Bioinformatics: Model Selection and Scientific Visualization

Metodi Traykov[1(✉)], Radoslav Mavrevski[2], Slav Angelov[1], and Ivan Trenchev[3]

[1] New Bulgarian University, Sofia, Bulgaria
mtraykov@nbu.bg, slav_angelov@abv.bg
[2] South-West University "Neofit Rilski", Blagoevgrad, Bulgaria
radoslav_sm@abv.bg
[3] University of Library Studies and Information Technologies, Sofia, Bulgaria

Abstract. Bioinformatics is an interdisciplinary field that develops methods for understanding biological data and solving different bioinformatics problems. Therefore, mathematical models and computer modeling are widely used tools in bioinformatics. Using them, we can analyze different biological systems, predict the outputs from biological processes, or develop new experimental cases for some bioinformatics problems. Therefore, the choice of the optimal mathematical model is essential. The main feature of the selected model must be to provide a balance between the goodness of the data fitting and the model complexity. This article aims to summarize and show the basic criteria for model selection in bioinformatics to develop a reliable approach for predicting different relationships in bioinformatics. In addition, we will briefly describe the application of computer modeling in the analysis of the results obtained by optimal mathematical models for problems in bioinformatics, such as protein folding problems, which is an analysis of biological structures, model selection, bioinformatics. (Review article).

Keywords: Computer modeling · Mathematical models · Model selection · Bioinformatics

1 Introduction

The basics of bioinformatics have been laid in the early 1960s of the last century when the scientists try to study the properties of the protein sequences through complex computational methods, such as de novo sequence assembly or substitution models. The main aim of bioinformatics researches is to propose different methods and models for describing different processes in living organisms, but how to determine which method or model is optimal and how to represent its results. The answers to these questions are in statistics (model selection) and computer modeling (optimizations and computer graphics) [1–4].

The application of nonlinear regression techniques to describe experimental data is widespread across wide areas of bioinformatics. The problem of selecting an "optimal"

T. Zlateva and R. Goleva (Eds.): CSECS 2022, LNICST 450, pp. 92–101, 2022.
https://doi.org/10.1007/978-3-031-17292-2_8

model is a fundamental problem in analyzing experimental data in bioinformatics. If we have a set of competitive mathematical models, we may use different statistical information criteria to find the model that approximates the data better, the so-called model selection process, and data fitting in the statistic. These statistical information criteria are an attractive way for model selection [1, 3–10].

To find an "optimal" model in a set of candidate models, we can use different fitting approaches, such as the least-squares (LS) method and robust (stable) regression (RR), available in more statistical software packages [7]. Akaike's information criterion (AIC) and Bayesian Information Criterion (BIC) are well-known in the literature, as criteria for evaluating models from different classes [7, 10].

Therefore, using the mentioned above criteria we may determine the optimal model in a set of mathematical models for a specific task. For example, there are many mathematical models to solve the Protein folding problem, i.e. to find the protein's tertiary structure, using its primary structure (the sequence of amino acids) [11–16]. Using model selection criteria we may find the optimal model for this problem, but the question with the visualization of the results that the model will generate is still open. The visualization of the results for the protein folding problem is not so easy, because there are many approaches and models to solve the problem, such as the HP protein folding model, and they lead to different kinds of results. This is the reason to have little information about the visualization of these results in model selection [17].

2 Methods

2.1 Fitting Experimental Data

The finding of the individual "optimal" model for a specific class of tasks, can be made using the least-squares method or robust regression fitting by GraphPad Prism, Origin, SPSS, Matlab, or other statistical software (see Fig. 1) [7, 9, 10].

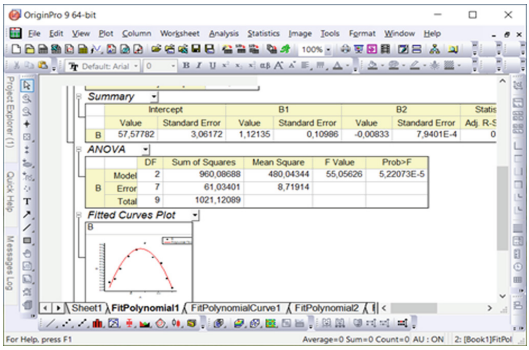
The purpose of the least-squares method is to minimize the sum of squares of the deviations between the points and the curve [5]. The deviations are the distances between Y-values.

