Amino Acids (2015) 47:329–333
DOI 10.1007/s00726-014-1862-4

ORIGINAL ARTICLE

Identification of mitochondrial proteins of malaria parasite using analysis of variance

Hui Ding · Dongmei Li

Received: 29 September 2014 / Accepted: 27 October 2014 / Published online: 11 November 2014

Abstract As a parasitic protozoan, *Plasmodium falciparum* (*P. falciparum*) can cause malaria. The mitochondrial proteins of malaria parasite play important roles in the discovery of anti-malarial drug targets. Thus, accurate identification of mitochondrial proteins of malaria parasite is a key step for understanding their functions and finding potential drug targets. In this work, we developed a sequence-based method to identify the mitochondrial proteins of malaria parasite. At first, we extended adjoining dipeptide composition to *g*-gap dipeptide composition for discretely formulating the protein sequences. Subsequently, the analysis of variance (ANOVA) combined with incremental feature selection (IFS) was used to pick out the optimal features. Finally, the jackknife cross-validation was used to evaluate the performance of the proposed model. Evaluation results showed that the maximum accuracy of 97.1 % could be achieved by using 101 optimal 5-gap dipeptides. The comparison with previous methods demonstrated that our method was accurate and efficient.

Keywords Malaria parasite · Mitochondrial protein · Analysis of variance · *g*-Gap dipeptide

H. Ding (✉)
Key Laboratory for Neuro-Information of Ministry of Education, Center of Bioinformatics, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China
e-mail: hding@uestc.edu.cn

D. Li (✉)
College of Sciences, Inner Mongolia University of Technology, Hohhot 010051, China
e-mail: lidongmeid@imut.edu.cn

Abbreviations

ANOVA	Analysis of variance
auROC	Area under the receiver operating characteristic
IFS	Incremental feature selection
MCC	Matthews correlation coefficient
ROC	Receiver operating characteristic
Sn	Sensitivity
Sp	Specificity
Acc	Overall accuracy
SVM	Support vector machine

Introduction

Malaria is one of the serious infection diseases caused by *Plasmodium falciparum* (*P. falciparum*). According to the report from World Health Organization (WHO) in 2014, over 207 million peoples were infected by malaria and 627,000 cases died in 2012. Commonly, female anopheles mosquito transmits this disease. Finding drug targets from *P. falciparum* proteome is a key step for the treatment of malaria. The data from GenBank shows that the *P. falciparum* genomic data can translate more than 5,000 proteins (Coordinators 2014). It is a challenge for all wet-experiment scholars to find potential targets by using biochemical experiments. Thus, computational methods are widely used.

As one of the important organelles in a cell, mitochondrion produces energy and regulates cellular metabolism. It has been reported that there are no dramatic similarities between mitochondrial proteins of *P. falciparum* and human homologs (Vaidya and Mather 2009). Thus, proteins located in the mitochondrion of *P. falciparum* have been considered as potential drug targets. Accurate identification of mitochondrial proteins of *P. falciparum* is a key step for screening anti-malaria drug targets.

In the past 10 years, many methods have been proposed to predict mitochondrial proteins (Guda et al. 2004; Kumar et al. 2006; Lin et al. 2013). In 2003, a neural network-based method was developed for the prediction of mitochondrial transit peptides in *P. falciparum* (Bender et al. 2003). By using amino acid composition in the N-terminal regions, the method achieved a Matthews correlation coefficient (MCC) of 0.74. Subsequently, Verma et al. (2010) constructed a web server called PFMpred to predict the mitochondrial proteins of *P. falciparum*. An MCC of 0.81 was achieved with an accuracy of 92 %. Jia et al. (2011) extracted the features from the N- and C-terminal sequences of proteins to improve the successful predictive rate of mitochondrial proteins. Recently, Chen et al. (2012) developed an increment of diversity-based method to predict the mitochondrial proteins of malaria parasite. Although the aforementioned methods could yield encouraging results, the accuracies of these methods are still far from satisfactory.

