12-2017

Development of New Bioinformatic Approaches for Human Genetic Studies

Jose Andres Guevara Coto
Clemson University

DEVELOPMENT OF NEW BIOINFORMATIC APPROACHES FOR HUMAN GENETIC STUDIES

A Dissertation
Presented to
the Graduate School of
Clemson University

In Partial Fulfilment
of the Requirements for the Degree
Doctor of Philosophy
Genetics

---

by
Jose Andres Guevara Coto
December 2017

Accepted by:
Dr. Liangjiang Wang, Committee Chair
Dr. Weiguo Cao
Dr. Charles Schwartz
Dr. Michael Sehorn

ABSTRACT

The development of bioinformatics methods for human genetic studies utilizes the vast amount of data to generate new valuable information. Machine learning and statistical coupling analysis can be used in the study of human diseases. These diseases include intellectual disabilities (ID), prevalent in 1-3% of the population and caused primarily by genetics. Although many cases of ID are caused by mutations in protein-coding genes, the possible involvement of long non-coding RNAs (lncRNAs) in ID due to their role in gene expression regulation, has been explored. In this study, we used machine learning to develop a new expression-based model trained using ID genes encoded with the developing brain transcriptome. The model was fine-tuned using the class-balancing approach of synthetic over-sampling of the minority class, resulting in improved performance. We used the model to predict candidate ID-associated lncRNAs. Our model identified several candidates that overlapped with previously reported ID-associated lncRNAs, enriched with neurodevelopmental functions, and highly expressed in brain tissues. Machine learning was also used to predict protein stability changes caused by missense mutations, which can lead to disease conditions including ID. We tested Random Forests, Support Vector Machines (SVM) and Naïve Bayes to find the best-performing algorithm to develop a multi-class classifier. We developed an SVM model using relevant physico-chemical features after feature selection. Our work identified new features for predicting the effect of amino acid substitutions on protein stability and a well-performing multi-class classifier solely based on sequence information. Statistical approaches were used to analyze the

association between mutations and phenotypes. In this study, we used statistical coupling analysis (SCA) to cluster disease-causing mutations and ID phenotypes. Using SCA we identified groups of co-evolving residues, known as protein sectors, in ID protein families. Within each distinct sector, mutations associated with different phenotypic manifestations associated with a syndromic ID were identified. Our results suggest that protein sector analysis can be used to associate mutations with phenotypic manifestations in human diseases. The bioinformatic methods developed in this dissertation can be used in human genetic research to understand the role of new genes and proteins in human disease.

DEDICATION

I dedicate this thesis to my parents, siblings, close family and dear friends Shenise, Jake, Aaron and my partner Sarah. The love and support given by them made this possible.

ACKNOWLEDGMENTS

I wish to thank my advisor Dr. Wang for his support and guidance throughout my research. His knowledge and teachings have provided me with the foundations to grow into a better bioinformatics researcher every day. I appreciate Dr. Cao for his support and providing the opportunity for exciting research collaborations. Also, Dr. Schwartz, for the opportunity to be part of the Greenwood Genetic Center and for providing valuable insight into human genetics. Dr. Sehorn, for his research and career advice. I want to extend my appreciation to Dr. Marcotte and Dr. Frugoli for their support and assistance. A special acknowledgment to my group members Brian Gudenas, Jun Wang and Shuzhen Kuang. Special recognition to the assistance provided by Austin Gorman, Eda Ozyesilpinar, and Kalee Lineberger at the Writing Center, and Priscilla Harrison from the Office of Access and Equity. Finally, a warm thank you to the staff and faculty of Clemson University that supported and assisted me in obtaining my PhD.

TABLE OF CONTENTS

Table of Contents (Continued)

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER I – LITERATURE REVIEW

1.1 Introduction to Bioinformatics

The term bioinformatics was coined in the 1970s by Paulien Hogewegen and Ben Hesper to define the use of informatic technologies in the study and understanding of biological processes [1]. This definition can be further expanded to state that bioinformatics is the science that looks to understand and organize, via computational approaches, the function, role, and information associated with biological macromolecules [2,3]. Although seemingly distant at first glance, at a closer look, life can be described as an information system. The fact that there is terminology such as genetic code, used to refer to the genetic material and how it is stored, and that this crucial information for survival needs to be encoded, stored, and eventually transferred, shows the close ties to information systems [1,2,4].

Originally envisioned as an approach for sequence analysis [2,4], bioinformatics evolved into a more complex field due to the increasing power of computers. Previous reports [5,6] describe that the goal of bioinformatics is to organize and understand biological data. In its current form, it is understood that bioinformatics has a broader scope, including the development of methods for data analysis and knowledge discovery from biological datasets [7]. The wide array of tasks undertaken by a bioinformatician can be summarized in three main areas:

1. Data management: This area focuses on accessing, combining, converting, manipulating, storing, and annotating data. Data management requires routine quality checks, automatization of existing methods, and summarization of large datasets.
2. Data processing: This area requires running a variety of tools that include but are not limited to aligners, variant callers, RNA-Seq quantifiers, transcript analyzers, etc. During the preparation for the analysis, the bioinformatician must anticipate and prepare for potential pitfalls that could present themselves, and develop alternative solutions/analyses that fit the data.

3. Knowledge discovery: Knowledge discovery has a crucial function in the interpretation of the data. The data a bioinformatician gathers through data management and processing would remain uninformative without insightful interpretation. By using biological data and computational methods, knowledge discovery can extract insightful information from the data, providing new and valuable information.

Bioinformatics has three main aims. The first, which is highly associated with one of the areas, is organization and management of biological data. This allows the scientific community access to existing information as well as submit new entries in various repositories such as GenBank [8], Protein Data Bank [9] or UniProtKB [10]. Data management not only includes the incorporation of new data entries but also the process of data curation, which refers to the manual validation of the quality of the entries or the datasets [1–4]. The second aim is to develop tools and resources for the analysis of data that has been stored in the repositories. This aim is closely related to the second and third areas that are essential in bioinformatics, as such tools are necessary for data analysis and data interpretation. To accomplish this aim, computational and biological knowledge and expertise are required [2,3,7]. The third aim is the use of the tools and methods developed for data analysis and their implementation to extract new information from biological datasets to further the understanding of biological systems. It is possible to use the methods to understand biological systems in a single organism view, where the system is seen as an independent entity, or a multi-organism approach where it is possible to uncover shared relationships and interactions amongst different biological systems [1–3,5–7,11].

Currently, bioinformatics research has grown and evolved to cover the fields of structural biology, genomics, and evolutionary biology [4–6]. Interestingly, someone working in the field of bioinformatics must have analytical skills and knowledge in statistics and computer sciences, making it an interdisciplinary field that requires competency in various areas. The ability of bioinformatics to incorporate knowledge and methods, initially developed to analyze data from other fields into the biological context, has made it invaluable in the current research environment.

Bioinformatics has been enriched by a community that continuously provides support and advancement of the available methods. This in turn allows the integration of the various methods for pipeline construction, exemplified by programming languages such as R or Python which provide multiple packages or libraries with supported and integrated modules that have been developed by community members. In most cases, these packages or modules have been published in peer-reviewed journals [12–14]. Published materials tend to be hosted, maintained, and updated by their developers, providing the basis for reproducibility, provenance, and upgradability of the methods [11,15,16]. In return, a competent bioinformatician can use the programming language and the supporting packages/modules to develop methods that can generate novel information from the data.

The importance of this field can be seen in the increasing number of publications related to this area. The impact of bioinformatics as a research field was shown in a study by [17], where the number of publications from 1991-2010 for different fields and sub-fields in the Web of Science for the biological sciences were compared. Subsequently, the number of publications in the different sub-fields were quantified for the periods encompassing the last five and three years of the study. The results indicate that the growth of bioinformatics papers per annum appears to follow an exponential pattern, and that the amount of bioinformatics-related journal articles increased by 65% in the five-year period of the study, only third behind synthetic biology and systems biology. The data from this study has been adapted and summarized in Table 1.1.

Table 1.1 Growth and change of major sub-fields in biological research focused on a publication-based search of the Web of Science (WoS) spanning from 1991-2010. The table indicates the number of publications, the growth for each sub-field per year as exponential or linear based on the number of publications recorded, the changes observed per sub-field as percentages of publications and the percentage of published papers for the sub-filed in the period of 2006-2010, data from Pautasso (2012) [17].

Sub-field	Publications in WoS 1991-2010	Publication Growth/Year	Change in Number of Publications (%) 1991-2010	Percentage of Publications 2006-2010
Synthetic Biology	8.50E+02	Linear	0.7	94
Systems Biology*	6.20E+03	Exponential	3.7	84
Bioinformatics	1.70E+04	Exponential	6.8	65
Structural Biology	2.20E+03	Linear	0.6	54
Cell Biology	1.10E+04	Exponential	1.2	50
Ecology and Evolution*	2.40E+04	Exponential	1.6	48
Genetics	8.50E+04	Linear	-7.2	40
Molecular Biology	2.20E+04	Linear	-5.3	31
Biochemistry	2.50E+04	Exponential	-10.1	31

In summary, the role of bioinformatics has seen a change from the sequence-analysis oriented field in the 1970s to a more multi-disciplinary area in the 1990s. However, the major shift in the paradigm appeared to come around the late 1990s or early 2000s. According to Pop and Salzberg [18] and Mardis et al. [19] emphatic change can be attributed to the impact that the Human Genome Project had on influx of data and the need of new methods to analyze it. Bioinformatics has also seen a change from a support role in research to a more central one. Currently, bioinformatics has become a key component in modern research, from candidate gene prediction to drug target discovery. In this dissertation, using bioinformatical approaches, such as supervised machine learning and statistical analysis of multiple sequence alignments for functional analysis of protein families, new knowledge has been generated for human genetic studies.

1.2 Overview of Machine Learning

In its essence, machine learning refers to the field concerned with the development and application of computer algorithms that improve with knowledge or experience [20]. The learning process usually utilizes a set of example data to teach an algorithm in a manner that it will eventually achieve the optimal performance, or criterion [20,21]. This criterion can be represented as a series of measurements—for example, the accuracy of a model in a classification problem, or the value of a fitness function in an optimization problem [21]. Because of their predictive capabilities, machine learning models have become commonplace in different fields, from research in the sciences and engineering, such as fuel efficiency prediction in automobiles [22], to the business environment [23,24], with stock price prediction. In the field of biological sciences and more specifically in genetics and genomics research [21,25], machine learning has been of great importance to model problems using large datasets, which, due to their size, cannot be modelled using traditional statistical methods. In studies where machine learning has been used, it has provided valuable information from these large and complex datasets, such as the function of open reading frames in yeast using mutant phenotype data, or pathway identification in plants from large transcriptomics sets [25–28]. Machine learning has also been used to address tasks deemed time-consuming to solve using

other approaches. By efficiently harnessing the potential of big data [28], machine learning approaches have been able to generate classifiers that have provided new insight into the possible genetic causes associated with human diseases such as cancer [29] and autism [30], as well as novel biomarker development [31] and drug target discovery [32].

The machine learning process consists of transforming data into knowledge. This can be done in both the construction and validation stages [21,33,34]. The construction part of the process is composed of a sequential series of steps, starting with dataset pre-processing. In the pre-processing step, the data is curated, and inconsistencies are identified and resolved [20,21,33,34]. It is during data pre-processing that normalization strategies such as z-score standardization, or missing data handling techniques, like imputation, are applied to the dataset [20].The outcome of this step is a high-quality dataset defined as the training set. Subsequently, dimensionality reduction can be performed. This process includes feature extraction and feature selection. Feature selection is focused on selecting the subset of important features, whereas feature extraction creates new features by transforming raw data using methods such as principal component analysis. This is followed by training a predictive model, where machine learning methods and frameworks are used to model and analyze the data. The idea of training a model aims to develop a learner capable of extracting significant information from the data (knowledge discovery) [21,28,33–35]. The built models need to be validated and evaluated using performance metrics. Further interpretation from domain experts should be done to provide an additional layer of validation of the significance of the information extracted from the data, if any, that the model can provide. If inconsistencies are found at any point of model construction, it is possible to revert to a prior step in order to retrace and troubleshoot. However, if the model is deemed satisfactory, in terms of performance metrics or validity of biological information, it can provide new insight into the biological system [20,21,29].

1.2.1 Supervised Machine Learning

In a simple overview, a supervised machine learning problem can be a classification task that consists of a dataset that is divided into different classes, each with a known label. The label represents the relationships between the inputs (data instances) and the outputs (predictions) [33]. Supervised machine learning evolved from pattern recognition, and aims to identify hidden signals that are commonly shared within a class [24,36]. The presence of the output or response variable is what marks a clear distinction between supervised learning and unsupervised learning, where the instances are unlabeled [33,37].

In supervised learning, the algorithm analyzes the inputs in the labelled training set and generates a function which can predict the class label for each data instance [33,36]. These predictions are based on supervised learning's pattern recognition abilities, which identify common or shared signals between data instances that have the same label [38,39].

It is important to determine the generalization capabilities of a model. This is usually determined through validation methods, a process designed to evaluate the model's overall quality using accuracy and other metrics [21,28], or to detect over-fitting. There are multiple approaches to evaluate the model. These include a holdout set, where the training data is separated into two splits; one is used for training and the other for validation [21,40]. Also in cross-validation is 10-fold or 5-fold, where the dataset is split into n= 5 or 10 subsets of the same size, and the training set is n-1, while the remainder is used for validation. The leave-one-out (LOO) cross-validation is included within the leave-p-out cross-validation, where p instances are left-out for validation while the rest are used for training. In LOO cross-validation, p = 1, meaning that for each round of cross-validation, statistics are calculated for the left-out sample and the remaining samples are used for training [21,40–42].

After model construction and validation, it is a recommended practice to determine the model's generalization capabilities and performance with a test set. The test set has to be independent from the training data, thus remaining "unseen." Although the test set is independent from the training set, and as it has never been used for training purposes, both the training and test sets should have a similar probability distribution for the feature values. Otherwise, results would not be informative, as the differences in training

and test sets make prediction impossible [35,41]. If the training set would have included the test data, then the assumption of independence would no longer be valid. Another important aspect of this separation between training set and test set is that if model performance is high on the training set but significantly lower on the test set, poor model generalization or over-fitting can be potentially identified [33], as the model might be finding certain relationships that are not general but found only in the training set.

In supervised learning, a labelled test set is used to represent the relationship between the input attributes and the output [35,39]. The labelled test set allows for the construction of the confusion matrix [40]. The confusion matrix is a table that describes a model's performance providing the classes and how each instance in the test set was classified. This matrix allows for the calculation of measures such as true positive, false positive, true negative, and false negative, required to calculate the different performance metrics such as sensitivity, specificity, and accuracy [43]. The measurements derived from the confusion matrix are also used to calculate classifier metrics such as the receiver operating characteristic curve (ROC), which represents the diagnostic ability of a classifier. It also allows for the calculation of the area under the ROC curve (ROC-AUC). The ROC-AUC is defined as the probability that a randomly selected instance from the positive class will be ranked higher than one from the negative class, and represents the percentage of randomly selected instances correctly classified by a model. The ROC-AUC can be used to compare classifier performance for model selection [38,41,43]. Another used metric is the Matthews correlation coefficient (MCC) which allows the measure of quality of a binary classifier, with values ranging from -1 (disagreement between predicted and observed) to +1 (perfect prediction), with 0 representing that prediction is not better than a random guess [44]. It is important to note that these various performance measurements of classifier performance can also be calculated during validation, using the different folds that were split from the training set, depending on the cross-validation method.

In supervised machine learning, one potential issue is model over-fitting. The effects of model over-fitting are generally observed when comparing the performance of the model with the training set and the test set. An over-fitted model performs well on the training set but poorly on the test set [34]. In other

words, the model has learned the training data well, but when presented with an unseen test set, its performance will be below expected [33,34,36]. Thus, the difference in model performance between training and test sets can be used to identify over-fitting. If significantly higher performance is observed in the training set than in the test set, it is possible that the model might be over-fitted [40]. Model over-fitting can be caused by background or noisy data during model construction [41]. Another possible source is the absence of data points which can result in the model attempting to classify instances by using uninformative features. These possible causes of over-fitting can be handled using methods such as feature selection [45,46] and class-balancing approaches which have been reported to address the drawback of model over-fitting and thus improve classifier performance [47,48].

1.2.2 Feature Selection in Machine Learning

Feature selection for classifier construction is the process by which attributes used to encode data instances are selected to generate an optimal feature subset [27]. This process is undertaken in the following manner: given a dataset with $\boldsymbol{A}$ features, the goal is to generate a new set $\boldsymbol{B}$, where $\boldsymbol{B} \subset \boldsymbol{A}$, meaning that $\boldsymbol{B}$ is a subset of $\boldsymbol{A}$ [45]. The idea of feature selection is based on the concept of dimensionality reduction. Dimensionality reduction removes irrelevant features and reduces the feature set. A reduced feature set can lead to reduced computational time for classifier construction, reduction in computational cost for the calculation of the function between the input attributes and the class [49], and possible improvement of classifier performance [35,50].

In a dataset with a large number of attributes, it is possible that certain attributes are uninformative. This can lead to over-fitting [35,51]. By selecting the attributes that provide valuable information and discarding those that are background noise [45,52], it is possible to increase model performance [46] and prevent over-fitting, as irrelevant features could lead to spurious conclusions [50,51]. The incorporation of irrelevant features may create irrelevant relationships between features and response variables, leading to models that have reduced generalization.

Feature selection can be used to extract important information from the dataset, providing significant insights into the data [53]. This is of special interest in knowledge discovery where identification of significant features within a dataset can also be useful to further the understanding of what attributes could be of importance in the association, causality, or development of a biological process; for example, it can identify important developmental stages associated with diseases such as autism, or identify risk factors associated with cancer [30,35,54,55].

Feature selection methods can also be used in knowledge discovery to extract new and valuable information from biological datasets. By applying feature selection methods, it is possible to use the variables selected as a source of new biological knowledge [53]. For example, Cogill and Wang [30] used feature selection to identify spatiotemporal expression features from the developing brain associated with autism spectrum disorder (ASD), providing new insight into critical stages of development for ASD. It is also possible to compare feature selection methods to identify overlapping or commonly identified features. The overlapping variables identified by different methods can provide important insight regarding the dataset, such as association with a disease. For example, Liu et al. [56] used the overlapping features from proteomic data in an ovarian cancer association study.

1.2.3 Handling Class Imbalance in Machine Learning Problems

Imbalanced datasets are common for machine learning, and more often occur in biological cases [20,57]. The presence of class-imbalanced datasets in the biological context can be found in target gene prediction, disease gene discovery, binding site prediction, and pre-miRNA prediction [47,57], where the targets are generally an underrepresented class. To address this potential problem, several possible solutions have been proposed at the data-level as well as at the algorithm level [33,58,59]. The sampling techniques used to address the class imbalance problem or the class distribution problem in a dataset involve artificial re-sampling of the dataset. This can be achieved in two ways: under-sampling of the majority class, or over-

sampling of the minority class [33,58–60]. It is also possible to combine both the re-sampling approaches to pre-process the data to handle class imbalance in the dataset.

Algorithmic approaches include parameter modifications, such as the class weight adjustment to counter class imbalance, decision threshold adjustment, and recognition-based learning instead of discrimination-based learning [58,60]. Algorithmic approaches can be used in combination with data-based approaches. The combination of algorithmic and data-level class-balancing methods is known as mixture-of-experts. Mixture-of-experts approaches combine the results from multiple classifiers using over-sampling, under-sampling, or fine-tuning of the class weights. We will next briefly discuss those methods at the data level used in handling imbalanced datasets [33,58–60].

1.2.3.1 Under-sampling Techniques

Under-sampling methods randomly select instances from the majority class to balance the dataset. By reducing the instances in the majority class, random under-sampling improves classifier performance, reducing the classification bias introduced when a class is over-represented [33,58]. It has been reported [57] that the use of under-sampling techniques with multiple machine learning algorithms has been successful in pattern recognition and classification tasks. However, one potential problem associated with random under-sampling techniques is the loss of significant data instances, which can lead to loss of information. To address this potential issue, more advanced under-sampling methods have been developed, including Tomek links [33,58–60].

The under-sampling technique consisting of forming Tomek links between pairs of data instances identifies the minimally distanced nearest neighboring instances from different classes. The majority class instances are subsequently removed to balance the classes. The definition of a Tomek link is given by [33,61]: two data instances, E_i and E_j, each belonging to a different class, have a distance defined by $d(E_i,E_j)$. The pair of data instances (E_i,E_j) will be identified as a Tomek link pair unless, another instance, E_l, changes $d(E_i,E_j)$. If $d(E_i,E_l) < d(E_i,E_j)$ or $d(E_j,E_l) < d(E_i,E_j)$, then E_j and E_i will no longer be considered Tomek links.

For class balancing, the data instance from the majority class forming a Tomek link pair is removed from the balanced dataset. This strategy can be effective in addressing the over-representation of the majority class, and has been reported to handle the issue of loss of informative data instances associated with random under-sampling [57,61]. However, previous studies report that Tomek links as with other under-sampling strategies are more effective when combined with over-sampling approaches, and their model performance improvement capabilities have been observed when combined with synthetic over-sampling of the minority-class technique (SMOTE) [33,61].

