

Data Sets. We have obtained 1093 RNA structures deposited in protein databank with resolution <3.0 Å and RNA chains longer than 32 bases (March 2015). To avoid over training, we have removed redundant chains by the program cd-hit-est with the lowest allowed sequence-identity cut-off at 80%[23]. A total of 142 RNA chains were obtained. We randomly divided these chains into roughly 75% for training and cross validation (108 chains, 26764 bases) and 25% for test (34 chains, 5989 bases). The B-factor for each nucleotide is represented by atom type C1.

B-factor Normalizations. B-factors are not consistently measured across different structures because different refinement procedures and different temperatures are used for structure determination[24]. Thus B-factor normalization is necessary so that the relative flexibility for a given RNA sequence is used for method development[25]. Here, the data was normalized according to the method of Smith et al[26]. In this method, outliers were first detected and removed by using a median based approach. Next, the mean and standard deviation of B-factors in an RNA structure was calculated. The normalized B-factors for a given RNA structure are the deviations of the raw B-factors from the mean divided by the standard deviation. The normalized B-factor profile for the training set falls roughly between -3.00 and 4.00.

Single-sequence Input Features. A simple method to represent the RNA sequence is to use vector-based orthogonal codes[27] in which A, U, G, and C are represented by four-dimensional vectors of (1000), (0100), (0010), and (0001), respectively.

Evolution-based sequence profile. Evolution-based sequence profiles have been found useful in predicting protein temperature B-factors (e.g. [20]). This is because sequence conservations in regions with different flexibility have different patterns. In a previous study, we have obtained evolution-based sequence profiles[28] by querying the RNA sequences against RNA sequence library using BLASTN[29] with E-value < 0.001 and maximum of 50,000 homologous sequences[30]. The *j* base probability (*j* = A, T/U, G, C) in multiple aligned homologous sequences at a given position *i*, $P_{i,j}$ was calculated as $P_{i,j} = -\log[(N_{i,j})/\Sigma_j(N_{i,j})]$, where $N_{i,j}$ is the number of observed base type *j* at position *i*. In order to avoid zero values, a small number correction $s(b_i)$ was used in $N_{i,j}$ based on the normalized expected average occurrences for the types (native base and other types). $s(b_i)$ was set to 0.3 for the other base type b_i and 9.0 for the query base type. The obtained sequence profiles were normalized to a range of (-1, 1) before employed for training and test[28].

Window-based features. To predict the B-factor of a given RNA base, we also input the sequence or sequence-profile information of neighboring RNA bases. The size of the sequence window is optimized for training and cross validation.

Support Vector Machine (SVM). We employed LibSVM[31] to build the predictive SVM models based on radial basis function (RBF) kernel. Support Vector Regression (SVR)[32] was employed to predict the real value of B-factors. The two parameters of RBF (*gamma* and *C)* were optimized by a grid search to find the best model that produced the best performance for five-fold cross validation. The optimal values for the gamma and C parameters were 0.03125 and 1, respectively.

Artificial Neural Network (ANN). ANNs are made of interconnected multi-layer units to learn non-linear relations between input and desired output. Deep learning or structured learning is ANNs with three or more hidden layers. We employed a rectified linear unit (ReLU) as the activation function except that tanh was utilized for the activation in the output layer. Stochastic gradient descent algorithm (Adam) was used for weight optimization. We also examined the effect of different number of hidden layers (1 to 4), number of neurons (100 and 800), number of epochs (10 and 100), and the size of mini-batches within an epoch which is, the number of samples over which we average to find the updates to weights/biases needed to descend the gradient (50 and 300). A grid search was implemented to find the best parameters above. The final model employed 1 hidden layer, 600 neurons, 50 epochs, and 50 bases per Mini batch. The learning rate was set to 0.001. Here we have used ANN implemented in the Keras software package which is a high-level neural networks API[33].

Random Forest (RF). Random forests or random decision forests[34,35] are ensemble learning methods for classification, regression and other tasks. RF creates a multitude of decision trees when training the model and outputs the classes (classification) or mean prediction (regression) of the individual trees. Each individual tree is trained using a subset of the train set and is evaluated on the test set. The final prediction output is then calculated by combining the output of all trees. The RF model parameters are optimized by setting the minimum number of samples at a leaf node, the number of trees in the forest, the maximum number of features in each tree and a function to measure the quality of split. A grid search was implemented to find the optimized values for each parameter. Thus, the best model is trained using 1 minimum leaf size and 1000 trees with the maximum number of features in each tree set to 'Auto'. The RF was implemented using the Scikit-Learn machine learning library[36].

