UNIVERSITY OF BELGRADE
FACULTY OF MATHEMATICS

Samira A. Alshafah

Data mining on protein sequences: n-gram analysis of ordered and disordered protein regions

Doctoral Dissertation

Belgrade, 2018

UNIVERZITET U BEOGRADU
MATEMATIČKI FAKULTET

Samira A. Alshafah

Istraživanje podataka na proteinskim niskama: n-gramska analiza uređenih i neuređenih regiona proteina

Doktorska disertacija

Belgrade, 2018

Mentor

dr Nenad Mitić, vanredni profesor
Univerzitet u Beogradu, Matematički Fakultet

Članovi komisije

dr Saša Malkov, vanredni profesor
Univerzitet u Beogradu, Matematički Fakultet

dr Miloš Beljanski, naučni savetnik
Institut za opštu i fizičku hemiju

DISSERTATION DATA

Doctoral dissertation title: Data mining on protein sequences: n-gram analysis of ordered and disordered protein regions

Abstract: Proteins with intrinsically disordered regions are involved in large number of key cell processes including signaling, transcription, and chromatin remodeling functions. On the other side, such proteins have been observed in people suffering from neurological and cardiovascular diseases, as well as various malignancies. Process of experimentally determining disordered regions in proteins is a very expensive and long-term process. As a consequence, a various computer programs for predicting position of disordered regions in proteins have been developed and constantly improved.

In this thesis a new method for determining Amino acid sequences that characterize ordered/disordered regions is presented. Material used in research includes 4076 viruses with more than 190000 proteins. Proposed method is based on defining correspondence between n-grams (including both repeats and palindromic sequences) characteristics and their belonging to ordered/disordered protein regions. Positions of ordered/disordered regions are predicted using three different predictors.

The features of the repetitive strings used in the research include mole fractions, fractional differences, and z-values. Also, data mining techniques association rules and classification were applied on both repeats and palindromes. The results obtained by all techniques show a high level of agreement for a short length of less than 6, while the level of agreement grows up to the maximum with increasing the length of the sequences. The high reliability of the results obtained by the data mining techniques shows that there are n-grams, both repeating sequences and palindromes, which uniquely characterize the disordered/ordered regions of the proteins. The obtained results were verified by comparing with the results based on n-grams from the DisProt database which contains the positions of experimentally verified disordered regions of the protein. Results can be used both for the fast localization of disordered/ordered regions in proteins as well as for further improving existing programs for their prediction.

Keywords
n-gram, data mining, ordered/disordered regions, association rules, proteins

Scientific field
Computer Science

Scientific subfield
Data Mining

Podaci o doktorskoj disertaciji

Naslov doktorske disertacije: Istraživanje podataka na proteinskim niskama: n-gramska analiza uređenih i neuređenih regiona proteina

Rezime: Proteini koji imaju neuređene regione učestvuju u velikom broju ćelijskih procesa kao što su prenos signala, transkripcija i remodelovanje funkcija hromatina. Sa druge strane, pojava takvih proteina je uočena kod osoba koje boluju od neuroloških i kardiovaskularnih bolesti, kao i različitih oblika maligniteta. Eksperimentalno određivanje neuređenih regiona protiena je vrlo skup i spor proces. Zbog toga su razvijeni i stalno se usavršavaju različiti računarski programi za predviđanje pozicija neuređenih regiona u proteinu.

U radu je prikazana nova metoda za određivanje niski amino kiselina koje karakterišu neuređene i uređene regione proteina. Materijal nad kojim je vršeno istraživanje obuhvata 4076 virusa sa preko 190000 proteina. Metoda je zasnovana na ispitivanju osobina n-grama (koji obuhvataju ponavljajuće i palindromske niske) i njihove pripadnosti uređenim i neuređenim regionima proteina. Pozicije neuređenih /uređenih regiona u proteinima su određene korišćenjem tri programa za predviđanje. Osobine ponavljajućih niski koje su korišćene u istraživanju uključuju molske frakcije, frakcijske razlike i z-vrednost. Takođe, na ponavljajuće niske kao i na palindromske niske primenjene su određivanje pravila pridruživanja i klasifikacija, kao tehnike istraživanja podataka. Rezultati dobijeni svim tehnikama pokazuju visok nivo saglasnosti, za niske dužine manje od 6, dok nivo saglasnosti rezultata raste sve do maksimalnog sa porastom dužine niski. Visoka pouzdanost rezultata dobijenih tehnikama istraživanja podataka, pokazuje da postoje n-grami, kako ponavljajuće sekvence tako i palindromi, koji jednoznačno karakterišu neuređene/uređene regione proteina. Dobijeni rezultati su provereni upoređivanjem sa rezultatima zasnovanim n-gramima iz DisProt baze koja sadrži pozicije eksperimentalno verifikovanih neuređenih regiona proteina, i mogu da budu korišćeni kako za brzu lokalizaciju neuređenih/uređenih regiona u proteinima tako i za dalje poboljšanje postojećih programa za njihovo predviđanje.

Ključne reči
n-gram, istrživanje podataka, uređeni/neuređeni regioni, pravila pridruživanja, proteini

Naučna oblast
Računarstvo

Naučna podoblast
Istraživanje podataka

Table of Content

1 Introduction

1.1 Bioinformatics

From its very beginning until the fourth quarter of the 20th century biology has been observational and experimental science. Recent development and using computers not altered completely this orientation, but introduce new methods and algorithms in processing of biological material. Nature of data has changed - data have become discret and more precise. The quantity of available data grow rapidly bringing to the scene a new discipline capable to provide efficient processing of data in the new conditions - Bioinformatics. Bioinformatics has a lot of subdisciplines and research directions [1]. Most pressing task of bioinformatics has moved to analyze and interpret various types of data, including nucleotide and amino acid sequences, protein structures and interactions, and so on. To meet the new requirements arising from the new tasks, researchers in the field of bioinformatics are working on the development of new algorithms (mathematical formulas, statistical methods, etc) and software tools which are designed for assessing relationships among large data sets stored, such as methods to locate a gene within a sequence, predict protein structure and/or function, understand diseases at gene expression level and etc.

A particular active area of research in bioinformatics is the application and development of data mining techniques to solve biological problems. Analyzing large biological data sets requires making sense of the data by inferring structure or generalizations from the data. Examples of this type of analysis include protein structure prediction, gene classification, cancer classification based on microarray data, clustering of gene expression data, statistical modeling of protein-protein interaction, etc.

1.2 Proteins

Proteins are biological macromolecules, of polymeric nature, that are built by forming of so called "peptide bond" (polypeptides) between their basic constituents amino acids (Figure 1). Amino acids (AA) are organic molecules that posses at least one amino ($-NH_2$) and carboxyl (-COOH) group. There are 20 (+2) amino acids that constitute all, so far, known proteins. Protein structure and function are mainly determined by so called "*protein primary structure*", which represents amino acid content of protein molecule, its number and sequence (Figure 1). In bioinformatics amino acids are represented by one or three letter code as shown in Appendix table A1.

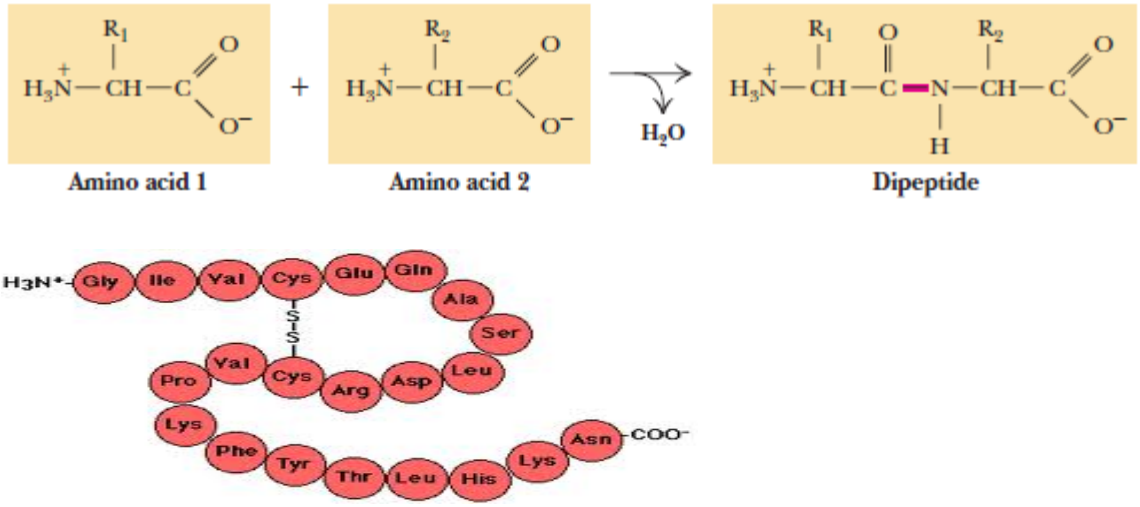


Figure 1. Forming of peptide bond between two amino acids (a), schematic representation of protein primary structure (b). Primary structure is "reed" from N- to C-terminal of polypeptide (protein) chain.

Protein "*secondary structure"* may be defined by so called "torsion angles", (φ and ψ), from Ramachandran diagram [2], between successive amino acids, that forms backbone of polypeptide chain (Figure 2.A and 2.B). If three or more pairs of torsion angles are the same, than there is a regular secondary structure. Secondary structure results from forming secondary, noncovalent H-bonds between C=O and H-N groups; the exact pattern of them is different in different forms of secondary structure. Two of most represented secondary structures in proteins are **alpha (α) helix** structure and the **beta (β) pleated sheet** (Figure 3.).

