Detection of Sequential Outliers using a Variable Length Markov Model

Cécile Low-Kam, Anne Laurent, Maguelonne Teisseire

Detection of Sequential Outliers using a Variable Length Markov Model

Cécile Low-Kam
Institut de Mathématiques
et Modélisation de Montpellier
Univ. Montpellier 2 - CNRS
Pl. Eugène Bataillon, Montpellier, France
clowkam@math.univ-montp2.fr

Anne Laurent Maguelonne Teisseire
Laboratoire d'Informatique de Robotique
et de Microélectronique de Montpellier
Univ. Montpellier 2 - CNRS
161, rue Ada, Montpellier, France
laurent,teisseire@lirmm.fr

Abstract

The problem of mining for outliers in sequential datasets is crucial to forward appropriate analysis of data. Therefore, many approaches for the discovery of such anomalies have been proposed. However, most of them use a sample of known typical sequences to build the model. Besides, they remain greedy in terms of memory usage. In this paper we propose an extension of one such approach, based on a Probabilistic Suffix Tree and on a measure of similarity. We add a pruning criterion which reduces the size of the tree while improving the model, and a sharp inequality for the concentration of the measure of similarity, to better sort the outliers. We prove the feasability of our approach through a set of experiments over a protein database.

1 Introduction

Lately, mining for outliers in large sequential databases has risen as an active domain of research. By outlier, we mean "*an observation that deviates so much from others as to arouse suspicion that it was generated by a different mechanism*" [12]. Indeed, the potential applications of anomaly detection are numerous and diverse, as this is a matter of interest for areas such as knowledge discovery (Web Usage Mining) and biostatistics (DNA mutations detection). Even so, it remains a challenging problem as outliers are rare by definition [18]. Furthermore, they have to be separated from noise which arises unavoidably in a dataset: considering noise as an anomaly wrongly raises suspicion. Within the last few years, detection of outliers has been investigated for all kind of datasets. Among those approaches, some were based on discordance tests under assumptions of one probabilistic distribution of the observations, in the cases of univariate or multivariate data [4]. Others were based on a distance, allowing the detection of anomaly when confronted to multidimensional data [14].

Meanwhile, bases of DNA or protein sequences have been studied to understand biological mechanisms, for example through the extraction of motifs in [10]. Those datasets are challenging because of their structure and their size, demanding appropriate model selection. Therefore, a very efficient approach for mining for outliers in such bases is proposed in [18]. However, some issues are problematic: first, this method remains very memory-demanding, and second, it involves some previous knowledge of typical sequences to build a model. Consequently, we propose here to extend this approach to overcome both of those challenges: we use an information criterion to reduce the size of the model, while making it more accurate. We also add refined bounds for the similarities of the sequences. This leads to a better detection of anomalies among the database.

The rest of the paper is organised as follows: in Section 2 is presented an accurate method for the detection of anomalies. Our improvements are proposed in Section 3, resulting in an adequate and parsimonious method for mining for outliers in sequential databases. The pertinence of our approach is showed through experiments in Section 4, followed by a detailed analysis of all our propositions. We summarise our work and evocate potential further research in Section 5.

2 The Basic Method

In this Section we recall an existing approach for the extraction of outliers in sequential databases [18]. This work of Sun *et al.* is based on a natural assumption for sequences such as DNA or proteins: they have a property of "*short memory*" [15], that is, a variable length Markov condition: for a sequence $s = s_1...s_\ell$, if for $2 \leq i \leq \ell$, $P^T(s_i|s_1...s_{i-1})$ is the probability that s_i follows $s_1...s_{i-1}$, then there is an integer $1 \leq K \leq i-1$ such that

$$P^T(s_i|s_1...s_{i-1}) = P^T(s_i|s_{i-K}...s_{i-1}). \quad (1)$$

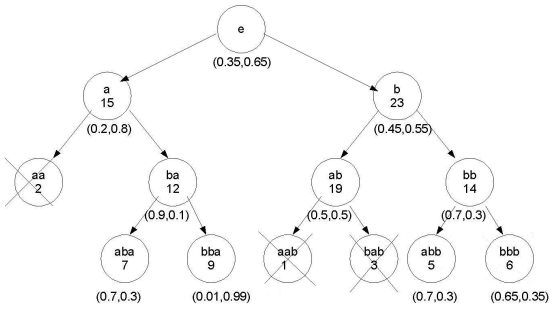


Figure 1. An example of PST on the binary alphabet.