This paper aims to improve the prediction accuracy of mitochondrial proteins of the malaria parasite. Firstly, the *g*-gap dipeptide composition was introduced to formulate the protein samples. Secondly, ANOVA was proposed to optimize the features. Finally, the support vector machine (SVM) was used to perform the prediction. The jackknife cross-validation was used to evaluate the anticipated accuracy of the predictor. Moreover, we compared the performance of our method with other published methods.

Materials and methods

Benchmark dataset

The benchmark dataset used in this study was obtained from (Bender et al. 2003) and can be expressed as

$$S = S_{\text{mito}} \cup S_{\text{non-mito}}, \tag{1}$$

where S_{mito} contains 40 mitochondrial proteins and $S_{\text{non-mito}}$ contains 135 non-mitochondrial proteins. The subcellular localizations of the above 175 proteins have been confirmed by experiments.

General dipeptide composition

Generally, one of the key steps in protein prediction is to use an effective mathematical expression to formulate proteins. In a sequence-based predictor, the most important issue is the way in which to extract features from primary sequences of proteins. Currently, various features have been proposed to reflect the correlation between the intrinsic features of the sequence and the protein types to be predicted. A common strategy is to use amino acid compositions as inputting features (Lin and Chen 2011). Unfortunately, the sequence order information is lost. To overcome the shortcoming, the adjoining dipeptide composition was proposed to represent protein sequences (Ding et al. 2014a). However, the adjoining dipeptide composition cannot reflect the correlation between two amino acids with an interval of g residues. Based on the above analysis, we defined a general dipeptide composition called g-gap as

$$P = [f_1^g, f_2^g, \ldots, f_\varepsilon^g \ldots f_{400}^g]^{\mathrm{T}}, \tag{2}$$

where f_ε^g is the frequency of the ε-th ($\varepsilon = 1, 2, \ldots, 400$) g-gap dipeptide and can be calculated by

$$f_\varepsilon^g = n_\varepsilon^g/(L - g - 1), \tag{3}$$

where n_ε^g is the occurrence number of the ε-th g-gap dipeptide; L is the length of the protein. It should be noted that $g = 0$ represents the adjoining dipeptide composition.

Analysis of variance (ANOVA)

Theoretically, if all features were selected to predict mitochondrial proteins of malarial parasite, the prediction model would not be the best one with maximum accuracy because of overfitting, information redundancy and dimension disaster in the model (Ding et al. 2012). To overcome these disadvantages, it is necessary to develop feature selection techniques to pick out the optimal features. Recently, the ANOVA has been proposed for feature selection in protein prediction (Ding et al. 2013, 2014b). Thus, the ANOVA was adopted in this study.

Based on the theory of ANOVA statistics, the score (F) of the ε-th g-gap dipeptide is defined as

$$F(\varepsilon) = \frac{\sum_{i=1}^{K} m_i \left(\frac{\sum_{j=1}^{m_i} f_\varepsilon^g(i,j)}{m_i} - \frac{\sum_{i=1}^{K}\sum_{j=1}^{m_i} f_\varepsilon^g(i,j)}{\sum_{i=1}^{K} m_i} \right)^2 / df_{\mathrm{B}}}{\sum_{i=1}^{K}\sum_{j=1}^{m_i} \left(f_\varepsilon^g(i,j) - \frac{\sum_{j=1}^{m_i} f_\varepsilon^g(i,j)}{m_i} \right)^2 / df_{\mathrm{W}}}, \tag{4}$$

where $df_{\mathrm{B}} = K-1$ and $df_{\mathrm{W}} = M-K$ are the degree of freedom for the sample variance between groups and the degrees of freedom for the sample variance within groups, respectively; $K = 2$ represents the number of groups; $M = 175$ denotes the total number of samples; $f_\varepsilon^g(i, j)$ is the frequency of the ε-th g-gap dipeptide of the j-th sample in the i-th group, defined in Eq. 3; m_i denotes the number of samples in the i-th group (here $m_1 = 40$, $m_2 = 135$). According to the theory of statistics, the $F(\varepsilon)$ in Eq. (4) obeys an F sampling distribution with df_{B} and df_{W} degrees of freedom under the null hypothesis.

It is obvious that the larger the value of $F(\varepsilon)$, the better the discriminative capability of the ε-th feature. Therefore, we ranked the features according to their F values. Subsequently, we used the incremental feature selection (IFS) to determine the optimal features (Ding et al. 2014b).