1.2.3.2 Over-sampling Techniques

Over-sampling methods randomly create new instances from the minority class [33,58,59,62]. Over-sampling may lead to model over-fitting due to over-representation of data instances. Over-representation refers to creation of new data instances from a single data instance [57,62]. The new data instances will have the same attributes as the data from which they were generated. Although the data set is balanced, it does not represent the distribution of the attributes of the minority class. The use of over-sampling can result in model over-fitting [57,62]. To prevent over-fitting, over-sampling should not be used on the validation split, and should only be used in the training set. The use of over-sampling methods can be considered a data pre-processing step, which, depending on the size and dimensionality of the dataset, could prove a computationally exhaustive task [33,62]. However, over-sampling does have its advantages over under-sampling methods. Contrary to under-sampling, over-sampling does not reduce the majority class, ensuring no loss of informative instances [47,48,63]. Over-sampling methods can also generate new samples not only from existing data instances but also from the lines joining multiple data instances in the dataset. This can be used to handle the over-representation issue that can lead to over-fitting. This is implemented in the over-sampling technique, Synthetic Minority Over-Sampling Technique (SMOTE). The SMOTE technique balances a dataset by selecting a data instance or point and creates synthetic or new minority class data though interpolation between an instance and its neighbors [48].

SMOTE over-sampling occurs when new instances are generated from the line segments that connect a selected data instance to *k* elements of the minority class [47,48,63], where *k* is the number of neighboring elements used to interpolate the new synthetic samples. Figure 1.1 illustrates the process of generating synthetic samples, where x_i is the selected data instance and x_{i1} to x_{i4} are the nearest neighbors. In the figure, r_1 to r_4 are the randomly generated synthetic data points created from interpolation. SMOTE handles possible over-fitting by generating new samples from the line between the data point x_i and randomly selected nearest neighbors, thus avoiding over-representation.

In over-sampling, the creation of exact replicates from data instances lead to the decision boundary to tighten, which can result in over-fitting. Tightening refers to the boundary being close to the minority class. Thus, the decision boundary is far from the ideal boundary expected for class separation. In contrast, the synthetic data instances generated from SMOTE are not exact copies of any existing instance; this leads to a softer decision boundary. In this case, the decision boundary is closer to the ideal boundary for class separation. This reduces the effects of over-fitting [48,61,63]. In this dissertation, SMOTE was used on a dataset comprising intellectual disability (ID) genes and non-ID disease genes. The use of this method improved model performance.

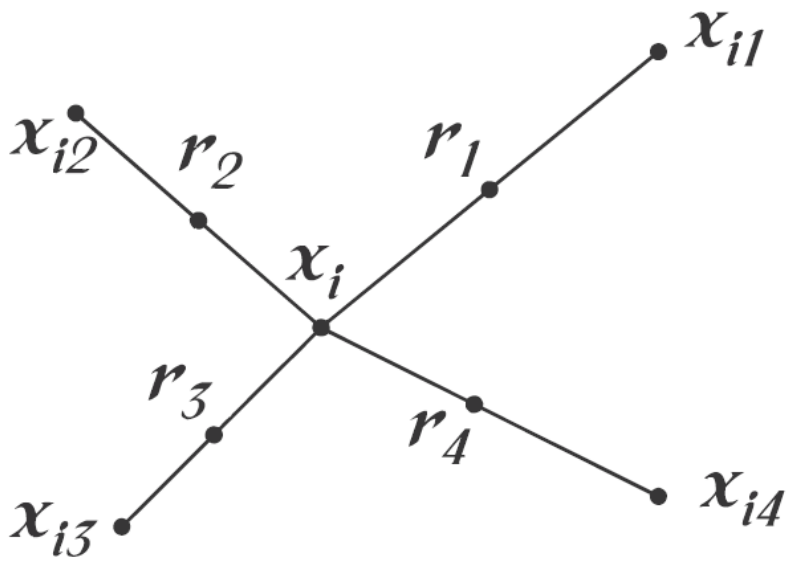

Figure 1.1 Illustration of how SMOTE creates new synthetic instances. The initial selected data point x_i has a set of neighbors defined as x_{i1}-x_{i4}. The algorithm creates lines that join the initial data point with the neighbors and within these segments the new synthetic samples are interpolated. The image was used from Lopez et al. (2009) [63] with permission.

1.3 Intellectual Disabilities and Long Non-coding RNAs

1.3.1 The Etiology of Intellectual Disabilities

Intellectual Disability (ID), formerly known as Mental Retardation (MR), refers to a large heterogenous group of disorders that have the impairment in intellectual abilities as a common trait. ID has a prevalence in the population of 1-3% [64,65], with severely affected individuals representing 0.3%, and approximately 85% classified as mild ID patients. Individuals with ID are usually diagnosed prior to the age of 18 years with an intellectual quotient (IQ) of below 70 [65,66], and limited adaptive behaviors such as self-care, communication, and social skills [65,67]. ID can be classified using the level of impairment on the IQ: profound (IQ<20), severe (IQ 20–34), moderate (IQ 35–49) and mild (IQ 50–69). A more simplified classification scale can be used, where severe ID includes all IQ values below 50 and mild ID

refers to all IQ values between 50 and 70 [65]. Although ID represents a lifelong impairment, there have been studies reporting the use of cognitive training in individuals with mild ID [67], stimulated their working memory and attention, improving their abilities in tasks such as self-care or speech. However, the challenges for individuals afflicted with severe ID remains, as they will require constant care for the rest of their lives.

ID disorders can also be classified as syndromic or non-syndromic [66]. This distinction is based on the presence of other manifestations associated with ID. In syndromic ID, individuals diagnosed with ID commonly present dysmorphic traits or metabolic disorders. In non-syndromic ID, the intellectual impairment is the only manifestation [64–66]. It is possible for a non-syndromic ID to be re-classified as syndromic; for example, *OPHN1*, a non-syndromic ID gene located on the X-chromosome (XLID), was later identified to be associated with cerebellar hypoplasia through magnetic resonance imaging, causing it to be re-classified as a syndromic ID gene due to the presence of an associated malformation [65]. Both syndromic and non-syndromic ID can be caused by genetic and exogenous factors. Exogenous or external elements have an impact on the development of ID, with some factors like fetal alcohol syndrome, pre- and postnatal infections and peri- and postnatal asphyxia, amongst others, causing approximately 25-50% of reported mild ID [64,65]. However, for moderate to severe ID, approximately 65% of ID cases have a genetic etiology [65,68]. In cases where genetics is the main cause of ID, the genomic abnormalities leading to ID are chromosome aneuploidy as in Down syndrome, or structural abnormalities, such as chromosomal deletions in *cri du chat* syndrome, or monogenic (single-gene) ID, as in fragile X syndrome [65].

Previous studies have identified structural variations as a cause of ID [69]. These structural variations refer to chromosomal regions larger than 1 kilobase (kb) that include inversions, translocations, and copy number variants (CNVs) [65,66]. Chromosomal translocations are exemplified by the Robertsonian translocation, associated with chromosomes 14 and 21, where the chromosomes break at the centromeres and the *q* (long) arms of both chromosomes fuse to form a larger chromosome. Individuals with this kind of unbalanced translocation carry a trisomy at chromosome 21, that leads to Down syndrome.

CNVs are defined as insertions or deletions in a stretch of DNA that are >1 kb, and up to several megabases (Mb) in length [66]. It has been reported that for ID, the association of CNVs with ID causality is estimated to be ~14%, with some of these CNVs being >400 kb [69]; however, as more *de novo* CNVs are identified this number could change. CNVs have been associated with both syndromic and non-syndromic ID, and have been identified through whole genome sequencing in multiple chromosomes, including chromosomes 4, 15, 17, and X [70].

Monogenic ID have been mapped across the human genome in both autosomal and the X chromosome [65]. Notably, X-linked ID (XLID) genes have been identified to represent 10-15% of known protein-coding ID genes in the genome [68]. There are ~100 known XLID genes, representing approximately 10% of the known protein-coding genes (800) for the X chromosome. Overall, the number of ID genes in the genome represents ~4% of all protein-coding genes, based on the estimate that there are ~19,000 protein-coding genes in the human genome [68,71]. Currently, the number of known protein-coding ID genes is 818, including genes that overlap with other cognitive disorders [65] .

Previous studies have reported that ~40% of individuals affected by an ID are also affected by another cognitive disorder or intellectual impairment [65,66]. The report by van Bokhoven [65] suggests that there is evidence of comorbidity between ID and cognitive disorders such as autism spectrum disorder (ASD). However, not all ID genes overlap with cognitive disorders; ~450 genes have been identified to uniquely manifest ID. The importance of monogenic ID was identified to be the second cause of ID after Down syndrome in a study of 4,730 children affected by intellectual and cognitive impairments [72].

1.3.2 The Study of Intellectual Disability-Causing Mutations and Phenotypes

The identification of disease-causing mutations is the result of genotype-phenotype association studies, where the effect of a mutation on the clinical manifestation is analyzed [65]. These studies are focused on analyzing individuals affected by a disease, such as a syndromic or non-syndromic ID, to discover the mutation that caused the development of the phenotype being characterized. Mutations at

different positions in a gene can result in very similar phenotypic manifestations; for example, in Snyder-Robinson Syndrome (*MRXSSR*), an XLID caused by mutations in the spermine synthase (*SMS*) gene, where two identified mutations (V132G and Y328C) led to mild forms of ID with similar phenotypic manifestations [73,74]. In other cases, mutations in the same position can generate different phenotypes, such as in Alpha Thalassemia Mental Retardation X-Syndrome (*ATRX*), where the R246C amino acid change, which represents 36% of all *ATRX* mutations, causes the development of diverse phenotypes including various degrees of ID (moderate to severe) and urogenital defects [75]. The examples of *MRXSSR* and *ATRX* demonstrate how it is possible to associate mutations in ID genes with clinical manifestations of a disorder. However, these examples show that it is also possible for other factors, including the environment, to affect the phenotypic manifestations of ID.

The identification of the mutations associated with the phenotype include familial studies that focus on the individual's relatives as well as other affected families. Other analyses are histochemical and biochemical assays on the tissues, which are followed by molecular genetics analyses. Familial studies are used to identify the possible mode of inheritance of a disease, but also to study multiple affected families to determine possible environmental effects on disease manifestation. The importance of familial studies can be exemplified in the identification of *ATRX* where 80 unrelated families with affected children were studied to identify the mutations associated with the disorder [76].

The use of histochemical, biochemical, and molecular genetics assays can identify the presence of ID-causing mutations. Each approach generates different information about the mutation, which when combined can provide a better understanding of the cause of the ID phenotype. For example, histochemical reactions have shown that phenotype-associated mutations were identified in mitochondria and not the nucleus. Subsequently, molecular genetics was used to identify the mutation as an insertion within in the mitochondrial gene *COX3* [77]. More recently, techniques such as whole genome sequencing, which have become more accessible, have been used in the identification of CNVs and their association with ID

phenotypes [70]. This approach has made it possible to identify disease-causing mutations previously undetected by other methods.

Computational methods can be used to facilitate the identification of ID-causing mutations. Statistical coupling analysis (SCA) is a method based on multivariate statistics, and can be used to find amino acid positions in a protein that are associated with ID phenotypes [78]. SCA identifies groups of residues within a multiple sequence alignment that share an underlying relationship [79,80]. This relationship is defined as co-evolution, and has been described as coordinated changes in pairs or groups of residue positions [80]. The groups identified by SCA are denominated sectors and represent clusters of amino acids that are functionally important, and if mutations occur within sector residues, protein identity, such as activity, substrate affinity, or stability, may be affected [79]. Protein sectors differ from functional units such as domains, as protein sectors are not necessarily found in sequential amino acid positions, or in continuous blocks of residues [78]. Although residues in a sector might not be continuous, when a protein folds, the seemingly distant positions may interact with each other [79].These interactions in the protein structure support that co-evolution between residues is necessary to maintain protein function.

The effect of mutations in sector residues was reported by Halabi et al. [79]. The authors analyzed the results of site-directed mutagenesis on sectors associated with catalytic activity and thermal stability. These experiments resulted in reduced enzymatic activity and thermal stability. The study also reported that each sector was functionally independent, meaning that mutations in residue positions of one sector caused one phenotype, and mutations on another sector caused a different phenotype [79]. Protein sectors can identify new sites of functional importance on the protein. Previous studies [79,81,82] have reported that residues in a sector represent functional positions, with characterized phenotypes when mutated. Further insight can be obtained by analyzing the uncharacterized residues within a sector. These positions, when mutated, are expected to have similar effects on protein function. This characteristic of SCA can provide valuable information into the study of mutations and phenotypes in ID and other diseases. SCA can be used to identify new functionally important sites in a sector associated with distinct ID phenotypes. These sites

are expected to have similar phenotypic manifestations as the known sites clustered within the sector. The use of SCA to cluster disease-causing mutations in an ID protein [83] has shown that it is possible to associate different phenotypic manifestations in syndromic ID with different protein sectors.

1.3.3 Intellectual Disability-Causing Genes

Previous studies have suggested that there may be ~800 protein-coding ID genes [71]. Of those genes, it was reported that ~450 were ID-causing genes with no overlap with other intellectual or cognitive impairments, with 400 being associated with syndromic disorders and ~50 causal of non-syndromic ID [65]. More recently, the Intellectual Disability Gene Database (http://gfuncpathdb.ucdenver.edu/iddrc/home.php) annotated 484 ID-causing genes. Notably, 95 of the protein-coding genes, representing ~19% of the list, were located on the X chromosome, a percentage similar to the number of known XLID (~15% of all protein-coding ID genes) [68].

The entries in the ID Gene Database have been curated to provide functional information about the observed manifestations that have been described through association studies and various mutations reported for the phenotypes. The effect of mutations on phenotypes can be accessed through the provided Online Mendelian Inheritance in Man (OMIM) entry number [84]. For example, *SMS* has the OMIM entry 300105 in the ID Gene Database, which can be used to access OMIM to retrieve information about mutations and the phenotypes causing Snyder-Robinson syndrome, as well as other valuable information such as literature and molecular genetics of the disease.

In the last decade, the regulatory functions of non-coding RNA genes in gene expression have been analyzed in many studies [85–87], expanding our knowledge on the importance of what was previously described as “junk DNA”, due to our lack of understanding of its role and potential functions [88]. Previous studies have described the importance of non-coding RNAs, especially long non-coding RNAs (lncRNAs) by reporting their regulatory roles [89], tissue specificity [87], and evolutionary importance in mammalians, as well as their potential involvement in ID [90]. The works by Loohuis et al. [91] and van de Vondervoort

et al. [90] also have been of importance in understanding how lncRNAs are potential ID candidates. In their studies, the authors report that a significant number of lncRNAs are found in the brain, although for the exception of a few examples (UBE3A in Angelman syndrome, BC1 in fragile X syndrome amongst others), the functions of most lncRNAs remain to be elucidated. A recent study by Chiurazzi and Perozzi [71] emphasized the importance of lncRNAs and other non-coding elements as ID candidates [92]. In this section, we have described ID-causing genes and focused on the role of protein-coding genes in ID causality. In section 1.3.6, we will discuss the potential involvement of lncRNAs in ID pathogenesis.

1.3.4 Overview of Long Non-Coding RNAs

Long non-coding RNAs (lncRNAs) are non-coding transcripts, >200 nucleotides, with relatively low conservation [85,86,92]. Previous studies [85,86,93,94] reported the structural similarities that lncRNAs share with messenger RNAs, with most of these non-coding transcripts having a 5' cap, being polyadenylated, and spliced. However, lncRNAs do not have open reading frames with protein coding potential [85,87,95,96]. Another interesting aspect of lncRNAs is their low abundance, high tissue specificity, and low evolutionary conservation when compared to protein-coding mRNAs. Because of their lower overall expression, it is more difficult to identify and quantify lncRNAs than protein-coding mRNAs [87,97,98].

The annotation of genomic elements has been a task endeavored by the GENCODE project [99]. This has included the identification of lncRNAs. Although lncRNAs, as with other non-coding elements, were initially part of what was considered "junk DNA" or termed the "dark matter" of the genome due to lack of knowledge associated to its function [97], it was later revealed that lncRNAs as well as other non-coding elements are essential to the regulation of developmental processes in complex organisms [68,87,89]. The annotation of lncRNAs in GENCODE 7 (2010), manually curated using the HAVANA manual curation protocol, identified 9,640 lncRNAs within the human genome. However, it was estimated that this number would only increase as more lncRNAs were expected to be discovered in subsequent

releases of GENCODE. For the 2011 release of GENCODE, identified as GENCODEv10, the number of lncRNAs was 10,840, corroborating their initial assessment that the number of such elements was only to become greater with each subsequent release. The current version of GENCODEv21 (2014) reports 15,877 lncRNAs (http://www.gencodegenes.org/stats.html). The increase in the number of lncRNAs demonstrates that the actual number of lncRNAs in the human genome could surpass the number of protein coding genes.

When GENCODEv7 was released, lncRNAs were classified as belonging to a specific biotype. This classification was based on the lncRNAs' positions with respect to protein-coding genes on the genome [99]. These biotypes were defined as antisense RNAs, lincRNAs, sense overlapping, sense intronic, and processed transcripts [99,100]. Antisense lncRNAs refer to transcripts that intersect either introns or exons of protein-coding genes or have evidence of antisense regulation of a coding gene. LincRNAs are defined as long intervening non-coding RNAs, with an evolutionarily conserved genomic architecture in vertebrates that consists of 2-3 exons that are slightly longer than protein-coding genes, with most of them polyadenylated. Sense intronic RNAs are defined as transcripts within introns of a coding gene found not overlapping any known exons. Another biotype of lncRNAs is the processed transcripts, which according to the GENCODE biotype definition [99] do not contain an open reading frame and cannot be placed in any other biotypes. The biotypes have continuously been expanded to include more lncRNAs as they are identified [100]. Amongst the new lncRNA biotypes are: macro lncRNAs, which are mostly unspliced transcripts that are >10 kb in length, and bidirectional promoter lncRNAs, referring to those originating within the promoter region of a protein-coding gene. These biotype definitions were used in this dissertation to generate a list of lncRNAs to identify candidate intellectual disability-associated long non-coding RNAs using an expression-based machine learning classifier [99,100].

1.3.5 Mechanisms of LncRNA Function

The functional characterization of lncRNAs is still a major challenge. However, based on those whose function has been elucidated, it has been suggested that they can interact with DNA, RNA, and

proteins (Figure 1.2). Previous studies [95,96,101] have identified that the regulatory interactions of lncRNAs occur through base-pairing with short stretches of RNA, structural domains in the chromosome, in the case of chromatin remodeling, and also by acting as an allosteric regulator in regulation of transcription factor binding to DNA [86]. The possible mechanisms of action of lncRNAs include molecular scaffolds, which bring two or more proteins into a complex or physical proximity. These protein complexes can have *trans* regulatory effects on the DNA, which can lead to chromatin remodeling. These chromatin remodeling complexes can have important effects on gene activation or silencing, modifying the expression landscape, especially during developmental stages [86]. LncRNAs can act at the transcriptional level as RNA decoys, which act as target mimics, diminishing the ability of transcription factors to bind to their DNA-binding sites; for instance, it has been noted that splicing events can be regulated by the action of lncRNAs [86,96]. Also, lncRNAs can regulate mRNA translation. A previous study [102] reported that lincRNA-p21 is involved in translational regulation of mRNAs JUNB and CTNNB1 by associating with translational machineries [102]. LncRNAs can also have an effect on the levels of microRNAs (miRNA), as miRNA sponges. The miRNA sponge acts as a decoy for miRNA target sites, reducing the effects of the miRNA complexes on their target mRNA [86,103,104]. The effect of miRNA sponges on miRNA complexes is known as titration away, or a reduction of the miRNA complexes that can degrade the mRNA, resulting in an upregulation of the levels of mRNA [86,104]. This titration effect seems to be not only limited to miRNA, but can affect other targets such as transcription factors, as reported by Gaiti et al. [104]. LncRNAs have been shown to act in both *cis* and *trans* regulatory mechanisms, making them an important part in developmental processes as well as transcription and translation [103,104].

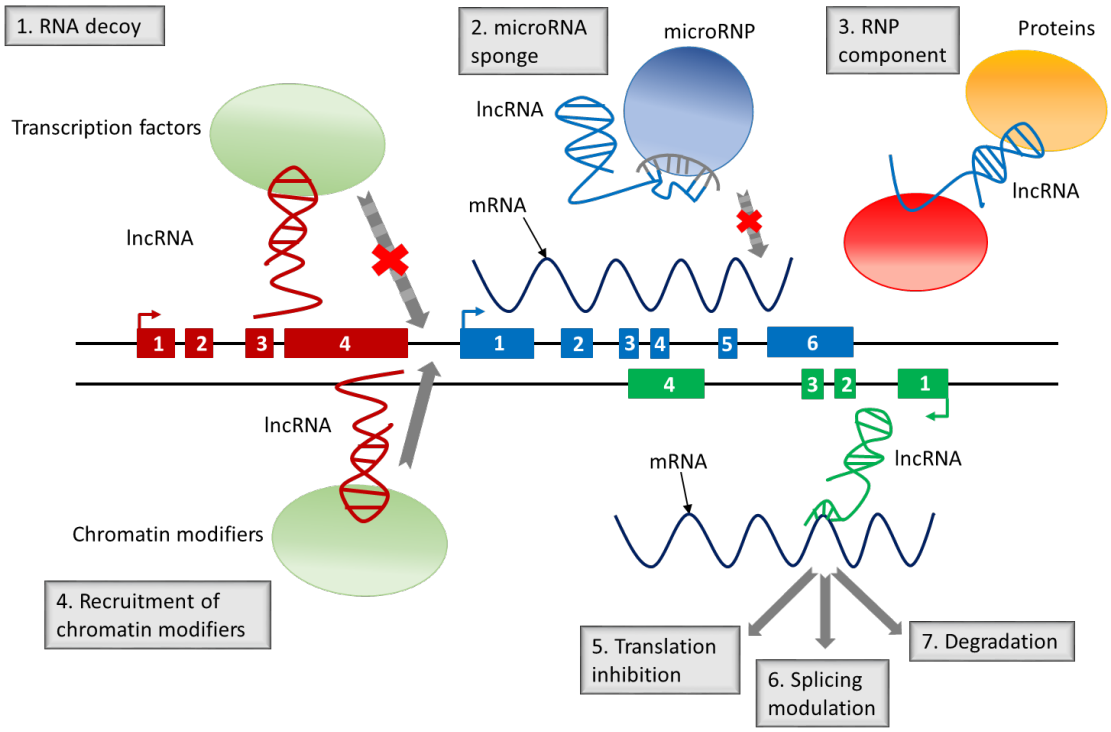

Figure 1.2 Schematic overview of LncRNA regulatory mechanisms. LncRNAs regulate gene expression through various mechanisms that include: (1) RNA decoys, by directly binding to the transcription factors. (2) LncRNAs can also act as microRNA sponges, to titrate away microRNA complexes from their mRNA targets. (3) LncRNAs can act as scaffolds, within ribonucleoprotein complexes or by binding to specific proteins. (4) LncRNAs can also recruit chromatin-modifying complexes. Other regulatory processes include: translational regulation of mRNA (5), splicing (6), and degradation of mRNA (7). The figure was adapted from Hu et al. (2012) [86].