It is a Markov property of order K. It is said to be variable as K is not fixed. An usual representation of such a model is a *suffix tree*,where each node has its greater suffix as a parent. Each leaf represents a memory of the associated Markov chain. This model is for example used for the classification of protein sequences in families [6]. Indeed, it allows to estimate the probability of each sequence in the set.

2.1 Probabilistic Suffix Tree (PST)

Let us consider a set S of sequences over a finite alphabet Σ. The probability of a sequence s is denoted by $P^T(s)$. A PST is a classic suffix tree, provided with conditional probabilities for each node. More precisely each node contains the count of its associated sequence s in the database and a vector of length $|\Sigma|$ of the conditional probabilities $P^T(\sigma|s)$ for each $\sigma \in \Sigma$. As the size of the tree grows exponentially with the length of the memory, it is pruned. To this end, a maximum length L for the tree is fixed. And nodes with low frequency, considered as negligible, are removed.

Example 2.1 *Figure* 1 *shows a PST on the alphabet* $\{a, b\}$. *A maximum length* $L = 3$ *is fixed.* 4 *is the minimal frequency for the sequences. The nodes* aa, aab *and* bab *have been pruned because they are too rare.*

Once the PST has been constructed, the conditional probabilities associated to its nodes are used to estimate the distribution of each sequence, as for any sequence $s = s_1...s_\ell$:

$$P^T(s_1...s_\ell) = P^T(s_\ell|s_1...s_{\ell-1})..P^T(s_2|s_1) \times P^T(s_1) \quad (2)$$

Example 2.2 *Using the former PST,*

$$\begin{aligned} P^T(aabb) &= P^T(b|aab) \times P^T(b|aa) \times P^T(a|a) \times P^T(a) \\ &= P^T(b|ab) \times P^T(b|a) \times P^T(a|a) \times P^T(a) \\ &= 0.5 \times 0.8 \times 0.2 \times 0.35. \end{aligned} \quad (3)$$

Thus, the probability of each sequence is computed according to the PST. Afterwards, a measure of similarity is introduced, in order to discreminate between typical and non-typical observations.

2.2 Measure of Similarity and Information Theory

In [18], for each sequence $s = s_1...s_\ell$, the measure of similarity is defined as:

$$SIM_N(s) = \frac{1}{\ell} \log P^T(s_1...s_\ell). \quad (4)$$

To allow the computation of any new sequence's similarity, the conditional probabilities of the PST are smoothed, thus, infinity is avoided. The measure SIM_N is normalised, and therefore it is not biased by the length of the sequence. Moreover, when under appropriate assumptions, it has an interesting aymptotic property. First, the sequences are supposed to be generated by an information source. This implies that they take their values in a finite alphabet, and that their distribution is stationary, *i.e.* it does not change over time. Second, let us remind that the *entropy* of a random variable is a measure of regularity [3]. This notion can easily be extended to two or more variables, through the concept of joint and conditional distribution, leading to joint and conditional entropies. The uncertainty of an information source is then defined as the limit of the conditional entropy.

Let us suppose that the sequences of the base S have been generated by an unique ergodic information source, then Shannon-McMillan theorem [16] states that $-SIM_N$ is consistent towards the uncertainty of the source. A proof of this result can be found in [3]. Therefore, when ℓ is large, $SIM_N(s)$ should be close to minus the uncertainty of the source, if s has indeed been generated by it. Otherwise, $SIM_N(s)$ will lay far from the similarity of other sequences. Consequently, a concentration inequality such as Chebyshev's is used to determine bounds outside which sequences are likely to be anomalies. However, this inequality is known to be ill-fitted for points far from the mean, which actually are the potential outliers.

In [18], experiments on a protein database are led with success. But they lay on the prior knowledge of which sequences are typical, as only those are used to build the PST. Indeed, first, a model is built over the typical sequences, and second, the authors determine whether new sequences are anomalies with respect to the given model. But we wish to directly extract the outliers from the set of sequences. For all that, however well-fitted the method presented in this section may be to point out structural differences between families of proteins, it fails when one wishes to truly mine for outliers. Moreover, though partially pruned, the tree remains space-consuming. Therefore, the aim of the present

paper is the improvement of the approach proposed in [18]. We determine whether a sequence s is an outlier given a base S and a threshold t. (Our approach may be easily extended to find the top-n outliers in S.)