2.2 Model Selection Criteria

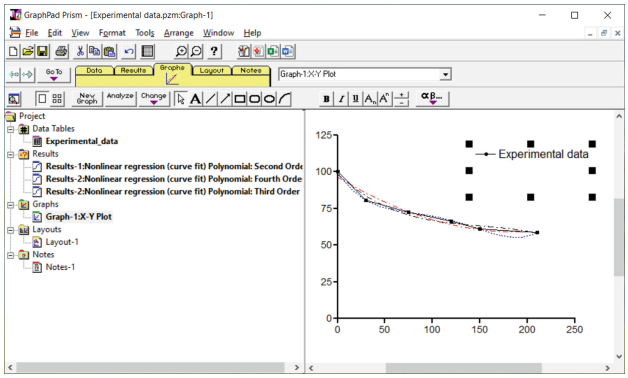
AKAIKE'S Information Criterion (AIC). The goal of this criterion is to select a model that minimizes the negative probability of penalizing. To achieve this the criterion uses the number of the parameters in the models. It is one of the most commonly used criterion:

$$AIC = \begin{cases} n * \ln\left(\frac{RSS}{n}\right) + 2 * k, \frac{n}{k} \geq 40 \\ n * \ln\left(\frac{RSS}{n}\right) + 2 * k + \frac{2*k*(k+1)}{n-k-1}, \frac{n}{k} < 40 \end{cases} \quad (1)$$

- Residual Sum of Squares (RSS) – the sum of the squares of deviations of each data point from the curve of the selected "optimal" model;
- k – the number of the parameters mapped by the regression plus 1;



(*a*)



(*b*)

Fig. 1. Examples for fitting experimental data by least squares method using Origin (a) and GraphPad Prism (b)

- n – the size of the sample.

Bayesian Information Criterion (BIC). BIC has the highest posterior probability. It is similar to AIC [5, 6]. The main difference between AIC and BIC is in the coefficient multiplied by the number of parameters. This coefficient determines how strongly the criteria will penalize large models:

$$BIC = n * \ln\left(\frac{RSS}{n}\right) + k * \ln(n). \tag{2}$$

The meaning of *RSS*, *n*, and *k* is the same as in the previous model (AIC).
Therefore, based on the above definitions, we may say that:

1. The AIC criterion does not depend directly from the sample size.
2. The model that minimizes the BIC criterion will has the highest posterior probability.
3. The BIC criterion penalizes the studied models more than AIC at an increasing number of parameters.

We may conclude that the models selected through BIC will be more parsimonious than those selected through AIC.

Software for Calculating the AIC and BIC Criteria. Below you may see the graphical user interface of a program for the calculation of both criteria (AIC and BIC), according to the mentioned above formulas. We developed this program, and it is described in details in [7]. Figure 2 shows the option in the program for calculating the AIC criterion.

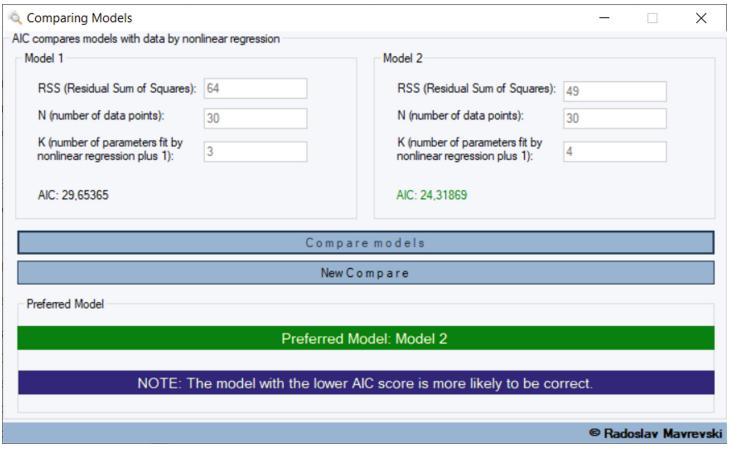


Fig. 2. Calculating the AIC criterion, using the "Comparing Models" software.

2.3 HP Protein Folding Model

The simplest and most used model for protein folding problem is the Hydrophobic-Hydrophilic/Polar model. This model divides the 20th amino acids in the human body into two groups, namely Hydrophilic (H) or Hydrophilic/Polar (P). The process of folding an amino acids sequence in a 2D or 3D lattice leads to a self-avoiding walk, where the main aim is to maximize the number of neighboring H amino acids that are not adjusted in the primary sequence. This is also known as optimal conformation.