Support vector machine (SVM)

The support vector machine is a popular machine learning method widely used in bioinformatics (Ding et al. 2012, 2013, 2014a, b; Guo et al. 2014; Hayat et al. 2014; Jia et al. 2011, 2014; Lin and Chen 2011; Nanni et al. 2014; Saha et al. 2014). Many soft packages of SVM algorithm have been launched. In the current study, the LibSVM (version 3.18) was adopted to implement SVM (Fan et al. 2005). A grid search method was used to optimize the regularization parameter C and kernel parameter γ through jackknife cross-validation. The search spaces for C and γ were $[2^{15}, 2^{-5}]$ and $[2^{-5}, 2^{-15}]$ with the steps of 2^{-1} and 2, respectively.

Performance evaluation

In statistical prediction, three cross-validation methods, namely, independent dataset test, sub-sampling (e.g., 2-, 5- or 10-fold cross-validation) test, and jackknife cross-validation, are often used to evaluate the performances of the predicted methods (Ding et al. 2014a; Guo et al. 2014; Hayat et al. 2014; Qiu et al. 2014). The jackknife cross-validation can always yield a unique result for the given data. During the process of jackknife cross-validation, each protein is singled out in turn as a testing sample, and the remaining proteins are used as the training set to calculate the testing sample's membership and predict the class. The jackknife cross-validation was used in this study.

The four metrics, namely, sensitivity (Sn), specificity (Sp), overall accuracy (Acc) and Mathew's correlation coefficient (MCC), were used to measure the performance of the predictor and, respectively, defined as follows:

$$\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{5}$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}, \tag{6}$$

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}, \tag{7}$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{(\mathrm{TP} + \mathrm{FN}) \times (\mathrm{TN} + \mathrm{FN}) \times (\mathrm{TP} + \mathrm{FP}) \times (\mathrm{TN} + \mathrm{FP})}, \tag{8}$$

where TP denotes the number of correctly predicted mitochondrial proteins; FN denotes the number of mitochondrial proteins predicted as non-mitochondrial proteins; FP denotes the number of non-mitochondrial proteins predicted as mitochondrial proteins; and TN denotes the number of correctly predicted non-mitochondrial proteins.

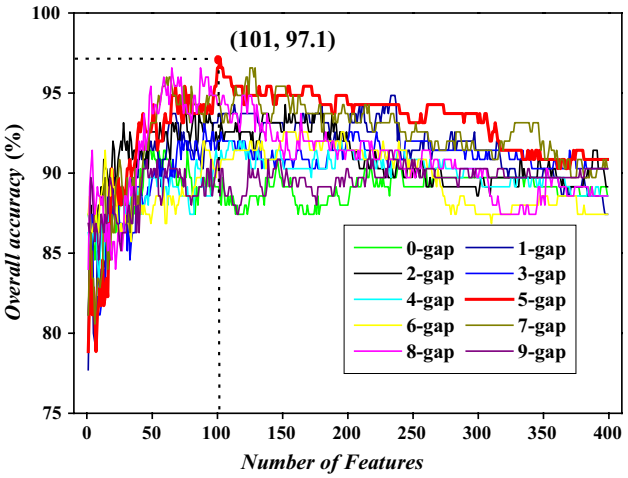

Fig. 1 A plot showing the ten IFS curves for $g = 0$–9 dipeptides. The number of features is set as x axis and the overall accuracy as y axis. When the *top* 101 5-gap dipeptides were used to perform prediction, the overall success rate reached an IFS peak of 97.1 % (color figure online)

The receiver operating characteristic (ROC) curves were plotted to study the performance of models across the entire range of SVM decision values. The area under the receiver operating characteristic curve (auROC) was calculated for objectively evaluating the performance of the proposed method.