3 Our Extension

In this section, we detail our approach. Namely, we introduce:

- A further pruning of the tree, with an information criterion, inducing systematic discovery of anomaly, thanks to an adequate and reduced model.
- An exponential concentration inequality for the measure of similarity, leading to sharper bounds and therefore better discrimination of outliers.

We adopt the same hypotheses of short memory property and stationarity, and also consider the case of a finite alphabet Σ.

3.1 Pruning the PST with Akaike's Information Criterion

We have seen in Section 2 that a PST is pruned in two steps: a maximal length L is fixed. Any node deeper than L is pruned. Any node with a lower frequency in the dataset than a given threshold is pruned. In addition to the two pruning procedures we have seen so far, in [15] a PST is built as follows: a node is added only if it differs statistically from its parent. To this end, a criterion based on the Kullback-Leibler information is used. The Kullback-Leibler information or distance is sometimes called relative entropy, and denotes the information lost when a distribution is used to approximate another [8]. The statistic of error for a letter σ and a sequence s, denoted by $Err(\sigma s, s)$, is defined in [15] by the Kullback-Leibler distance between the distributions $P^T(.|\sigma s)$ and $P^T(.|s)$, pondered by the probability of σs. If $Err(\sigma s, s)$ is less than a given threshold, the node σs is pruned. Nodes corresponding to strings which observation probabilities are weak are also pruned, no matter how they differ from their parent: being rare, they are considerated as negligible. In [18], only this last criterion is applied. But there could be cases where the compared distributions differ at a deeper level. Therefore, in [15], all the potential descendants of each pruned node are also tested.

Example 3.1 *Let us consider again the binary suffix tree of Figure* 1. *Suppose the threshold is* 0.1.
$Err(abb, bb) = 0.001$*, and* $Err(bbb, bb) = 0$*. Nodes* abb *and* bbb *are thus pruned, because their vector of conditional probabilities is similar to their parent's: therefore no additional knowledge is gained using a memory of order* 3 *instead of one of order* 2*.*

This last pruning method is not used in [18]. For this purpose, we recall a well-known criterion Akaike called An Information Criterion (AIC) and introduced in [2]. It allows to balance between the fit of a model and its complexity. Let $\mathcal{L}$ be the likelihood function of a model, and k its number of parameters. Then the AIC is:

$$AIC = 2k - 2\log \mathcal{L}. \tag{5}$$

The AIC is related to the Kullback-Leibler information. It is based on the maximum likelihood estimate of the model. The correcting term asymptotically unbiases the estimate. This criterion allows to compare the distance of two possible models from the "true" unknown one, then choose the closest (see [8] for details). Therefore, over a set of candidate models, we should select the one with the lowest AIC. In practice, the AIC may perform poorly when the number of parameters is important with respect to the size of the dataset. This problem was lifted in [17]. Therefore, the Second-Order Information criterion was defined in [13] by:

$$AIC_c = AIC + \frac{2k(k+1)}{n-k-1}, \tag{6}$$

where n is the length of the data. The AIC_c performs well, no matter how many parameters are in the model. For all that, we always use this corrected version in our experiments.

We apply this criterion in two steps. First, let us denote by M_L the markovian model of fixed order L. We select the "best" global model among the set $\{M_L, L \geq 0\}$, according to the criterion. Therefore, a maximal length for the tree is fixed. In the second step, the same criterion is applied locally at the parent-son level: let M_p be the model based on a parent and M_s the model based on its sons, then,

$$\Delta AIC_c = AIC_c(M_p) - AIC_c(M_s) \tag{7}$$

expresses the difference between the two models. We add all its sons to a parent if ΔAIC_c is greater than 0. Otherwise, we do not add any of the parent's descendants. Here the AIC differs from the Kullback-Leibler criterion, as nodes are added one by one for the latter.

Thus, we obtain a variable order Markov model. We introduce the corrected Akaike's Criterion in the algorithm proposed in [18]. Our experiments show that the size of the tree reduces drastically with this new criterion, and that the quality of the prediction is improved. In summary, besides

getting a statistical-based choice of model, we are able to lower the number of nodes of the PST, in order to gain on its size. Once the model has been selected, the conditional probabilities of the PST are used to calculate $SIM_N(s)$ for each sequence s in S.