The model can be summarized as follow:

Maximize

The contacts between the Hydrophilic amino acids (H-H contacts).

subject to

1. (Connectivity) Every two amino acids that are neighboring in the protein's sequence must occupy neighboring cells in the lattice.
2. (Non-overlapping) Two amino acids cannot share the same cell in the lattice.
3. (Assignment) Each amino acid must occupy exactly one cell in the lattice.

It is proven (using different criteria, such as biological, mathematical, and statistical) that the HP protein folding model is an optimal model and leads to an optimal solution for protein folding problem in 2D or 3D lattices [13, 18, 19].

3 Results and Discussion

The least-squares method is most widely used in different fields of bioinformatics. The model selection criteria can ignore both the statistical adequacy and the reliability of the conclusions about the problem. For this reason, the assessment needs to be made, based on more than one model selection criterion. The next figure (see Fig. 3) shows an example of curves of fitting models (polynomial curves, from 2 to 5°) and experimental data (30 points).

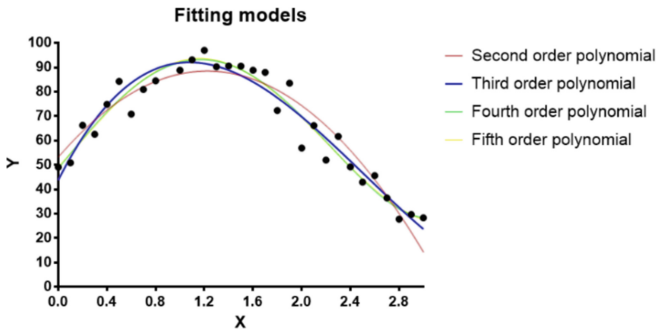


Fig. 3. Curves of fitting models (polynomials curves, from 2 to 5°) and experimental data (30 points)

Generally, the AIC and/or BIC criteria are very appropriate methods in order to select an "optimal" model with the smallest mean square error. Both criteria will choose the optimal model that has the relatively same error for each experimental data point. The criteria will return as result a value that represents a compromise between the complexity of the model (the number of parameters) and the accuracy (Table 1).

Table 1. Model selection (assessment) from different classes by **AIC** and **BIC**

Polynomial class model	Number of data points	Number of parameters	AIC value	BIC value
Second degree	30	3	128.93	132.94
Third degree		4	118.47	122.97
Fourth degree		5	**117.08**	**121.84**
Fifth degree		6	120.35	125.07

3.1 HP Protein Folding Model

There are many models to solve the protein folding problem, as we mentioned above. We may find the optimal model for this problem, using different criteria (such as AIC or BIC) and approaches (statistically, mathematical, biological, and so on.). Once we identify the optimal model, we may solve it, using optimization software (such as CPLEX or GUROBI) or implementing it in an application. In most cases the models for solving the protein folding problem generate the results as numbers, 2D or 3D coordinates of the amino acids, after that we need any additional software to draw the obtained protein conformation. The figure (see Fig. 4) below shows software with the name "HP Folding Visualization" (developed and presented by us) for visualization of a solution for the protein folding problem in the 2D HP lattice model [17].

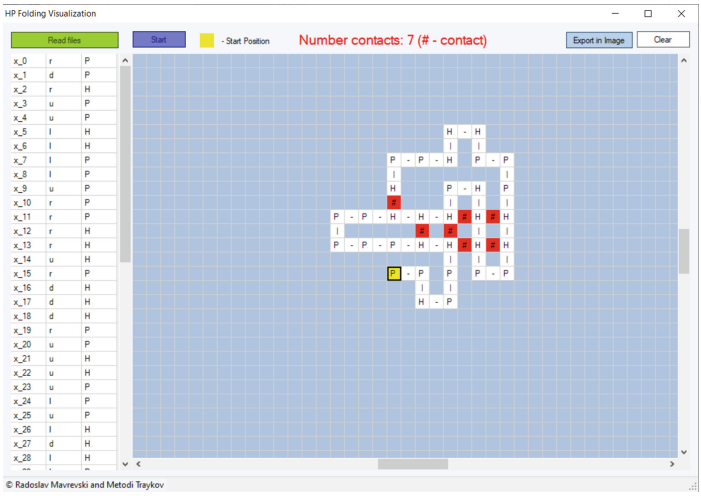


Fig. 4. The main window of "HP Folding Visualization" software.

Figure 5 shows the results (the input data for the "HP Folding Visualization" software), obtained by the selected model (2D HP lattice model) for the problem.