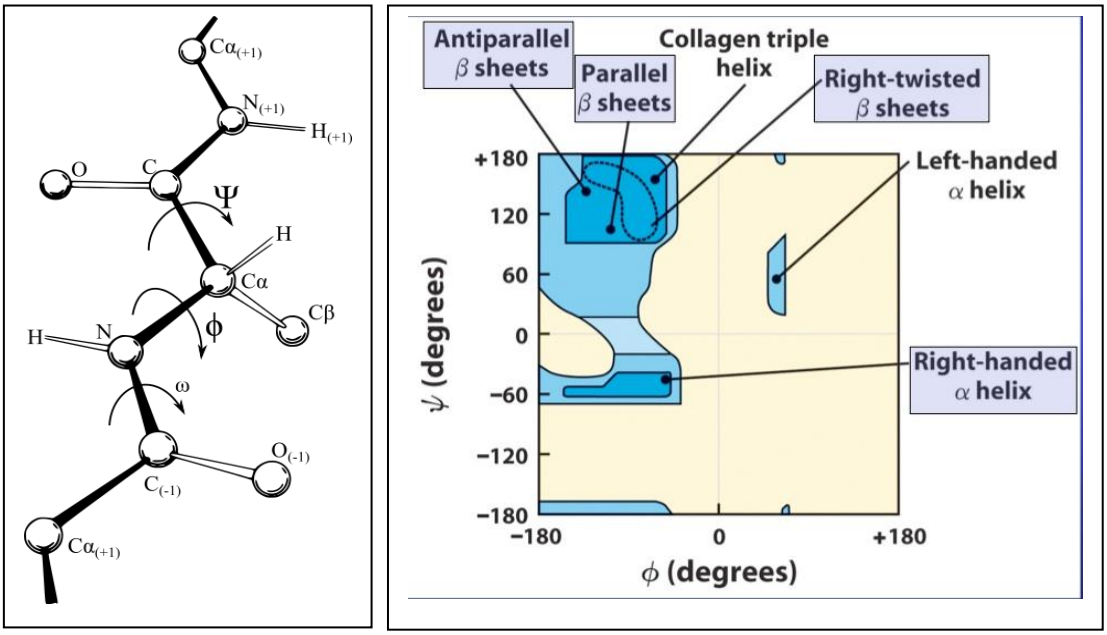


Figure 2. **A**: Part of polypeptide chain with φ and ψ torsion angles between C_α and N atom (from amino group) and C_α C atom from carboxyl group, ω torsion angle that corresponds to peptide link is small and usually neglected. **B**: Ramachandran diagram with marked areas that correspond to certain secondary structures.
Source: A - Jane S. Richardson, The Anatomy and Taxonomy of Protein Structure. In Advances in protein chemistry, Vol. 34 (1981)
B - https://www.studyblue.com/notes/note/n/protein-structure/deck/7778686

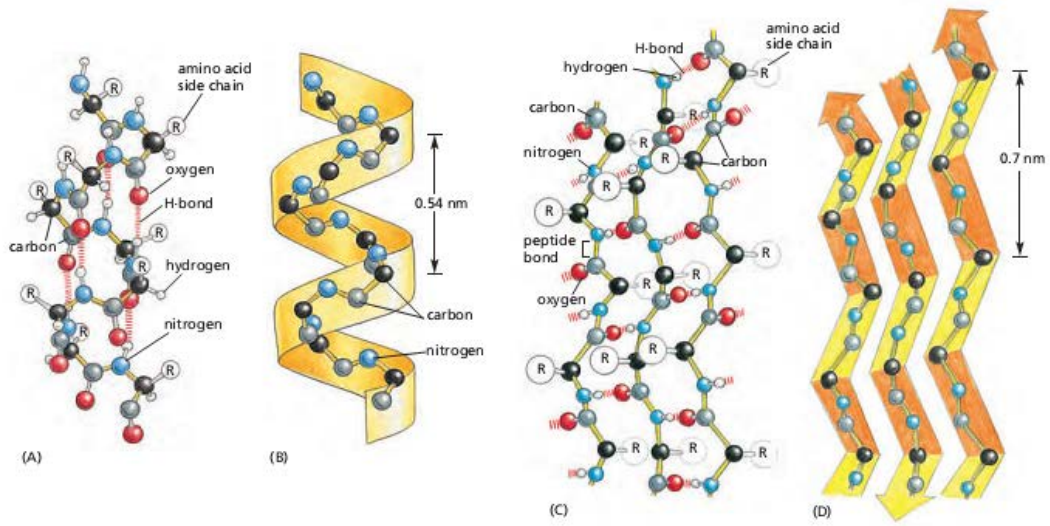


Figure 3. The *α*-helix (A, B) and the β-pleated sheet (C, D) are the two principal secondary structures found in protein. C, N, O and H atoms involved in polypeptide chain forming.
Source: Bruce Alberts, Alexander Johnson, Julian Lewis, David Morgan, Martin Raff, Keith Roberts, Peter Walter, Molecular biology of the cell. 6 ed, (2015), Garland Science, Taylor & Francis Group, 711 Third Avenue, New York, NY 10017, US 3 Park Square, Milton Park, Abingdon, OX14 4RN, UK, ISBN 978-0-8153-4432-2

Protein *tertiary structure* refers to the spatial arrangement of a polypeptide chain through folding and coiling to produce a compact globular shape. It may be defined by knowing positions of all atoms that protein consists of [3, 4].

1.2.1 Intrinsically disordered proteins/protein regions (IDP/IDPR)

In last 15 years, it became more and more evident that a significant number of proteins, under physiological conditions, do not possess a well defined 3 dimensional ordered structure (Figure 4). They exhibit a variety of conformational isomers in which the atom positions and the polypeptide backbone torsion angles of the Ramachandran plot vary over time, with no specific equilibrium values, typically involving non-cooperative conformational changes [5]. They may be completely or partially disordered and may undergo a disorder-to-order, or vice versa, transition upon interaction with other molecules. Thanks to their high structural mobility they readily interact with other molecules/proteins and carry out mostly regulatory functions related to molecular recognition, signal transduction, protein-protein, and protein-nucleic acid interaction.

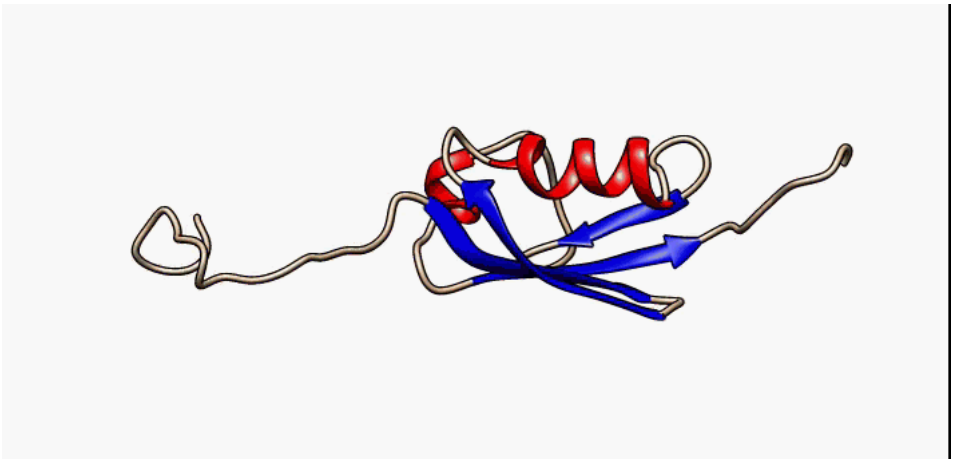


Figure 4. SUMO-1 protein (PDB:1a5r), with central part that shows relatively ordered structure. The N- and C-terminal regions (left and right, respectively) are intrinsically disordered (grey disordered regions). Secondary structure elements: α-helices (red), β-strands (blue arrows).
Source: http://www.rcsb.org/pdb/explore/explore.do?structureId=1a5r

In accordance to arising function, they are classified into, at least, 16 structural/functional categories, as listed in the DisProt database, that currently contain 803 experimentally determined IDP/IDPRs [6]. Taxonomically, IDPs are represented in

the proteomes of all of the three superkingdoms (Archaea, Bacteria and Eukarya), as well as in viruses. Primary structure of IDP/IDPRs are characterized by low sequence complexity (i.e. often consist of repetitive short fragments) and are biased toward polar and charged, but against bulky hydrophobic and aromatic AA residues (Figure 5), i.e., they are enriched in Ala, Arg, Gly, Gln, Ser, Glu, Lys and Pro and depleted in order-promoting Trp, Tyr, Phe, Ile, Leu, Val, Cys, Asn AAs [7]. Experimentally, IDP/IDPRs may be detected by more than 20 various biophysical and biochemical techniques such as: x-ray diffraction crystallography, heteronuclear multidimensional NMR, circular dichroism, etc. Since IDP/IDPRs experimental study is costly and difficult (because of the lack of unique structure in the isolated form), a number of prediction tools have been developed [8].

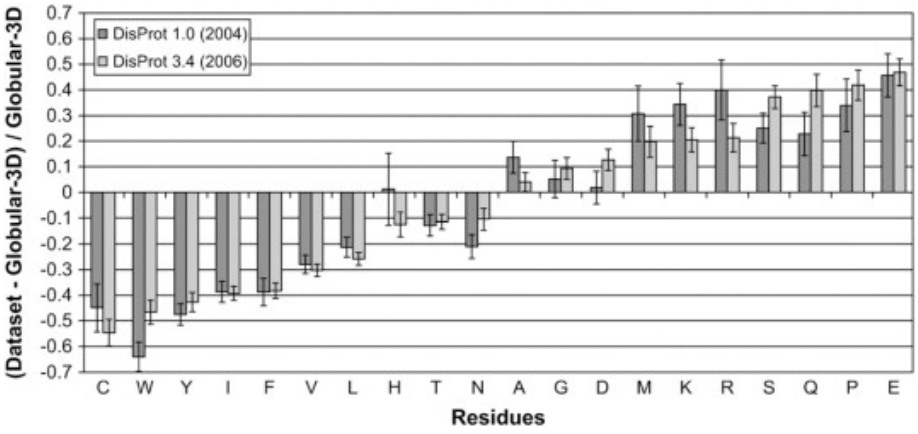


Figure 5. Fractional differences in composition between disordered and ordered sets of regions**,** calculated on the basis of data from DisProt DB. On right part of diagram are amino acids with higher propensity to disorder.
Source: Predrag Radivojac, Lilia M. Iakoucheva, Christopher J. Oldfield, Zoran Obradovic, Vladimir N. Uversky, and A. Keith Dunker, Intrinsic Disorder and Functional Proteomics. Biophysical Journal Volume 92 March 2007 1439–1456. doi: 10.1529/biophysj.106.094045

Disorder prediction

Disordered regions of the protein chain are important for the protein function. Today there are special programs (disorder predictors) that can predict them. IDP/IDPRs predictors can be grouped according characteristics or methods used for prediction. For example, one group include those that use physico-chemical properties of amino acids in proteins (PONDR, FoldUnfold, IUPred, GlobPlot, PreLINK, and FoldIndex), the second one those that use alignment of homologous protein sequences (Ronn, Disopred), etc. [9]. A summary of these methods can be found in Appendix Table A2.

