Disease Gene Identification Using Reliable Robust Classifier

V. Murugesan , P. Balamurugan

Abstract: ***Identification of genes causing the diseases is a major challenging problem towards diagnosing and providing treatment in a earlier manner. Many motivating methodologies are being proposed for the identification of disease genes. Generally, the unique variation among the previously proposed methodologies depend on the prior knowledge, also machine learning methodologies utilized for identifying. Identification of disease gene is normally observed as two class classification issue. Nature of information generates a key issue which can have an effect on results. In this research work, reliable robust classifier (RRC) based on dual simplex concept has been proposed to allocate a genes to a single disease class. RRC classifies the genes of M classes into M vertices of*** $(M-1)$ ***dimension dual simplex which results in M-class classification turn out to be*** $(M-1)$ ***class task. Since there exist no benchmark method to characterize the genes that have-diseases and not-have-diseases, this research work utilizes support vector machine to predict it. The results of experiments clearly demonstrate the effectiveness of the method with better precision, recall, and F-measure respectively.***

Keywords: Classification, Disease, Gene, Mining, SVM, simplex

I. INTRODUCTION

Identification of disease genes has become an major problem to improve the awareness regarding diseases mechanisms, and also to get better clinical methodologies. Conventional relationship based studies were conceded to find a huge amount of candidate genes that are interrelated by means of diseases. While utilizing the investigational based approach to recognize disease that connected with disease genes in enormous counts of candidates genes seems to be a costly mission, need of different approaches are in for this purpose. Hence, multiple machine learning methodologies are being introduced to identify and detect the features that are similar between different genes that are known and unknown. The methods vary in two different ways. Firstly, a kind of genomic data is utilized for producing the vector that are based on feature, which are interaction between proteins, profiles of gene expression, and ontology of gene. Additional methods incorporate various sources of data to give priority to candidates disease gene. Secondly, special type of algorithms are being utilized to train the model of prediction. Hence, most studies take this problem as the classification problem of two-class. Few studies suggest the positive set as recognized disease gene and negative set as unrecognized disease gene. But, unrecognized genes set is frequently holds different disease genes, where further methodologies made an attempt to minimize the misunderstanding in the process of classification, that is by choosing a tiny portion of unrecognized gene as a whole negative set. However, the methods seems to be non-robust and non-reliable which attains from unrecognized genes and data that are noisy.

The above discussed methodologies seems to be not implemented in a proposer manner due to relying on the protein information attained from preceding knowledge, holding errors. Furthermore, it lacks due to its incompleteness. Hence, a global preceding knowledge is necessary to handle this issue. Sequence-of-proteins is considered as powerful tool to resolve many issues towards the prediction of disease genes, and there exist null information about unrecognized genes. In addition, it has no use of utilizing the unrecognized genes as negative data. So, utilization of two-class classification algorithm will not be suitable for disease gene prediction.

II. LITERATURE REVIEW

A positive unlabeled learning method [1] based on graph method was proposed with 3 types of attributes of biological terms namely (i) gene ontology, (ii) domains of protein, and (iii) interaction networks between protein-to-protein. In this graph based method, a consistent set of true positive and true negative genes were extracted by utilizing co-training schema, where the performance got down when implementing with unlabelled gene dataset. A methodology [2] was proposed to identify vulnerable disease genes, where this methodology makes classification by means of using two classes namely non-sense and mis-sense by combining many single nucleotide polymorphisms functional annotation databases. But after performing functional enrichment analysis process the misclassification got raised. A single class classification strategy based on sequence [3] was proposed allot genes to a specific disease class that is, either yes or no. Initially features of the vectors and proteins sequences were generated to transform the vector by utilizing the properties of amino acid namely physicochemical property. But the results came with weak classification accuracy and f-measure. A fusion method based on sequence [4] aimed to classify the disease genes, where more priority was given to candidate gene to choose true negatives from it. In this method 4 sets were created to hold true negative by utilizing distance method. C4.5 methodology was applied to increase classification accuracy, but the results had gone inversely. A study [5] was made with the objective to look for

Revised Manuscript Received on March 20, 2019.

V. Murugesan, Department of Computer Science, VLB Janakiammal College of Arts and Science, Coimbatore, Tamil Nadu, India.