Cross-Validation Test. We performed the 5-fold cross validation test on the training set. Here the dataset was randomly divided into five folds with similar number of chains. Each fold was selected as the test set while the other folds were used as the training set. This process was repeated five times so that every fold was tested once.

Independent Test. Once the best model was identified, the trained model was tested with our independent test set which was not used when training the model. Similar performance in 5-fold cross validation and independent test would indicate robustness of the model trained.

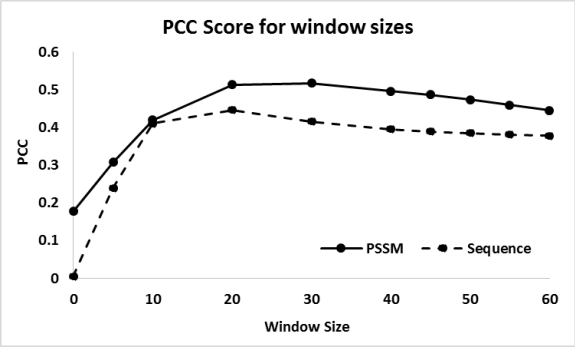

Figure 1 Pearson's correlation coefficient (PCC) between predicted and actual temperature B-factors as a function of window sizes for five-fold cross validation on the train set by using SVM models. Single-sequence and sequence-profile-based models are shown in dashed and solid lines, respectively.

Performance Evaluation Criteria. The overall performance of the SVM model is assessed by Pearson's Correlation Coefficient (PCC) between predicted and actual values of temperature B-factors[37].

Results

Figure 1 displays PCC as a function of the window size based on a SVM model and five-fold cross validation on the train set. There are two curves: one is based on single-sequence features only and the other is based on evolution-derived sequence profile. There is a fast increase in PCC as the window size increases and the change is small after the window size is greater than 20. We found that the best window sizes are 20 and 30 for SVM models based on single sequence and sequence profiles, respectively.

Table 1 compares the performances of SVM models based on single sequence and sequence profile, respectively. The performances in 5-fold cross validation and independent test are similar for both single-sequence and sequence-profile based techniques, suggesting the robustness of the model in its application to unseen data. The performance by the profile-based method is significantly better than the single-sequence based method: the PCC value increases from 0.4 to 0.5. This highlights that the sequence conservation plays a significant role in chain flexibility.

Table 1: Performance by using different features and models.

Input /Model	5-fold PCC	Test set PCC
Sequence/SVM	0.4467	0.4640
PSSM/SVM	0.5176	0.5028
PSSM/RF	0.4908	0.5036
PSSM/DNN	0.4791	0.4439

In addition to SVM models, we also employed Neural Networks and Random Forests based on sequence profiles as input. We also found a window size of 30 as the optimal widow size for NN and RF models. Results for five-fold cross validation and independent tests are shown in Table 1. The performances of the three models are similar with SVM having the best performance.

Figure 2 compares predicted to actual normalized temperature B-factors given by single-sequence-based and profile-based SVM models for the test set for all bases in the test set. It is clear that the profile-based method is substantially more accurate than the single- sequence-based method based on the spread of the distribution around the overall regression line.