Programs of the first group differ by the property of amino acids in proteins used for prediction of disordered regions. For example, PONDR uses local amino acid composition and hydrophobicity, FoldUnfold uses number of expected contacts, PreLINK uses propensity of a chain region to form a hydrophobic cluster, and IUPred uses estimation of the energy interaction between neighbouring amino acids. In the second group, the RONN program uses a neural network and compares the given sequence with a number of sequences whose structure can be a priori determined (ordered/disordered/mixture), while DISOPRED uses the network trained to distinguish regions that are missed in the structure obtained by x-ray analysis [10 , 11, 12, 13].

1.3 Viruses

Viruses are small infectious agent that proliferates only inside the cells of all life forms: Archaea, Bacteria and Eukaryote. Outside of a cell viruses exist in the form of a virion, that consist of two, or three parts: (i) the genetic material made from either DNA or RNA; (ii) a protein coat, called the capsid, which surrounds and protects the genetic material; and in some cases (iii) an lipid envelope that surrounds the protein coat when they are outside a cell [14].

Genomic organization of viruses shows an enormous variety (as a group, they contain more structural genomic diversity than in all of three superkingdoms). Genome size varies greatly: in general, RNA viruses have smaller genome sizes than DNA viruses, although the smallest viral genome is that of ssDNA circoviruses (family *Circoviridae*), have a genome size of only two kilobases and code for only two proteins. The largest–genome size is that of the pandoraviruses of around two megabases, which code for about 2500 proteins. Virus genes are often arranged so that they overlap and rarely have introns.

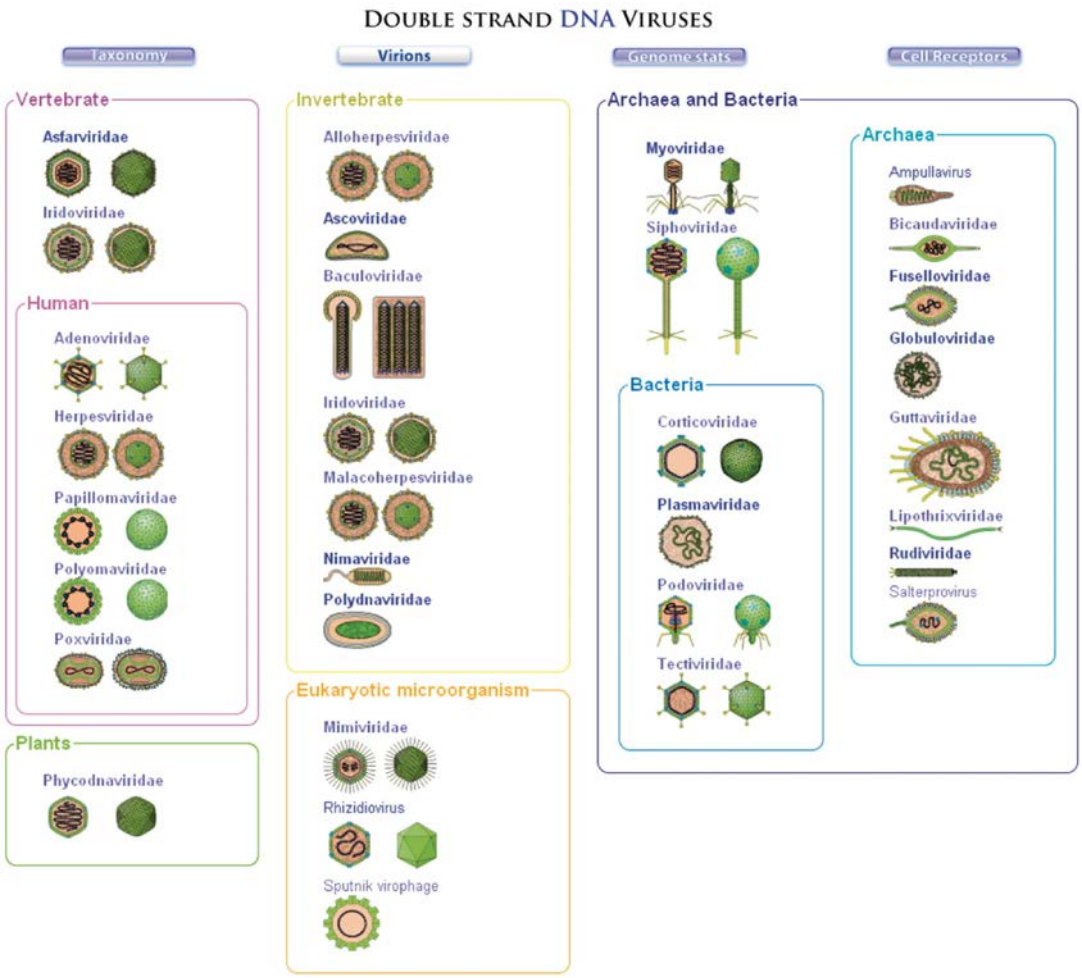


Figure 6. Viral classification according to host and morphology.
(Source: Nucleic Acids Research 2011;39 (Database issue) :D576–D582)

Viral proteins exhibit distinct and structural features than the host proteins. There are several potentially unique characteristics of viral proteins, that include (a) the low contact densities, (b) the high occurrence of random coil segments and short disordered regions and (c) the lower destabilizing effects of mutations [15]. It has been shown that viruses have the largest variation range of the disordered residue fractions in their proteomes (human coronavirus NL63 has only 7.3% disordered residues, while Avian carcinoma virus proteome has 77.3% disordered residues). Also, some viral species are highly enriched in intrinsic disorder. With the increase of proteome size, the fractions of disordered residues seem to converge to a range between 20 and 40%. IDP/IDPRs help viruses to deal with their hostile habitats, in managing of their gene expression and generally, better adaptability and functioning of their proteins [16].

There are probably millions of different types of viruses, although only about 5,000 species have been described in detail. At the beginning of 2017 year, the NCBI Virus genome database has more than 7000 complete virus genomes. Viruses may be classified according to different criteria: their host and morphology (as shown on Figure 6), their morphology (symmetry and possession of envelope), genome organization (ds or ss; DNA or RNA) and in the case of Baltimore classification on mechanisms of viral genome replication (i.e., mechanism of viral mRNA production). This classification places viruses into seven groups as shown on Figure 7.

1.4 Topic of the dissertation

Because of importance of disorder regions for protein function, the research topic in this dissertation is to find amino-acids strings that characterize ordered/disordered protein regions. The aim is not to produce new disorder predictor, but to discover are there any AA (or series of AAs) that can be used as 'indicators' of region type, without pretension to determine exact boundaries of such regions.

The characteristics of AA can be mapped to the problem of finding characteristics of sequence of AAs (called n-gram) where the length of the sequence can be 1, 2, ..., N. There are different methods for characterization such n-grams in some environment (e.g. string), but no one can, in advance, determine characteristics that can be used as indicators, with high accuracy. This research will use set of viral proteins as material. Viruses from different phyla are used as material to minimize potential influence of group of specific phyla.

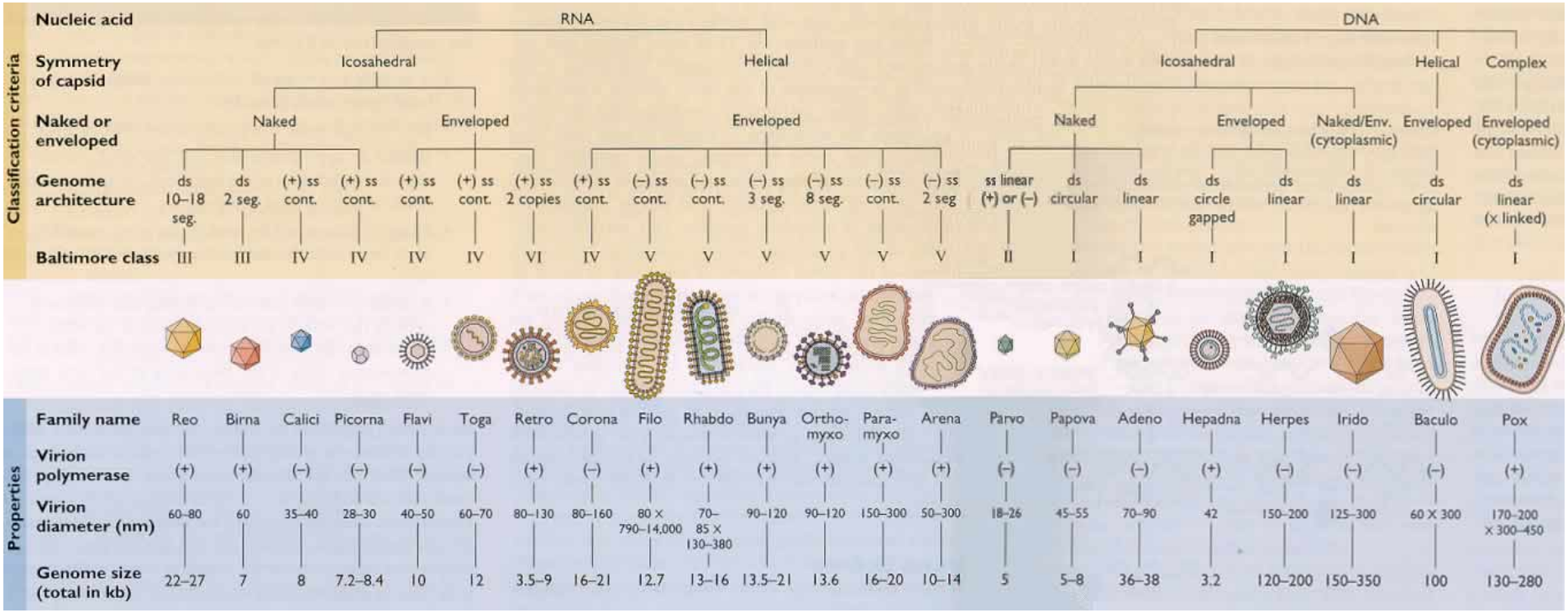


Figure 7. Baltimore classification of viruses (Source: http://www.nlv.ch/Virologytutorials/Classification.htm)

2 Methods for determining characteristics strings in protein regions

In this chapter idea for construction of specific model for determining n-grams that characterize disorder/order (D/O) protein regions is described. First part of the chapter includes description of methods for n-gram characterization in the specific string. In the second part, idea for discover n-gram characteristics that are related to ordered/disordered regions in proteins is described. Second part also includes discussion about quality of proposed model.