1.3.6 Possible Involvement of LncRNAs in Intellectual Disabilities

It has been proposed that there is a correlation between the diversity of non-coding elements and the evolutionary divergence and complexity of organisms [87,90]. Evolutionary studies have identified that lncRNAs have low conservation at the sequence level. However, studies indicate that for higher animals (metazoans), lncRNAs have been important in the development of cognitive complexity, especially in the

mammalian and primate brains [87,89,90]. In fact, it has been reported that spatiotemporal expression patterns of lincRNAs in the prefrontal cortex of the brains of macaques and humans in comparative studies showed to be as conserved as those of protein-coding genes [89]. This has suggested that lncRNAs could be playing an important role in neurodevelopmental processes, and perhaps in cognitive development. Previous studies [65,89] have identified that lncRNAs could be associated with neuronal diversification as well as specification of individual neuronal subtypes. It is also possible that lncRNAs could be associated with disease in the human brain due to their role in gene regulation [89], and neurodevelopmental disorders, which could include intellectual disabilities amongst such diseases.

This possible association was explored in previous studies [90,91], which analyzed the effect that non-coding genes, such as lncRNAs, had on the deregulation of transcription and translation and their impact on ID gene expression and phenotype manifestations. Deregulation of non-coding transcripts in neurodevelopmental disorders was supported by the observation in Loohuis et al. [91], where miRNA were identified to regulate the expression of mRNAs in axosomal and the synaptodendritic compartments. The possible association between deregulation of miRNA and lncRNAs can be established due to the role of lncRNAs as miRNA sponges, titrating away the miRNA complexes that regulate mRNA translation. A study by van de Vondervoort et al. [90] reported that a major portion of the non-coding transcripts in the brain are lncRNAs, with a large fraction of brain-specific non-coding genes having altered levels in ID brain tissues. Deregulation of lncRNA expression can impact ID-risk gene expression. According to the study by van de Vondervoort et al. [90] as well as other authors [65,91], deregulation of lncRNAs can lead to abnormal transcription and translation of ID-risk genes, potentially resulting in various neurodevelopmental disroders, including fragile X syndrome, Down syndrome, and Angelman syndrome.

In chapter II of this dissertation, we developed an expression-based machine learning classifier that used ID and non-ID disease genes to identify novel ID-associated lncRNAs using the developing brain transcriptome data. Our goal was to identify novel candidates that could be used in human genetics research. In Chapter III, a three-state SVM model was developed for predicting the effect of amino acid substitutions on protein stability. We show the effect of feature selection on model performance. In Chapter IV of this

dissertation, we demonstrated the use of statistical coupling analysis to cluster amino acid residues in protein sectors and determine the effect of mutations in these groups in the development of phenotypic manifestations in ID. In Appendix A, we show our efforts in a collaborative study that focused on the use of mutual information to understand the functional divergence of the UDG protein family.

References

1. Hogeweg P. The roots of bioinformatics in theoretical biology. PLoS Comput. Biol. 2011;7:1–5.

2. Luscombe NM, Greenbaum D, Gerstein M. What is bioinformatics ? An introduction and overview. Yearb. Med. Inform. 2001;1:83–100.

3. Isea R. The Present-Day Meaning Of The Word Bioinformatics. Glob. J. Adv. Res. 2015;2:70–3.

4. Fenstermacher D. Introduction to Bioinformatics. J. Assoc. Inf. Sci. Technol. 2005;56:440–6.

5. Welch L, Lewitter F, Schwartz R, Brooksbank C, Radivojac P, Gaeta B, et al. Bioinformatics Curriculum Guidelines : Toward a Definition of Core Competencies. PLoS Comput. Biol. 2014;10:10.

6. Welch LR, Schwartz R, Lewitter F. Message from ISCB A Report of the Curriculum Task Force of the ISCB Education Committee. PLoS Comput. Biol. 2012;8:1–2.

7. Altman RB, There R, States U. A Curriculum for Bioinformatics: The Time is Ripe. Bioinformatics. 1998;14:549–50.

8. Benson DA, Cavanaugh M, Clark K, Karsch-mizrachi I, Lipman DJ, Ostell J, et al. GenBank. Nucleic Acids Res. 2017;41:36–42.

9. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000;28:235–42.

10. Boutet E, Bairoch A, Boutet E, Lieberherr D, Tognolli M, Bairoch A. Uniprotkb / Swiss-prot. Plant Bioinforma. methods Protoc. 2007;406:89–112.

11. Nekrutenko A, Taylor J. Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. Nat. Publ. Gr. [Internet]. Nature Publishing Group; 2012;13:667–72. Available from: http://dx.doi.org/10.1038/nrg3305

12. Pedregosa F, Weiss R, Brucher M. Scikit-learn : Machine Learning in Python. J. Mach. Learn. Res. 2011;12:2825–30.

13. Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A. and Huber W. BioMart and Bioconductor: A powerful link between biological databases and microarray data analysis. Bioinformatics. 2005;21:3439–40.

14. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor : open software development for computational biology and bioinformatics. Genome Biol. 2004;5:R80.

15. Giardine B, Riemer C, Hardison RC, Burhans R, Shah P, Zhang Y, et al. Galaxy : A platform for interactive large-scale genome analysis. Genome Res. 2005;15:1451–5.

16. Afgan E, Baker D, Beek M Van Den, Blankenberg D, Bouvier D, Chilton J, et al. The Galaxy platform for accessible , reproducible and collaborative biomedical analyses : 2016 update. Nucleic Acids Res. 2016;44:3–10.

17. Pautasso M. Publication Growth in Biological Sub-Fields: Patterns, Predictability and Sustainability. Sustainability. 2012;4:3234–47.

18. Pop M, Salzberg SL. Bioinformatics challenges of new sequencing technology. Trends Genet. 2008;

19. Mardis ER. The impact of next-generation sequencing technology on genetics. Trends Genet. 2008;24:133–41.

20. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat. Rev. Genet. [Internet]. Nature Publishing Group; 2015;16:321–32. Available from: http://dx.doi.org/10.1038/nrg3920

21. Larranaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., Lozano, J.A., Armañanzas, R., Santafé, G., Pérez, A. and Robles V. Machine learning in bioinformatics. Brief. Bioinform. 2006;7:86–112.

22. Wong KI, Wong PK, Cheung CS, Vong CM. Modeling and optimization of biodiesel engine performance using advanced machine learning methods. Energy [Internet]. Elsevier Ltd; 2013;55:519–28. Available from: http://dx.doi.org/10.1016/j.energy.2013.03.057

23. Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. Science (80-.). 2015;349:255–60.

24. Mjolsness E, Decoste D. Machine Learning for Science : State of the Art and Future Prospects. Science (80-.). 2001;293:2051–6.

25. Leung BMKK, Delong A, Alipanahi B, Frey BJ. Machine Learning in Genomic Medicine : A Review of Computational Problems and Data Sets. Proc. IEEE. 2016;104:176–97.

26. Clare A, King RD. Machine learning of functional class from phenotype. Bioinfor. 2002;18:160–6.

27. Blum AL, Langley P. Selection of relevant features and examples in machine. Artif. Intell. 1997;245–71.

28. Ma C, Zhang HH, Wang X. Machine learning for Big Data analytics in plants. Trends Plant Sci. [Internet]. Elsevier Ltd; 2014;19:798–808. Available from: http://dx.doi.org/10.1016/j.tplants.2014.08.004

29. Kourou K, Exarchos TP, Exarchos KP, Karamouzis M V., Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput. Struct. Biotechnol. J. [Internet]. Elsevier B.V.; 2015;13:8–17. Available from: http://dx.doi.org/10.1016/j.csbj.2014.11.005

30. Cogill S, Wang L. Support Vector Machine Model of Developmental Brain Gene Expression Data for Prioritization of Autism Risk Gene Candidates. Bioinformatics [Internet]. 2016;btw498. Available from: http://bioinformatics.oxfordjournals.org/lookup/doi/10.1093/bioinformatics/btw498

31. Kenny LC, Dunn WB, Ellis DI, Myers J, Baker PN. Novel biomarkers for pre-eclampsia detected using metabolomics and machine learning. Metabolomics. 2005;1:227–34.

32. Burbidge R, Trotter M, Buxton B, Holden S. Drug design by machine learning : support vector machines for pharmaceutical data analysis. Comput. Chem. 2001;26:5–14.

33. Kotsiantis S, Kanellopoulos D, Pintelas P. Handling imbalanced datasets : A review. GESTS Int. Trans. Comput. Sci. Eng. 2006;30:25–36.

34. Kotsiantis SB, Zaharakis ID, Pintelas PE. Machine learning : a review of classification and combining techniques. Artif. Intell. Rev. 2007;26:159–90.

35. Liu B. Web Data Mining. Second. Carey M., Ceri S, editors. Springer Science & Business Media;

36. Ben-Hur A, Ong CS, Sonnenburg S, Schölkopf B, Rätsch G. Support vector machines and kernels for computational biology. PLoS Comput. Biol. 2008;4.

37. Jain AK, Murty MN, Flynn PJ. Data Clustering : A Review. ACM Comput. Surv. 1999;31:264–323.

38. Gentleman R, Huber W, Carey VJ. Supervised Machine Learning. Bioconductor Case Stud. Springer New York; 2008. p. 121–36.

39. Dutton DM, Conroy GV. A review of machine learning. Knowl. Eng. Rev. 1996;12:341–67.

40. Inza I, Calvo B, Armañanzas R, Bengoetxea E, Larrañaga P, Lozano JA. Machine Learning : An Indispensable Tool in Bioinformatics. Bioinforma. methods Clin. Res. Springer Science & Business Media; 2010. p. 25–49.

41. Alpaydin E. Introduction to Machine Learning Second Edition. Second Edi. Dietterich T, editor. Cambridge, Massachusetts: The MIT Press; 2010.

42. Tarca AL, Carey VJ, Chen X, Romero R, Dra S. Machine Learning and Its Applications to Biology. PLoS Comput. Biol. 2007;3:e116.

43. Fawcett T. An introduction to ROC analysis. Pattern Recognit. Lett. 2006;27:861–74.

44. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim. Biophys. Acta (BBA)-Protein Struct. 1975;405:442–51.

45. Guyon I, Elisseeff A. An Introduction to Variable and Feature Selection 1 Introduction. J. Mach. Learn. Res. 2003;3:1157–82.

46. Janecek, A., Gansterer, W., Demel, M. and Ecker G. On the Relationship Between Feature Selection and Classification Accuracy. New Challenges Featur. Sel. Data Min. Knowl. Discov. 2008. p. 90–105.

47. Blagus R, Lusa L. Class prediction for high-dimensional class-imbalanced data. BMC Bioinformatics [Internet]. 2010;11:523. Available from: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-523

48. Chawla N V, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE : Synthetic Minority Over-sampling Technique. J. Artif. Intell. Res. 2002;321–57.

49. Genuer R, Poggi J, Tuleau-Malot C. Variable selection using Random Forests. Pattern Recognit. Lett. 2010;31:2225–36.

50. Haury A, Gestraud P, Vert J. The Influence of Feature Selection Methods on Accuracy , Stability and Interpretability of Molecular Signatures. PLoS One. 2011;6:1–12.

51. Wang L. Feature selection in bioinformatics. Indep. Compon. Anal. Compressive Sampling, Wavelets, Neural Net, Biosyst. Nanoeng. [Internet]. Baltimore,, MD: EEE Conference Papers; 2012. p. 6. Available from: http://dx.doi.org/10.1117/12.921417

52. Larran P, Saeys Y. Gene expression A review of feature selection techniques in bioinformatics. Bioinformatics. 2007;23:2507–17.

53. Sutha K. A Review of Feature Selection Algorithms for Data Mining Techniques. Int. J. Comput. Sci. Eng. 2015;7:63–7.

54. He Z, Yu W. Stable feature selection for biomarker discovery. Comput. Biol. Chem. [Internet]. Elsevier Ltd; 2010;34:215–25. Available from: http://dx.doi.org/10.1016/j.compbiolchem.2010.07.002

55. Mhamdi H. Feature Selection Methods for Biological Knowledge Discovery : A survey. 25th Int. Work. Database Expert Syst. 2014. p. 46–50.

56. Liu H, Li J, Wong L. A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. Genome Informatics. 2002;13:51–60.

57. Anand A, Pugalenthi G, Fogel GB, Suganthan PN. An approach for classification of highly imbalanced data using weighting and undersampling. Amino Acids. 2010;39:1385–91.

58. Longadge R, Dongre S. Class Imbalance Problem in Data Mining : Review. Int. J. Comput. Sci. Netw. 2013;2.

59. He H, Garcia EA. Learning from Imbalanced Data. IEEE Trans. Knowl. Data Eng. 2009;21:1263–84.

60. Japkowicz N, Stephen S. The class imbalance problem : A systematic study. Intell. Data Sustems. 2002;6:429–49.

61. GE B, RC P, MC M. A study of the behavior of several methods for balancing machine learning training data. ACM Sigkdd Explor. Newsl. [Internet]. 2004;6:20–9. Available from: http://portal.acm.org/citation.cfm?doid=1007730.1007735

62. Guo X, Yin Y, Dong C, Yang G, Zhou G. On the Class Imbalance Problem *. Fourth Int. Conf. Nat. Comput. 2008. p. 192–201.

63. López V, Fernández A, García S, Palade V, Herrera F. An insight into classification with imbalanced data : Empirical results and current trends on using data intrinsic characteristics. Inf. Sci. (Ny). [Internet]. Elsevier Inc.; 2013;250:113–41. Available from: http://dx.doi.org/10.1016/j.ins.2013.07.007

64. Vissers LELM, Gilissen C, Veltman JA. Genetic studies in intellectual disability and related disorders. Nat. Rev. Genet. [Internet]. 2015;17:9–18. Available from: http://www.nature.com/doifinder/10.1038/nrg3999

65. Van Bokhoven H. Genetic and Epigenetic Networks in Intellectual Disabilities. Annu. Rev. Genet. 2011;45:81–104.

66. Mefford H, Batshaw M, Hoffman E. Genomics, Intellectual Disability, and Autism. N. Engl. J. Med. 2012;366:733–43.

67. Kirk HE, Gray K, Riby DM, Cornish KM. Cognitive training as a resolution for early executive function difficulties in children with intellectual disabilities. Res. Dev. Disabil. [Internet]. Elsevier Ltd.; 2015;38:145–60. Available from: http://dx.doi.org/10.1016/j.ridd.2014.12.026

68. Gécz J, Shoubridge C, Corbett M. The genetic landscape of intellectual disability arising from chromosome X. Trends Genet. 2009;25:308–16.

69. Cooper GM, Coe BP, Girirajan S, Rosenfeld JA, Vu TH, Baker C, et al. A copy number variation morbidity map of developmental delay. Nat. Genet. [Internet]. 2011;43:838–46. Available from: http://www.nature.com/doifinder/10.1038/ng.909

70. Gilissen C, Hehir-Kwa JY, Thung DT, van de Vorst M, van Bon BWM, Willemsen MH, et al. Genome sequencing identifies major causes of severe intellectual disability. Nature [Internet]. Nature Publishing Group; 2014;511:344–7. Available from: http://www.nature.com/doifinder/10.1038/nature13394

71. Chiurazzi P, Pirozzi F, Chiurazzi P, Pirozzi F. Advances in understanding – genetic basis of intellectual disability. F1000Research [Internet]. 2016;5:599. Available from: http://f1000research.com/articles/5-599/v1

72. Hou JW, Wang TR, Chuang SM. An epidemiological and aetiological study of children with intellectual disability in Taiwan. J. Intellect. Disabil. Res. 1998;42:137–43.

73. Becerra-Solano LE, Butler J, Castañeda-Cisneros G, McCloskey DE, Wang X, Pegg AE, et al. A missense mutation, p.V132G, in the X-linked spermine synthase gene (SMS) causes Snyder-Robinson syndrome. Am. J. Med. Genet. Part A. 2009;149:328–35.

74. Zhang Z, Norris J, Kalscheuer V, Wood T, Wang L, Schwartz C, et al. A Y328C missense mutation in spermine synthase causes a mild form of snyder-robinson syndrome. Hum. Mol. Genet. 2013;22:3789–97.

75. Villard L, Fonte M. Alpha-Thalassemia X-Linked Mental Retardation Syndrome. Eur. J. Hum. Genet. 2002;10:223–5.

76. Gibbons RJ, Higgs DR. Molecular – Clinical Spectrum of the ATR-X Syndrome. Am. J. Med. Genet. 2000;97:204–12.

77. Tiranti V, Corona P, Greco M, Taanman JW, Carrara F, Lamantea E, et al. A novel frameshift mutation of the mtDNA COIII gene leads to impaired assembly of cytochrome c oxidase in a patient affected by Leigh-like syndrome. Hum. Mol. Genet. 2000;9:2733–42.

78. Teşileanu T, Colwell LJ, Leibler S. Protein Sectors : Statistical Coupling Analysis versus Conservation. PLoS Comput. Biol. 2015;11:e1004091.

79. Halabi N, Rivoire O, Leibler S, Ranganathan R. Protein sectors: evolutionary units of three-dimensional structure. Cell. 2009;138:774–86.

80. Juan D De, Pazos F, Valencia A. Emerging methods in protein co-evolution. Nat. Rev. Genet. [Internet]. Nature Publishing Group; 2013;14:249–61. Available from: http://dx.doi.org/10.1038/nrg3414

81. Smock RG, Rivoire O, Russ WP, Swain JF, Leibler S, Ranganathan R, et al. An interdomain sector mediating allostery in Hsp70 molecular chaperones. Mol. Syst. Biol. [Internet]. Nature Publishing Group; 2010;6:1–10. Available from: http://dx.doi.org/10.1038/msb.2010.65

82. Xu F, Du P, Shen H, Hu H, Wu Q, Xie J, et al. Correlated Mutation Analysis on the Catalytic Domains of Serine / Threonine Protein Kinases. PLoS One. 2009;4:e5913.

83. Guevara-Coto J, Schwartz CE, Wang L. Protein sector analysis for the clustering of disease-associated mutations. BMC Genomics [Internet]. BioMed Central Ltd; 2014;15 Suppl 1:S4. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25559331

84. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 2005;33:514–7.

85. Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M.C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C. and Kodzius R, Lenhard B, Wells C, Kodzius R. The Transcriptional Landscape of the Mammalian Genome. Science (80-.). 2005;309:1559–63.

86. Hu W, Alvarez-Dominguez JR, Lodish HF. Regulation of mammalian cell differentiation by long non-coding RNAs. EMBO Rep. 2012;13:24–6.

87. Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, et al. The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature. 2014;505:635.

88. Ma L, Bajic VB, Zhang Z. On the classification of long non-coding RNAs. RNA Biol. [Internet]. 2013;10:924–33. Available from: http://www.tandfonline.com/doi/abs/10.4161/rna.24604

89. He Z, Bammann H, Han D, Xie G, Khaitovich P. Conserved expression of lincRNA during human and macaque prefrontal cortex development and maturation. Rna [Internet]. 2014;20:1103–11. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4074677/

90. van Devondervoort IIGM, Gordebeke PM, Khoshab N, Tiesinga PHE, Buitelaar JK, Kozicz T, et al. Long non-coding RNAs in neurodevelopmental disorders. Front. Mol. Neurosci. [Internet]. 2013;6:53. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3874560&tool=pmcentrez&rendertype=abstract

91. Loohuis NO, Kos A, Martens GJM, Van Bokhoven H, Kasri NN, Aschrafi A. MicroRNA networks direct neuronal development and plasticity. Cell. Mol. Life Sci. 2012;69:89–102.

92. Wang L, Huang C, Yang MQ, Yang JY. BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features. BMC Syst. Biol. 2010;4 Suppl 1:S3.

93. The GTEx Consortium, Welter D, MacArthur J, Morales J, Burdett T, Hall P, et al. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science (80-.). [Internet]. 2015;348:648–60. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25954001%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4547484

94. Kapranov, P., Cheng, J., Dike, S., Nix, D.A., Duttagupta, R., Willingham, A.T., Stadler, P.F., Hertel, J., Hackermüller, J., Hofacker, I.L. and Bell I. RNA Maps Reveal New RNA Classes and a Possible Function for Pervasive Transcription. SKapranov, P., Cheng, J., Dike, S., Nix, D.A., Duttagupta, R., Willingham, A.T., Stadler, P.F., Hertel, J., Hackermüller, J., Hofacker, I.L. Bell, I.cience. 2007;316:1484–9.