Results and discussion

Feature selection for improving accuracy

We initially investigated the performance of adjoining dipeptide composition (0-gap dipeptide composition) with jackknife cross-validation and obtained an overall accuracy of 88.6 %. The 400 features would not only lead to the overfitting problem, but also bring about information redundancy or noise. Thus, we used the ANOVA to rank the 400 adjoining dipeptides. According to the IFS approach, the feature subset started from a feature with the highest F value in the ranked feature set. Subsequently, the feature with the second highest F value was added into the feature subset to generate a new feature subset. This process was repeated until 400 feature subsets were obtained. We examined the performances of the 400 feature subsets for finding the optimal features with jackknife cross-validation. For the convenience of observation, we plotted an IFS curve in Fig. 1. The peak (the maximum overall accuracy)

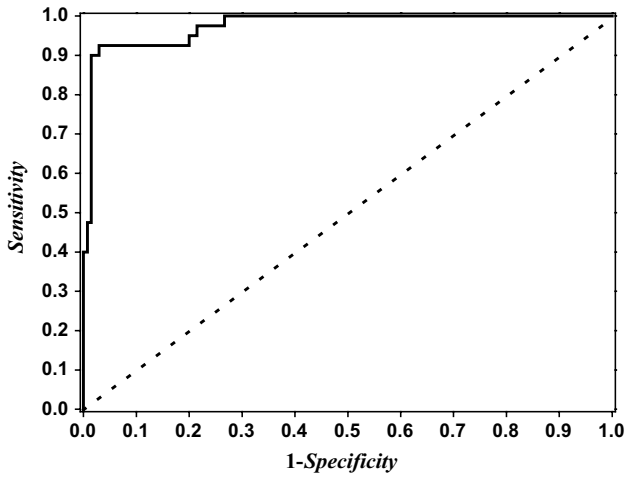

Fig. 2 The ROC curve obtained by using 101 optimal 5-gap dipeptides. The auROC of 0.975 was obtained in jackknife cross-validation. The diagonal *dotted line* denotes a random guess with an auROC of 0.5

can be found in this curve. As shown in Fig. 1, by using 47 optimal 0-gap dipeptides, the accuracy improves from 88.6 to 91.4 %, suggesting that the ANOVA-based technique is effective for feature selection.

It is necessary to investigate whether other g-gap feature subsets can obtain the higher accuracies or not. The g was varied from 0 to 9. The feature selection process was repeated for finding the maximum accuracy. The results in Fig. 1 show that the 101 optimal 5-gap dipeptides ($P < 0.001$) can produce a maximum accuracy of 97.1 % in jackknife cross-validation. Using these features, 90.0 % (36/40) of mitochondrial proteins and 99.3 % (134/135) of non-mitochondrial proteins can be correctly identified. To study the performance of the model across the entire range of SVM decision values, we plotted ROC curve in Fig. 2. It shows that the auROC reaches 0.975. It should be noticed that the number of features (101) are dramatically lower than the number of samples (175), suggesting that the model is robust.

Comparison with other methods

Because the benchmark dataset is not balanced in terms of positive and negative samples, we firstly investigated the results obtained from weight random guess. Results show that the Acc is $[40 \times (40/175) + 135 \times (135/175)]/175 = 64.7$ %, which is lower than that of our method. Subsequently, we compared the prediction performance of the proposed method with those of previous methods. All results on this benchmark dataset are listed in Table 1. As we can see from Table 1, although the Sn obtained in this paper is lower than that of previous methods, the Sp, Acc, and MCC of our method are the highest among the four methods.

Table 1 Comparison of the proposed method with other methods

Classifier	Sn (%)	Sp (%)	Acc (%)	MCC
This paper	**90.0**	**99.3**	**97.1**	**0.92**
PlasMit (Bender et al. 2003)	94.0	89.0	90.0	0.74
PFMpred (Verma et al. 2010)	97.5	90.4	92.0	0.81
ID (Chen et al. 2012)	100.0	89.6	92.0	0.82

The best results are shown as bold values

Sn, sensitivity; Sp, specificity; Acc, accuracy; MCC, Matthews correlation coefficient