2.1 N-gram analysis

There are many definitions of n-gram. In this research we used the following one [17]:

Definition 1: Given a sequence of letters $S=s_1s_2...s_{N+(n-1)}$ over the alphabet A, with N and n a positive integer, an n-gram of the sequence S is an n-long subsequence of consecutive letters. The i-th n-gram of S is the sequence $s_is_{i+1}...s_{i+n-1}$.

There are N such n-grams in S. For an alphabet A with $|A|$ distinct letters, there are $|A|^n$ possible unique n-grams. *Gram* is a Greek word; depends on value of n, n-grams are denoted as monograms ($n=1$), bigrams ($n=2$), trigrams ($n=3$), tetragrams ($n=4$), pentagrams ($n=5$), hexagrams ($n=6$), etc. Some authors prefer using names unigram, bigram, trigram, quadrigram..., etc.

Simple n-gram analysis includes counting of specific n-gram occurrences in observed (analyzed) areas, as well as calculating the difference and, if applicable, the standard deviation of its occurring in those areas compared to the whole material. In this research the n-gram analysis for the occurrence of amino acids in the ordered/disordered

regions of proteins has been performed. *N-Grams* belong to any of the three regions including: disordered region (D), ordered region (O) and borderline transition from ordered to disordered region or vice versa (N) in the proteins, whereas monograms can belong to either D or O region only. For example, the amino acids in the sequence RAVERSQVSEN in a protein may correspond to the following ordered/disordered regions: OODODDDOOOO. The set of monograms in the sequence is {R A V E R S Q V S E N} and their corresponding disordered/ordered characteristics are {O O D O D D D O O O O}. The set of bigrams for the above amino acids sequence is {RA AV VE ER RS SQ QV VS SE EN}, while corresponding ordered/disordered regions characteristics are {O N N N D D N O O O}. Analogously, the set of the trigram representations of the above amino acids sequence is {RAV AVE VER ERS RSQ SQV QVS VSE SEN}, with the corresponding ordered/disordered region characteristics {N N N N D N N O O}.

N-gram analysis has also been performed at the level of nucleotide sequence. Because nucleotide sequences are widespread across whole genome sequence, there are four (compared to three in the case of proteins) possible regions: disordered regions (D) which corresponds to the positions (in the genome sequence) of the disorder regions in proteins, ordered region (O) which corresponds to the positions (in the genome sequence) of the order regions in proteins, intergenic regions (I) which corresponds to the parts of the genome sequence that did not corresponds to any of the proteins, and borderline transition (N) between some of the previous three kinds of regions. In this research, the objects of n-gram analysis are nucleotide sequences that correspond to amino-acid sequences in proteins, so they belong to D, O or N regions only.

2.2 Repeats

Repeats can be considered as a special type of n-grams. Various kinds of repeats can be defined based on underlying n-gram characteristics. The following definition of repeats is taken from [18]:

Definition 2: Let *A= { a, b, c, d, ...}* denote an alphabet with arbitrary symbols and L={l_1, l_2, ..., l_n} is a language over alphabet *A* which includes strings over *A* with an arbitrary length, including empty string, and let |s| denote length of string $s \in L$, which is equal to the number of symbols (letters) from alphabet *A*.

An ordered triplet (x,s,p_x) denotes a substring $x \in L$ of string $s \in L$ at the position $p_x \geq 1$ if $\exists y, z \in L : s = yxz \wedge |s| = |x| + |y| + |z| \wedge |x| \geq 1.$ where $|y|=p_x$

Let the following functions be defined as:

(a) $f : L \rightarrow L$

$$f(x) = z, \quad \text{if } |x| = 1 \quad \text{for some } z \in A$$
$$f(x_1) f(x_2), \quad \text{if } x = x_1 x_2 \in L \wedge |x| > 1$$

(b) $g : L \rightarrow L$

$$g(xy)=yx, \quad \text{if } |x| = 1 \wedge |y| = 1$$
$$g(xy)=yg(x), \quad \text{if } |x| > 1 \wedge |y| = 1$$
$$g(xy)=g(y)x, \quad \text{if } |x| = 1 \wedge |y| > 1$$
$$g(xy)=g(y)g(x), \quad \text{otherwise}$$

then, for all string $s \in L$ the following four types of repeats can be defined (Figure 8):

1) The substring pair (a,s,p_a) and (b,s,p_b) is a direct *non-complementary repeat* (*DN*) if and only if $a = b \wedge p_a < p_b$
2) The substring pair (a,s,p_a) and (b,s,p_b) is a *inverse non-complementary repeat* (*IN*) if and only if $a = g(b) \wedge p_a \leq p_b$
3) The substring pair (a,s,p_a) and (b,s,p_b) is a *direct complementary repeat* (*DC*) if and only if $a = f(b) \wedge p_a < p_b$
4) The substring (a,s,p_a) and (b,s,p_b) is a *inverse complementary repeat* (*IC*) if and only if $a = f(g(b)) = g(f(b)) \wedge p_a \leq p_b$

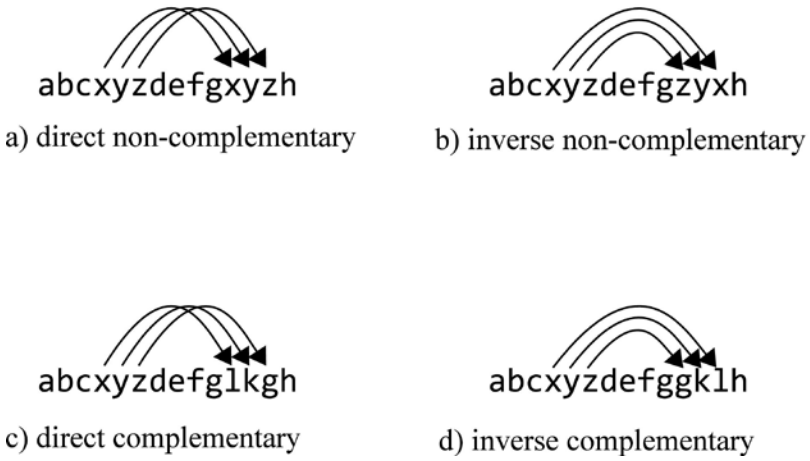


Figure 8. Graphical presentation of repeat types. In the examples, f(x)=l, f(y)=k, f(z)=g is used for complementary mapping.

Extracting repeats from protein sequences is done using StatRepeats program [18]. Two different alphabets were used:

- A={A, C, G, T}, when extracting repeats from (protein) nucleotide sequences, and
- A={A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y, U, O}, when extracting repeats from (protein) amino-acids sequences.

By using nucleotide complementary characteristics (A↔T, G↔C) in definition of functions *f* and *g*, it is possible to obtain all four types of repeats for nucleotide sequences. For amino-acid sequences only non-complementary repeats are correct. Although, on first sight, looks that, except for monograms, set of direct non-complementary repeats is equal to set of n-grams, this is not correct because StatRepeats extracts maximal repeats (i.e. repeats that not belongs to longer one). On the other side, StatRepeats can extract all (maximal) repeats or just subset of statistically significant repeats which can be used for additional checking of results.

2.3 Mole Fractions and fractional difference

Mole fractions are one way of representing the *concentrations* of the various chemical elements. In chemistry, mole fraction *x* is a way of expressing the composition of a mixture. The mole fraction x_i of each component *i* is defined as its amount of substance k_i divided by the total amount of substance in the system, k_{sum}:

$$x_i = \frac{k_i}{k_{sum}} \quad \text{where} \quad k_{sum} = \sum_{i=1}^{N} k_i$$

k_{sum} is calculated over all components, including the solvent in the case of a chemical solution. Consequence of such definition is that the sum of all the mole fractions is equal to 1.

$$\sum_{i=1}^{N} x_i = \sum_{i=1}^{N} \frac{k_i}{k_{sum}} = \frac{\sum_{i=1}^{N} k_i}{k_{sum}} = 1$$

In our research the mole fractions of amino-acid and nucleotide *n*-grams in regions was used as additional method for discovering *n*-grams that characterize specific type of regions. Mole fraction of specific *n*-gram in some region is calculated as quotient of number of *n*-gram occurrences in region and region length.

As a measure for difference of occurrences the same n-gram in different regions, the **fractional difference** (**FD**) is used. Fractional difference of occurrences of n-gram *ngr* in region reg_1 related to the region reg_2 can be defined as

$$FD(ngr, reg_1, reg_2) = (x_{ngr\text{-}reg1} - x_{ngr\text{-}reg2}) / x_{ngr\text{-}reg2}$$

where $x_{ngr\text{-}regi}$ denotes mole fraction of the n-gram *ngr* in the region reg_i. Thus a negative value for FD indicates a poorer concentration of n-gram *ngr* in the region reg_1, while a positive value of FD indicates a richer concentration of n-gram *ngr* in the region reg_1 then in the region reg_2.

2.4 Z-score

A z-score (also known as z-value, standard score, or normal score) is a measure of the divergence of an individual experimental result from the most probable result, the mean. Z-Score is a statistical measurement of a score's relationship to the mean in a group of scores, and is expressed in terms of the number of standard deviations from the mean value. A Z-score of 0 denotes that the score is equal to the mean. A Z-score can also be positive or negative, indicating how many standard deviations it is above or below the mean [18]. Prerequisites for applying z-score test are normal (or approximately normal) distribution of data and existence of standard deviation. In general, z-values are calculated according to the following formula:

$$z = \frac{X - \mu}{\sigma}$$

where X is experimentally observed mean in N items, μ is the mean value, and σ is the standard deviation.

Most statistical tests begin by identifying a null hypothesis. The Z score is a test of statistical significance that helps to decide whether or not to reject the null hypothesis. P-value, or probability value, is a statistical measure that also helps to decide if hypotheses are correct. It is directly related to the significance level, which is an important component in determining whether the data obtained from scientific research is statistically significant. In the other words, the p-value is the probability of incorrectly rejecting the null hypothesis. Z-score and p-value are connected. The judgment of rejecting the null hypothesis is often connected to some confidence levels. Typical confidence levels are 90%, 95%, or 99%. A confidence level of 99% indicates that null hypothesis will not be rejected unless the probability that the pattern was created by random chance is less than a 1% probability. The Table 1 shows the critical p-values and z-scores for different confidence levels.