95. Wang KC, Chang HY. Molecular mechanisms of long noncoding RNAs. Mol. Cell. 2012;43:904–14.

96. Rasool M, Malik A, Zahid S, Abdul M, Ashraf B, Husain M, et al. Non-coding RNAs in cancer diagnosis and therapy. Non-coding RNA Res. [Internet]. Elsevier Ltd; 2016;1:69–76. Available from: http://dx.doi.org/10.1016/j.ncrna.2016.11.001

97. Derrien T, Guigó R, Johnson R. The long non-coding RNAs : a new (p) layer in the “ dark matter .” Front. Genet. 2012;2:1–6.

98. Rinn JL, Chang HY. Genome regulation by long noncoding RNAs. Annu. Rev. Biochem. 2013;81:145–66.

99. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE : The reference human genome annotation for The ENCODE Project. Genome Res. 2012;22:1760–74.

100. Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, et al. Ensembl 2015. Nucleic Acids Res. 2015;43:662–9.

101. Wilusz JE, Sunwoo H, Spector DL. Long noncoding RNAs : functional surprises from the RNA world. Genes Dev. 2009;23:1494–504.

102. Yoon JH, Abdelmohsen K, Srikantan S, Yang X, Martindale JL, De S, et al. LincRNA-p21 Suppresses Target mRNA Translation. Mol. Cell [Internet]. Elsevier; 2012;47:648–55. Available from: http://dx.doi.org/10.1016/j.molcel.2012.06.027

103. Militello G, Weirick T, John D, Do C, Dimmeler S, Uchida S. Screening and validation of lncRNAs and circRNAs as miRNA sponges. Brief. Bioinform. 2017;bbw053.

104. Gaiti F, Fernandez-Valverde SL, Nakanishi N, Calcino AD, Yanai I, Tanurdzic M, et al. Dynamic and Widespread lncRNA Expression in a Sponge and the Origin of Animal Complexity. Mol. Biol. Evol. 2015;32:2367–82.

CHAPTER II – EXPRESSION-BASED PREDICTION OF CANDIDATE LONG NON-CODING RNAS ASSOCIATED WITH INTELLECTUAL DISABILITY

Jose A. Guevara-Coto[1], Brian L. Gudenas[1], Charles E. Schwartz[2], and Liangjiang Wang[1]

[1]Department of Genetics and Biochemistry, Clemson University, Clemson, SC 29634, USA

[2]J.C. Self Research Institute of Human Genetics, Greenwood Genetic Center, Greenwood, SC 29646, USA

To be submitted: *BMC Bioinformatics*

Abstract

Background: Intellectual disabilities (ID) are neurodevelopmental disorders with an estimated prevalence in the population of 1-3%. These disorders have a complex genetic etiology, with possible involvement of not only protein-coding genes but also non-coding elements, including long non-coding RNAs (lncRNAs). However, the role of lncRNAs in ID is still unclear although evidence supports possible associations of lncRNAs with disease causality and development. To identify candidate lncRNAs associated with ID, predictive methods need to be based on gene expression patterns.

Results: We have developed an expression-based classification model using Support Vector Machines. The model was trained using a developmental brain transcriptome dataset, with known ID-causing genes as the positive instances and non-ID disease genes as the negatives. We used Random Forest-based methods for

feature selection, however, this did not significantly improve model performance when compared with the full feature set. To address the class imbalance problem, we used synthetic oversampling of the minority class, which resulted in the best-performing model with the overall accuracy of 80.22% (76.34% sensitivity and 83.95% specificity) in tenfold cross-validations. This model was then used to predict ~760 candidate lncRNAs associated with ID. These candidate lncRNAs were analyzed for co-enrichment with known ID genes in co-expression modules. We identified the co-enrichment in modules related to neurodevelopmental and synaptic processes.

Conclusions: A new gene expression-based classifier has been constructed and used to identify candidate lncRNAs associated with ID. The results from our analyses suggest a potential role of these candidate lncRNAs in development and etiology of ID.

2.1 Background

Intellectual disabilities (ID) are neurodevelopmental disorders, which are estimated to be prevalent in 1-3% of the general population [1], and are commonly diagnosed prior to 18 years of age [2]. ID patients are associated with intelligence quotient (IQ) sores <70 and diminished cognitive abilities [1,2]. ID can be syndromic, meaning that there are also morphological abnormalities or metabolic disorders, or can be non-syndromic, in which the intellectual impairment is the only clinical manifestation [3]. It is also possible for ID to be associated with other conditions such as autism spectrum disorders (ASD) [2]. This complexity of ID is also represented in the possible causes. Although exogenous factors can influence ID causality, genetic factors remain the major cause of ID [1–4]. To further our understanding of ID, we need to expand the knowledge of ID-causing genes. Currently, almost all of the known ID-causing genes are protein-coding, whereas the numerous long non-coding RNAs (lncRNAs) have not been thoroughly examined for their possible roles in the etiology of ID [1,2,4].

The importance of non-coding transcripts and their roles in gene expression regulation have been expanded since their discovery, with special interest in lncRNAs and their tissue-specific expression [5,6]. These transcripts, which are more than 200 nucleotides long and are a heterogeneous group, are involved in a variety of biological processes [7–10]. In recent years, more and more lncRNAs have been identified and studied becoming interesting candidates that could potentially be inovlved in diseases such as neurodevelopmental disorders or cancer. Additionally, lncRNAs have known functions such as X-chromosome inactivation in mammals [8] or interacting with chromatin remnodelling complexes [5]. A recent study used an integrative approach to link lncRNAs with neurodevelopmental functions, and to examine their possible associations with ID [2]. However, the extent and number of lncRNAs associated with ID remains to be discovered, and methods capable of identifying and providing new insights into the possible roles and number of genes involved in the causality and development of ID are necessary to further our understanding of disease etiolog

By harnessing information from known ID-causing genes it is possible to use machine learning algorithms to discover novel ID-associated lncRNAs. For a classification task such as this, Support Vector Machines (SVMs) can provide accurate model building and efficient computational cost [11]. Besides the success in predicting DNA/RNA binding residues [12] and splice sites [11,13], recent research has reported a number of efficient and accurate SVM classifiers using gene expression data [14–16]. Moreover, SVMs have the advantage of combining parameter optimization and feature selection methods [17, 18], and using approaches to deal with class imbalance such as oversampling and under-sampling [19, 20]. In this study, we developed a new SVM classifier based on developmental brain gene expression patterns to identify candidate lncRNAs associated with ID. The idea behind this was due to the non-coding nature of lncRNAs and thus to avoid using protein sequence-based features. We also performed gene co-expression network analysis to further understand the possible function of these candidate lncRNAs.

2.2 Methods

Data preparation

The developmental brain gene expression dataset was obtained from the BrainSpan website at the Allen Institute [21]. From this dataset, the expression profiles of ID-causing genes (IDGs) and non-ID disease genes (NIDGs) were extracted and used for classifier construction. The filtering was done based on their known identity as either ID-causing or disease genes that are not causal of IDs. Classification of ID genes was based on previous studies [22–24], which provided functional annotation and association of these genes with ID. We also used the information available from the ID Gene Database Project (http://gfuncpathdb.ucdenver.edu/iddrc/iddrc/home.php), which compiled a comprehensive list of known ID genes. The training dataset consisted of 423 IDGs and 1830 NIDGs. A second set of expression profiles was also extracted for brain-expressed lncRNAs. This second set was used by the classifier to identify new candidate lncRNA associated with ID. For each dataset, the gene expression values were log2-transformed to make the variation magnitudes amongst individual gene expression profiles similar.

Classifier training and testing

Prediction of ID-associated lncRNAs can be regarded as a binary classification problem, for which a classifier is constructed to discriminate the belonging of an object to one of two classes – positive or negative [11, 25]. In this case, genes are classified into either of two classes (ID or non-ID disease causing). We used Support Vector Machines (SVMs) to develop the classifier. SVMs use two key ideas, a separation margin and a kernel function, to solve the classification task [11, 25]. The use of SVMs in this study was based on the reported success of SVMs in developing expression-based classification models. Previous studies developed high-performing SVM models capable of predicting gene associations with cancer, ASD, and neurodegenerative diseases [14, 16, 26–28]. In addition, we also tested two other machine learning algorithms, Random Forests and naïve Bayes, for classifier construction.

In this study, the SVM classifier was developed using the radial basis function kernel. The misclassification cost (C) and gamma parameters were optimized for classifier training. We used a two-

step grid search to find the optimal *C* and gamma combination. The preliminary search tested the possible combinations of *C* values ranging from 0.5 to 10 with different gamma values from 1e-6 to 1. Once this step was completed, a refining grid search was done to determine the best-performing combination of *C* and gamma values. The SVM classifier was trained using the package e1071 [29] for R version 3.3.3 [30].

Model performance was evaluated initially with fivefold cross-validation and subsequently with tenfold cross-validation. The following performance metrics were used to evaluate the models:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (1)$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}} \quad (2)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} \quad (3)$$

$$\text{F-Measure} = 2\times\frac{\text{precision} \times \text{recall}}{\text{precision+recall}} = \frac{2\text{TP}}{(2\text{TP} + \text{FP} + \text{FN})} \quad (4)$$

$$\text{MCC} = \frac{\text{TP}\times\text{TN-FP}\times\text{FN}}{\sqrt{(\text{TP+FP})(\text{TP+FN})(\text{TN+FP})(\text{TN+FN})}} \quad (5)$$

In the above equations, TP = true positive, TN = true negative, FP = false positive, and FN = false negative. The F-measure, also referred as the F1 score, is often used to evaluate models constructed from an imbalanced dataset, and has the range from 0 (worst) to 1 (best). In this study, the overall accuracy, sensitivity, specificity and F-measure were calculated using the performanceEstimation package [31]. Model selection was based on the Matthews Correlation Coefficient (MCC) and the area under the receiver operating characteristic (ROC) curve or ROC-AUC. The ROC was plotted using the ROCR package [32].

Feature selection

The reduction of redundant and noisy features to improve classifier performance were done using two feature selection algorithms: Random Forests with the Boruta R package [33, 34] and recursive feature elimination with the caret package [18, 35]. Each method generated a list of selected features, and the effect of reducing the feature space on model performance was determined. For each selected feature subset, the classifier was re-trained and tested using tenfold cross-validation to determine the performance metrics, including sensitivity, specificity, MCC, and ROC-AUC.

Minority class oversampling

In this study, the training dataset is imbalanced, with many more NIDGs than IDGs. When models are trained using an imbalanced dataset, there are strategies to counter the difference between the under- and over-represented classes and thus increase the model's ability to discriminate between the classes. These strategies include the adjustment of class weights, under-sampling of the majority class, and over-sampling of the minority class. Although under-sampling was shown to be effective [20], this could reduce the information for model construction. The reduction of the majority class (NIDGs) could mitigate the model's ability to recognize the hidden patterns that differentiate between IDGs and NIDGs. To maximize the amount of information for classifier construction, we used synthetic minority over-sampling technique (SMOTE) [19, 36]. The approach computes synthetic examples of the minority class by interpolation of the *k* nearest neighbors [36]. With the over-sampling percentage set to 325 and the *k* nearest neighbors to 5, this operation resulted in 1835 positive instances (470 IDGs and 1365 synthetic samples) and 1825 NIDGs in the balanced dataset for model training. The synthetic over-sampling was done with the DMwR package [37].

Prediction of ID-associated lncRNAs

Once the model was trained, it was used to predict candidate lncRNAs associated with ID. The list of brain-expressed lncRNAs was composed of ~9500 transcripts (SI Table 1), of potential association with

ID. The expression dataset was log2-transformed in R. The list of lncRNA biotype definitions was obtained using the biomaRt package in R [38] and matched to the BrainSpan gencode v10 release [39]. The lncRNAs were identified using the Ensembl IDs [40, 41] in the further analysis.

Gene co-enrichment analysis

The co-enrichment analysis utilized a human brain gene co-expression network, which was generated in a previous study [24] with the WGCNA package [42]. Using the module information, we analyzed for co-enrichment of candidate lncRNAs with known ID-causing genes. For the enriched modules, gene over-representation values were calculated using Fisher's exact test. The resulting P-values were adjusted to calculate the false discovery rate. To be considered as statistically significant enrichment of a gene list, the adjusted P-value should be <0.05 and the odds ratio (calculated with Fisher's exact test) >1. We also compared the ID-associated lncRNAs identified in this study with a set of prioritized candidate lncRNAs from the previous study [24]. The resulting matches were tested for statistical significance by constructing a contingency table and using Fisher's exact test (true odds ratio >1 and P-value <0.05).

Gene expression pattern analysis

The analysis of lncRNA expression in different tissue types was performed to determine if the candidates were specifically expressed in the brain. We used the Genotype-Tissue Expression (GTEx) dataset with 53 human tissue types [43]. Heatmaps were generated to compare the expression of candidate lncRNAs in the brain and other tissues.

2.3 Results and discussion

Development of an expression-based classifier for ID gene prediction

Since the performance of machine learning algorithms may be influenced by the type of problem, we used three algorithms, Support Vector Machine (SVM), naïve Bayes and Random Forest, to develop a

model to discriminate between ID-causing genes and non-ID disease genes using the BrainSpan expression data [21]. The performance metrics shown in Table 2.1 suggested that the model trained with SVM was better suited for further optimization given that MCC and ROC-AUC were significantly higher for the SVM. The other metrics such as the F-measure, commonly used for imbalanced datasets, as well as the sensitivity and specificity (Table 2.1) also supported our conclusion regarding the SVM's ability to predict ID genes based on developmental brain gene expression profiles.

Feature selection was previously shown to improve model performance by reducing redundancy and background noise [17, 33]. We thus compared the performance of classifiers constructed using the full feature set and feature subsets selected by Random Forests. However, for the prediction of ID-causing genes using expression profiles, feature selection did not significantly improve the predictive power (Table2.2 and Figure 2.1). When comparing the full feature set model (SVM_Full) and the selected feature model (SVM_176), they had similar MCC and ROC-AUC, whereas the measurements for accuracy, sensitivity and F-measure were better for the full feature set model (Table 2.2). The results suggest that developing an accurate model for ID gene prediction may need the full feature set with 524 features representing gene expression at various developmental stages and brain regions. This is consistent with ID as complex genetic disorders and that ID diagnosis can occur until 18 years of age [1, 2].

Balancing the dataset by synthetic oversampling of the minority class improves model performance

Imbalanced datasets are commonplace for machine learning problems in biology [11, 44]. The effect of having differentially represented classes on model performance is well known [44]. Because of this, various strategies have been developed to counter the bias of an over-represented class in the dataset during model construction. Initially, we used the adjustment of the class weight parameter for training SVM models based on the ratio between IDGs and NIDGs (~1:4), assigning more weight to the under-represented class. We expected this to improve model performance. However, that was a rather simple technique, and we explored the possibility of applying a more advanced method to our imbalanced dataset.

We selected the synthetic minority oversampling technique (SMOTE) [19, 36] to balance the training dataset in this study. By using the available instances of the under-represented class, IDGs in this case, SMOTE could derive synthetic samples from the existing IDGs. This approach was deemed appropriate as there was no loss of information about either class, and we expected that new ID genes could have similar expression patterns as the synthetic samples. The ability of SMOTE to maximize the use of class information was a critical factor in balancing method selection when comparing it to other approaches, especially with under-sampling, where selection of a subset of NIDGs would have allowed to balance the dataset with the possible outcome of losing some information about the hidden pattern in the NIDGs.

The model constructed using the SMOTE-balanced dataset with the full feature set performed better than the classifier built through the class weight adjustment (Table 2.2 and Figure 2.1). The performance metrics such as MCC and ROC-AUC indicate that class balancing provided the better approach for constructing an accurate model. Implementation of SMOTE was also found to be more effective for model performance improvement than feature selection in this study. However, the use of oversampling alone did not ensure high model performance, as observed when comparing the full and reduced feature sets with balanced classes (Table 2.2 and Figure 2.1). Therefore, the model constructed using the balanced dataset with all 524 spatiotemporal expression features achieved the best performance, and was used to predict new candidate lncRNAs associated with ID.

Identification of new candidate lncRNAs associated with ID

The fine-tuned SVM model was used to predict 767 possible ID-associated lncRNAs from a list of ~9500 lncRNAs within the BrainSpan dataset [21]. These candidate lncRNAs were subsequently analyzed for co-enrichment with known ID genes based on the results from a previous study [24], in which a brain gene co-expression network was created to identify the ID gene-enriched modules. We examined whether the possible ID-associated lncRNAs were co-enriched in any of these modules that were associated with neurodevelopmental functions. The analysis revealed statistically significant co-enrichment in two

modules, the purple module and the greenyellow module (Figure 2.2). These two modules were enriched for Gene Ontology [45, 46] terms associated with biological processes including neuron differentiation and synapse organization (Table 2.3). We also compared the list of candidate lncRNAs predicted in this study with the list of ID-associated lncRNAs previously published by Gudenas et. al. [24], identified using co-expression networks to determine if there was a significant overlap between the two lists. We identified 55 common ID-associated lncRNAs (SI Table 2) across multiple modules, including the purple module and the greenyellow module. Representative overlapping lncRNAs from the co-enriched modules are shown in Table 2.3. Although gene co-expression analysis has been used to predict the possible lncRNA functions/roles, it is important to note that co-expression is not synonymous to causality [47]. There are some caveats with the guilty-by-association approach; however, it may provide some insights about the candidate lncRNAs and pathways associated with ID.

Next, we utilized the GTEx dataset [43] to examine the expression patterns of the candidate lncRNAs. Notably, the possible ID-associated lncRNAs overlapped with those from the previous study [24] were expressed in various brain regions (Figure 2.3). We also observed a higher expression of lncRNAs in testes, which is not unexpected, as it has been reported that testes followed by neural tissues, and in some species the ovary, are where the largest amount of lncRNAs are expressed [48]. We also identified expression of lncRNAs in the pituitary gland, which is expected as these genes have a regulatory role in the endocrine system [49]. Primate brain studies have discovered significant numbers of lncRNAs expressed in different developmental stages in brain regions like the prefrontal cortex. The degree of specificity in the expression patterns appears similar to those of protein coding genes [50]. The relatively high expression of the ID-associated lncRNAs in brain regions (Figure 2.3) is in accordance with their possible function in neurodevelopmental processes.

2.4 Conclusions

We have developed a new machine learning model for genome-wide identification of candidate ID-associated lncRNAs using brain gene expression profiles. Three different learning algorithms were examined for model construction. Using a training dataset of ID genes and non-ID disease genes encoded with spatiotemporal expression features, we developed and compared the performance of these models. The SVM model was selected due to its high performance, and then used for fine-tuning to further improve its accuracy. Next, feature selection and class balancing methods were implemented to improve model performance. For ID gene prediction using expression data, synthetic oversampling of the minority class to address class imbalance, but not feature selection, was found to significantly improve model performance. Finally, the best-performing SVM model was used to predict 767 possible ID-associated lncRNAs, which were then analyzed for co-enrichment with known ID genes in co-expression modules. We identified 55 candidate lncRNAs that were co-expressed with known ID genes and showed relatively high expression in various brain regions. These lncRNAs could serve as high-priority targets for experimental investigation into the roles of lncRNAs in brain development and ID pathogenesis.

Tables

Table 2.1 Performance of the SVM, Naïve Bayes, and Random Forest models based on tenfold cross-validation. The models were constructed using the full feature set composed of 524 spatiotemporal expression features.

	Accuracy	Sensitivity	Specificity	F-Measure	MCC	AUC
SVM	70.61%	70.48%	70.65%	0.8046	0.3562	0.7630
Naïve Bayes	64.98%	64.62%	65.12%	0.4286	0.2582	0.6573
Random Forest	79.34%	10.88%	96.93%	0.1756	0.1376	0.5414

Table 2.2 Performance of the SVM models after feature selection and synthetic oversampling of the minority class (SMOTE). The models were constructed using the full set of 524 features or the reduced set of 176 features as indicated.

	Accuracy	Sensitivity	Specificity	F-Measure	MCC	AUC
SVM_Full	71.66%	73.42%	64.64%	0.8046	0.3562	0.7630
SVM_Full_SMOTE	80.22%	76.34%	83.95%	0.7902	0.6466	0.8747
SVM_176	70.61%	70.48%	70.65%	0.4947	0.3630	0.7651
SVM_176_SMOTE	75.62%	82.74%	68.21%	0.7765	0.5662	0.8362

Table 2.3 List of selected candidate lncRNAs from this study. These ID-associated lncRNAs were also identified in a previous study [24], and co-enriched with ID genes in brain co-expression modules. Additional information from [24] includes the type of copy number variant (CNV), the highest correlated ID gene, and the enriched GO term.