P. Balamurugan, Department of Computer Science, Government Arts College, Coimbatore, Tamil Nadu, India.

defects of molecular causative in PKD1 & PKD2 genes. For the study 18 patients were selected and performed the following for disease genes, namely (i) sanger sequence, (ii) probe augmentation method for multiplex ligation dependent. But, unexpected results of false negative came with prior value. A method [6] for identifying the gene by integrating the multiple classification algorithms. Initially, priority was given to select the features in order to choose the preferable attribute, then classifier was developed to classify the alzheimers disease genes. Normally, feature selection would be used to improve the classification accuracy, but in this research the false positives were mostly identified which reduces the classification accuracy. A classification method [7] with the base concept of identifying the genes and segregating it into only one class was proposed. The procedures includes grouping of data which have positive values, utilizing the single-class model for classification, and selection of negative data that are common in multiple sets. The procedures were aimed to give prioritization, but the prioritization didn't showed the expected result. An algorithm for gene selection [8] was proposed termed to identify the genes that are disease. It combines the details that are gathered from protein interaction network and the profiles of gene expression. Also, it picks out the genes through micro array as the confirmed disease genes, that is by increasing the consequence and efficient resemblance. It works with the concept of measuring the similarity and results decreased the true positives. A classification algorithm [9] with the concept of multi-category was proposed to classify the disease gene. It uses the support vector machine concept with linear kernel for testing the classification. The results indicate that the proposed method was not fit for identifying the tumor genes, due to low f-measure value in result. Bu utilizing the different updates from gene databases, 2 human protein interaction network based on large scale was developed [10], for the identification of disease genes. The statistical investigation shows that house-keeping and tissue-enriched networks were having high density, which reduces the classification accuracy. An analysis [11] has made on diseasome network by utilizing a strategy of detecting the disease-gene algorithm with the base of PCA. It is utilized to make investigation between more than one nodes in the same group. The results demonstrated that the algorithm is not fit for real-time datasets.

III. RELIABLE ROBUST CLASSIFIER

In the initial stage of RRC, it is necessary to think about the following training for the dataset for L-class classification problem:

$$T = ([w^1, z^1], [w^2, z^2], \dots, [w^m, z^m]) \quad (1)$$

where $w^j = \{w^{j1}, w^{j2}, \dots, w^{jm}\}^S$ is the element vector of j^{th} test and $z^j \approx (1,2, \dots L)$ is the number of classification

Consider any two vertex of a customary simplex are at equal distance, and establish 1-to-1 map between more than one vertex and class. Hence, every class can be addressed by equivalent vertex's coordinating vector.

Linearity in RRC to discover a $(L+1)-$dimensional point e^w, of which every segment have the following format:

$$e^i[w] = x * \frac{S}{i} + w - a^i \quad (2)$$

by resolving the consecutive problem:

$$\max_{x,a} \prod_{i=1}^{l+1} \frac{1}{2} \times \frac{x^{\frac{S}{i}} x^i}{a^{\frac{i}{2}}} - D \prod_{j=1}^{M} \prod_{k==d^j} \omega^{j,l}$$
$$subject\ to$$
$$\prod_{i=1}^{l+1} \left(2 \times U^{d^j,i} + U^{l,i}\left[x^{\frac{S}{i}} w^j - a^i\right] - U^{l,\frac{i}{2}} + U^{d^i,\frac{i}{2}}\right) > (\theta + \omega^{j,l}, \{j = 1,2, \dots M\}) \quad (3)$$

wherever $D > 0, \theta > 0$ are the parameters. RRC utilizes an threshold of $x^{\frac{S}{i}} w - a^i [i \approx (1,2, \dots, l+1)]$ to replace. The initial expression of Eq. (3) $\prod_{i=1}^{l+1} (0.5 \times x^{\frac{s}{i}} x^i - a^{\frac{i}{2}})$ is the customary expression, it expects to ensure the assumption capacity of the representation. Clearly, issue Eq. (3) is an derived enhancement with linear imbalance limitations.