Table 1. Values of z-score and p-value for some confidence levels

z-score (standard deviations)	p-value (probability)	Confidence level
< -1.65 or > +1.65	< 0.10	90%
< -1.96 or > +1.96	< 0.05	95%
< -2.58 or > +2.58	< 0.01	99%

In analysis the null hypothesis is that all n-grams have the (almost) similar number of occurrence in all region types. Because we work with n-grams (e.g. sequences), the meaning of p-value can be stated as the probability that at least one sequence will produce the same score by chance, while z-value for some n-gram measures how much standard deviations above the mean of the score distribution is number of its occurrences. In this research for evaluation of the results obtained from n-gram extracted, the statistic z-score with p-value 0.01 has been used. Presumptions of normal distributed data and existence of standard deviations for n-grams hold. Z-value for n-gram $X=L_1L_2....L_n$ where L_i denotes amino acid or nucleotide is calculated as following [29]:

$$X_Z(L_1L_2...L_n) = \frac{N(L_1L_2...L_n) - \mu}{\sigma}$$

where $N(L_1L_2...L_n)$ denotes the number of occurrences of n-gram X. The mean value μ is equal to

$$\mu = \frac{N(L_1L_2...L_{n-1}) \times N(L_2...L_n)}{N(L_2...L_{n-1})}$$

and the standard deviation σ is equal to

$$\sigma = \frac{\sqrt{\mu} \times \sqrt{[N(L_2...L_{n-1}) - N(L_1...L_{n-1})] \times [N(L_2...L_{n-1}) - N(L_2...L_n)]}}{N(L_2...L_{n-1})}$$

2.5 Data mining techniques

Data mining is the process of extracting interesting information or patterns from large information store such as: relational database, data warehouses, XML repository, etc. Also data mining is known as one of the core processes of Knowledge Discovery in Database (KDD). There are various types of data mining techniques such as association rules, classifications and clustering, etc. In this research two methods were used: association rules and classification.

Classification is a data mining technique which uses input data to build classification model. Classification uses a learning algorithm to identify model that best fits the relationship between attribute set and class label of the input data [20, 21]. Direct application of classification (for example, tree based algorithm) on complete material used in this research do not bring satisfactory results. Quality of such model is between 50% and 60%, which can not guarantee correct results of prediction. Instead of that classification was applied on parts of material (more precisely on groups of organisms that belong to the same family). Corrections and accuracy of model obtained can be measured with different measures, depends on applied classification algorithm. Detailed information about different classification algorithms and appropriate measures can be found in [20, 21, 22].

Association rules are relationships between seemingly unrelated data in a relational database or other information repository, with aim to extract interesting correlations [20, 21]. An association rule is an implication expression of the form of $X \rightarrow Y$, where X and Y are disjoint sets of items called itemsets. X is called the body (or the antecedent) of the rule, and Y the head (or the consequent) of the rule.

There are two important basic measures for association rules quality, support (denoted as s) and confidence (denoted as c). Support is defined as the percentage/fraction of records that contain $X \rightarrow Y$ to the total number of records in the database. Support reflects frequency of a set of items. Confidence is defined as the percentage/fraction of the number of transactions that contain $X \rightarrow Y$ to the total number of records that contain X. Confidence is a measure of strength of the association rules, The higher the confidence and support, the rule is more significant [20, 21].

The formal definition of support and confidence are:

Support $s(X \rightarrow Y) = \frac{\sigma(X \bigcup Y)}{N}$

Confidence $c(X \rightarrow Y)) = \frac{\sigma(X \cup Y)}{\sigma(X)}$

where $\sigma(X \rightarrow Y)$ denotes number of occurrences of an item $X \rightarrow Y$, N is the total number of items and $\sigma(X)$ denotes number of occurrences of an item X.

Using support and confidence as a measure for quality of association rules in some cases can give wrong result [20]. The reason is the fact that the confidence ignores the support of the itemset appearing in the rule consequent. One way to overcome this pitfall is to use *lift* as a metric. Lift is evaluated as the ratio between rule's confidence and the support of the itemset in the rule consequent: Lift = $c(X \rightarrow Y)/s(Y)$

The lift is a value between 0 and infinity:

(a) A lift value greater than 1 indicates that the rule body and the rule head appear more often together than expected, which makes such rule interesting.

(b) A lift smaller than 1 indicates that the rule body and the rule head appear less often together than expected. This means that the occurrence of the rule body has a negative effect on the occurrence of the rule head. Such rule can be interesting as indicate absence of rule body constituents in the case of rule head occurring.

(c) A lift value near 1 indicates that the rule body and the rule head appear almost as often together as expected, so such rule will not be considered in this research.

In this research rule head can contain only two possible forms (including "order" and "disorder"). From this point of view association rules can be considered as auxiliary method for classification.

2.6 Disorder prediction

In this research the IUPred-long [10, 23], VSL2b [13, 24] and IsUnstruct [12, 25] predictors have been used for predicting ordered and disordered level for each protein in all dataset. Three predictors with different prediction algorithms have been used in order to minimize influence of prediction algorithm to results of prediction.

Disorder predictors are very complex programs. For example, architecture of VSL2b consists of three component predictors in two-level (VSL2B-M1 and VSL2B-M2) architectures (Figure 9). At the first level, there are two specialized predictors: a short disorder predictor, VSL2b-S, for disordered regions of ≤30 residues, and a long disorder predictor, VSL2b-L, for disordered regions of >30 residues. At the second level, there is a metapredictor that combines outputs of the two specialized predictors into the final prediction. All component predictors are built as binary classifiers that approximate the posterior class probability p(c=1|x), where x is the feature (input) vector and c is the class label [24].

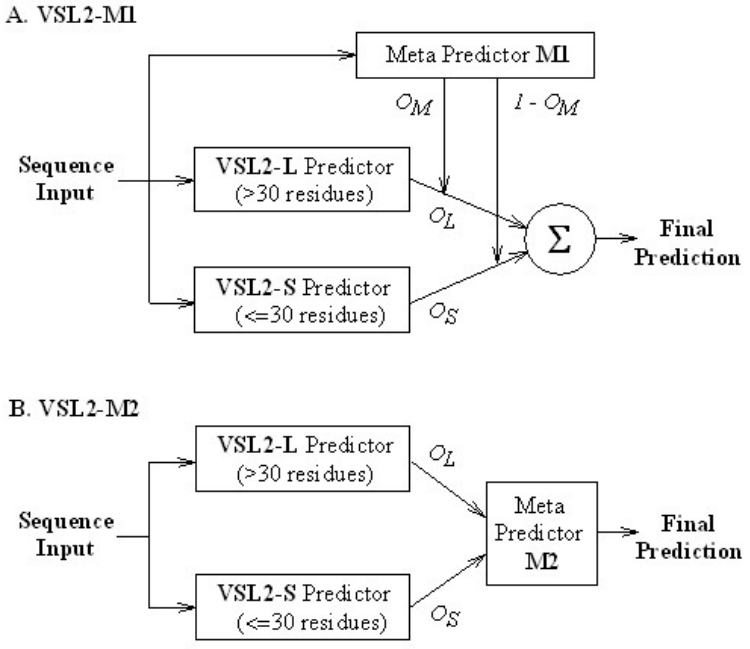


Figure 9. VSL2b predictor architectures (taken from [24])

2.6.1 IUPred predictor

IUPred assumes that globular proteins have larger numbers of effective inter-residue interactions (negative free energy) than disordered proteins due to the different types of amino acids involved in possible residue contacts. The core of IUPred is a method that enables the direct estimation of the interaction energies using the protein sequence alone. The estimated energy for each residue depends on the amino acid type but also on the amino acid composition in the neighbourhood. Generally, residues with less favourable predicted energies are more likely to be disordered [10].

The IUPred server takes a single amino acid sequence as an input and calculates the pairwise energy profile along the sequence. The energy values are then transformed into a probabilistic score ranging from 0 (complete order) to 1 (complete disorder). Residues with a score above 0.5 can be regarded as disordered. Optional is the prediction of long disorder, short disorder, and structured domains, each using slightly different parameters. The main profile is to predict context-independent global disorder that encompasses at least 30 consecutive residues of predicted disorder [23].

2.6.2 VSL2b predictor

VSL2b predictor is a combination of neural network predictors for both short and long disordered regions. It marks residues of length at least 30 as long disordered regions; otherwise regions are marked as short. Each individual predictor is trained by the dataset containing sequences of that specific length. The final prediction is a weighted average determined by a second layer predictor. VSL2b applies not only the sequence profile, but also the result of sequence alignments from PSI-blast and secondary structure prediction from PHD and PSI-pred.

2.6.3 IsUnstruct predictor

IsUnstruct is program based on the Ising model for prediction of disordered residues from protein sequence. IsUnstruct searches not only for disordered regions but also for individual disordered residues in a protein chain. It takes an amino acid sequence in the FASTA format as an input and calculates probabilities for each residue. A residue is considered as disordered if the probability is larger than 0.5. In IsUnstruct, the interaction term between neighbours has been replaced by a penalty for a state change (the energy of border). This allows applying dynamic programming to the Ising problem.

The energy of each residue in one state or the other depends on the type of residue in our model. To estimate the energy of any state we introduce the energy of the border between ordered and disordered residues and the energies of initiation of disordered state at the ends [12, 25]. The energy of the *j*-th state of a protein chain is calculated according the following formula:

$$E_j = \sum_{i=1}^{L} \omega(a_i, s_{ij}) + k_j . \omega_g + \delta_{N,j} . \omega_N + \delta_{C,j} . \omega_c$$

where a_i is the type of amino acid residue, s_{ij} describes the state of the i residue in the j conformation (1 in the case of disordered residue and 0 in the case of ordered state), ω_g is the energy of border, k_j is the number of borders between ordered and disordered residues in the j conformation, ω_N, ω_c are the energies of initiation of disordered state at the ends, and δ_{Nj} , $\delta_{C,j}$ are equal to 0 if the corresponding terminal residue is in the ordered state and to 1 in the opposite case, and L is the length of protein chain.