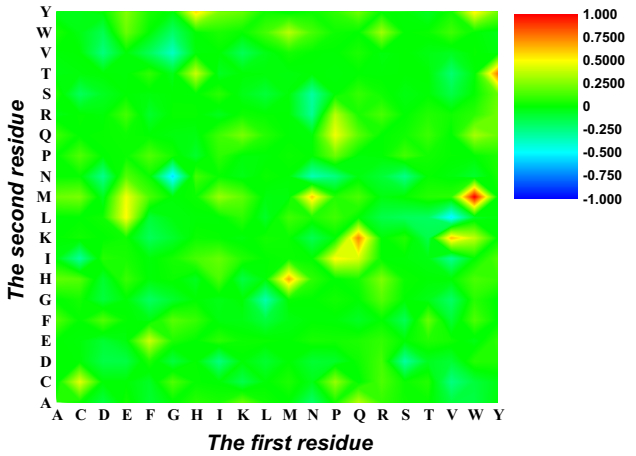

Fig. 3 A heat map for the 400 $F^0(\varepsilon)$ of the 5-gap dipeptides. The *red elements* indicate f_ε^5, mito $>$ f_ε^5, non-mito, whereas the *blue elements* indicate f_ε^5, mito $<$ f_ε^5, non-mito (color figure online)

Feature analysis

Feature analysis is very important for understanding the relationship between the optimal features and mitochondria proteins. Results in Fig. 1 and Table 1 showed that the correlation between two residues with five-residue interval ($g = 5$) plays an important role in the identification of mitochondria proteins of malaria parasite. For the convenience of observation, we used the following function to scale the $F(\varepsilon)$ of the ε-th 5-gap dipeptide:

$$F^0(\varepsilon) = \frac{F(\varepsilon) - F_{\min}}{F_{\max} - F_{\min}} \times \text{sgn}\left[\overline{f_\varepsilon^5}, \text{ mito} - \overline{f_\varepsilon^5}, \text{ non-mito}\right], \quad (9)$$

where $F_{\min}$ and $F_{\max}$ are the minimum and maximum F values of all the 400 5-gap dipeptides, respectively; $\overline{f_\varepsilon^5}$, mito and $\overline{f_\varepsilon^5}$, non-mito are the average frequencies of the ε-th 5-gap dipeptide in mitochondria proteins and non-mitochondria proteins, respectively; sgn is called the sign function.

For the purpose of providing an overall and intuitive view, a heat map with the first residue as column and the second residue as row was drawn in Fig. 3. The color of each element was marked based on its $F^0(\varepsilon)$. As shown in Fig. 3, most of the elements are green, indicating that these 5-gap dipeptides are maybe redundant information, which are irrelevant to the mitochondria protein prediction. Further observation showed that some 5-gap dipeptides [WM, YT, QK, MH, VK, NM, and HY (red-yellow)] often appear in mitochondria proteins, whereas other 5-gap dipeptides [GN, VL, GV (blue)] do not occur in mitochondria proteins.

In Fig. 3, one may also notice that the number of elements with red-yellow is more than that of elements with blue. The reason of this phenomenon is that non-mitochondria proteins consist of different types of proteins in malaria parasite. The features of different non-mitochondria proteins annihilated each other. For improving prediction performance, in the future, we should use multi-negative sets, in which each negative set has its given type to train and test the model.

Conclusion

The knowledge about mitochondrial proteins of malaria parasite is helpful for discovering anti-malaria drug targets. This study focused on the development of predictor for identifying the mitochondrial proteins of the malaria parasite. The ANOVA was proposed to screen the optimal g-gap dipeptides. A series of experiments demonstrates that the proposed method is powerful. The good results demonstrate that the correlation of amino acids contains the important information for predicting mitochondrial proteins of the malaria parasite. This promising method may have broad applications in protein classification and DNA motif identification.

Acknowledgments We would like to thank the anonymous reviewers for their valuable suggestions. This work was supported by the National Nature Scientific Foundation of China (No. 61301260) and the Fundamental Research Funds for the Central Universities (No. ZYGX2012J113).

Conflict of interest The authors declare that there is no conflict of interest.