Ensembl ID	lncRNA Gene Symbol	Type of CNV	Module	Highest Correlated ID Gene	GO Enrichment Term
ENSG00000245864	CTC-467M3.1	de Novo	purple	MEF2C	cell part morphogenesis; synapse organization; axonogenesis
ENSG00000246477	AF131216.6	Decipher	purple	PAK3	cell part morphogenesis; synapse organization; axonogenesis
ENSG00000224855	OPA1-AS1	Decipher	purple	BRAF	cell part morphogenesis; synapse organization; axonogenesis
ENSG00000236850	LL22NC03-80A10.6	Decipher	purple	RPGRIP1L	cell part morphogenesis;

					synapse organization; axonogenesis
ENSG00000183308	AC005037.3	Decipher	purple	PAK3	cell part morphogenesis; synapse organization; axonogenesis
ENSG00000257769	CTB-193M12.1	Decipher	purple	USP9X	cell part morphogenesis; synapse organization; axonogenesis
ENSG00000228794	RP11-206L10.11	de Novo	greenyellow	MBD5	neuron projection development; neuron differentiation; neuron development
ENSG00000093100	XXbac-B461K10.4	Decipher	greenyellow	RALGDS	neuron projection development; neuron differentiation; neuron development

ENSG00000233058	AC046143.7	Decipher	greenyellow	MBD5	neuron projection development; neuron differentiation; neuron development
ENSG00000254635	RP11-164A7.1	de Novo	greenyellow	CEP290	neuron projection development; neuron differentiation; neuron development
ENSG00000188242	CTD-2228K2.5	Decipher	greenyellow	PPOX	neuron projection development; neuron differentiation; neuron development

Figures

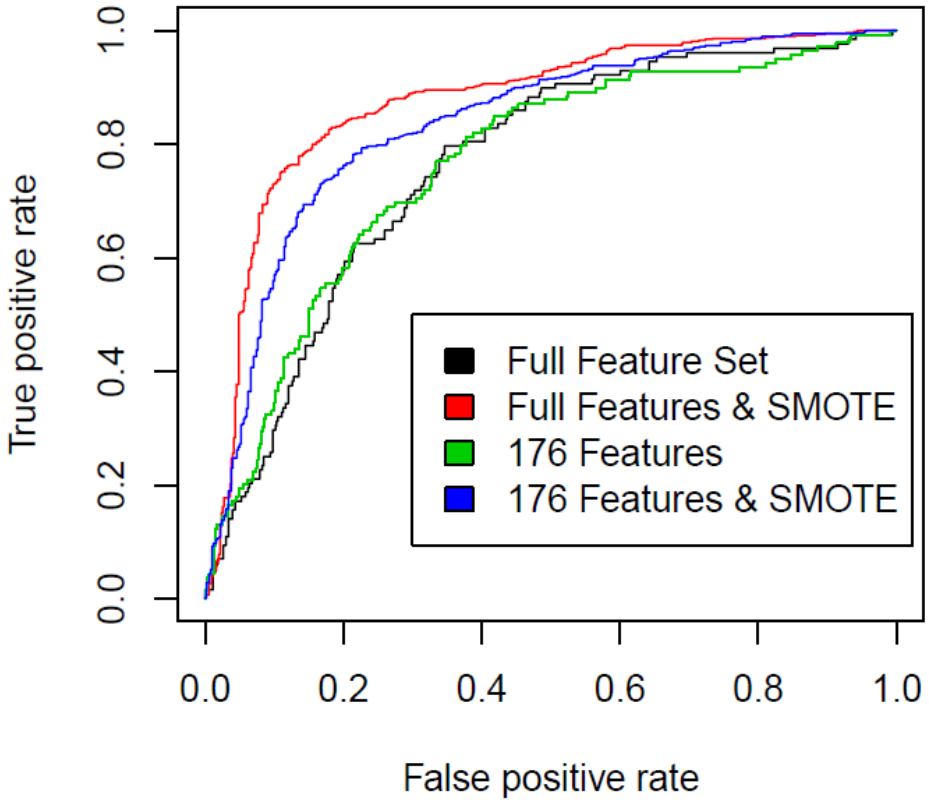

Figure 2.1 ROC curves of the SVM models constructed using the full and reduced feature sets in combination with the SMOTE method to overcome class imbalance.

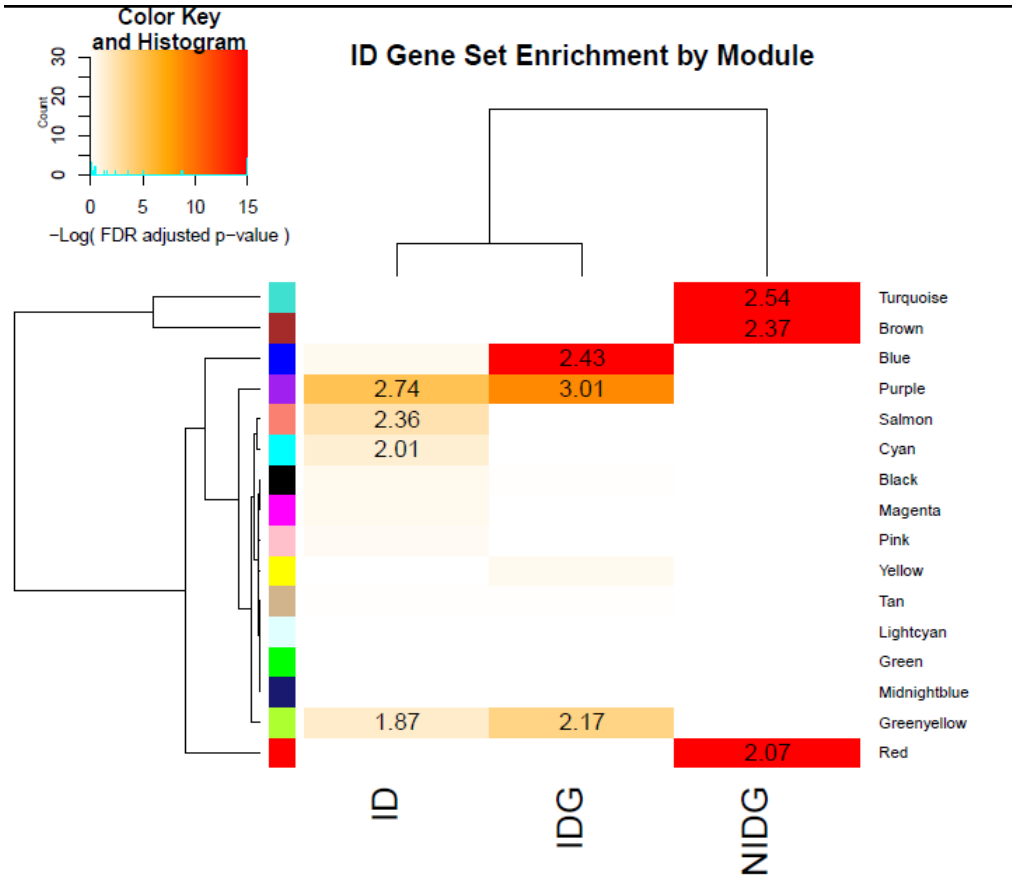

Figure 2.2 Co-enrichment of lncRNAs and ID genes in brain co-expression modules. The modules were obtained from a previous study [24], and the co-enrichment was identified for two modules, purple and greenyellow. IDG stands for the possible ID-associated lncRNAs, ID for the known ID genes, and NIDG for non-ID disease genes. The numbers in the boxes represent the odds-ratio from Fisher's exact test.

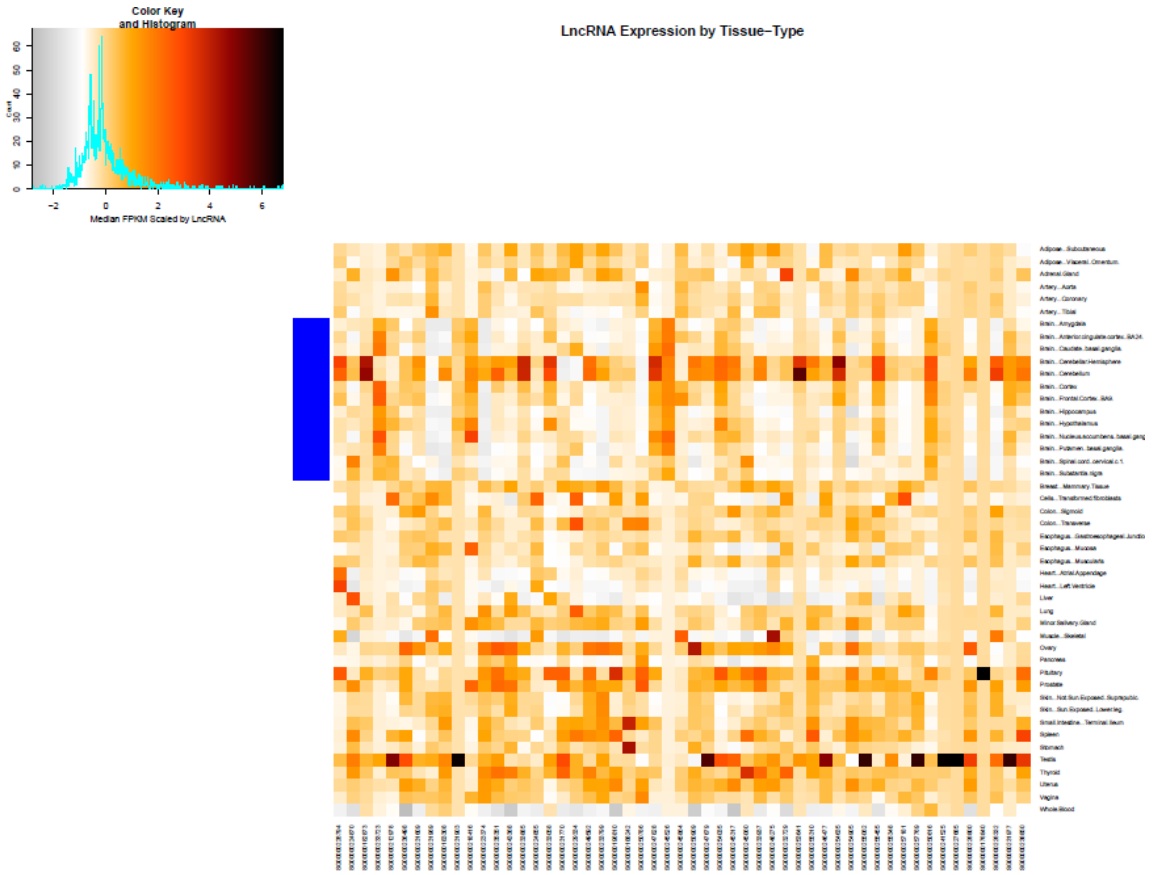

Figure 2.3 Expression patterns of the ID-associated candidate lncRNAs. The lncRNA candidates were co-enriched with ID genes in brain co-expression modules from a previous study [24]. The GTEx dataset [43] was used to examine lncRNA expression in different tissues, including brain (highlighted in blue).

References

1. Vissers LELM, Gilissen C, Veltman JA. Genetic studies in intellectual disability and related disorders. Nat. Rev. Genet. 2015;17(1):9–18.

2. D'haene E, Jacobs EZ, Volders PJ, De Meyer T, Menten B, Vergult S. Identification of long non-coding RNAs involved in neuronal development and intellectual disability. Scientific Reports. 2016;6:28396.

3. Van Bokhoven H. Genetic and Epigenetic Networks in Intellectual Disabilities. Annu. Rev. Genet. 2011;45:81–104.

4. Gécz J, Shoubridge C, Corbett M. The genetic landscape of intellectual disability arising from chromosome X. Trends Genet. 2009;25:308–16.

5. Quinn JJ, Chang HY. Unique features of long non-coding RNA biogenesis and function. Nat. Rev. Genet. 2015;17:47–62.

6. Li L, Chang HY. Physiological roles of long noncoding RNAs: insight from knockout mice. Trends Cell Biol. 2014;24:594–602.

7. Huarte M. The emerging role of lncRNAs in cancer. Nat. Med. 2015;21:1253–61.

8. van Devondervoort IIGM, Gordebeke PM, Khoshab N, Tiesinga PHE, Buitelaar JK, Kozicz T, et al. Long non-coding RNAs in neurodevelopmental disorders. Front. Mol. Neurosci. 2013;6:53.

9. Han P, Chang C-P. Long non-coding RNA and Chromatin Remodeling. RNA Biol. 2015;12(10):1094–98.

10. Zhang F, Lupski JR. Non-coding genetic variants in human disease. Hum. Mol. Genet. 2015;24:R102–10.

11. Ben-Hur A, Ong CS, Sonnenburg S, Schölkopf B, Rätsch G. Support vector machines and kernels for computational biology. PLoS Comput. Biol. 2008;4(10). e1000173.

12. Wang L, Huang C, Yang MQ, Yang JY. BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features. BMC Syst. Biol. 2010;4 Suppl 1:S3.

13. Sonnenburg S, Schweikert G, Philips P, Behr J, Rätsch G. Accurate splice site prediction using support vector machines. BMC Bioinformatics 2007;8:S7.

14. Bastani M, Vos L, Asgarian N, Deschenes J, Graham K, Mackey J, et al. A machine learned classifier that uses gene expression data to accurately predict estrogen receptor status. PLoS One. 2013;8.

15. Vanitha CD, Devaraj D, Venkatesulu M. Gene expression data classification using Support Vector Machine and mutual information-based gene selection. Procedia Comput. Sci. 2014;47:13–21.

16. Guan P, Huang D, He M, Zhou B. Lung cancer gene expression database analysis incorporating prior knowledge with support vector machine-based classification method. J. Exp. Clin. Cancer Res. 2009;28:103.

17. Genuer R, Poggi J, Tuleau-Malot C. Variable selection using Random Forests. Pattern Recognit. Lett. 2010;31:2225–36.

18. Díaz-Uriarte R, De Andrés SA. Gene selection and classification of microarray data using random forest. BMC Bioinformatics. 2006;7:3.

19. Chawla N, Bowyer K, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. J. Artif. Intell. Res. 2002;16:321–57.

20. Anand A, Pugalenthi G, Fogel GB, Suganthan PN. An approach for classification of highly imbalanced data using weighting and undersampling. Amino Acids. 2010;39:1385–91.

21. Hawrylycz MJ, Lein ES, Guillozet-Bongaarts AL, Shen EH, Ng L, Miller JA, et al. An anatomically comprehensive atlas of the adult human brain transcriptome. Nature Nature Publishing Group; 2012;489:391–9.

22. Chiurazzi P, Pirozzi F, Chiurazzi P, Pirozzi F. Advances in understanding – genetic basis of intellectual disability. F1000Research 2016;5:599.

23. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 2005, 33;514–7.

24. Gudenas BL, Wang L. Gene coexpression networks in human brain developmental transcriptomes implicate the association of long noncoding RNAs with intellectual disability. Bioinform. Biol. Insights. 2015;9:21–7.

25. Noble WS. What is a support vector machine? Nat. Biotechnol. 2006;24:1565–7.

26. Krishnan A, Zhang R, Yao V, Theesfeld CL, Wong AK, Tadych A, et al. Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. Nat. Neurosci. 2016;19:1454–62

27. Cogill S, Wang L. Support Vector Machine Model of Developmental Brain Gene Expression Data for Prioritization of Autism Risk Gene Candidates. Bioinformatics 2016;32(23):3611–18.

28. López MM, Ramírez J, Górriz JM, Álvarez I, Salas-Gonzalez D, Segovia F, et al. SVM-based CAD system for early detection of the Alzheimer's disease using kernel PCA and LDA. Neurosci. Lett. 2009;464:233–8.

29. Dimitriadou E, Hornik K, Leisch F, Meyer D, Weingessel A, Leisch MF. Package "e1071" R Softw. Packag. Avaliable http//cran. rproject. org/web/packages/e1071/index. html. 2009, 1–62.

30. R Core Team. R: A Language and Environment for Statistical Computing Viena, Austria: R Foundation for Statistical Computing. 2017. Available from: https://www.r-project.org/

31. Torgo L. Infra-Structure for Performance Estimation and Experimental Comparison of Predictive Models in R. Comput. Res. Repos. 2014;abs/1412.0:1–40. Available from: http://arxiv.org/abs/1412.0436

32. Sing T, Sander O, Beeren N, Lengauer T. ROCR: visualizing classifier performance in R. Bioinformatics. 2005;20(20):7881

33. Kursa MB, Rudnicki WR. Feature Selection with the Boruta Package. J. Stat. Softw. 2010;36:1–13.

34. Kursa MB, Rudnicki WR. Wrapper Algorithm for All Relevant Feature Selection. R Softw. Packag. Available from: https://m2.icm.edu.pl/boruta/. 2016;1-14

35. Kuhn M. Package "Caret." J. Stat. Softw. 2008;1–26.

36. Lusa L, Blagus R. SMOTE for high-dimensional class-imbalanced data. BMC Bioinformatics 2013;14:106. Available from: http://www.biomedcentral.com/1471-2105/14/106

37. Torgo L, Torgo M. Package "DMwR." Compr. R Arch. Netw. 2015;1–106.

38. Durinck, S, Moreau, Y, Kasprzyk, A, Davis, S, De Moor, B, Brazma, A, Huber W. BioMart and Bioconductor: A powerful link between biological databases and microarray data analysis. Bioinformatics. 2005;21(16):3439–40.

39. Allen Institute for Brain Science. Technical White Paper: Transcriptome Profiling by RNA Sequencing and Exon Microarray. BrainSpan Atlas Dev. Brain 2013;1–15.

40. Zeeberg BR, Riss J, Kane DW, Bussey KJ, Uchio E, Linehan WM, et al. Mistaken identifiers: gene name errors can be introduced inadvertently when using Excel in bioinformatics. BMC Bioinformatics 2004;5(1):80.

41. Ziemann M, Eren Y, El-Osta A, Zeeberg B, Riss J, Kane D, et al. Gene name errors are widespread in the scientific literature. Genome Biol. 2016;17(1):177.

42. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559.

43. The GTEx Consortium, Welter D, MacArthur J, Morales J, Burdett T, Hall P, et al. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science (80-.). 2015;348(6235):648–60.

44. Blagus R, Lusa L. Class prediction for high-dimensional class-imbalanced data. BMC Bioinformatics 2010;11:523.

45. Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. 2004;32:258D–261.

46. Ashburner, M, Ball, CA, Blake, JA, Botstein, D, Butler, H, Cherry, JM, Davis, AP, Dolinski, K, Dwight, SS, Eppig, JT, Harris MA. Gene Ontology: tool for the unification of biology. Nat. Genet. 2000;25(1):25–9.

47. De Smet R, Marchal K. Advantages and limitations of current network inference methods. Nat. Rev. Microbiol. 2010;8(10):717–29.

48. Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, et al. The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature. 2014;505(7485):635–40.

49. Knoll M, Lodish HF, Sun L. Long non-coding RNAs as regulators of the endocrine system. Nat. Rev. Endocrinol. 2015;11:151–60.

50. He Z, Bammann H, Han D, Xie G, Khaitovich P. Conserved expression of lincRNA during human and macaque prefrontal cortex development and maturation. Rna. 2014;20(7):1103–11.

51. Aprea J, Calegari F. Long non-coding RNAs in corticogenesis: deciphering the non-coding code of the brain. EMBO J. 2015;34(23):2865–84.

52. Piro, RM, Molineris, I, Ala, U, Provero, P, Di Cunto F. Candidate gene prioritization based on spatially mapped gene expression: an application to XLMR. Bioinformatics. 2010;26(18):i618–24.

53. Lombard Z, Park C, Makova KD, Ramsay M. A computational approach to candidate gene prioritization for X-linked mental retardation using annotation-based binary filtering and motif-based linear discriminatory analysis. Biol. Direct. 2011;6(1):30.

CHAPTER III – THREE-STATE PROTEIN STABILITY PREDICTION FROM SEQUENCE-BASED FEATURES

Jose Guevara-Coto[1], Charles E. Schwartz[2] and Liangjiang Wang[1]

[1]Department of Genetics and Biochemistry, Clemson University, Clemson, SC 29634, USA

[2]J.C. Self Research Institute of Human Genetics, Greenwood Genetic Center, Greenwood, SC 29646, USA

Published: Proceedings of the 2017 International Conference on Bioinformatics and Computational Biology, pp. 45-49

Abstract

Amino acid substitutions can have significant and deleterious effects on proteins. Prediction of the effects of substitutions on protein stability has been explored, but many studies make use of structure-based features, which are not available for all proteins. In this study, we have developed a sequence-based SVM model for three-state protein stability prediction. This model used features extracted from the primary sequence, and feature selection identified the most informative feature set for model construction. We evaluated this model with an independent test dataset, and obtained the accuracy of 70.52% with 61.20% sensitivity and 79.84% specificity. Our results suggest that sequence features contain sufficient information for accurate prediction of three-state protein stability changes caused by amino acid substitutions.

Keywords: Amino acid substitutions, three-state protein stability prediction, sequence features, support vector machines

3.1 Introduction

Disease-causing sequence changes have been identified for many human genes [1–3]. The type of changes varies in nature, affecting multiple mechanisms from RNA processing to post-translational modifications. However, it has been identified that the most common effect of these changes is on protein stability [1–3]. An analysis of human single nucleotide polymorphisms (SNPs) revealed that ~80% of disease-causing amino acid substitutions affect protein stability [3]. In humans, ~60% of single nucleotide missense mutations in coding regions account for monogenic diseases based on the Human Gene Mutation Database [4]. Thus, understanding the effects of sequence variations on proteins is an important task.

Variations in the amino acid sequence can impact the physicochemical characteristics of a polypeptide. These sequence changes may alter the biochemical properties of the protein, resulting in modified structural characteristics with deleterious effects. The notion of predicting the effect of amino acid substitutions on protein stability has been previously explored [5,6]. However, our approach, similar to [7] and more recently [8], proposes to focus on sequence-based features and avoid using any features based on the protein structure.

The development of a three-state protein stability predictor was previously reported [9]. However, the approach made use of structural features which require a known protein structure. This can be a limiting factor as not all proteins have structural information available. Although computational methods are available for protein structure prediction [10], they are not as accurate in resolution as the experimental determination of a protein structure. The limitation of previous works in the use of structural features may be circumvented by encoding the instances with features derived completely from amino acid sequence.