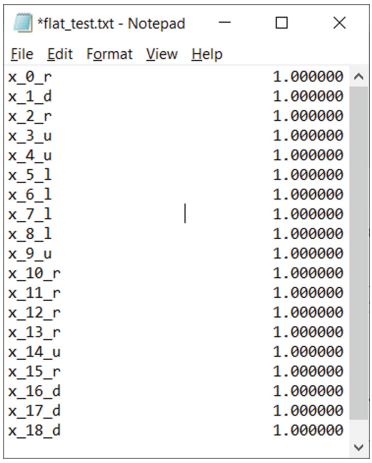


Fig. 5. Input file "flat_test.txt" for the "HP Folding Visualization" software.

Using our visualizing software, we expect to be able to shed light on the nature of these various conformational states (see Fig. 6).

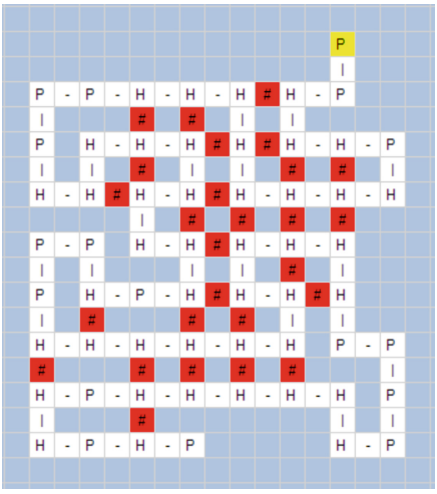
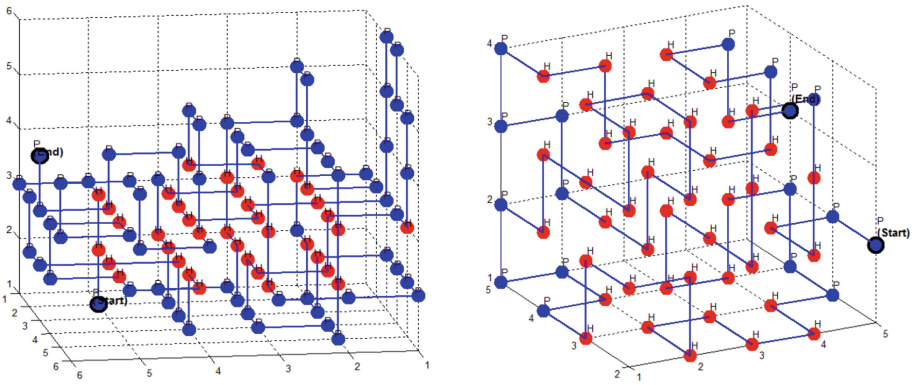


Fig. 6. Visualization of obtained solution for protein with a length of 60 amino acids (27 H-H contacts).

The next figure (see Fig. 7) shows optimal solutions in 3D. The results were obtained using an algorithm, based on HP folding model and described in [20]. The 3D visualization software follows the same approach as in the "HP Folding Visualization" software.



(*a*) (*b*)

Fig. 7. 3D Visualization of optimal solutions for protein folding problem: (a) 102 amino acids, (b) 60 amino acids (MATLAB).

Figure 8 shows a visualization of results obtained by the 3D HP off-lattice model for the problem.

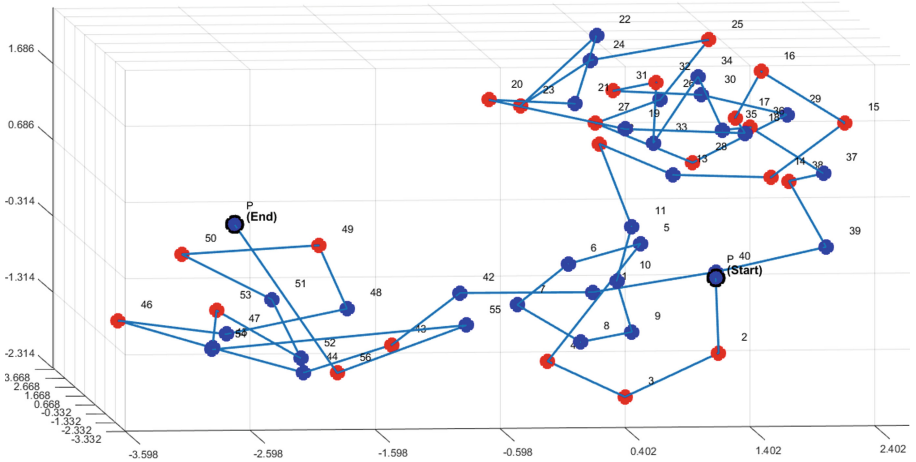


Fig. 8. Visualization of results in 3D HP off-lattice model (MATLAB).