It's necessary to develop a Lagrangian based function for Eq. (3) as pursues:

$$K\{x, a, \omega, \beta, \alpha\} = \prod_{i=1}^{l+1} \frac{1}{2} \times x^{\frac{S}{i}} x^i - a^{\frac{i}{2}} - \prod_{i=1}^{l+1} -D \prod_{j=1}^{M} \sum_{l=d^j} \omega^{j,l}\, \beta^j, l \left\{ \prod_{i=1}^{l+1} 2 \times U^{a^{i,j} + U^{l,i} \times x^{\frac{S}{i}} w^j} - \theta - \omega^{j,l} \right\} \quad (4)$$

where $\beta^{j,l} \equiv 0\ and\ \alpha^{j,l} \equiv 0$ are Lagrangian multipliers. As indicated by the Kuhn Tucker's method [27], we have the following:

$$\partial^{x^i} K = x^i + \prod_{j=1}^{M} \prod_{l=d^j} \beta^{j,l} \times 2 \left\{ \frac{U^{d^i,i}}{U^{l,i}} \right\} w^j = 0 \quad (4)$$

$$\partial^{a^i} K = a^i + \prod_{j=1}^{M} \prod_{l=d^j} \beta^{j,l} \times 2 \left\{ \frac{U^{d^i,j}}{U^{l,i}} \right\} = 0 \quad (5)$$

furthermore,

$$\partial^{\omega^{j,l}} = K = D + \beta^{j,l} + \alpha^{j,l} = 0 \quad (6)$$

At that point, of alternative Eq. (4) and Eq. (5) into Eq. (6) and by setting the following conditions

Eqs. (7)– (9) into Eq. (10)

$$\beta^j = [\beta^{j,1}, \dots, \beta^{j,d^{j+1}}, \beta^{j,d^{j-1}}, \dots, \beta^{j,l}]^S \quad (7)$$

$$\tilde{\beta} = [\beta^S, \beta^{\frac{S}{2}}, \dots, \beta^{S/M}]^S \quad (8)$$

$$F^{j,i} = 2 \times \{U^{d^{j,i}} + U^{1,i}, \dots, U^{d^j,i} + U^{d^{j+1},i}, U^{d^j,\frac{i}{2}}, \dots., U^{d^j,\frac{i}{2}} + U^{L,i/2}\}^S \quad (9)$$

$$\tilde{F} = \begin{pmatrix} F^{1,i} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & F^{M,i} \end{pmatrix} \quad (10)$$

$$E^{j,i} = \{U^{d^{j,\frac{i}{2}}} + U^{1,\frac{i}{2}}, \dots, U^{d^{j,\frac{i}{2}}} + U^{d^j + \frac{1}{2}}, U^{d^{i,\frac{j}{2}}} + U^{d^{j-1,\frac{i}{2}}}, \dots., U^{d^{j,i/2} + U^{L,i/2}}\}^S \quad (11)$$

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

while performing the classification for disease genes, it is necessary to perform the following

$$\tilde{E}^i = \{E^{\frac{S}{\tilde{J}}}, E^{S,\frac{i}{2}}, \dots E^{\frac{S}{M,i}}\}^S \quad (12)$$

From Eq. (12) we can acquire the double problem's Eq. (3) as:

$$\max_{\beta} \frac{1}{2}\tilde{\beta}^S [\prod_{i=1}^{l+1} \tilde{F}^i \{BB^S - ff^S\} \tilde{F}^{\frac{S}{i}}] \tilde{\beta} + \tilde{\beta}^S \prod_{i=1}^{L+1} [\tilde{E}^i - \theta f] \quad (13)$$

Subject to

$$0 \leq \widetilde{\beta} \leq Df$$

where f denotes the vector of $[1,1 \dots 1]^S$ by means of suitable dimensionalities. We additionally have the accompanying from Eq. (14) and Eq. (15) for making the classification as optimum towards prediction of disease genes with the support vector:

$$x^i = B^S \tilde{F}^{S/i} \tilde{\beta} \quad (14)$$

$$a^i = f^S \tilde{F}^{S/i} \tilde{\beta} \quad (15)$$

IV. ABOUT DATASET AND PERFORMANCE METRICS

A. Experimental data

The dataset utilized in [12] is employed in this research work. The dataset holds 5,405 recognized (known) genes traversing 2,751 diseases phenotype. The genes were made to extract by joining OMIM[13] and GENECARD [14] gene data. Also, 16,000 genes were chosen as the unidentified gene set from ENSEMBL [15].

B. Performance Metrics

To measure the prediction performance of existing and proposed classification algorithms, this research work utilizes the traditional performance metrics classification accuracy and F-measure for the evaluation purpose.