In this study, we have developed a new three-state protein stability predictor based on sequence features. The importance of the sequence features in model performance was examined using two feature selection methods capable of reducing and ranking variables, random forests and recursive feature elimination. Our three-state protein stability predictor based on Support Vector Machines (SVMs) showed performance comparable to the currently available methods using structural information. The results suggest that sequence features can provide useful information for predicting the effect of amino acid substitutions on protein stability.

3.2 Methods

Data acquisition

The protein dataset was derived from that used by Capriotti et al [9] and later modified by Folkman et al [11]. We modified this dataset to contain 68 unique protein entries whose sequences were obtained in FASTA format from the Protein Data Bank (PDB) [12]. The sequences were subsequently subjected to redundancy reduction with a threshold of 85% sequence similarity using BlastClust. This assured the uniqueness of each sequence and eliminated the presence of multiple chains for a single entry. This process also resulted in the condensation of small sequences into larger clusters. The resulting dataset contained 1,332 non-redundant instances (henceforth s1332) with information about the PDB identifier, the wild-type position, the mutated amino acid, the substituted position, and the free energy change ($\Delta\Delta G$), which was used to determine the class label of the instance. In this study, we had three classes or states: decreased stability (DS), increased stability (IS), and no significant change (NC). The $\Delta\Delta G$ thresholds used in this study were as follows: DS if the effect of an amino acid substitution on the protein stability change $\Delta\Delta G \leq$ -0.5 kcal/mol, IS if $\Delta\Delta G \geq 0.5$ kcal/mol, or NC if -0.5 kcal/mol $\leq \Delta\Delta G \leq 0.5$ kcal/mol. For building an independent test dataset, we selected the s238 sequence set, and subjected it to the same process as the s1332 dataset. This test dataset contained unique entries that were not included in the training dataset s1332, and consisted of variants representing the three different states.

Sequence-based features

The instances in the s1332 dataset were encoded with a total of 31 sequence features using the R package "Peptides" [13]. This study used various physicochemical and biochemical features as defined by ExPASy in ProtParam (http://web.expasy.org/protparam/protparam-doc.html). These features include molecular mass, amino acid composition, charge, and aliphatic index, defined as the relative volume occupied by aliphatic side chains (Alanine, Valine, Isoleucine, and Leucine). Other features include: the Kidera factors, which comprise 10 features derived from 188 physical properties; three different hydrophobicity indices, obtained with three different scales (Fasman, Bull, Chothia); instability index, based on dipeptide composition; and the Boman index, which indicates the potential of a peptide to bind to the membrane or other proteins.

Model construction

In this study, the multiclass protein stability model was constructed using Support Vector Machines (SVMs) with sequence-based features. SVMs were shown to produce well-performing models for protein stability prediction in previous studies [7,8]. In its essence, an SVM model defines a separation hyperplane that divides the space into two distinct halves. Based on the sign given by the f(x), a point will be assigned to a given side of the hyperplane. If $f(x)>0$, a point will be assigned to the positive side of the hyperplane. The soft margin can increase the performance of the classifier when compared to hard margins, and this is achieved by allowing misclassification of some points [15]. The use of soft margins in SVMs comes as a response to the fact that not all data is linearly separable, which is especially true with biological data [15]. The third component refers to kernels, which can be used to make the calculation process more efficient, especially in feature spaces of high dimensionality. The radial basis function (RBF) kernel was used to construct the SVM model in this study. RBF is one of the most widely used kernel functions in model development and often performs well on biological data [15].

The caret and e1071 packages [16,17] available in R were used to construct the SVM model. We examined multiple options for model construction, including multiclass SVM models, kernel functions, and

performance metrics. The Receiver Operating Characteristic (ROC) curve, specifically the area under the curve (ROC-AUC), was calculated using the package pROC [18]. The SVM model was also compared with two different Random Forest (RF) classifiers, a single RF and an ensemble of RFs. The RF learning algorithm was previously shown to perform well for protein stability prediction [19].

Feature selection

Feature or variable selection for a classification problem can have two main objectives: (1) to identify highly important variables that are related to the response variable, and (2) to reduce the feature space to improve the prediction of the class label [20]. An efficient feature selection method can not only achieve these objectives, but may also combine important components such as determining importance thresholds, variable ranking and stepwise introduction of variables into the feature set [20,21]. In this study, we implemented a Random Forest (RF) based feature ranking method [21]. RFs are built upon decision trees, with every node being a condition on a variable. We also tested the recursive feature elimination method [22], in which the lowest 20% features were eliminated in each round to determine the most important sequence-based features.

Model performance evaluation

Model performance was evaluated using the R package performanceEstimation [23]. We used the tenfold cross-validation method. Briefly, this method consists of partitioning the data into different folds where one-fold is used for testing and the remaining folds are used for training. We also used the independent test dataset s238 to evaluate model performance. The following metrics were used in this study to measure model performance:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad (2)$$

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (3)$$

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (4)$$

The accuracy provides the information regarding the true positives (TP) and true negatives (TN) in the total of the dataset, which also includes the identified false positives (FP) and false negatives (FN). Sensitivity or true positive rate refers to the proportion of positive instances that were properly identified in each class, whereas specificity or true negative rate is the number of negative instances identified as such. The Matthews Correlation Coefficient (MCC) was also used in this study to measure the performance of multi-class models. For a model with three classes (A, B and C), the values for TP, TN, FP and FN can be calculated as described previously [23]:

$$TP = TPA + TPB + TPC \quad (5)$$

$$TN = TNA + TNB + TNC \quad (6)$$

$$FP = FPA + FPB + FPC \quad (7)$$

$$FN = FNA + FNB + FNC \quad (8)$$

3.3 Results

Prediction of three-state protein stability changes

To develop an accurate model for three-state protein stability prediction, we tested two widely used machine learning algorithms, Support Vector Machine (SVM) and Random Forest (RF). The SVM and RF models were constructed using 31 sequence features. As shown in Table 3.1, the SVM method outperformed the RF learning algorithm for predicting protein stability changes. The SVM model that was

fine-tuned with the training parameters (C = 15, gamma = 0.40) achieved higher performance measures than the RF models in the tenfold cross-validation. In this study, we constructed two RF models: a single RF model and an ensemble comprising of three RFs (RF-E). It was previously shown that an ensemble could result in improved model performance [24]. However, the RF ensemble achieved similar performance measures as the single RF model for protein stability prediction (Table 3.1). We thus selected the SVM model for further analyses.

Selection of relevant sequence features for model construction

Feature selection was conducted to determine the impact of sequence features in the model's ability to discriminate the three protein stability states, and to potentially enhance the model performance. The first feature selection method ranked the importance of the variables (sequence features) using the Gini index [22,26]. The approach based on Boruta [25] sets the variable selection threshold based on the value of shadow attributes, which are shuffled copies of all attributes to create randomness, culminated in the reduction of the feature set from 31 to 12 features. The second method, recursive feature elimination, identified a set of 7 features. Interestingly, the two feature selection methods identified several common features, including the hydrophobicity indices of Fasman and Chothia, some of the Kidera factors, and the aliphatic index. The first method also identified additional features associated with the putative overall effect of amino acid substitutions on the peptide, such as the isoelectric point.

However, the SVM models constructed with the selected features did not show improved performance in the tenfold cross-validation (Table 3.2). The SVM model using all the 31 sequence features (SVM_Full) achieved higher accuracy, ROC-AUC and MCC than the two models after feature selection (SVM_12 and SVM_7). One possibility was that the model SVM_Full might be slightly overfitted. To examine this possibility, we further compared the model performance using an independent test dataset.

Model performance evaluation using an independent test dataset

When compared using an independent test dataset (s238), the SVM models constructed with the selected features appeared to give slightly better performance measures than the model with the full feature set (Table 3.3). In particular, the SVM model using the 12 selected features (SVM_12) achieved slightly better ROC-AUC and MCC than the other two models (SVM_Full and SVM_7). The results suggest that the 12-feature set might be optimal for predicting three-state protein stability changes. Using the full set of 31 sequence features could cause model overfitting, whereas the 7-feature set might not provide sufficient information for prediction.

3.4 Conclusions

In this study, we have developed a new model for three-state prediction of protein stability changes caused by amino acid substitutions. The SVM model was built with sequence-based features. Feature selection identified 12 features for accurate protein stability prediction. We further validated the predictive performance of the model using an independent test dataset. Our results suggest that sequence features can provide sufficient information for predicting the effect of amino acid substitutions on protein stability.

Tables

Table 3.1 Performance of the SVM and RF models based on tenfold cross-validation. The models were constructed using 31 sequence features.

Model	AUC	MCC	Accuracy	Sensitivity	Specificity
SVM	0.8972	0.7038	0.8659	0.8411	0.8907
RF	0.7654	0.5360	0.7938	0.6907	0.8453
RF-E	0.7666	0.5163	0.7826	0.6814	0.8348

Table 3.2 Performance of the SVM models after feature selection based on tenfold cross-validation. SVM_Full was constructed using all the 31 sequence features; SVM_12 was constructed with 12 selected features; and SVM_7 was constructed with 7 selected features.

Model	AUC	MCC	Accuracy	Sensitivity	Specificity
SVM_Full	0.8972	0.7038	0.8659	0.8411	0.8907
SVM_12	0.8827	0.6115	0.8204	0.7830	0.8578
SVM_7	0.8439	0.5529	0.7902	0.7445	0.8360

Table 3.3 Predictive performance of the SVM models on an independent test dataset. SVM_Full was constructed using all the 31 sequence features; SVM_12 was constructed with 12 selected features; and SVM_7 used 7 selected features.

Model	AUC	MCC	Accuracy	Sensitivity	Specificity
SVM_Full	0.7423	0.3682	0.6647	0.5731	0.7563
SVM_12	0.7477	0.4555	0.7052	0.6120	0.7984
SVM_7	0.7367	0.4452	0.7092	0.6235	0.7950

References

1. Alber T. Mutational effects on protein stability. Annual Review Biochemistry. 1989;58:765–798.

2. Ramensky V, Bork, P and Sunyaev, S. Human non-synonymous SNPs : server and survey. Nucleic Acids Research. 202;30(17) 3894–00.

3. Wang Z and Moult, J. SNPs , Protein Structure , and Disease. Human Mutation. 2001;270:263–70.

4. Stenson PD, et al. "The Human Gene Mutation Database: 2008 update. Genome Medicine. 2009;1(1) 1–6.

5. Huang L., Gromiha MM, and Ho S. Structural bioinformatics iPTREE-STAB : interpretable decision tree based method for predicting protein stability changes upon mutations. Bioinformatics. 2007;23(10) 1292–93.

6. Capriotti E, Fariselli P, Calabrese R, and Casadio R. Protein Structure and Function Predicting protein stability changes from sequences using support vector machines. Bioinformatics. 2005;21:54–8.

7. Cheng J, Randall A, and Baldi P. Prediction of Protein Stability Changes for Single-Site Mutations Using Support Vector Machines. Proteins: Structure, Function, and Bioinformatics. 2006;62(4) 1125–32.

8. Teng S, Srivastava AK, and Wang L. Sequence feature-based prediction of protein stability changes upon amino acid substitutions. BMC Genomics. 2010;11(1):S5.

9. Folkman L, Stantic B, and Sattar A. Sequence-only evolutionary and predicted structural features for the prediction of stability changes in protein mutants. BMC Bioinformatics. 2013;14(2):S6.

10. Capriotti E, Fariselli P, and Casadio R. I-Mutant2 . 0 : predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Research. 2005;33(2):W306–10.

11. Raman S et al. Prediction Report Structure prediction for CASP8 with all-atom refinement using Rosetta. Proteins: Structure, Function, and Bioinformatics. 2009;77(S9) 89–99..

12. Folkman L, Stantic B, and Sattar A. Feature-based multiple models improve classification of mutation-induced stability changes. BMC Genomics. 2014;15(4):S6.

13. Berman HM et al. The Protein Data Bank. Nucleic Acids Research.2000l28(1):235–42.

14. Osorio D, Rondon-Villarreal P, and Torres R. Peptides: A Package for Data Mining of Antimicrobial Peptides. The R Journal. 2015;7(1): 4–14.

15. Ben-Hur A,, Ong CS, Sonnenburg S, Schölkopf B, and Rätsch G. Support vector machines and kernels for computational biology. PLoS Computational Biology. 2008;4(1):p.e1000173.

16. Kuhn M et al. Caret package. Journal of Statistical Software. 2008;28(5) 1–26.

17. Dimitriadou E, Hornik K, Leisch F, Meyer D, Weingessel A, and Leisch MF. Package e1071. R Software Packag, avaliable at http://cran.rproject.org/ web/packages/e1071/index. html.2009;1–62

18. Robin X et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011;12:77.

19. Li Y, and Fang J. PROTS-RF : A Robust Model for Predicting Mutation- Induced Protein Stability Changes. PLoS One. 2012;7(10):p. e47247.

20. Zhang L and Nagaratnam P. Random Forests with ensemble of feature spaces. Pattern Recognition. 2014;47(10):3429–37.

21. Genuer R, Poggi J, and Tuleau-Malot C. Variable selection using Random Forests. Pattern Recognition Letters. 2010;31(14): 2225–36.

22. Kursa MB and Rudnicki WR. Feature Selection with the Boruta Package. Journal of Statistical Software. 2010;36(36):1–13.

23. Díaz-Uriarte R and De Andrés SA. Gene selection and classification of microarray data using random forest. BMC Bioinformatics. 2010;7(1):3.

24. Torgo L. An infra-structure for performance estimation and experimental comparison of predictive models in R. 2015. arXiv Prepr. arXiv1412.0436.

25. Sebastiani F. Machine Learning in Automated Text Categorization. ACM Computational Surveys. 2002;34(1): 1–47.

26. Kursa MB and Rudnicki WR. Wrapper Algorithm for All Relevant Feature Selection. R Software Package. 2016. available at https://m2.icm.edu.pl/ borutapage/

CHAPTER IV – PROTEIN SECTOR ANALYSIS FOR THE CLUSTERING OF DISEASE-ASSOCIATED MUTATIONS

Jose Guevara-Coto[1], Charles E. Schwartz[2] and Liangjiang Wang[1]

[1]Department of Genetics and Biochemistry, Clemson University, Clemson, SC 29634, USA

[2]J.C. Self Research Institute of Human Genetics, Greenwood Genetic Center, Greenwood, SC 29646, USA

Published: BMC Genomics. 2014;15(Suppl. 11):S4.

Support : This work was supported by CSREES/USDA under project number SC-1000675

Abstract

Background: The importance of mutations in disease phenotype has been studied, with information available in databases such as OMIM. However, it remains a research challenge for the possibility of clustering amino acid residues based on an underlying interaction, such as co-evolution, to understand how mutations in these related sites can lead to different disease phenotypes.

Results: This paper presents an integrative approach to identify groups of co-evolving residues, known as protein sectors. By studying a protein family using multiple sequence alignments and statistical coupling analysis, we attempted to determine if it is possible that these groups of residues

could be related to disease phenotypes. After the protein sectors were identified, disease-associated residues within these groups of amino acids were mapped to a structure representing the protein family. In this study, we used the proposed pipeline to analyze two test cases of spermine synthase and Rab GDP dissociation inhibitor.

Conclusions: The results suggest that there is a possible link between certain groups of co-evolving residues and different disease phenotypes. The pipeline described in this work could also be used to study other protein families associated with human diseases.

Keywords: Co-evolving residues, statistical coupling analysis, protein sectors, disease phenotype.

4.1 Background

The role of mutations in disease phenotypes is an important focus for human genetics [1-3]. The identification of mutations and their link to a disease has generated numerous data entries which should optimally be accessible for ongoing research efforts. This has resulted in the establishment of databases such as the Online Mendelian Inheritance in Man (OMIM) [4], where it is possible to access data regarding a gene or gene product and browse through the information of how different mutations are associated with reported phenotypes. Even though this represents a valuable resource for clinical and molecular studies, the challenge of determining if differences in the phenotypes (severity, expansion of symptoms) within a syndrome can be associated with changes in specific groups of residues remains to be fully resolved. However, recent approaches in the study of proteins and their evolution have opened a door to analyze proteins from a different view [5-9] and attempt to associate reported clinical phenotypes to groups of correlated residues.

The techniques such as statistical coupling analysis (SCA) and direct coupling analysis (DCA) amongst others [5-7, 10] have approached the study of proteins by focusing on the idea of

residue co-evolution within the protein super-family [6, 8, 11]. These residues, whose interaction is not hindered by their spatial distribution, are organized within groups that have a seemingly underlying evolutionary relationship amongst them. Such groups of residues have been termed protein sectors and have been identified in protein super-families that comprise various lineages [6, 10]. These protein sectors have been characterized as important to the protein by either contributing to its biological identity or its function [5, 9].

The statistical coupling analysis requires the use of a large number of sequences as well as structural models for a representative member of the protein family to be analyzed [6, 7]. Although some protein structures are still not available, the advancement in crystallography accessibility as well as sequencing technologies coupled with increasing computational power and protein modeling accuracy and efficiency, has made it possible to obtain large sets of data in order to explore these approaches in multiple protein families, including those of clinical significance in humans [12-14].

Previous studies have focused on identifying co-evolving residues, and therefore understanding how these correlated amino acid units, and the mutations within these regions, could affect the protein. The most common approaches consist of biochemical assays that provide experimental support to the link between correlated residue units and measureable characteristics such as stability or catalytic activity [6, 11]. However, it is important to note that most of these studies have focused primarily on the analysis of regions with an associated function, such as domains, thus concentrating only on a segment of the totality of residues that comprise the proteins.

Although the study of specific regions provides valuable information to further understand how changes in amino acids can affect the overall function of the protein, the exclusion of residues outside the defined region could however lead to a possible loss of valuable information regarding

residue distribution in functional sectors. This is of special importance in human genetic disorders, where diseases have multiple phenotypes ranging in severity, which can be associated with the location of a mutation within the protein.

We propose an integrative approach that consists of analyzing full-length sequence alignments from proteins and the subsequent identification of protein sectors using statistical coupling analysis [6]. The identified sectors and known mutation data from OMIM as well as other information resources were then used to determine a possible link between the location of an amino acid residue change and disease phenotype. Our results based on test case proteins known to be associated with X-linked intellectual disabilities revealed that it might be possible to associate disease variants to protein sectors.

4.2 Methods

As shown in Figure 4.1, the pipeline used in our work consisted of three different stages: (1) dataset acquisition, multiple sequence alignment and curation; (2) identification of protein sectors and mapping of disease-causing mutations; and (3) association between protein sectors and disease phenotypes.

Dataset acquisition

The first step was to obtain a reference protein sequence. Selection of the sequence and its organism is determined based on the type of dataset to be analyzed. The retrieval of homologous sequences belonging to the protein family to be analyzed was done using one iteration of PSI-BLAST with a maximum number of target sequences set to 1000. This allowed us to obtain multiple sequences representing diverse evolutionary lineages. The motivation behind this was to enrich sequence diversity, thus allowing for the possible identification of groups of conserved co-evolving residues. Another reason to attempt to identify various sequences was to diminish the possible bias towards an organism that has overrepresentation of entries in the database. The approach of protein

sector analysis may not be applied to disease genes with few homologous sequences in the database. However, this is not supposed to be a common situation with more and more genomes being sequenced.

BlastClust [15] was subsequently used to reduce sequence redundancy within the dataset. The parameters used were: similarity threshold of 85% and 100% coverage. This helped diminish the over-representation of highly similar proteins with different entry numbers. Clustering of sequences further reduces the number of partial sequences that could cause alignment discordances due to their limited size. Further manual curation of the sequences was done in order to generate a high-quality sequence dataset for the alignment. In our pipeline, we implemented conditions to maintain dataset quality which primarily consisted of elimination of low quality predicted proteins and discarding of partial sequences.

Protein structure retrieval

Protein structure files for the reference sequences used in the test cases were obtained from the Protein Data Bank [16]. The entry numbers were obtained by running a BLAST search with the reference sequence and setting the PDB as the search database. Highest scoring matches were then selected and downloaded from the PDB website. The PDB file was used to map the multiple sequence alignment to the residue positions defined in the structure file. This information was required for analyzing each site to identify residue co-evolution.

An alternative for proteins without a 3D structure is the use of homology modeling. We generated such homology structure files for our test sets by using the freely available software such as CPHmodels 3.2 [17] or the Swiss-Model server [18]. The resulting PDB files could be used in

the pipeline. This alternative requires the use of a separate numbering file for the residues, which can be included in the pipeline.

Multiple sequence alignment

Each of the working datasets was subsequently aligned using ClustalW [19] incorporated in MEGA5 [20]. The parameters were set as default for gap penalties and the substitution matrix selected was BLOSUM62 [21]. After alignment, manual adjustments were undertaken to further curate the alignment file. These modifications included the elimination of high gap-creating sequences and trimming of large sequences (based on comparison of starting and ending residues of the reference or "seed" sequence). After this was accomplished, gaps were eliminated and sequences were realigned using the same algorithm and parameters. The final alignment files varied in the number of sequences, but for each test case there were more than 250 amino acid sequences in each file.

Protein sector identification

The statistical coupling analysis was undertaken to identify the groups of coevolving residues. The SCA toolbox for MATLAB was obtained from the Rama Ranganathan laboratory at the Green Center for Systems Biology, UT Southwestern Medical Center. The MATLAB script was modified and adapted to analyze the test cases.

The identification of protein sectors was accomplished using the methodology previously described [6, 11]. This consists of the calculation of the degree of conservation for each amino acid position. Subsequently, the statistical coupling analysis generates a positional correlation and a sequence correlation matrix. This was followed by a spectral decomposition and analysis of the statistically significant eigenvalues, and an independent component analysis, which further defines

the identified protein sectors. Finally, the protein sectors are mapped onto the primary structure of the protein based on their position in the structure file. This information can also be used to generate a 3D structure with different protein sectors (For further information refer to Halabi et al [6]).