3.2 Sharper Bounds for the Concentration of SIM_N

In the previous subsection, we have seen how similarity measures are computed. Then, in order to find the outliers among the base, an inequality of concentration is used in [18]. Let $\mathbb{E}(SIM_N)$ and $\mathbb{V}(SIM_N)$ be the expectation and the variance of the random variable SIM_N. Then Chebyshev's inequality states that

$$\mathbb{P}\left\{|SIM_N - \mathbb{E}(SIM_N)| \geq t\right\} \leq \frac{\mathbb{V}(SIM_N)}{t^2}. \quad (8)$$

The outliers should be those which similarity lay far from the mean, out of the bounds defined in (8). But, although satisfying for points close to the mean, this inequality performs poorly for observations far from it, *i.e.* potential outliers. For these points, exponential concentration inequalities are known to be more accurate. Among them is Bennett's inequality [7]:

Theorem 3.1 *Let $X_1, ..., X_\ell$ be independant real-valued random variables with zero mean, and assume that $|X_i| \leq c$ with probability one. Let $S_\ell = \sum_{i=1}^{\ell} X_i$ and $\sigma^2 = \frac{1}{\ell}\sum_{i=1}^{\ell} \mathbb{V}(X_i)$. Then for any $t > 0$,*

$$\mathbb{P}\left\{S_\ell > t\right\} \leq \exp\left(-\frac{\ell\sigma^2}{c^2}\, h\left(\frac{ct}{\ell\sigma^2}\right)\right), \quad (9)$$

where the function h is defined by $h(u) = (1+u)\log(1+u) - u$ for $u \geq 0$.

A proof of this result can be found in [9]. We consider the random variables $X_i = \log P^T(s_i|s_1...s_{i-1})$, $1 \leq i \leq \ell$. They are bounded as the conditional probabilities of the tree are smoothed. The exponential type of Bennett's inequality assure that sharper concentration results are thus obtained. Indeed, our experiments show that as far as outlier detection is concerned, bounds obtained with Bennett's perform better than those induced by Chebyshev's inequality.

4 Experiments and Analysis

In order to prove the efficiency of our approach, we have led some experiments on the Pfam database [5], which contains about 9300 families of proteins, on the alphabet of amino acids of size 20. Pfam is known to cover many protein families [10]. We use the R software [19], and its Bio3D package [11] to read data in the FASTA format. In [18], it has been observed that a good similarity measure should be able to detect the difference of structure between two families. Therefore, it has been proposed to build a PST over one family, then compute the similarities of each sequence to obtain bounds. This tree has then be used to calculate similarity measures of members of others families, in order to know how many of them fall out of bounds. We have led similar experiments, comparing results obtained when pruning or not with the corrected Akaike's Criterion, and when using Bennett's or Chebyshev's inequality. All methods have given similar satisfying results, suggesting that all Markov models of reasonable order work well for this purpose. However, our aim here is to detect what are the outliers among a set of sequences, without knowledge about which members are typical, and should thus be used to build the tree.

Table 1. $\mathrm{AIC_c}$ for Markov Models of order 0 to 3.

Model	AIC_c
M_0	10.2×10^5
M_1	6.4×10^5
M_2	1.6×10^5
M_3	3.2×10^5

Therefore, we consider the HCV_core family of the Pfam database, containing over 3000 members, to which we add a few sequences belonging to the NADHdh family. In this paper, we present the results obtained for two such datasets. The first one, denoted by D_1, contains 30 sequences from the NADHdh family, that is, about 1% of outliers. The second set D_2 contains 300 sequences from the NADHdh family, representing 10% of the total. We build a PST over those datasets, and check how well the similarity measure picks out the members of the NADHdh family).

First, we select the global model (the maximal order of the Markov chain) using the AIC_c. We consider four Markov models of increasing order $\{M_L, 0 \leq L \leq 3\}$. Table 1 summarises the results obtained for D_1. The criterion select the model of a Markov chain of order 2 corresponding to the lowest score. Let us consider the histograms of similarities obtained with M_0, M_1, M_2 and M_3: Figure 2 shows an estimation of the distribution of the similarities of both typical sequences and outliers, given one model. M_2 discriminates best between the two groups of similarities, as they are clearly separated, therefore an adequate inequality of concentration should be able to pick up well the outliers within the data, as it will be seen later. On the contrary, the other models allow similarities for both groups to overlap on each other, making the distinction hard to figure out. We see that the most complex model of the list is not ade-