2.7 Model for determining region-characteristic n-grams in proteins

The basic idea for model construction is a very simple but effective: to combine the results of previously described methods. Using the n-gram analysis, repeat analysis and z-score technique we determine sets S_n, S_r, and S_z of n-grams which have, in some region, peak values (for example, the number of occurrences) either below or above

mean value of other n-grams. Additionally, set of n-grams S_{FD} is defined based on fractional difference. Finally, applying association rule mining methods on the sets S_n and S_{rz} (intersection of the sets S_r, and S_z), the additional set S_{AR} will be obtained. Appropriate quality depends on the following factors:

1) Confidence. Only rules that have confidence of at least 50% can provide support for determining n-grams that characterize regions in general. If intention is to find some n-grams that are close to "absolute" (>50%) confidence in the set of three possible values ('O', 'D', 'N) association rules with confidence lower than 50% can be searched in material[1].

2) Support. Only rules with sufficient support will be taken. Sufficient support depends on body-n-gram length and n-gram constituents and is equal to the probability of the n-gram occurrence. The initial probability ('weight') for monograms (individual AA) is equal to the probability of occurrence of AA in the analyzed material. Probability for single AA occurrence in some region(s) is[2]

$$w_{AA} = x_{AA} = \frac{n_{AA}}{reg\_len} \quad \text{where} \quad \sum_{i=1}^{20} w_{i(AA)} = 1$$

where $\mathbf{w}_{AA}$ denotes probability ('weight') of AA. In calculation probability for n-grams ($n \geq 2$) the model assumed that n-gram constituents are independent. Thus, probability for the *n*-gram of length *n* is equal to the product of probabilities of its monograms. For the *n*-gram *i* in some region the probability of its occurrence is

$$w_i = \prod_{j=1}^{n} w_{AAj}$$

where w_{AAj} denotes probability of the AA in the *j*-th position in the *n*-gram *i*. If for specific n-gram calculated probability is lower than support obtained from association rule mining where this n-gram occurs in the body of the rule, then such rule is preserved, otherwise rejected. Association rules selected in this

[1] This can be important for 'N' regions.
[2] Probability of occurrence some AA in region is equal to mole fraction of this AA (taken as a monogram) in this region.

process give information that some dependency between region type and n-gram in specific region exists.

3) Lift. Only rules with lift ≥ 1.05 or lift ≤ 0.95 are considered [20].

4) Only rules with 'unique' both left and right sides are considered. 'Unique' means that do not exist two or more rules with the same body that cover all types of regions. For example, none of the rules $ABC \rightarrow D$ and $ABC \rightarrow O$ is considered if both are suggested. Using threshold of 50% for confidence automatically reject all such rules.

5) Rules with body that is extension of the body of some other rule are rejected. For example, rule $ABC \rightarrow R$ is rejected if exist rule $B \rightarrow R$ with similar support, confidence and lift.

Sets S_{FD}, S_z and S_{AR} are determined for each type of region. Their intersection will give the set S which include n-grams that characterize regions type. N-grams are characteristic n-grams for such region type if they:

- are rare or frequent in this type of region (from FD)
- have very high confidence (from Z-score), and
- their statistically significant occurrence (from association rules mining or from statistically significant repeats) is connected only to specific type of region.

Although n-grams have been already used in research for finding some genome characteristics [26, 27, 28, 29], the presented approach is new and original and, according to available literature in the time of doing research described in this thesis, not previously used for determining characteristic regions in protein.

3 Material

Viral genomes material used in this research was downloaded from NCBI site: ftp://ftp.ncbi.nlm.nih.gov/genomes/Viruses/. Material includes amino acids and nucleotide sequences of viral proteins and, among others, taxonomic information. During research, different versions of data have been used. New versions commonly represent extension of the old ones with addition on several new genomes. Corrections of obtained results have been checked on sets of such new genomes. Results presented in this thesis are produced on data downloaded at January 2017.

After downloading, data were passed through the process of checking and cleansing. Incomplete and duplicate genomes and their proteins have been removed as well as individual proteins with some failure or incompatibility (for example, proteins with non-continual code, proteins with different length of amino acid and nucleotide codes, etc.). In order to eliminate influence of possible noise and outliers, classes with small number (<10) of genomes were eliminated from further processing. Finally the set of 190626 proteins is used as research material. Proteins are sourced from 4076 viruses which belong to 8 phyla and 31 different classes. Proteins in selected sets were coded with two translation tables[3]: 11 (190493 proteins) and 4 (133 proteins). As these translation tables differ only in TGA nucleotide triplet (coded as stop codon in translation table 11 and amino acid W (Tryptophan) in the translation table 4), and because only 583 W amino acid (coded with TGA or TGG) exists in the proteins that have translation table 4 (which is 0.09% of total occurrence of amino acid W in the dataset), all proteins in the dataset are considered as they have translation table 11.

For additional verification of obtained results the proteins from DisProt database (http://www.disprot.org/) have been used. Total of 803 proteins with 2167 disorder regions was used from DisProt (Version 7.03, September 2016). DisProt database includes proteins with experimentally verified disordered regions. Because there is no guarantee that the rest of the protein (not belongs to verified disordered region) is completely order, data from DisProt database can be used primarily for verifying

[3] https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi?

characteristics related to disordered regions[4], while verification related to ordered regions can be taken with some caution[5].

3.1 Determining threshold for n-grams

Observing individual n-gram mole fractions can not give satisfactory (strong) prediction of n-grams which characterize (dis)order regions. The reason is meaning of mole fraction (concentration of object, i.e. n-gram) which can not adequately cover object probability of occurrences and its uniqueness or majority. For example, hypothetically, if some n-gram **N** occurs once in ordered region(s) and ten times in disordered region(s), and if length of these disordered regions is 9.99 times larger than length of the ordered region, then mole fraction of **N** is lower in disordered than in ordered region. On the other side, this single occurrence of **N** in ordered region can have smaller probability than (single) occurrence o **N** in disordered region. The similar situation is with fractional difference (in previous example $\mathbf{FD(N)}_{D\_O}$ will be negative but close to zero, so it can be concluded that **N** can not characterize neither ordered nor disordered region). Additionally, large number of n-grams with small numbers of occurrences produce a noise that, although not affect the results, can make data mining process significantly slower. Parameters related to material used in the research are presented in Table 2 (number of AA n-grams) and Table 3 (number of regions, in material and in DisProt).

Table 2. Number of AA n-grams in material

N-gram length	Total number of AA n-grams in material	Number of unique AA n-grams in material	
		Present	Missing
1	46,413,638	20	0
2	46,223,012	400	0
3	46,032,386	8,000	0
4	45,841,760	159,988	12
5	45,651,134	2,886,848	313,152
6	45,460,508	17,821,832	46,178,168
7	45,269,882	27,652,049	1,252,347,951

[4] For example, in protein DP01070 (P42568) only positions 490-567 are annotated as disorder. On the other hands, all predictors listed in MobiDb (http://mobidb.bio.unipd.it/) recognized region 137-475 as disorder. This is in accordance with content of the region which includes mainly disorder promoting AAs, among others sequence of 42 consecutive Serine AAs. In previous version of DisProt database (up to version 6.0.2) some proteins include information about experimentally verified ordered regions, but in the new version (from 7.03) such explicit information are removed.

[5] Maybe some region of protein that is predicted (by predictor) as disorder is not annotated as disorder in current version of DisProt, and is count as order in verification process.

8	45,079,256	29,360,800	25,570,639,200
9	44,888,630	29,907,712	511,970,092,288
10	44,698,004	30,278,252	10,239,969,721,748

Table 3. Number of regions in material

For length >20 average protein length (AA) is shown in brackets

Region		Number of regions			
Type	Length	DisProt	IUPred-L	IsUnstruct	VSL2b
Disordered	1	1	230.615	20.424	60.776
	2		118.577	39.258	49.362
	3		67.201	47.393	60.373
	4		42.984	44.031	82.299
	5	31	31.454	40.605	104.052
	6	38	24.789	35.228	90.843
	7	40	18.406	31.011	70.737
	8	29	14.730	27.388	55.485
	9	26	12.197	24.160	42.595
	10	32	10.191	21.030	32.208
	11	38	8.595	19.158	27.730
	12	36	7.780	17.315	23.215
	13	19	6.936	15.847	19.227
	14	28	6.227	14.440	16.171
	15	27	5.472	13.197	13.640
	16	29	5.224	12.216	12.027
	17	23	4.367	10.901	10.527
	18	15	4.122	10.248	9.319
	19	15	3.908	9.241	7.847
	20	20	3.534	8.430	7.259
	>20	761 [114,59]	60.842 [47,77]	140.415 [44,86]	141.647 [53,83]
Ordered	1	27	140.217	12.015	30.914
	2	5	76.420	7.012	15.230
	3	8	45.108	5.947	10.497
	4	8	29.669	5.331	10.082
	5	5	23.450	5.435	11.158
	6	3	19.612	5.113	10.969
	7	10	15.517	4.876	11.318
	8	11	13.105	4.591	12.140
	9	7	11.163	4.557	12.232
	10	8	9.321	4.095	11.905
	11	5	8.538	4.090	12.497
	12	12	7.863	3.818	12.529
	13	8	7.489	3.882	11.763
	14	3	6.845	3.640	11.885
	15	7	6.203	3.719	12.226
	16	15	5.498	3.615	11.884
	17	9	4.898	3.631	11.253
	18	12	4.757	3.387	10.921
	19	14	4.408	3.408	10.327
	20	15	3.869	3.341	9.616
	>20	1.148 [272,06]	368.846 [105,72]	315.851 [112,86]	508.965 [60,69]

That problem can be alleviated by observing only those n-grams that appear (in disordered or ordered regions) more times than a predefined threshold. Threshold must be defined to eliminate n-grams with very small probability (i.e. can occur by chance). Also, threshold must not be too strong to eliminate n-grams that include possibly important information. Based on n-grams distribution show in Table 4, the following rule is used to define threshold: *All n-grams that appear once in the complete material will not be taken into account in the research*.