Mapping of mutations within protein sectors

The information associated with mutations in a protein family can be obtained from databases such as OMIM as well as publications related to the protein family, their clinical importance and reported phenotypes known to be associated with the mutations. The clustering or grouping of mutations can be based on their location, within one of the major regions: N-terminal, central/other, and C-terminal. This allows for the preliminary classification of the mutations as well as the latter characterization of the residues identified within the protein sectors.

The labeling and numbering of residues and their corresponding sectors was done using a PDB file obtained by BLAST search using the reference sequence for each family as a query. The full-length sequence was used to maximize the number of possible residues belonging to protein sectors. By increasing the length and coverage of the alignment, it is possible to identify co-evolving residues outside the main protein domain, which could be associated with a disease phenotype.

Visualization of sectors in the protein 3D structure

The program Jmol (http://jmol.sourceforge.net/) was used to visualize the 3D structures in PDB files. Protein domains were subsequently identified and colored using the information available from PDB. Each domain was assigned an arbitrary color, based on visualization purposes (not functional association). For the visualization of protein sectors, the residues identified were

assigned the color that corresponded to the group of residues to which they belonged based on the SCA analysis.

4.3 Results and discussion

Identification of protein sectors

The selection of spermine synthase (SMS) and the Rab GDP dissociation inhibitor 1 (GDI1) as test cases was based on the available information for both proteins. SMS has been widely studied, with multiple mutations and the associated phenotypes comprehensively described as well as a vast amount of information on the importance of this molecule in brain development [22-25]. These characteristics as well as the fully resolved crystal structure (PDB: 3C6K) made SMS a good candidate as a test case.

GDI1 has been associated with non-specific X-linked intellectual disability, with different mutations being reported [26, 27]. This represents an interesting study case to determine if co-evolving residues could possibly be associated with disorders that appear to have clinical and genetic heterogeneity [26]. This protein also has an available crystal structure (PDB: 1LV0), which made it a suitable test case to analyze with the proposed approach.

The test cases of SMS and GDI1 were analyzed using full-length sequences and their corresponding structures. The results reported in this paper were obtained using a personal computer (Intel Core i7 Q720 at 1.60 GHz, 4.00 GB RAM) installed with the MATLAB 2012a and SCA packages. The analysis covered 90% of the residues represented in the crystal structures of SMS [28] and GDI1 [29]. This allowed for further identification of sector-grouped residues, located across the covered area of the protein. A possible issue is the need of protein structural data for the analysis. Such data may not be always available. However, this can be addressed using methods such as homology or ab initio modeling. Since the analysis of residue co-evolution is based on the multiple sequence alignment, the accuracy of homology modeling does not represent an issue to

the approach. The structure model is only used to map the positions of the multiple sequence alignment to the primary structure of the protein. Thus, the modeled structure file can be used in the pipeline in the same manner as the files obtained from the Protein Data Bank.

The protein sector analysis of both SMS and GDI1 revealed the distribution of groups of co-evolving residues (Figure 2a and 2b) that present a degree of spatial separation which differs from that found in protein domains (Figure 3a and 3b). When analyzing the sectors of SMS, we were able to identify three groups classified as the red, blue and green sector. Our results indicated that the red and blue sectors appear to be in the N-terminal region and to some extent the long loop and the beta strand domain (Figure 2a). However, the presence of residues of the red sector appears to be limited to these three regions, whereas the blue and green sectors span the C-terminal region, with no red sector residues identified in this region.

For GDI1, our analysis identified three distinct sectors, red, blue and green. The FAD/NAD(P) binding domain, located in the N-terminal region of GDI1, had amino acid residues classified within the blue and green sector (Figure 2b), with only three positions (82, 125 and 283) identified as belonging to the red sector. When analyzing the FAD-linked reductase and the FAD/NAD(P) binding domain located in the central and C-terminal regions of the protein, the residues in both are mostly classified as being part of the red sector (Figure 2b), with reduced presence of residues classified into the green and blue sectors. Although sector distribution of residues appears to follow a pattern where one group of residues is predominant over the others in a specific region, our analysis revealed that even within these regions, residues belonging to different sectors were identified (Figure 2a and 2b). This finding supports a degree of differentiation between protein sectors and domains.

Our results indicate that although protein sectors may be distributed across the proteins, there is a concentration of these co-evolving residues in certain areas of the proteins (Figure 2a and 2b), with stretches of residues that were not identified to be part of any sector. This observation could be attributed to the presence of conserved catalytic centers or active sites, necessary for protein function. One example was residue 276 in SMS (Table 1), which is a known active site that was not assigned to any sector. For GDI1, residue 92 (Table 1) has been reported to be necessary for the binding and recycling of RAB3 [27], and mutations in this site could lead to a reduction in its activity. This residue was also not classified in any of the three identified sectors. Because such important residues tend to be highly conserved, it is possible that these sites are not prone to undergo the co-evolutionary process

.

Disease-associated residues in protein sectors

The protein sector analysis of our test cases revealed a possible link between different disease phenotypes and the locations of the mutations within groups of co-evolving residues. We were able to identify multiple residues, in which disease-causing mutations have been identified (Table 1). As stated in Halabi et al. [6], groups of co-evolving residues seem to be associated with specific functions. In SMS, we identified the presence of a sector within the N-terminal region, which is necessary for protein dimerization. It has been reported that the ability to form dimers is necessary for full protein functionality. In our analysis, we found that the most severe phenotypes for the Snyder-Robinson Syndrome were predominantly found in the red sector, with residues mapped not only to the N-terminal region but also in the long loop and beta-strand domains (Figure 2a and 3a).

The results suggest that protein sectors, when compared to domains, allow for the clustering of mutations based on the underlying process of residue co-evolution instead of grouping

them by sequential or spatial locations. The protein sector analysis of SMS has shown that the known Snyder-Robinson syndrome phenotypes, including expanded phenotypes [30], are mainly related to the red sector (Table 1), which comprises the N-terminal region as well as the long loop and beta strand domain. Thus, a protein sector can be associated with diverse functions as previously reported [6]. It appears that the red sector in SMS comprises regions associated with both dimerization and protein stability (Table 1), and these spatially distant regions may have undergone the co-evolutionary process because of their roles in the function of the protein. Interestingly, another mutation in SMS has been mapped to the green sector. This residue at position 328 (Table 1) has been associated with a milder form of the Snyder-Robinson syndrome [31]. It is possible that the green sector in the C-terminal region comprises residues associated with the enzyme's substrate recognition [28]. This may lead to a decrease in enzymatic activity, thus possibly accounting for the milder form of the syndrome [28].

The analysis of GDI1 identified two known mutations located in the green and blue sectors (Table 1). These amino acid positions (70 and 423) are associated with non-specific X-linked intellectual disability [26, 27]. The two residues are located in spatially separated regions, with the residue at position 70 in the N-terminal FAD/NAD(P)-binding domain and the residue at position 423 in the C-terminal region (Figure 2b). It is still unknown whether these two mutations, distributed in different sectors, may cause any difference in disease phenotypes. Further studies are needed to explain our findings and expand the knowledge about this protein, including the possible role of the co-evolving resides in disease. The third residue at position 92 has been proposed as an important substrate binding site [26, 27], and appears not to be undergoing co-evolution, probably owing to its significance in stabilizing the Rab-binding region or recognition of the C-terminal prenyl group in substrate recycling [27]. This may explain why our analysis did not identify it in any of the three sectors.

Our results from the protein sector analysis of SMS suggest a possible link between groups of co-evolving residues and disease severity. However, to fully understand the role of protein sectors, the use of mutagenic analysis can provide further information. Experimental validation of our results would present further evidence to support the approach of clustering disease-associated residues with their corresponding clinical manifestation. In addition, although we have presented a pipeline useful for clustering mutations, it is possible that other factors, such as the type of amino acid substitutions, could be the underlying cause of the differences in disease severity. Nevertheless, by clustering residues into different sectors, our approach can be used as a valuable aid in further characterizing the possible causes of disease such as intellectual disability. Protein sector analysis may also provide useful information for understanding the effect of mutations at spatially distant positions on disease phenotypes.

4.4 Conclusions

We have proposed an integrative approach that makes use of the statistical coupling analysis method for the study of disease-causing mutations in proteins. Our test cases provided information regarding the role of protein sectors and how they could be associated to disease variants. We were also able to identify the effect of mutations in distinct sectors with variations in the clinical signs of a disease. These findings support the possible role of the protein sectors in specific functions that when affected could lead to variable phenotypes associated with a complex syndrome.

Figures

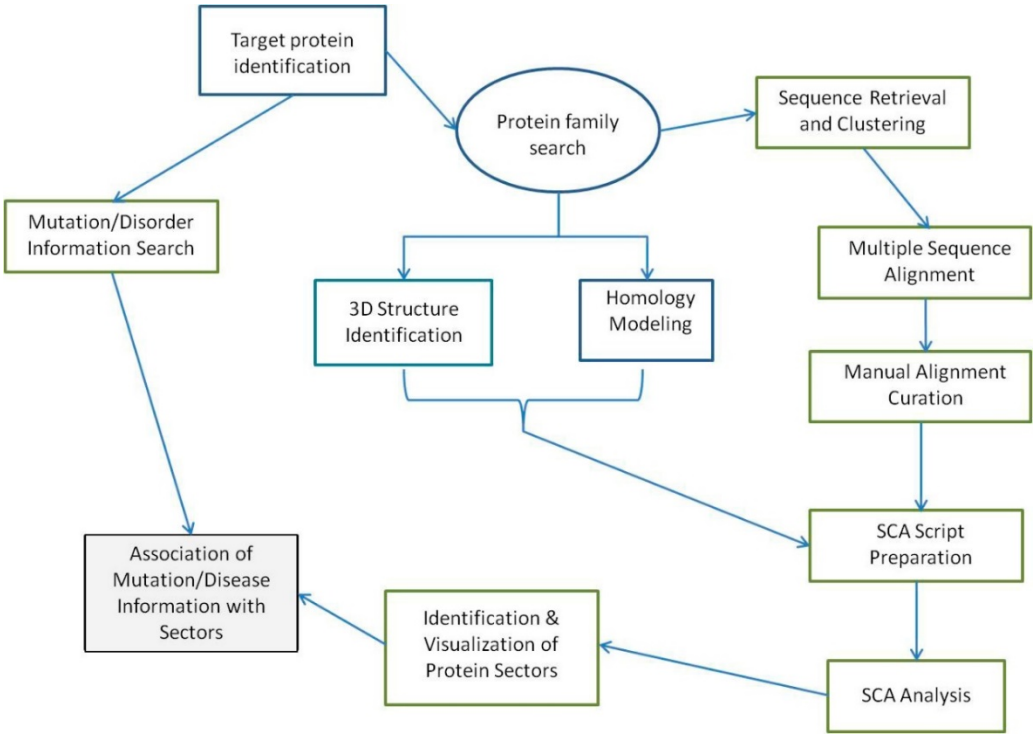

Figure 4.1 Proposed pipeline for analyzing co-evolving residues from protein families associated with human diseases. The flow chart includes the alternative of using homology modeling or other protein modeling methods if the 3D structure is not available. The association of disease information with protein sectors is in a grey shaded box because it is the end result of the analysis.

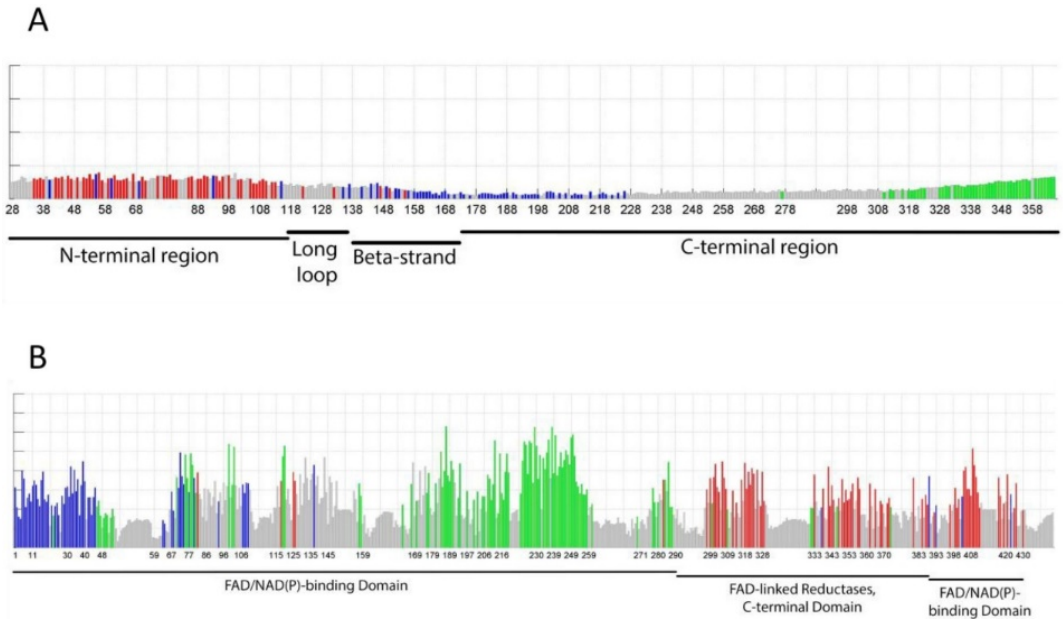

Figure 4.2 Identification of protein sectors in the primary structure of SMS and GDI1. A) Distribution of the protein sectors within the different regions of human spermine synthase. The analyzed residues correspond from position 28 to 365 based on the structure (PDB: 3C6K). The pattern observed appears to correspond to a stratification of sectors where the N-terminal region associated with dimerization appears to dominate the red sector, whereas the C-terminal region, ranging from residues 173 to 366, is composed of the blue and green sectors. B) Distribution of the protein sectors in Rab GDP dissociation inhibitor. The analyzed residues correspond from positions 1 to 447 based on the PDB structure file (1LV0). The N-terminal FAD/NAD(P) binding domain appears to be predominantly composed of residues belonging to the green and blue sectors. The FAD-linked reductase and the C-terminal FAD/NAD(P) binding domains appear to have a majority of residues classified within the red sector with a small presence of amino acids belonging to the blue and green sectors.

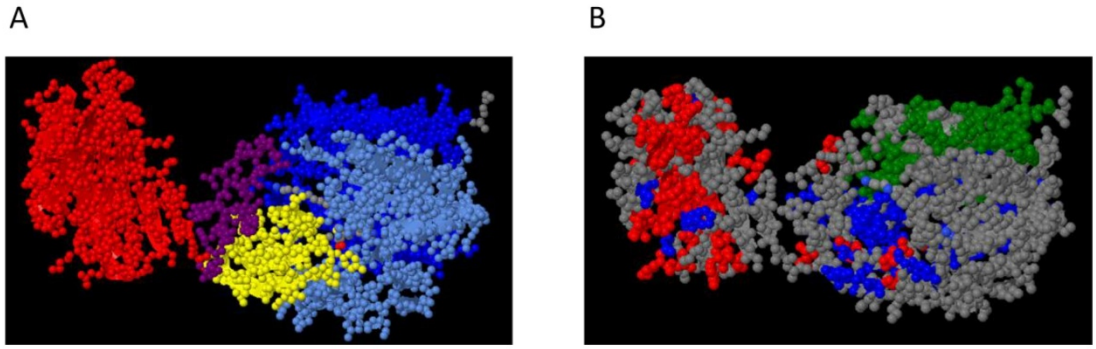

Figure 4.3 Representation of protein domains and identified sectors in the 3D structure of spermine synthase (PDB: 3C6K). A) Protein domains were colored in one of the chains of the homodimer. Red corresponds to the N-terminal domain, the long loop is in purple, the beta strand domain is in yellow, and the C-terminal region is in blue, with the catalytic center shown in light blue. B) Identified protein sectors in spermine synthase. Each sector is represented with a different color, and appears to have a distribution that is not limited to the described domains.

Table

Table 4.1 List of the residues associated with diseases in spermine synthase and Rab GDP dissociation inhibitor. The disease-associated sites were obtained from OMIM and publications, and the list is organized according to the residue location in the protein domains.

	Residue	Sector	Domain/Region	Effect/Associated Phenotype
Spermine Synthase	G56A	Red	N-terminal	Snyder-Robinson syndrome. Expanded phenotype of disease reported in [23]
	G67A	Red	N-terminal	New Snyder-Robinson syndrome phenotype. Expanded phenotype of disease reported in [30]
	V132A	Red	N-terminal/long loop	Snyder-Robinson syndrome. Expanded phenotype of disease reported in [24]
	I150A	Red	Central (beta-strand)	Decrease the stability of the C-terminal region, Snyder-Robinson syndrome
	Y328A	Green	C-terminal/catalytic	Mild mental retardation, Snyder-Robinson syndrome
GDI1	L92P	-	FAD/NAD(P) binding domain	Non-specific mental retardation

R70TER	Green	FAD/NAD(P) binding domain	Non-specific mental retardation
R423P	Blue	FAD/NAD(P) binding domain	Non-specific mental retardation

References

1. Lahiry P, Torkamani A, Schork NJ, Hegele RA. Kinase mutations in human disease: Interpreting genotype–phenotype relationships. Nat Rev Genet. 2010;11:60–74.

2. McClellan J, King M. Genetic heterogeneity in human disease. Cell. 2010;141:210–217.

3. Alexov E, Sternberg M. Understanding molecular effects of naturally occurring genetic differences. J Mol Biol. 2013;425:3911–3913.

4. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 2005;33:D514–D517.

5. Fuchs A, Martin–Galiano AJ, Kalman M, Fleishman S, Ben–Tal N, Frishman D. Co-evolving residues in membrane proteins. Bioinformatics. 2007, 23:3312–3319.

6. Halabi N, Rivoire O, Leibler S, Ranganathan R. Protein sectors: Evolutionary units of three–dimensional structure. Cell 2009;138:774–786.

7. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc Natl Acad Sci USA. 2011;108:E1293–E1301.

8. McLaughlin Jr RN, Poelwijk FJ, Raman A, Gosal WS, Ranganathan R. The spatial architecture of protein function and adaptation. Nature. 2012;491:138–142.

9. Bartlett GJ, Taylor WR. Using scores derived from statistical coupling analysis to distinguish correct and incorrect folds in de-novo protein structure prediction. Proteins. 2008;71:950–959.

10. de Juan D, Pazos F, Valencia A. Emerging methods in protein co–evolution. Nat Rev Genet. 2013;14:249–261.

11. Smock RG, Rivoire O, Russ WP, Swain JF, Leibler S, Ranganathan R, Gierasch LM. An interdomain sector mediating allostery in Hsp70 molecular chaperones. Mol Syst Biol. 2010;6:414.

12. Joachimiak A. High–throughput crystallography for structural genomics. Curr Opin Struct Biol. 2009;19:573–584.

13. Metzker ML. Sequencing technologies – the next generation. Nat Rev Genet. 2010; 11:31–46.

14. Yates JR, Ruse CI, Nakorchevsky A. Proteomics by mass spectrometry: Approaches, advances, and applications. Annu Rev Biomed Eng. 2009;11:49–79.

15. Altschul S, Gish W, Miller W, Meyers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990, 215;403–410.

16. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res. 2000;28:235–242.

17. Nielsen M, Lundegaard C, Lund O, Petersen TN. CPHmodels–3.0—remote homology modeling using structure–guided sequence profiles. Nucleic Acids Res. 2010;38:W576–W581.

18. Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, Kiefer F, Cassarino TG, Bertoni M, Bordoli L, Schwede T. SWISS–MODEL: Modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res. 2014;[Epub ahead of print].

19. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position–specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994;22:4673–4680.

20. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: MEGA5. Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol. 2011;28:2731–2739.

21. Eddy SR. Where did the BLOSUM62 alignment score matrix come from? Nat Biotech. 2004;22:1035–1036.

22. Cason AL, Ikeguchi Y, Skinner C, Wood TC, Holden KR, Lubs HA, Martinez F, Simensen RJ, Stevenson RE, Pegg AE, Schwartz CE. X–linked spermine synthase gene (SMS) defect: The first polyamine deficiency syndrome. Eur J Hum Genet. 2003;11:937–944.

23. de Alencastro G, McCloskey DE, Kliemann SE, Maranduba CMC, Pegg AE, Wang X, Bertola DR, Schwartz CE, Passos–Bueno MR, Sertié AL. New SMS mutation leads to a striking reduction in spermine synthase protein function and a severe form of Snyder-Robinson X–linked recessive mental retardation syndrome. J Med Genet. 2008;45:539–543.

24. Becerra–Solano LE, Butler J, Castañeda–Cisneros G, McCloskey DE, Wang X, Pegg AE, Schwartz CE, Sánchez-Corona J, García-Ortiz JE. A missense mutation, p.V132G, in the X–linked spermine synthase gene (SMS) causes Snyder–Robinson syndrome. Am J Med Genet A. 2009;149A:328–335.

25. Kesler S, Schwartz C, Stevenson R, Reiss A. The impact of spermine synthase (SMS) mutations on brain morphology. Neurogenetics. 2009;10:299–305.

26. Bienvenu T, des Portes V, Saint Martin A, McDonell N, Billuart P, Carrié A, Vinet M, Couvert P, Toniolo D, Ropers H, Moraine C, van Bokhoven H, Fryns J, Kahn A, Beldjord C, Chelly J. Non–specific X–linked semidominant mental retardation by mutations in a rab GDP–dissociation inhibitor. Hum Mol Genet. 1998;7:1311–1315.