4 Conclusions

Usually, finding a specific model in a set of models can be done through various fitting methods, e.g. the least-squares method (the most widely used) or robust fitting. Based on the results described in Sect. 3 (section "Results and Discussion") we may conclude:

1. The AIC criterion leads to relatively well results for samples with small sizes. The criterion is inconsistent.
2. The BIC criterion has poor performance in samples with small sizes, but it is consistent. This criterion leads to better results when we increase the sizes of the samples.

Generally, the current results suggest the use of the AIC criterion in the model selection process for samples with small sizes and the BIC criterion for larger samples. However, we recommend using both criteria for a more accurate assessment.

About the visualization, in the protein folding problem, developed by us visualization software are potential tools that can be helpful to study in details the folding trajectories and the number of contacts between amino acids in protein folding.

References

1. Acquah, H.: Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of an asymmetric price relationship. J. Dev. Agric. Econ. **2**, 1–6 (2010)
2. Ahn, S.J.: Geometric fitting of parametric curves and surfaces. J. Inf. Process. Syst. **4**, 153–158 (2008)
3. Akaike, H.: A new look at the statistical model identification. IEEE Trans. Autom. Control **19**, 716–772 (1974)
4. Bickel, P., Zhang, P.: Variable selection in nonparametric regression with categorical covariates. J. Am. Stat. Assoc. **87**, 90–97 (1992)
5. Burnham, P., Anderson, D.: Model Selection and Multimodel Inference, 2nd edn. Springer-Verlag, New York (2002). https://doi.org/10.1007/b97636
6. Joseph, B., Nicole, L.: Methods and criteria for model selection. J. Am. Statist. Assoc. **99**, 279–290 (2004)
7. Mavrevski, R.: Selection and comparison of regression models: estimation of torque-angle relationships. C. R. Acad. Bulg. Sci. **67**, 1345–1354 (2014)
8. Dzimbova, T., Sapundzhi, F., Pencheva, N., Milanov, P.: Computer modeling of human delta opioid receptor. Int. J. Bioautom. **17**, 5–16 (2013)
9. Mavrevski, R., Milanov, P., Traykov, M., Pencheva, N.: Performance comparison of model selection criteria by generated experimental data. In: ITM Web of Conferences, vol. 16 (2018) https://doi.org/10.1051/itmconf/20181602006
10. Mavrevski, R., Traykov, M., Trenchev, I., Trencheva, M.: Approaches to modeling of biological experimental data with graphpad prism software. Wseas Trans. Syst. Control **13**, 242–247 (2018)
11. Ahn, N., Park, S.: Finding an upper bound for the number of contacts in hydrophobic-hydrophilic protein structure prediction model. J. Comput. Biol. **17**, 647–656 (2010)
12. Alberts, B., Bray, D., Johnson, A., et al.: Essential Cell Biology: An Introduction to the Molecular Biology of the Cell. Garland Science Publishing, New York (1998)
13. Berger, B., Leighton, T.: Protein folding in the hydrophobic-hydrophilic (HP) is NP-complete. J. Comput. Biol. **5**, 27–40 (1998)
14. Carr, R., Hart, W., Newman, A.: Discrete optimization models for protein folding. Technical report SAND2002. Sandia National Laboratories (2003)
15. Chen, M., Huang, W.: A branch and bound algorithm for the protein folding problem in the HP lattice model. Genomics Proteomics Bioinf. **3**, 225–230 (2005)

16. Chandru, V., Rao, M., Swaminathan, G.: Protein folding on lattices: an integer programming approach. IIM Bangalore Research Paper No. 199 (2004)
17. Mavrevski, R., Traykov, M.: Visualization software for hydrophobic-polar protein folding model. Sci. Vis. **11**(1), 11–19 (2019)
18. Dill, K.A.: Theory for the folding and stability of globular proteins. Biochemistry **24**, 1501–1509 (1985)
19. Dill, K.A., Bromberg, S., Yue, K., et al.: Principles of protein folding. A perspective from simple exact models. Protein Sci. **4**, 561–602 (1995)
20. Yanev, N., Traykov, M., Milanov, P., Yurukov, B.: Protein folding prediction in a cubic lattice in hydrophobic-polar model. J. Comput. Biol. **24**, 412–421 (2017)