Although in literature [26, 27, 28, 29] was found that AA n-grams with length less than four can not be used to give precise characterization, because the threshold is weak, all monograms, bigrams, trigrams and almost all tetragrams will be used in the research, while the number of eliminated n-grams increase (up to 55% for n-grams with length 10) as increase their length. This is especially important for data mining, because it decrease the number of different objects (here n-grams) used in the mining process. Regardless those n-grams which appear exactly twice can also be considered as object with low probability that holds approximately 6% of the material for longer ones, they are not eliminated from the research. One of the reason was that such (pair of) n-grams represents direct non-complementary repeats. The same principle is also applied on nucleotide n-grams. Nucleotide n-grams are calculated from length 1 up to the length of 30 which corresponds to the AA n-grams of length 10. Also, nucleotide n-grams that appear only once are eliminated from research (percents are similar to the percents in the case of AA n-grams. For example, from initial 133712762 n-grams of length 30, after eliminating 91903558 n-grams that appear only once (about 69%) in research remain 41809204 n-grams).

Table 4. Threshold for AA n-grams and percentage of eliminated n-grams

N-gram length	Number of n-grams			Percentage related to total number of n-grams	
	Total	Appear once	Appear twice	N-grams that appear once	N-grams appear less than three times
1	46,413,638	0	0	0%	0%
2	46,223,012	0	0	0%	0%
3	46,032,386	0	0	0%	0%
4	45,841,760	33	61	0%	0%
5	45,651,134	280,445	237,269	0.61%	1.13%
6	45,460,508	9,256,115	3,531,199	20.36%	28.12%
7	45,269,882	20,955,158	3,519,377	46.28%	54.06%

8	45,079,256	23,357,355	3,155,572	51.81%	58.81%
9	44,888,630	24,082,091	3,093,361	53.64%	60.53%
10	44,698,004	24,575,343	3,060,423	54.98%	61.82%

3.2 Repeats and data mining

Because set of direct non-complementary repeats is equal to set of n-grams (that appear at least twice), by nature, and complementary repeats are not applicable to AAs, only inverse non-complementary repeats are determined for protein AA codes. For protein nucleotide codes inverse non-complementary, direct complementary and inverse non-complementary repeats are determined. Repeats are calculated in two versions - all repeats and statistically significant repeats.

For data mining application (classification) both amino acids and nucleotide n-grams and repeats were divided into two parts: model and test. Proteins from each phylum was divided related to their number and length in proportion belongs to [68, 72] interval for model and [28, 32] for test. In the cases where proteins could not be divided according to both criteria, a division with a weaker proportion ([65, 75] for model and [25, 35] for test) was used. Distribution of proteins over groups and their phyla are shown in Table A3 in Appendix.

Number of determined amino-acids and nucleotide repeats from the used material and amino-acids repeats from DisProt are shown on Table 5. Determination of nucleotide repeats started with length 6 which corresponds to 2 AAs. It is interesting that numbers of all repeats and statistically significant repeats are the same for the lengths greater than 7 for AA repeats and 15 for nucleotide repeats (16 for *in* nucleotide repeats). Direct complementary repeats were not determined because they are included in the set of already determined n-grams.

Table 5. Determined amino-acid and nucleotide repeats

Legend: in - inverse non-complementary repeats
ic - inverse complementary repeats
dc - direct complementary repeats
all - all repeats
ssr - statistically significant repeats

Repeat length		Amino acids repeats		Nucleotide repeats					
		in		dc		ic		in	
		all	ssr	all	ssr	all	ssr	all	ssr
2	disprot	3,734,384	3,536,416						
	model	37,582,276	24,616,388						
	test	16,706,373	10,903,875						
3	disprot	327,782	272,105						
	model	4,523,944	2,620,175						
	test	2,005,362	1,158,337						
4	disprot	27,926	25,326						
	model	377,681	322,862						
	test	164,839	141,687						
5	disprot	6,387	6,008						
	model	195,007	154,603						
	test	86,328	68,538						
6	model	24,500	24,412	7,536,950	3,963,612	8,639,587	4,737,143	9,220,206	5,302,808
	test	10,231	10,185	7,536,950	3,963,612	8,639,587	4,737,143	9,220,206	5,302,808
	disprot	985	982						
7	model	22,884	22,817	2,108,140	1,200,145	2,324,517	1,480,539	3,098,382	1,820,441
	test	8,966	8,937	2,108,140	1,200,145	2,324,517	1,480,539	3,098,382	1,820,441
	disprot	604	604						
8	model	5,254	5,254	594,551	397,164	807,787	556,359	885,763	638,963
	test	2,021	2,021	594,551	397,164	807,787	556,359	885,763	638,963
	disprot	196	196						
9	model	4,464	4,464	338,002	232,807	394,544	312,527	789,931	510,954
	test	1,761	1,761	165,073	118,026	191,207	156,137	380,354	261,928
	disprot	218	218						
10	model	1,285	1,285	96,345	73,336	194,610	132,360	242,117	180,646
	test	525	525	48,652	38,456	94,035	66,837	117,226	90,903
	disprot	74	74						
11	model	8,314	8,314	28,645	24,284	34,828	30,926	143,070	98,384
	test	3,033	3,033	14,566	12,525	17,153	15,488	67,466	48,235
	disprot	612	612						
12	model			8,527	7,853	32,305	27,267	44,917	39,507
	test			4,291	4,014	15,863	13,577	21,928	19,607
13	model			2,533	2,470	3,410	3,362	36,610	32,984
	test			1,314	1,278	1,691	1,662	17,538	15,886
14	model			791	788	8,093	7,962	11,483	11,329
	test			430	428	3,680	3,607	5,088	5,020
15	model			262	262	415	415	10,761	10,655
	test			127	127	226	226	4,722	4,672
16	model			78	78	2,179	2,179	3,739	3,739
	test			40	40	1,108	1,108	1,629	1,629
17	model			34	34	105	105	3,459	3,459
	test			13	13	33	33	1,536	1,536

18	model		12	12	743	743	1,159	1,159
	test		9	9	367	367	773	773
19	model		12	12	17	17	1,187	1,187
	test		7	7	8	8	588	588
20	model		1	1	329	329	589	589
	test		2	2	146	146	202	202
21	model		1	1	4	4	510	510
	test		6	6	9	9	246	246
22	model				123	123	191	191
	test				55	55	91	91
23	model		2	2	12	12	320	320
	test		1	1	5	5	133	133
>23	model		3	3	170	170	1,258	1,258
	test				91	91	485	485

4 Results

Results will be presented for each of the previously described methods and their combinations, and compared with corresponding DisProt data.

4.1 Mole fractions

4.1.1 Mole fractions of AA n-grams

It is not expected that mole fractions (especially for longer n-grams) can be used for finding n-grams that characterize either order or disorder regions. But, mole fractions can be good markers for compatibility of material used in research with material in DisProt version 7.03. Comparison of mole fractions for monograms in complete material and DisProt is presented on Figure 10.

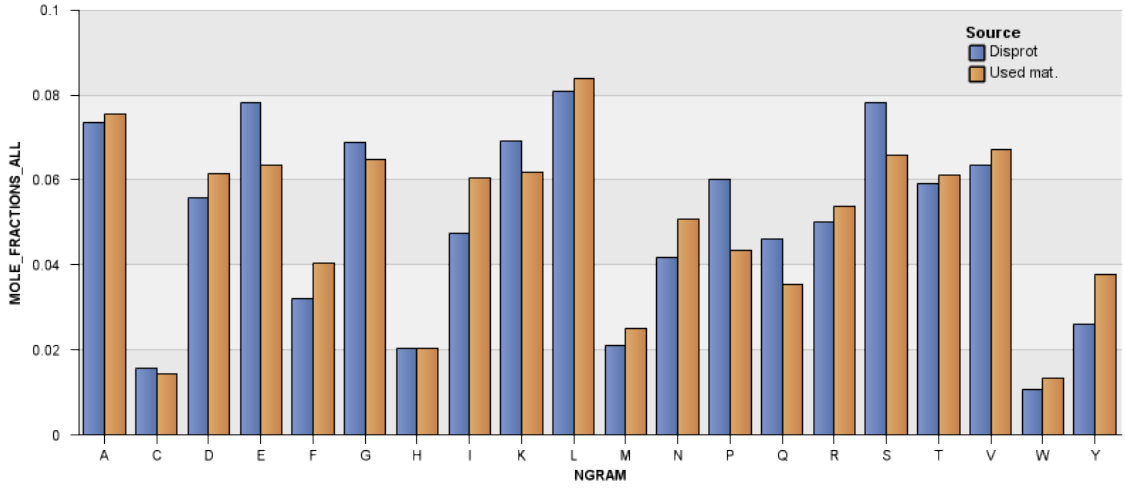


Figure 10. Comparison of mole fraction in used material and DisProt database V7.03

In general, most amino acids have similar mole fractions in both series. Larger differences exists for amino Glutamic acid (E), Proline (P) and Serine (S) (higher level in DisProt), and Phenylalanine (F), Isoleucine (I), Asparagine (N) and Tyrosine (Y) (higher level in our material). Number of significant differences became even smaller if mole fractions are compared separately in predicted disordered with experimentally found disordered regions from DisProt (Figure 11), and in predicted ordered regions with non-disordered regions from DisProt (Figure 12).

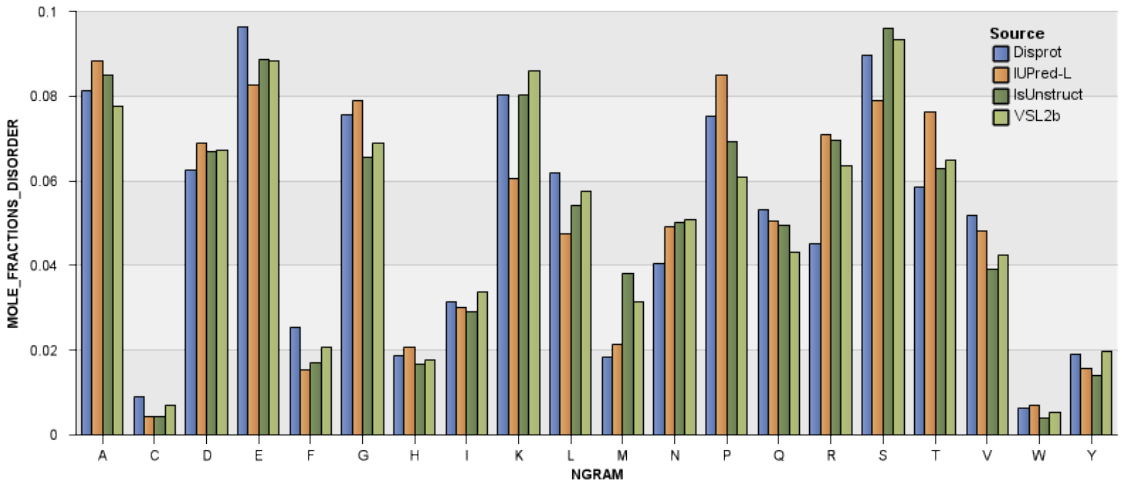


Figure 11. Comparison of mole fraction in disordered regions of used material (predicted by disorder predictors) and disordered regions from DisProt database

Different disorder predictors predict different regions and consequently have different mole fractions for individual AAs. But, from the Figures 11 and 12 it is evident that content of AAs in the predicted regions (later used in the research) have very similar behaviour (related to dis/ordered regions) to material in DisProt.