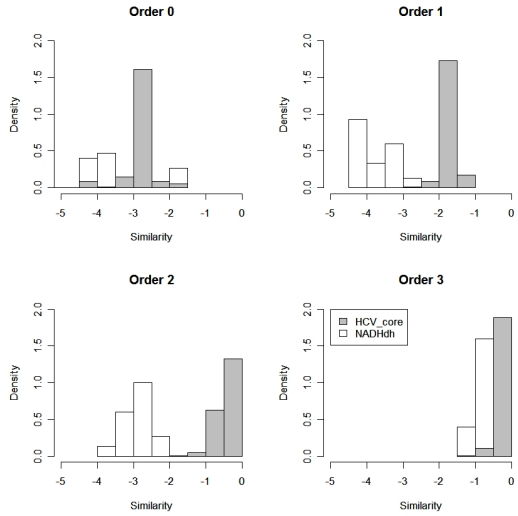
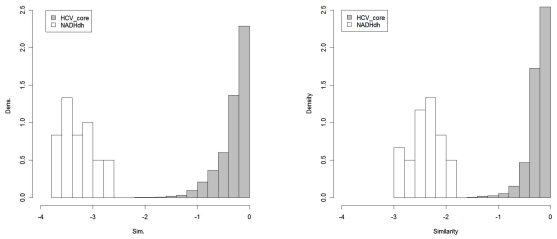


Figure 2. Comparaison of Markov models of order 0 to 3.



$L = 2$ L is not fixed

Figure 3. Histograms obtained when applying the criterion locally.

Table 2. Percentages of true and false outliers out of bounds.

Dataset	Inequality	Threshold	True	False
D_1	Chebyshev	0.11	100.0	0.3
D_1	Bennett	0.01	100.0	0.5
D_2	Chebyshev	0.11	5.0	0.7
D_2	Bennett	0.10	100.0	4.0

quate. A too simple model such as M_0 is not good either. The choice of the model should be grounded on an adequate criterion.

Let us now proceed with the second step of our pruning strategy. We have seen in Section 3 that once the maximal depth for the tree has been found, one can also use the criterion locally, for each node. We prune the PST according to the AIC_c at the local father and child level, and obtain a similar histogram. The tree has now 312 nodes instead of 368. The first histogram of Figure 3 shows a clear separation between the two groups. As the anomalies are detected all the same, and the cost of calculation for the local criterion is important (it is a sum on all the alphabet), we may wonder if it is worth it. But when dealing with large alphabets, one might wish to select a variable length Markov model level by level, without first having to fix a maximal length, then build all the tree, and eventually prune nodes that should be removed. Therefore, we build the PST on the same base of sequences only using the local ΔAIC_c criterion, that is, we do not fix a maximal depth for the tree. The PST thus obtained has only 515 nodes for a maximal depth of 3 and leads to comparable accurate results, as shown by the second histogram of Figure 3. However, it is usually advised to first select a global model [8], so this last approach must be employed with caution.

Once the model has been selected, we determine whether an observation is an outlier with respect to a given threshold. To this end, under the model M_2, we compute the bounds obtained with Bennett's inequality with a threshold correponding to the ratio of outliers introduced in the datasets. Then we compare our results to those obtained using Chebyshev's inequality with a threshold of 11% as it is recommended in [18]. This is the same as setting the bound at 3 standard-deviations from the mean. We get the results presented in Table 2: it shows what percentages of true or false outliers are extracted.

For the first dataset, both inequalities lead to comparable accurate bounds. However, the threshold for Bennett's inequality seems to make more sense, regarding the number of outliers introduced. Generally, one can not know in advance how many outliers are in the dataset. But checking the similarities on a histogram such as those presented in Figure 2 gives an intuition. For the second dataset, Chebyshev's inequality is clearly outperformed by Bennett's.

Eventually, pruning the PST using the AIC_c leads to satisfying results. But in [18], a pruning criterion solely based on frequency was proposed. One may thus wonder whether it would lead to similar detection. Table 3 summarises the results of a such pruning for D_1 and $L = 4$, using Bennett's inequality with a threshold of 1%. For thresholds lower than 15, all outliers are not detected. For a threshold greater than 15, all anomalies are out of bounds, but the tree may be larger than our optimal result. Thus, this method has comparable and sometimes even better results than our, regarding the size of the tree. However, no indication about the threshold is given, while it depends on the size of the dataset, on the number of outliers, and on the very structure of the sequences of the base. Indeed, for D_2, when fixing the minimal frequency at 15 and using Bennett's inequality

Table 3. Results with a frequency-based criterion for $\mathbf{D_1}$.