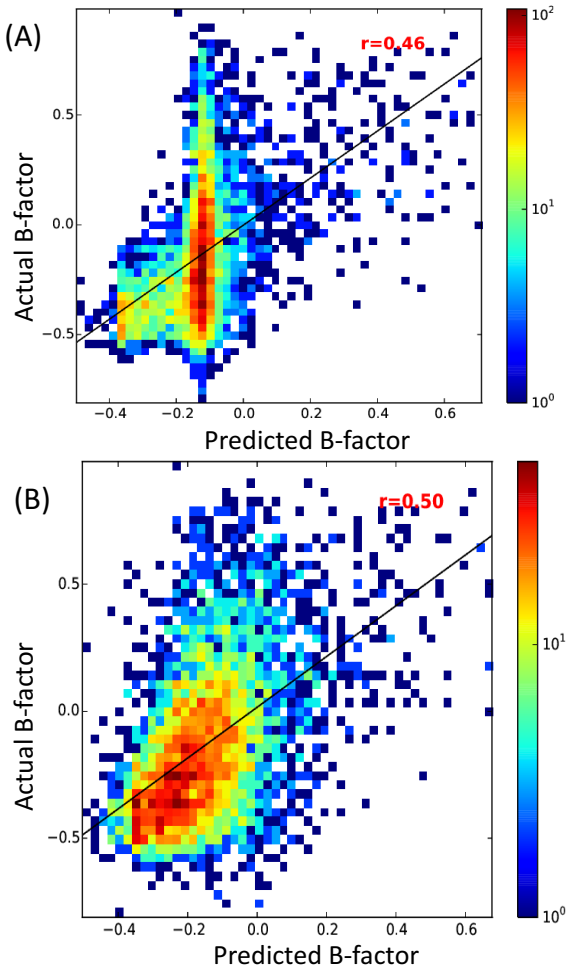

Figure 2 Density plots of actual normalized temperature B-factors against predicted temperature B-factor for the single-sequence-based (A) and profile-based (B) SVM models for the test set

To further understand the method performance for different types of RNAs, we separated RNAs into rRNA (39 chains), tRNA (46 chains), riboswitches (10 chains) and others (47 chains). Here we have combined cross-validated and test results so that we have a large dataset to analyse. We can do this because our method has similar performance in cross validation and independent test. We found that the overall PCC values are 0.58 for rRNA, 0.48 for tRNA, 0.05 for riboswitches, and 0.13 for others. Low accuracy of riboswitches is somewhat expected because they have more than one conformation. Temperature B-factors are fluctuations around a single conformation. We can also classify RNAs into those bound with proteins (93 chains) or not bound

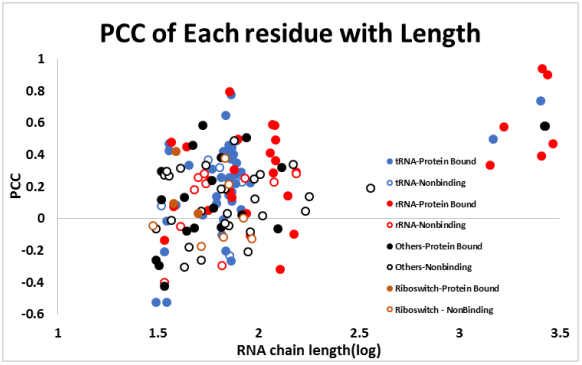

Figure 3 Pearson's correlation coefficients for different types of RNA chains as a function of chain length. Because of large size difference, X-axis is shown in a logarithmic scale

with proteins (49 chains). We found that overall PCC values are 0.56 for protein-bound RNAs and 0.11 for protein-free RNAs.

Figure 3 further shows the PCC values for individual chains as a function of RNA chain length. Longer chains have more accurately predicted temperature-B factors. The structures of all long chains are rRNA or tRNA complexed with proteins. For short chains, RNAs complexed with proteins are in general predicted more accurately than protein-free RNAs. Most difficult-to-predict RNAs are in other categories and not bound with proteins. Examples of those poorly predicted other RNAs are 1f1t (ligand-bound RNA aptamer), 1xjr (short virus RNA element) and 3p22 (short virus RNA element) with PCC of less than 0.01.

Figure 4 demonstrates one example of highly accurate prediction for a long chain. This is a 23S rRNA of thermophilic bacterium called thermus thermophiles. The peaks and valleys of thermal fluctuations were reproduced quite accurately with PCC =0.73.