- **Precision :** Percentage of true positives over the total of false positives and true positives, which is denoted as Eq. (1)

$$Precision = (TP)/(FP + TP) \quad (16)$$

- **Recall :** Percentage of true positives over the total of false negatives and true positives, which is denoted as Eq. (2)

$$Recall = (TP)/(FN + TP) \quad (17)$$

- **F-Measure :** Percentage of precision and recalls harmonic mean, which is denoted as Eq. (3)

$$F - Measure = \big(2 \times (Recall \times Precision)\big) \div (Recall + Precision) \quad (18)$$

V. RESULTS AND DISCUSSIONS

Initially, evaluation was made on the consequence of performance of RRC. Also, finest count of features were made to extract by principal component analysis algorithm for the proposed RRC classifier and GA-SVDD [3]. Following the feature selection process, it is made to assess the impact of every physicochemical property of amino acid in the process of identifying the disease genes. Lastly, a comparison was made to validate the effectiveness of RRC and GA-SVDD [3].

To reduce the identifying model's overfitting, five-fold cross validation was accomplished in experiments. 5,000 recognized and 5,000 unrecognized instance was employed. Disease-genes are seems to be present in the instances, so RRC is used for training purpose where it requires only the positive data. In order to test and estimate the error rate, few negative data is necessarily required. For this purpose, 2 decision approaches were employed. In the prior approach, few unrecognized data were chosen in a random manner and it is considered as a negative data which is utilized in the process of testing. In the final approach, positive-unrecognized strategy was used and it considers the furthermost unidentified instances.

To measure the reliability of RRC and GA-SVDD [3], this research work has attempt provide training and testing using RRC. The prediction of result is presented in two manner. Firstly, the result have been demonstrated by using all the features. Secondly, the result have been demonstrated after the features are reduced using PCA.

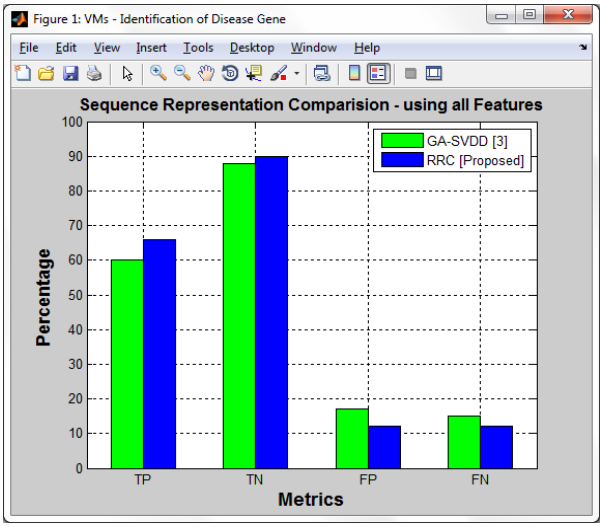

Figure 1. Positive-Negative Analysis vs All Features

In Fig. 1 to Fig. 4, the metrics are plotted in x-axis and percentage of results are plotted in y-axis. Fig. 1 and Fig. 3 demonstrates the results of GA-SVDD [3] and RRC towards identifying disease genes in terms of TP, TN, FP and FN. Fig . 1 shows the results while utilizing all the features, and Fig. 3 shows the results while utilizing the PCA extracted features. The results shows that the proposed classifier is performing better than GA-SVDD [3], it is because of dual classification by RRC where the GA-SVDD [3] perform sequence based one time classification. Fig. 2 and Fig. 4 demonstrates the results of RRC and GA-SVDD [3] before and after applying the PCA, which is a feature reducing technique. It is clear to note that the results are better when using PCA. In both aspects, that is while utilizing and not-utilizing the feature reduction techniques, the proposed classifier RRC outperforms GA-SVDD [3], it is due to performing the

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

classification in a dual manner.