Thresholds	Number of nodes	Non-detected outliers
5	1318	0.37
10	782	0.03
15	603	0.0

Table 4. Table of Notations

PST	Probabilistic Suffix Tree
SIM_N	Measure of similarity
Σ, σ	Alphabet, letter of the alphabet
S, $s = s_1...s_\ell$	Set of sequences, one sequence
$P^T(s)$	Probability of s
M_L	Markov Model of order L
H	Entropy or Uncertainty
AIC	Akaike's Information Criterion
$\mathcal{L}$	Likelihood
k	Number of parameters of the model
D_1, D_2	Datasets with 1%, 10% of outliers

at 10%, 66% of the outliers are not outlined. The quality of the detection may thus vary whereas pruning with an information criterion allows to systematically identify the outliers.

In this section, we have presented the results of our approach on bases of proteins. We also have led similar experiments on other families in the Pfam database, leading to the same accurate detection of outliers.

5 Conclusion and Further Work

In this paper, we have provided an approach for mining for outliers in sets of sequences of data. It is an extension of the one proposed in [18]: namely, the building of a PST from the dataset and the use of a measure of similarity. However, both the important size of the tree and the exact mining of outliers remained problematic issues. Therefore, we have improved this method through a further pruning of the PST, based on Akaike's Information Criterion, in order to reduce its size and to have an appropriate model, and through the use of the exponential inequality of Bennett to get more accurate bounds. Those additions have resulted in a more efficient mining of outliers, as the quality of prediction was improved while the size of the tree remained small. We have confirmed our conjectures through a whole set of experiments on a base of protein sequences. Future work will consist in extending our method to more complex structures of data, such as sequential patterns in sequences of sets [1].

References

[1] R. Agrawal and R. Srikant. Mining sequential patterns. In P. S. Yu and A. S. P. Chen, editors, *Eleventh International Conference on Data Engineering*, pages 3–14, Taipei, Taiwan, 1995. IEEE Computer Society Press.

[2] H. Akaike. Information theory as an extension of the maximum likelihood principle. In . F. Petrox, B.N. & Caski, editor, *Second International Symposium on Information Theory*, pages 267–281, 1973.

[3] R. Ash. *Information Theory*. Interscience publishers, New York, 1965.

[4] V. Barnett and T. Lewis. *Outliers in Statistical Data*. John Wiley, 1994.

[5] A. Bateman, E. Birney, R. Durbin, S. R. Eddy, K. L. Howe, and E. L. Sonnhammer. The pfam protein families database. *Nucleic Acids Res.*, 28:263–266, 2000.

[6] G. Bejerano and G. Yona. Modeling protein families using probabilistic suffix trees. In S. Istrail, P. Pevzner, and M. Waterman, editors, *Proceedings of the 3rd Annual International Conference on Computational Molecular Biology (RECOMB)*, pages 15–24, Lyon, France, 1999. ACM Press.

[7] G. Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57:33–45, 1962.

[8] K. P. Burnham and D. R. Anderson. *Model Selection and Inference: A Practical Information-Theoretic Approach*. Springer-Verlag Telos, 1998.

[9] L. Devroye and G. Lugosi. *Combinatorial method in density estimation*. Springer, 2001.

[10] P. G. Ferreira and P. J. Azevedo. Chapter vi: Deterministic motif mining in protein databases. In F. Masseglia, P. Poncelet, and M. Teisseire, editors, *Successes and New Directions in Data Mining*. 2007.

[11] B. Grant, A. Rodrigues, K. ElSawy, J. McCammon, and L. Caves. Bio3d: An r package for the comparative analysis of protein structures. *Bioinformatics*, 22:2695–2696, 2006.

[12] D. Hawkins. *Identification of Outliers*. Chapman and Hall, 1980.

[13] C. M. Hurvich and C. L. Tsai. Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307, 1989.

[14] E. M. Knorr and R. T. Ng. Algorithms for mining distance-based outliers in large datasets. In *Proc. 24th Int. Conf. Very Large Data Bases, VLDB*, pages 392–403, 24–27 1998.

[15] D. Ron, Y. Singer, and N. Tishby. The power of amnesia: Learning probabilistic automata with variable memory length. *Machine Learning*, 25(2-3):117–149, 1996.

[16] C. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 and 623–656, July and October 1948.

[17] N. Sugiura. Further analysis of the data by akaike's information criterion and the finite corrections. *Communications in Statistics: Theory and Methods*, 7:13–26, 1978.

[18] P. Sun, S. Chawla, and B. Arunasalam. Mining for outliers in sequential databases. In *SDM*, 2006.

[19] R. D. C. Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2006. ISBN 3-900051-07-0.