DISCUSSION

In this study, we have developed a method that predicts temperature B-factors of RNAs. By using SVM models and sequence profiles, we achieved PCC values at 0.52 for five-fold cross validation and 0.50 for the independent test set, respectively. Similar performance in cross validation and test sets and

similar performance to other machine learning

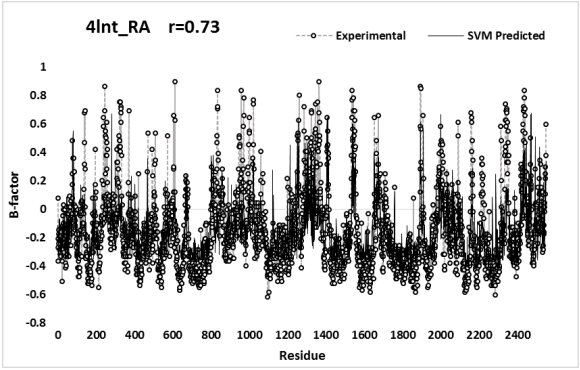

Figure 4 Comparison of predicted versus actual normalised B-factor profile of 23S rRNA of thermus thermophiles (Chain RNA in PDB 4lnt). Dashed lines with a node represent the actual temperature B-factor and solid line represents the predicted temperature B-factor. The PCC value between predicted and experimental B-factors is 0.73.

models (Table 1) confirm the robustness of the performance obtained. The performance is also similar to the published performance for a structure-based method in a prior work[21], which is limited to rRNA. Analysis of performance of our sequence-based method indicates that it provides reasonable accuracy in predicting temperature-B factors of rRNA, tRNA and protein-bound RNAs, for long chains in particular.

Predicting B-factor profiles is challenging. This is because of intrinsic noises residing within the experimental data. B-factors depend on the environment such as the temperatures, the stages of measurements and refinement methods[38-40]. The measured B-factor profiles should reflect the authentic fluctuation and static, dynamic and lattice disorders[16]. Radivojac et al. found a high correlation between homologous proteins (average PCC of 0.8)[5], which Yuan et al. believed to be the upper limit of predicting B-factor profiles[16]. All existing sequence-based techniques for predicting protein temperature B-factors yielded PCC around 0.5, regardless of the type of method used for prediction[16-18]. Similar performance achieved for RNA temperature B-factor in this study suggests that better features are required to further improve the overall performance of B-factor prediction. We have attempted to incorporate predicted secondary structures[41]. No significant improvement was observed.

Acknowledgments

This work was supported in part by National Health and Medical Research Council (1059775 and 1083450) of Australia and Australian Research Council's Linkage Infrastructure, Equipment and Facilities funding scheme (project number LE150100161) to Y.Z., National Natural Science Foundation of China (61671107) to Y.Y. We also gratefully acknowledge the support of the Griffith University eResearch Services Team and the use of the High Performance Computing Cluster "Gowonda", and the OpenEye to provide the academic license to complete this research. This research/project has also been undertaken with the aid of the research cloud resources provided by the Queensland Cyber Infrastructure Foundation (QCIF).

Keywords: RNA flexibility, temperature B-factor, Support Vectors Regression

References and Notes

1. Debye, P. Annalen der Physik 1913, 348(1), 49-92.
2. R.M. Daniel; R.V. Dunn; J.L. Finney; Smith, J. C. Annual Review of Biophysics and Biomolecular Structure 2003, 32(1), 69-92.
3. Levene, S. D. In eLS; John Wiley & Sons, Ltd, 2001.
4. Carugo, O.; Argos, P. Proteins 1998, 31(2), 201-213.
5. Radivojac, P.; Obradovic, Z.; Smith, D. K.; Zhu, G.; Vucetic, S.; Brown, C. J.; Lawson, J.