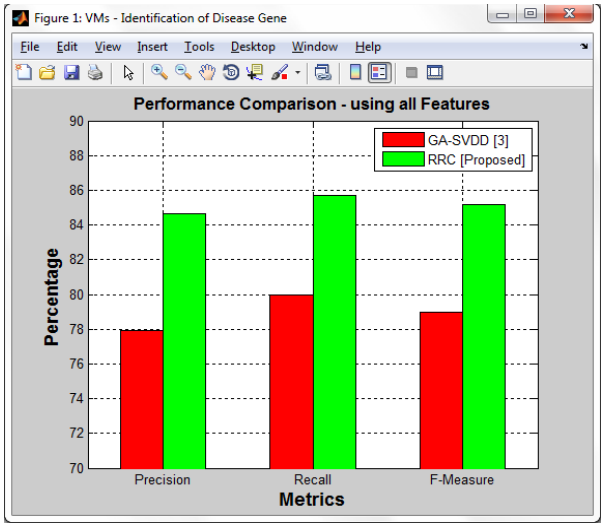

Figure 2. Performance vs All Features

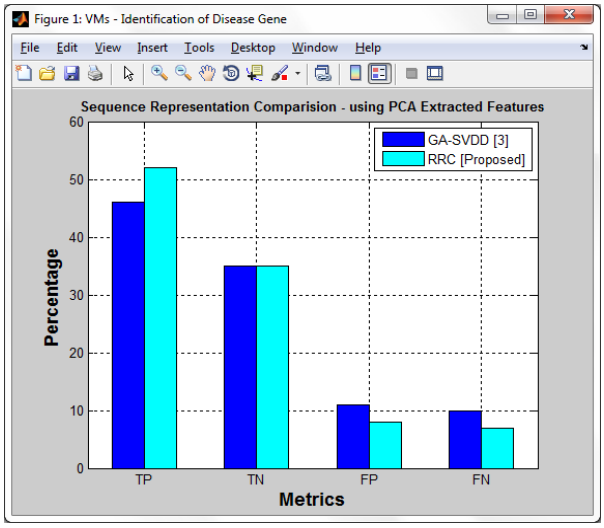

Figure 3. Positive-Negative Analysis vs PCA Extracted Features

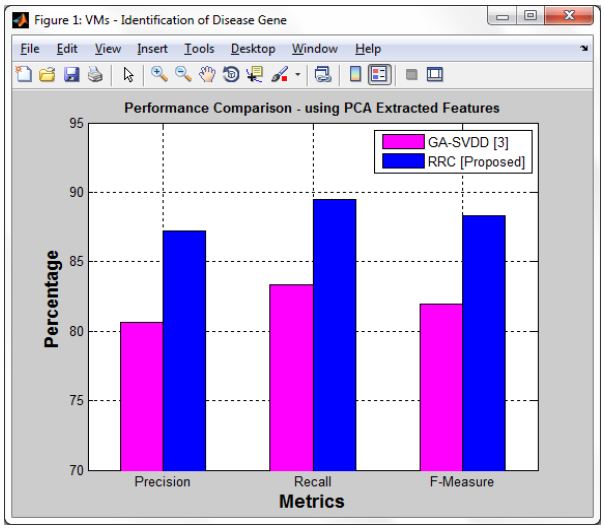

Figure 4. Performance vs PCA Extracted Features

While utilizing all the features (in Fig. 2) RRC attains Precision as 84.62%, Recall as 85.71%, and F-Measure 85.16%, where the GA-SVDD [3] attains Precision as 77.92%, Recall as 80%, and F-Measure 78.95%. While utilizing all the PCA extracted features (in Fig. 4) RRC attains Precision as 87.18%, Recall as 89.47%, and F-Measure 88.31%, where the GA-SVDD [3] attains Precision as 80.65%, Recall as 83.33%, and F-Measure 81.97%.

VI. CONCLUSION

The traditional classifier available are not suitable for predicting the disease gene in bioinformatics based dataset. This research work have proposed a robust classifier for towards predicting the disease gene, by allocating to a single disease class. It performs the classification by means of dual simplex concept , where the classes available are turned into vertices with $(M-1)$ dimension. The results of experiments clearly demonstrate the effectiveness of the method with better precision, recall, and F-measure respectively.