References

Bender A, van Dooren GG, Ralph SA, McFadden GI, Schneider G (2003) Properties and prediction of mitochondrial transit peptides from Plasmodium falciparum. Mol Biochem Parasitol 132:59–66. doi:10.1016/j.molbiopara.2003.07.001

Chen YL, Li QZ, Zhang LQ (2012) Using increment of diversity to predict mitochondrial proteins of malaria parasite: integrating pseudo-amino acid composition and structural alphabet. Amino Acids 42:1309–1316. doi:10.1007/s00726-010-0825-7

Coordinators NR (2014) Database resources of the national center for biotechnology information. Nucleic Acids Res 42:D7–17. doi:10.1093/nar/gkt1146

Ding C, Yuan LF, Guo SH, Lin H, Chen W (2012) Identification of mycobacterial membrane proteins and their types using over-represented tripeptide compositions. J Proteomics 77:321–328. doi:10.1016/j.jprot.2012.09.006

Ding H et al (2013) Prediction of Golgi-resident protein types by using feature selection technique. Chemometr Intell Lab 124:9–13. doi:10.1016/j.chemolab.2013.03.005

Ding H, Deng EZ, Yuan LF, Liu L, Lin H, Chen W, Chou KC (2014a) iCTX-Type: a sequence-based predictor for identifying the types of conotoxins in targeting ion channels. Biomed Res Int 2014:286419. doi:10.1155/2014/286419

Ding H, Feng PM, Chen W, Lin H (2014b) Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis. Mol BioSyst 10:2229–2235. doi:10.1039/c4mb00316k

Fan RE, Chen PH, Lin CJ (2005) Working set selection using second order information for training support vector machines. J Mach Learn Res 6:1889–1918

Guda C, Guda P, Fahy E, Subramaniam S (2004) MITOPRED: a web server for the prediction of mitochondrial proteins. Nucleic Acids Res 32:W372–W374. doi:10.1093/nar/gkh374

Guo SH, Deng EZ, Xu LQ, Ding H, Lin H, Chen W, Chou KC (2014) iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. Bioinformatics 30:1522–1529. doi:10.1093/bioinformatics/btu083

Hayat M, Tahir M, Khan SA (2014) Prediction of protein structure classes using hybrid space of multi-profile Bayes and bi-gram probability feature spaces. J Theor Biol 346:8–15. doi:10.1016/j.jtbi.2013.12.015

Jia C, Liu T, Chang AK, Zhai Y (2011) Prediction of mitochondrial proteins of malaria parasite using bi-profile Bayes feature extraction. Biochimie 93:778–782. doi:10.1016/j.biochi.2011.01.013

Jia C, Lin X, Wang Z (2014) Prediction of protein S-nitrosylation sites based on adapted normal distribution bi-profile Bayes and Chou's pseudo amino acid composition. Int J Mol Sci 15:10410–10423. doi:10.3390/ijms150610410

Kumar M, Verma R, Raghava GPS (2006) Prediction of mitochondrial proteins using support vector machine and hidden Markov model. J Biol Chem 281:5357–5363. doi:10.1074/jbc.M511061200

Lin H, Chen W (2011) Prediction of thermophilic proteins using feature selection technique. J Microbiol Methods 84:67–70. doi:10.1016/j.mimet.2010.10.013

Lin H, Chen W, Yuan LF, Li ZQ, Ding H (2013) Using over-represented tetrapeptides to predict protein submitochondria locations. Acta Biotheor 61:259–268. doi:10.1007/s10441-013-9181-9

Nanni L, Lumini A, Brahnam S (2014) An empirical study of different approaches for protein classification. Sci World J 2014:236717. doi:10.1155/2014/236717

Qiu WR, Xiao X, Lin WZ, Chou KC (2014) iMethyl-PseAAC: identification of protein methylation sites via a pseudo amino acid composition approach. Biomed Res Int 2014:947416. doi:10.1155/2014/947416

Saha I et al (2014) Ensemble learning prediction of protein-protein interactions using proteins functional annotations. Mol BioSyst 10:820–830. doi:10.1039/c3mb70486f

Vaidya AB, Mather MW (2009) Mitochondrial evolution and functions in malaria parasites. Annu Rev Microbiol 63:249–267. doi:10.1146/annurev.micro.091208.073424

Verma R, Varshney GC, Raghava GP (2010) Prediction of mitochondrial proteins of malaria parasite using split amino acid composition and PSSM profile. Amino Acids 39:101–110. doi:10.1007/s00726-009-0381-1

World Health Organization (2014). http://www.who.int/en/