D.; Dunker, A. K. Protein science : a publication of the Protein Society 2004, 13(1), 71-80.
6. Parthasarathy, S.; Murthy, M. R. Protein engineering 2000, 13(1), 9-13.
7. Klepeis, J. L.; Lindorff-Larsen, K.; Dror, R. O.; Shaw, D. E. Curr Opin Struc Biol 2009, 19(2), 120-127.
8. Bahar, I.; Atilgan, A. R.; Erman, B. Fold Des 1997, 2(3), 173-181.
9. Jacobs, D. J.; Rader, A. J.; Kuhn, L. A.; Thorpe, M. F. Proteins-Structure Function and Genetics 2001, 44(2), 150-165.
10. Micheletti, C.; Banavar, J. R.; Maritan, A. Phys Rev Lett 2001, 87(8).
11. Pandey, B. P.; Zhang, C.; Yuan, X. Z.; Zi, J.; Zhou, Y. Q. Protein Science 2005, 14(7), 1772-1777.
12. Kondrashov, D. A.; Van Wynsberghe, A. W.; Bannen, R. M.; Cui, Q.; Phillips, G. N. Structure 2007, 15(5), 637-637.
13. Sen, T. Z.; Feng, Y. P.; Garcia, J. V.; Kloczkowski, A.; Jernigan, R. L. J Chem Theory Comput 2006, 2(3), 696-704.
14. Riccardi, D.; Cui, Q.; Phillips, G. N. Biophys J 2009, 96(6), 2548-2548.
15. Lin, C. P.; Huang, S. W.; Lai, Y. L.; Yen, S. C.; Shih, C. H.; Lu, C. H.; Huang, C. C.; Hwang, J. K. Proteins-Structure Function and Bioinformatics 2008, 72(3), 929-935.
16. Yuan, Z.; Bailey, T. L.; Teasdale, R. D. Proteins 2005, 58(4), 905-912.
17. Pan, X.-Y.; Shen, H.-B. Protein and Peptide Letters 2009, 16(12), 1447-1454.
18. Pan, Y.; Lv, F.; Tian, F.; Luo, X.; Kong, X.; Li, Y.; Yang, Q. Molecular Informatics 2010, 29(3), 195-201.
19. de Brevern, A. G.; Bornot, A.; Craveur, P.; Etchebest, C.; Gelly, J. C. Nucleic acids research 2012, 40(Web Server issue), W317-322.
20. Yaseen, A.; Nijim, M.; Williams, B.; Qian, L.; Li, M.; Wang, J.; Li, Y. BMC bioinformatics 2016, 17 Suppl 8, 281.
21. Tian , F.; Zhang, C.; Fan, X.; Yang, X.; Wang, X.; Liang, H. Molecular Informatics 2010, 29(10), 707-715.
22. Ponting, C. P.; Oliver, P. L.; Reik, W. Cell 2009, 136(4), 629-641.
23. Li, W.; Godzik, A. Bioinformatics (Oxford, England) 2006, 22(13), 1658-1659.
24. Tronrud, D. Journal of Applied Crystallography 1996, 29(2), 100-104.
25. Karplus, P. A.; Schulz, G. E. Naturwissenschaften 1985, 72(4), 212-213.
26. Smith, D. K.; Radivojac, P.; Obradovic, Z.; Dunker, A. K.; Zhu, G. Protein science : a publication of the Protein Society 2003, 12(5), 1060-1072.
27. Jonsson, J.; Norberg, T.; Carlsson, L.; Gustafsson, C.; Wold, S. Nucleic acids research 1993, 21(3), 733-739.
28. Yang, Y.; Li, X.; Zhao, H.; Zhan, J.; Wang, J.; Zhou, Y. RNA (New York, NY) 2017, 23(1), 14-22.
29. Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Nucleic acids research 1997, 25(17), 3389-3402.
30. Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Journal of molecular biology 1990, 215(3), 403-410.
31. Chang, C.-C.; Lin, C.-J. ACM transactions on intelligent systems and technology (TIST) 2011, 2(3), 27.
32. Basak, D.; Pal, S.; Patranabis, D. C. Neural Information Processing-Letters and Reviews 2007, 11(10), 203-224.
33. Chollet, F.; GitHub: GitHub repository, 2015.
34. Ho, T. K. IEEE transactions on pattern analysis and machine intelligence 1998, 20(8), 832-844.
35. Ho, T. K. Document Analysis and Recognition, 1995, Proceedings of the Third International Conference on, 1995, pp 278-282.
36. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Journal of Machine Learning Research 2011, 12(Oct), 2825-2830.
37. Pearson, K. Proceedings of the Royal Society of London 1895, 58, 240-242.
38. Frauenfelder, H.; Petsko, G. A.; Tsernoglou, D. Nature 1979, 280(5723), 558-563.

39. Ringe, D.; Petsko, G. A. Progress in biophysics and molecular biology 1985, 45(3), 197-235.
40. Rudolph, L. Qualitative Mathematics for the Social Sciences: Mathematical models for research on cultural dynamics; Routledge, 2013.
41. Lorenz, R.; Bernhart, S. H.; Honer Zu Siederdissen, C.; Tafer, H.; Flamm, C.; Stadler, P. F.; Hofacker, I. L. Algorithms Mol Biol 2011, 6, 26.