REFERENCES

1. G. H. Jowkar, G. E. Mansoori, "Perceptron Ensemble of Graph-based Positive-Unlabeled Learning for Disease Gene Identification", *Computational Biology and Chemistry*, Vol. 64, 2016, Pages 263-270.
2. W. Li, L. Zhu, H. Huang, Y. He, J. Lv, L. Chen, W. He, "Identification of Susceptible Genes for Complex Chronic Diseases based on Disease Risk Functional SNPs and Interaction Networks", *Journal of Biomedical Informatics*, Vol 74, 2017, Pages 137-144.
3. *Yousef, N. M.Charkari, "A Novel Method based on Physicochemical Properties of Amino Acids and One Class Classification Algorithm for Disease Gene Identification", Journal of Biomedical Informatics, Vol 56, 2015, Pages 300-306.*
4. Yousef, N. M. Charkari, "SFM: A Novel Sequence-Based Fusion Method for Disease Genes Identification and Prioritization", *Journal of Theoretical Biology*, Vol 383, 2015, Pages 12-19.
5. M. Abdelwahed, P. Hilbert, A. Ahmed, H. Mahfoudh, S. Bouomrani, M. Dey, J. Hachicha, H. Kamoun, L. K. Ammar, N. Belguith, "Mutational analysis in patients with Autosomal Dominant Polycystic Kidney Disease (ADPKD): Identification of Five Mutations in the PKD1 gene", *Gene*, Vol 671, 2018, Pages 28-35.
6. Y. Miao, H. Jiang, H. Liu, Y. Yao, "An Alzheimers Disease Related Genes Identification Method based on Multiple Classifier Integration", *Computer Methods and Programs in Biomedicine*, Vol 150, 2017, Pages 107-115.
7. Vasighizaker, S. Jalili, "C-PUGP: A Cluster-Based Positive Unlabeled Learning Method for Disease Gene Prediction and Prioritization", *Computational Biology and Chemistry*, Vol 76, 2018, Pages 23-31.
8. P. Maji, E. Shah, "Significance and Functional Similarity for Identification of Disease Genes", *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 14, No. 6, pp. 1419-1433, 2017.
9. G. Ji, Z. Yang, W. You, "PLS-Based Gene Selection and Identification of Tumor-Specific Genes," IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), Vol. 41, No. 6, pp. 830-841, 2011.
10. P. Wang, Y. Chen, J. Lu, Q. Wang, X. Yu, "Graphical Features of Functional Genes in Human Protein Interaction Network", IEEE Transactions on Biomedical Circuits and Systems, Vol. 10, No. 3, pp. 707-720, 2016.
11. W. Liu, L. Chen, "Community Detection in Disease-Gene Network Based on Principal Component Analysis" Tsinghua Science and Technology, Vol. 18, N. 5, pp. 454-461, 2013.
12. P. Yang X. L. Li J. P. Mei, C. K. Kwoh, S. K. Ng, "Positive-Unlabeled Learning for Disease Gene Identification", Bioinformatics, Vol. 28, No. 20, pp. 2640–2647, 2012.
13. M. Safran, I. Dalah, J. Alexander, N. Rosen, T. I. Stein, M. Shmoish, N. Nativ, I. Bahir, T. Doniger, H. Krug, A. S. Madi, T. Olender, Y. Golan, G. Stelzer, A. Harel, D. Lancet, "GeneCards Version 3: The Human Gene Integrator. Database", Oxford, p. baq020, 2010.

14. V. A. McKusick, "Mendelian Inheritance in Man and Its Online Version, OMIM", The American Journal of Human Genetics, Vol. 80, No. 4, pp. 588-604, 2007.
15. P. Flicek, M. R. Amode, D. Barrell, K. Beal, S. Brent, Y. Chen, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, L. Gordon, M. Hendrix, T. Hourlier, N. Johnson, A. Kahari, D. Keefe, S. Keenan, R. Kinsella, F. Kokocinski, E. Kulesha, P. Larsson, I. Longden, W. McLaren, B. Overduin, B. Pritchard, H. S.Riat, D. Rios, G. R. S. Ritchie, M. Ruffier, M. Schuster, D. Sobral, G. Spudich, Y. A. Tang, S. Trevanion, J. Vandrovcova, A. J. Vilella, S. White, S. P. Wilder, A. Zadissa, J. Zamora, B. L. Aken, E. Birney, F. Cunningham, I. Dunham, R. Durbin, X. M. F. Suarez, J. Herrero, T. J. P. Hubbard, A. Parker, G. Proctor, J. Vogel, S. M. J. Searle, "Ensembl 2011", Nucleic Acids Research, Vol. 39, pp. D800–D806, 2011.