27. D'Adamo P, Menegon A, Lo Nigro C, Grasso M, Gulisano M, Tamanini F, Bienvenu T, Gedeon A, Oostra B, Wu S, Tandon A, Valtorta F, Balch W, Chelly J, Toniolo D. Mutations in GDI1 are responsible for X–linked non–specific mental retardation. Nat Genet. 1998;19:134–139.

28. Wu H, Min J, Zeng H, McCloskey DE, Ikeguchi Y, Loppnau P, Michael AJ, Pegg AE, Plotnikov AN. Crystal structure of human spermine synthase: Implications of substrate binding and catalytic mechanism. J Biol Chem. 2008;283:16135–16146.

29. An Y, Shao Y, Alory C, Matteson J, Sakisaka T, Chen W, Gibbs RA, Wilson IA, Balch WE. Geranylgeranyl switching regulates GDI–Rab GTPase recycling. Structure 2003; 11:347–357.

30. Peron A, Spaccini L, Norris J, Bova SM, Selicorni A, Weber G, Wood T, Schwartz CE, Mastrangelo M. Snyder–Robinson syndrome: A novel nonsense mutation in spermine synthase and expansion of the phenotype. Am J Med Genet A 2013;161:2316–2320.

31. Zhang Z, Norris J, Kalscheuer V, Wood T, Wang L, Schwartz C, Alexov E, van Esch H. A Y328C missense mutation in spermine synthase causes a mild form of Snyder–Robinson syndrome. Hum Mol Genet. 2013;22:3789–3797.

CHAPTER V – CONCLUSIONS

In this dissertation, we have presented several computational approaches, to analyze both genes and proteins associated with intellectual disability (ID). We have utilized various datasets, ranging from multiple sequence alignments to the developmental transcriptome of the brain to generate new knowledge that can be used in human genetics research.

We developed an expression-based machine learning classifier for prediction of ID-associated long non-coding RNAs (lncRNAs). By using the developing brain transcriptome to encode a list of ID-causing and non-ID disease genes, we developed a model using Support Vector Machines. The binary classifier performed well at discriminating between the two classes of genes and identified possible ID-associated lncRNAs. The use of feature selection did not improved model performance or provide new insight about ID development. Instead, use of a class-balancing strategy improved classifier performance. The identified lncRNAs were found to be expressed in modules co-enriched with known ID genes and associated with neurodevelopmental processes. These lncRNAs were also found to be expressed in brain tissues. These findings support the roles of the candidate lncRNAs in ID pathogenesis.

We constructed a three-state protein stability classifier using sequence-based features. By encoding each sequence in the training set with relevant physico-chemical features, a model was built using SVM. Through feature selection, we identified informative variables, important to discriminate between the three states (increase, decrease and no significant change in stability). Although the model was well-performing based on an independent test set, approaches such as semi-supervised learning might further improve the protein stability prediction, given the large number of unlabeled instances currently available.

We also used statistical coupling analysis (SCA) in the study of ID-causing mutations and their effect on disease phenotypes. By using the sequences from known ID proteins, we built a multiple sequence alignment that was used to identify groups of co-evolving residues denominated protein sectors. Protein sectors are functionally important, and when mutated, protein function can be affected. Using SCA we identified different mutations in distinct sectors, each associated with a different phenotypic manifestation for a syndromic ID. The sectors clustered mutations into mild and severe ID. Our results suggest that SCA can be used to identify functionally important sites and determine the impact of mutations in the clinical manifestations of disease.

The bioinformatics methods presented in this dissertation can be further improved in the future. The expression-based classifier used to identify candidate ID-associated lncRNAs can be improved with the incorporation of target prioritization approach, which would rank the candidate lncRNAs based on their possible importance to ID pathogenesis. For the protein stability classifier, the potential use of semi-supervised learning could provide model development with an improvement in classifier performance by using the numerous unlabeled data instances that are currently available. Our work on mutations and their effect on disease phenotypes using SCA can be expanded to study other diseases besides ID. The results from the ID test case suggest SCA can be used in the study of disease-causing proteins, as well as for studying functionally important sites in other protein families.

APPENDIX

APPENDIX A – MUTUAL INFORMATION TO IDENTIFY CORRELATED SITES IN URACIL DNA GLYCOSYLASE SUPERFAMILY

A.1 Introduction

When analyzing a multiple sequence alignment of a protein family to identify co-evolving residues, it is always reasonable to find three kinds of residues: highly or completely conserved, partially conserved, and non-conserved. Conserved residues are important because they tend to be associated with structural interactions. These interactions are usually necessary for structural integrity, as demonstrated by mutagenesis studies, where substitutions in these amino acid positions affect the interactions associated with protein folding or stability. However, partially conserved residues are also of special interest because of the potential information they can provide about the protein [1,2]. These positions are the result of divergence amongst species. This process leads to emergence of new sequences or amino acid sites with functional differentiation. Partially conserved residues appear conserved within a protein sub-family, which gives the residues new functions , such as substrate affinity or a specific catalytic action, characteristics that separate each sub-family from each other [1]. The identification of partially conserved residues is a challenge as their identification is necessary in evolutionary and functional studies. Currently, the task of identifying partially conserved residues requires the use of computational and experimental approaches in order to experimentally corroborate the sites predicted using computational methods [1,2].

Although approaches are available for the identification of partially conserved residues, it is important to select a method capable of properly identifying the coevolutionary signals that these residues share. The presence of functional association between amino acid pairs, the potential structural interactions or the phylogenetic signals in the multiple sequence alignment (MSA) can

affect the mutual information score [3,4]. This is of special importance because of the proper determination of partially conserved residues, these must share a coevolutionary signal based on a functional interaction, and not one defined by their phylogenetic or structural interactions. To address this, matter information theory, and more specifically mutual information (MI), has been used to analyze multiple sequence alignments (MSA) of different uracil DNA glycosylase (UDG) families.

The use of this approach allowed us to analyze the multiple sequence alignment (MSA) and identify two partially conserved co-evolving residues that appeared to have a significant level of conservation within each sub-family based on the MSA and the mutual information analysis. These residues were also analyzed for potential association with a function, an important trait of partially conserved residues. These amino acid positions, identified as pairs 41E-42G in family 4 and 63Q-64D in family 1, were confirmed to be functionally important through mutational and biochemical analysis. These results suggest that MI can provide valuable information in the functional studies of protein families and evolution [5].

A.2 Methods

Mutual information analysis

Residue co-evolution was determined using a mutual information-based tool, MISTIC (Mutual Information Server to Infer Co-evolution). This approach uses mutual information to determine the evolutionary relationship between two residue positions in a multiple sequence alignment. Prior to the MI analysis, dataset acquisition and construction, and multiple sequence alignment were performed as described in the following steps. The calculation of the MI co-evolution values on the MISTIC server was carried out as described [3]. This consisted of calculating the frequency of amino acid pairs by means of weighting and low count correction. The calculated frequency is then

compared with the expected frequency. MISTIC also assumes that mutations between amino acids are uncorrelated [3]. Afterwards, the MI scores for the protein family alignment were calculated. These scores were obtained by calculating a weighted sum of the log ratios of the expected and the observed frequencies from the amino acid pairs [3]. Mutual information background signal noise was corrected by implementing the Average Correction Product [3,6]. Subsequently, a z-score normalization was applied to the MI values. A threshold of 6.5 was used to report co-evolving residues identified by MI. This value has been reported to have significant values in specificity and sensitivity [3]. Identification and visualization of co-evolving residues were performed as described in Supplementary Information.

Dataset acquisition and construction for mutual information analysis

The first step was to obtain the available uracil DNA glycosylase (UDG) sequences. These were acquired from UniProtKB. A general search was done in order to find significant hits that could be used to generate a raw dataset, composed of representatives of all the UDG families. Subsequently, the raw dataset was sorted using a Perl script designed to separate sequences based on the presence of the distinct UDG family 4 and 1 motifs (GE[A/G][V/P]G and GQDPY, respectively), as reported in [7]. The output of the script was two distinct files, each containing approximately 1000 family 4 UDGa and family 1 UNG protein sequences.

The resulting sequences were then subjected to clustering analysis using BlastClust to reduce redundancy [8,9]. This is of special importance, as it reduces sequence redundancy as well as the bias effect that overrepresented sequences might have later on the residue co-evolution analysis [10,11]. The parameters used were: sequence similarity threshold of 75% and coverage percentage value of 85%. The selection of these parameters reduced the amount of highly similar sequences with different accession entries from each dataset. This resulted in a significant reduction

of sequence entries, with each data file containing approximately 200 sequence representatives for each family.

Multiple sequence alignment

The sequences in the dataset were then aligned using the ClustalW alignment tool incorporated in MEGA6 [12–14]. Preliminary alignments were performed using the default parameters followed by manual curation of the alignment. The process eliminated noise from outlier with no evolutionary, biological and functional relationship to family 4 UDGa and family 1 UNG. By manually examining the alignment, it was possible to eliminate these sequences to obtain and procure a high quality alignment dataset.

After the first alignment was completed, all gaps were removed from the datasets and a new alignment was performed, using MEGA6's implementation of ClustalW. The parameters were the following: the substitution matrix was BLOSUM62 [15,16], with gap opening penalty of 20 and gap extension penalty of 5. These stringent parameters were selected in order to reduce the number of gaps within the alignment. The resulting alignment was then subjected to a second round of manual curation, where any remaining outliers were eliminated. Subsequently, gaps were removed and sequences were re-aligned. The ClustalW alignment output was then subjected to a final manual inspection. Afterwards, a refining alignment was done using MUSCLE [17].

Identification and visualization of co-evolving residues

The curated sequence alignment files were uploaded into the MISTIC web server, and a reference sequence for each family was selected. The selection was based on two criteria: being the representative of the largest luster and second, having a similarity of more than 25% with the

family's canonic structure. This allows for the reference to be a significant representative of each of the UDG families studied.

After alignment file uploaded and reference sequence selected, the next step was modification of the default web server parameters. The protein structure file was left blank, as our analysis was intended to identify co-evolving relationships using only protein sequences. Within the advanced options the maximum fraction of gaps per column allowed in the calculations was set from 0.5 (default value) to 0.3. This number was selected in previous co-evolution analysis, and has yielded good results [11]. Finally, the file was submitted to the web server for analysis.

After the analysis was completed, results were visualized using the tools incorporated within the MISTIC web server. A sequential circular representation of the multiple sequence alignment, known as the Circos diagram, maps the amino acid positions to the reference sequence. In this diagram, there are three major components represented with a color scale. The first is the square boxes under each amino acid position, these boxes represent the conservation of the residue. The colors of these boxes range from red (highly conserved) to blue (less conserved). The second is the histograms associated with each position representing the cumulative mutual information, which illustrates the correlation a given residue has with other positions. The higher the histogram, the more residues that position is correlated with. Finally, the edges or lines connecting the co-evolving residues describe the relationship between positions in the multiple sequence alignment based on their mutual information. Red lines represent the top 5%; black lines refer to the MI relationships between 95% and 70%; and the gray lines denote the remaining MI relationships [3].

The MI analysis was also represented as a network, which was generated by Cytoscape [18]. Within the network representation, each node was an amino acid positions within the multiple sequence alignment, mapped to the reference sequence for each family. Residue conservations,

similarly to the Circos diagram, were represented by a color scale, where red represented highly conserved residues and blue less conserved ones. The lines or edges of the network represented significant MI values (those higher than the 6.5 threshold) [3].

A.4 Results and Discussion

Mutual Information (MI) was used to analyze a multiple sequence alignment from sub-families 1 and 4 UDGs to identify a doublet of partially coevolving residues. The positions identified in this doublet were considered specificity-determining sites (SDPs), identified as residues of importance to protein function [19]. The reason for defining the amino acids in the doublet as SDPs was their functional activity as part of motif 1 and their high conservation within the family (Figure A-1). These residues were confirmed as functionally important through mutational and biochemical analysis. MI also revealed that these residues are partially conserved sites. These positions presented high conservation in the multiple sequence alignment for residues 41E-42G in family 4 and 63Q-64D in family 1.

The use of MI is based on its ability to identify underlying relationships such as co-evolution between pairs of residues [19]. The two residues can be either in close distance (neighboring) or separated by large blocks of amino acids. In the case of the identified pairs 41E-42G in family 4 and 63Q-64D in family 1 (Figure A-1), these doublets are subjected to what MISTIC identifies as structural and functional linkage as both are coevolving residues that are neighboring positions [3]. The presence of structural and functional linkage results in the doublets 41E-42G in family 4 and 63Q-64D having a significant cumulative MI value (Figure A-1), represented as the outward bar in the Circos graphs. A closer inspection of the figure revealed that these neighboring sites have multiple correlations with other sites. Some of these interactions are ranked in the top 5% (red-colored lines) based on the MI score. Most of the correlations are within

the range of 95-70%. It is possible to trace some of these correlations to positions in motif 2. It appears that, based on Figure A-1, residues on motif 2 could be partially conserved as their amino acid identities, which change from family to family but appear highly conserved within each family. The correlation between the doublets 41E-42G in family 4 and 63Q-64D in family 1 and positions in motif 2 could be the result of potential structural and functional linkage. These could account for the observed formation of a pocket between motifs 1 and 2 that interacts with its substrate. The ability of MI to uncover this kind of associations is considered a robust approach to study co-evolution [4]. An advantage of MI over other approaches that identify coevolutionary interactions is the correction for phylogenetic bias in multiple sequence alignments, especially implemented in MISTIC [3,4]. The doubts 41E-42G in family 4 and 63Q-64D in family 1 were also correlated to non-conserved sites. Previous studies have proposed that the non-conserved sites could be positions where balancing mutations have taken place. It has been speculated that neutral sites could have developed correlations with functionally important sites to mitigate the effects of deleterious mutations [1,2].

A.5 Conclusions

We have demonstrated that mutual information (MI) can be used to uncover coevolution between residues in a multiple sequence alignment. In the study of the UDG superfamily, we focused on two families, 1 and 4, and initially identified two residue positions that had partial conservation. These were conserved within each sub-family but variable in the super-family. In the study, it was confirmed with biochemical assays that these sites were functionally important for the sub-families. These assays validated our results and supported the significance of MI in the identification and functional study of proteins.

In this research, we identified correlation in terms of possible coevolution between residues of different motifs (1 and 2) that have been identified to be of functional importance in the UDG family. The use of MI also revealed a possible role of non-conserved residues as possible sites where balancing mutations can occur. These non-specific sites would serve to minimize the effect of deleterious mutations that would disrupt protein function or structural stability. This could account for pairwise correlations between functionally important sites and positions with a seemingly randomly distributed across the protein with neither conservation or function. In a recent work, MI was also used to study UDG family 3 (SMUG1) and supported the functional characterization of the SMUG-1 glycosylase representative isolated from *Listeria innocua* [20].

Note

The MI analysis of UDG superfamily has been incorporated in two peer-reviewed publications:

Xia, B., Liu, Y., Guevara, J., Li, J., Jilich, C., Yang, Y., Wang, L., Dominy, B.N. and Cao, W., 2017. Correlated Mutation in the Evolution of Catalysis in Uracil DNA Glycosylase Superfamily. *Scientific Reports*, *7*.

Li J, Yang Y, Guevara J, Wang L, Cao W. Identification of a prototypical single-stranded uracil DNA glycosylase from Listeria innocua. DNA Repair (Amst). Elsevier; 2017;57:107–15.

Figures

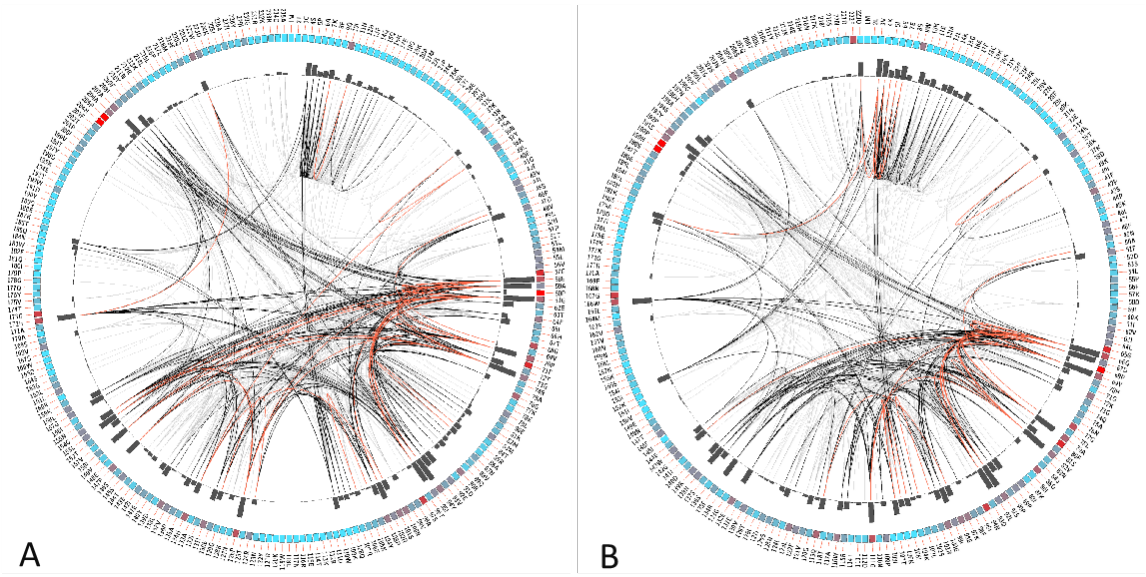

Figure A-1 Circos graph of the Mutual Information analysis of families 1 and 4. The graph represents the positions in the reference sequence for (A) family 1 and (B) family 4 of UDG. The lines represent the number of significant correlations, with the red lines representing the top 5%. The outward bars for each position are the cumulative mutual information, based on the MI score of each position and its neighbor residue. The external layer ranging from red to blue is the residue conservation, from highly conserved (deep red) to low conservation (blue) Adapted from Xia et al. 2017 [5].

References

1. Sandler I, Zigdon N, Levy E. The functional importance of co - evolving residues in proteins. 2014;673–82.

2. Sandler I, Abu-qarn M, Aharoni A. Protein co-evolution: how do we combine bioinformatics and experimental approaches? Mol. Biosyst. 2013;175–81.

3. Simonetti FL, Teppa E, Chernomoretz A, Nielsen M, Marino Buslje C. MISTIC: Mutual information server to infer coevolution. Nucleic Acids Res. 2013;41:8–14.

4. Gloor GB, Martin LC, Wahl LM, Dunn SD. Mutual Information in Protein Multiple Sequence Alignments Reveals Two Classes of Coevolving Positions †. Biochemistry. 2005;44:7156–65.

5. Xia B, Liu Y, Guevara J, Li J, Jilich C, Yang Y, et al. Correlated Mutation in the Evolution of Catalysis in Uracil DNA Glycosylase Superfamily. Sci. Rep. [Internet]. Nature Publishing Group; 2017;1–12. Available from: http://dx.doi.org/10.1038/srep45978

6. Dunn S, Wahl L, Gloor G. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics. 2008;24:333–40.

7. Lee HW, Dominy BN, Cao W. New family of deamination repair enzymes in Uracil-DNA glycosylase superfamily. J. Biol. Chem. 2011;286:31282–7.

8. Altschul SF, Gish W, Miller W, Myers EW LD. Basic Local Alignment Search Tool [Internet]. J. Mol. Biol. 2011. p. 403–10. Available from: http://www.biomedcentral.com/1471-2148/12/196%5Cnhttp://dx.plos.org/10.1371/journal.pone.0107169%5Cnhttp://doi.wiley.com/10.1111/j.1471-8286.2006.01657.x%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3316671&tool=pmcentrez&rendertype=abstract%5Cnht

9. Sikic K, Carugo O. Protein sequence redundancy reduction: comparison of various methods. Bioinformation [Internet]. 2010;5:234–9. Available from: http://www.bioinformation.net/005/005200052010.htm

10. Halabi N, Rivoire O, Leibler S, Ranganathan R. Protein sectors: evolutionary units of three-dimensional structure. Cell. 2009;138:774–86.

11. Guevara-Coto J, Schwartz CE, Wang L. Protein sector analysis for the clustering of disease-associated mutations. BMC Genomics [Internet]. BioMed Central Ltd; 2014;15 Suppl 1:S4. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25559331

12. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994;22:4673–80.

13. Thompson JD, Gibson TJ, Higgins DG. Multiple Sequence Alignment Using ClustalW and ClustalX. Curr. Protoc. Bioinforma. [Internet]. John Wiley & Sons, Inc.; 2002. Available from: http://dx.doi.org/10.1002/0471250953.bi0203s00

14. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular evolutionary genetics analysis version 6.0. Mol. Biol. Evol. 2013;30:2725–9.

15. Eddy SR. Where did the BLOSUM62 alignment score matrix come from? Nat. Biotechnol. [Internet]. 2004;22:1035–6. Available from: http://www.nature.com/doifinder/10.1038/nbt0804-1035

16. Styczynski MP, Jensen KL, Rigoutsos I, Stephanopoulos G. BLOSUM62 miscalculations improve search performance. Nat. Biotechnol. [Internet]. 2008;26:274–5. Available from: http://www.nature.com/doifinder/10.1038/nbt0308-274

17. Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32:1792–7.

18. Shannon P, Markiel A, Owen Ozier 2, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;2498–504.

19. De Juan D, Pazos F, Valencia A. Emerging methods in protein co-evolution. Nat. Rev. Genet. [Internet]. Nature Publishing Group; 2013;14:249–61. Available from: http://dx.doi.org/10.1038/nrg3414

20. Li J, Yang Y, Guevara J, Wang L, Cao W. Identi fi cation of a prototypical single-stranded uracil DNA glycosylase from Listeria innocua. DNA Repair (Amst). [Internet]. Elsevier; 2017;57:107–15.