BioSystems 172 (2018) 1–8

Contents lists available at ScienceDirect

BioSystems

journal homepage: www.elsevier.com/locate/biosystems

Disease Probability Index (DPI, χ): A new alignment-free scoring method to evaluate the propensities of polypeptide sequences leading to disease onset

Ananya Ali, Angshuman Bagchi*

Department of Biochemistry and Biophysics, University of Kalyani, Kalyani, Nadia, 741235, India

ARTICLE INFO

Keywords:
Single amino acid variants
Computational biology
Cross-validation
Disease probability index

ABSTRACT

The analyses of the amino acid sequences of proteins provide valuable information regarding the structure and function of the protein. A comparatively new approach is the alignment-free sequence comparisons. To-date most, if not all, sequence analysis techniques are used to find out the sequence homologies to measure the evolutionary relatedness among the species. However, a still untouched avenue in the field of sequence analyses is to build a comparative estimate of the sequence similarities between unrelated protein sequences from and within a single species.

In this work, we tried to develop an alignment-free scoring method to study sequences from different proteins belonging to humans to identify the disease-associations of the sequences. A total of 52 protein sequences were analyzed. There were 599 reported polymorphic sites and 802 (708 polymorphic and 94 disease-associated) Single Amino acid Variants (SAVs) in the training data set. For cross-validation purposes, another set of 62 protein sequences (26 enzymes, 16 Membrane-bound Enzymes and 20 Membrane-bound Proteins), with a total of 261 reported polymorphic sites and 799 (291 polymorphic and 508 disease-associated) SAVs, were used. A negative correlation was observed for both training and cross-validation data set between percentage of reported disease-associated SAVs with a ratio of (polymorphic site : protein length). A new scoring pattern was also developed that would take into account the ratio of polymorphic site and protein length by counting the number of polymorphic amino acids and the total numbers of amino acids in proteins.

1. Introduction

Sequence analyses, both for proteins and nucleic acids, have become a general practice both among molecular and computational biologists due to its versatile usefulness (Karlin and Altschul, 1990; Vinga and Almeida, 2003). Sequence analyses have been found to provide valuable results to identify biologically conserved domains, biological and functional constraints on molecular evolution, phylogenetic analyses to name a few (Pearson and Lipman, 1988; Valdar, 2002). A relatively unexplored avenue of sequence analyses is the development of new scoring systems to evaluate the properties of different mutations in protein sequences (Capra and Singh, 2017; Valdar, 2002; Vinga and Almeida, 2017). Thus, it has become important to build new scoring systems for mutational analysis purposes. There are a number of alignment free sequence comparison techniques available along-with traditional sequence alignment methods (Vinga and Almeida, 2003). Studies have shown that a vast range of theoretical background was employed for sequence comparison techniques that range from linear algebra to computation informatics and from applications of statistical methods to complex and non-linear system studies (Vinga and Almeida, 2003).

Most of these studies and methods were mainly aimed at analyzing the sequence conservation score to identify a conserved region of the sequences, by comparing homologous proteins (Capra and Singh, 2017; Karlint and Altschult, 1990; Lipman et al., 2002; Pearson and Lipmant, 1988; Valdar, 2002; Vinga and Almeida, 2017). Some studies also revealed that the amino acid sequence lengths of the proteins had good correlations with its' functions, conservational constraint and also to the evolutionary phylogeny (Lipman et al., 2002; Zhang, 2000).

Thus, in the present context, we made an attempt to find a new

* Corresponding author.
E-mail address: angshumanb@gmail.com (A. Bagchi).

https://doi.org/10.1016/j.biosystems.2018.06.001
Received 18 January 2018; Received in revised form 9 April 2018; Accepted 11 June 2018
Available online 09 August 2018

alignment free scoring system to analyze the disease propensity of a mutation by extracting the information from the amino acid sequences of the proteins. With the help of this newly developed scoring scheme, we could analyze the properties of the amino acid sequences from unrelated proteins from the same species. The built scoring scheme would depend only on the numerical information about the length of the protein and the number of polymorphic sites present in the sequence. The scoring scheme would be beneficial to analyze the properties of various types of mutations to categorize them as disease causing or polymorphic.

2. Materials and methods

2.1. Dataset

We used a dataset (provided in supplementary Table1), consisting of 52 amino acid sequences of different proteins from Uniprot (Leinonen et al., 2004) for our analyses. A total of 599 reported polymorphic sites and a total of 802 (708 polymorphic and 94 disease-associated) Single Amino Acid Variants (SAVs) were present in the aforementioned dataset used for the analyses. This dataset was used as the training dataset. In our analyses, we considered the following information:

Table 1
List of Data Set being used for both training and cross validation (Test data set).

Dataset	No	Protiens		Mutations			
		Uniprot ID	Length (N)	Number of polymorphic sites (i)	Total number of mutations	Number of Polymorphic mutations	Number of Disease causing Mutations
TRAINING DATA SET	1	P01891	365	13	17	17	0
	2	P01892	365	21	23	23	0
	3	P03989	362	13	16	16	0
	4	P04222	366	13	13	13	0
	5	P13746	365	15	15	15	0
	6	P13747	358	4	4	4	0
	7	P18464	362	5	5	5	0
	8	P18465	362	9	10	10	0
	9	P30443	365	16	16	16	0
	10	P30460	362	21	27	27	0
	11	P30464	362	8	9	9	0
	12	P30481	362	10	11	11	0
	13	P30504	366	15	15	15	0
	14	P30685	362	13	14	14	0
	15	Q95460	341	3	3	3	0
	16	Q96PC3	154	6	8	6	2
	17	Q9UBH0	155	1	5	1	4
	18	P25445	335	14	36	14	22
	19	P16410	223	1	2	1	1
	20	P01241	217	5	21	5	16
	21	P01589	272	1	3	1	2
	22	Q9NZK5	511	1	11	1	10
	23	P01903	254	2	2	2	0
	24	P01909	254	51	66	66	0
	25	P20036	260	19	19	19	0
	26	P28067	261	6	7	7	0
	27	P30481	362	10	11	11	0
	28	P01911	266	10	10	10	0
	29	P04229	266	27	31	31	0
	30	P13760	266	9	10	10	0
	31	P79483	266	30	43	43	0
	32	Q30154	266	25	38	38	0
	33	P01920	261	65	85	85	0
	34	P04440	261	41	56	56	0
	35	P28068	258	6	7	7	0
	36	P01730	458	3	4	3	1
	37	P01732	235	1	2	1	1
	38	P04234	171	1	3	1	2
	39	P05106	788	26	59	29	30
	40	P08571	375	2	2	2	0
	41	P08575	1304	7	7	7	0
	42	P08637	254	5	7	6	1
	43	P11215	1152	4	4	4	0
	44	P13591	858	4	4	4	0
	45	P15391	556	2	2	2	0
	46	P16284	738	4	5	5	0
	47	P20273	847	8	8	8	0
	48	P20701	1170	4	4	4	0
	49	P25063	80	2	2	2	0
	50	P28907	300	1	1	1	0
	51	P28908	595	5	6	6	0
	52	Q99062	836	11	13	11	2

(continued on next page)