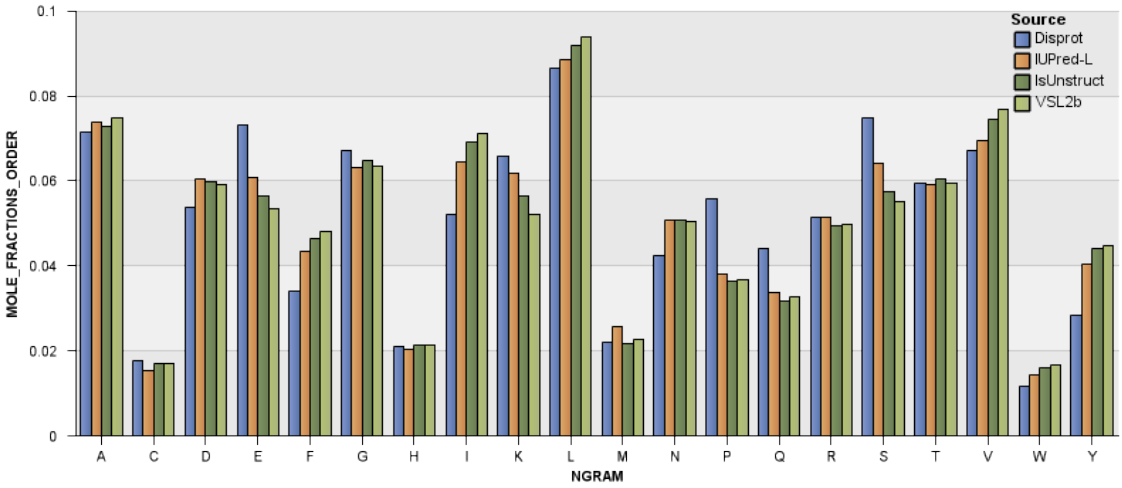


Figure 12. Comparison of mole fraction in ordered regions of used material (predicted by disorder predictors) and non-disordered regions from DisProt database

4.1.2 Mole fractions of nucleotide n-grams

Because nucleotide sequences are three times longer then corresponding amino-acid sequences and nucleotide n-gram can start at arbitrary positions, set of nucleotide

n-grams not completely correspond to the set of amino-acids n-grams. More precisely, only a third of nucleotide n-grams have their equivalent amino acid n-grams. Mole fractions of individual n-grams also depend on codon usage. For this reason, in some analysis, set of nucleotide n-grams is divided in three parts, according to their starting positions (i.e. relative offset to a closest codon starting position, takes a value from 0 to 2). Because of these dependences which can not guarantee correctness in the case of generalization to other material, the obtained mole fractions were used just as a method for sorting n-grams according to their abundance in the material. Nevertheless, some interesting information related to nucleotide n-grams mole fractions were found in the material. Mole fractions related to monograms for all material and grouped according their starting positions ("ORF", *open reading frame*) are shown in Table 6.

Table 6. Mole fractions of nucleotide monograms in all material and grouped according to their starting positions

ORF	n-gram	Mole fractions all	Mole fractions order	Mole fractions disorder
all	A	0.286286328169	0.279469194991	0.311333252030
	C	0.224269613742	0.217597454352	0.248783887397
	G	0.244571161031	0.239782427068	0.262165515114
	T	0.244871597151	0.263149516567	0.177716439107
1	A	0.296621113820	0.292812695421	0.310613676849
	C	0.190309343990	0.182180318583	0.220176309346
	G	0.332583474710	0.328734266434	0.346725904394
	T	0.180485787388	0.196272390646	0.122484008703
2	A	0.330734492305	0.321099981484	0.366132782567
	C	0.224520517008	0.209108422513	0.281146307835
	G	0.167926784795	0.165927007687	0.175274193692
	T	0.276818205890	0.303864588314	0.177446715904
3	A	0.231503378382	0.224494908069	0.257253296673
	C	0.257978980229	0.261503621959	0.245029045009
	G	0.233203223586	0.224686007081	0.264496447256
	T	0.277310798175	0.289311570740	0.233218592713

We can observe some characteristics of the n-grams:

- There are only 180 non-ACGT nucleotides in complete material (mole fractions 0.000001292720), so such nucleotide codes were not considered as separate group
- Percentage of GC nucleotides is almost half (51.09%) in complete material
- N-grams belong to ORF=3 set (i.e. are on the third position in the AA codon) have similar GC percent (50.95%); n-grams that belong to ORF=1 (first nucleotide in AA codon) are richer (56.69%) while nucleotides in the middle of AA codons are poorer (45.64%) in GC nucleotides

- As expected, n-grams that correspond to amino-acids homorepeats occur also in the nucleotide level. But, because individual amino acids have different corresponding codons at nucleotide level, the corresponding nucleotide repeats are not necessary homorepeats of appropriate trigrams. An interesting observation is that nucleotide sequences that are homorepeats or include homorepeats occur more often than sequences that are random sequences of (codon) trigrams. For example, on amino-acid level hexagram 'PPPPPP' occurs 1145 times in disorder regions. Corresponding nucleotide n-grams (for ORF=1) occurs in 285 variations; among them the most numerous groups are 'pure' homorepeats 'CCACCACCACCACCACCA' and 'CCGCCGCCGCCGCCGCCG' (each occurs 29 times), followed by tandem repeats ('CCTCCACCTCCACCTCCA' - 18 times and 'CCACCTCCACCTCCACCT'- -16 times, 'CCACCACCTCCACCACCT' - 8 times, 'CCGCCGCCACCGCCGCCG' - 8 times, 'CCACCACCTCCACCACCT' - 8 times, etc. This diversity of AAs translation into codons gives additional opportunity to more precisely describe characteristic n-grams.

4.2 Fractional difference

4.2.1 Fractional differences of AA n-grams

Fractional difference of some n-grams indicates their richer or poorer concentration in disordered or ordered regions. Fractional difference *disorder/order* of the n-gram N with length *n* (in the rest of the text FD_n(d_o, N) is positive if disorder region is richer of this n-gram, and negative if disorder region is poorer in this n-gram (i.e. order region is richer). If fractional difference *disorder/order* of some n-gram is positive, than this n-gram characterize disordered region; as opposite it characterize ordered region. Fractional difference for monograms is shown on Figure 13 together

with fractional differences of the monograms in DisProt database (version V7.03, September 2016).

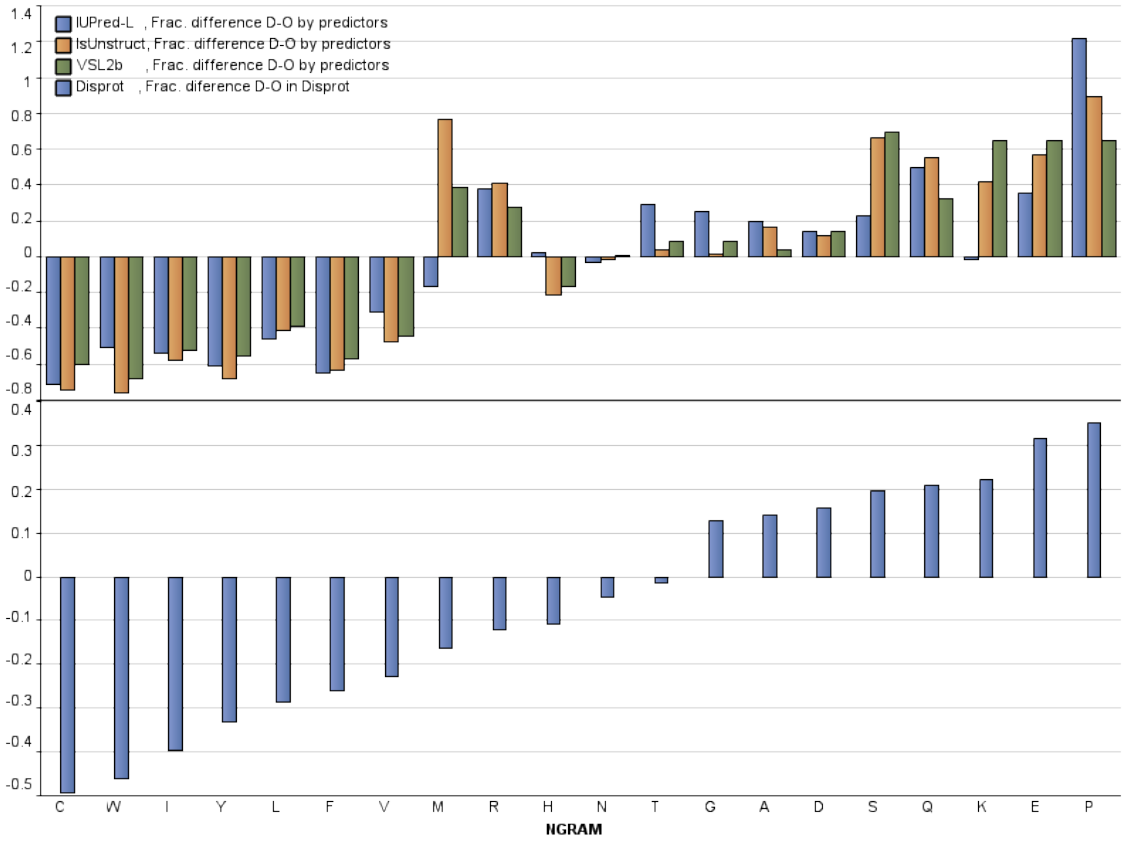


Figure 13. Comparison of *disorder/order* fractional difference of monograms from material used in research and material from DisProt database

There are some differences in predicted FD compared to the FD in DisProt material (VSL2b and IUPred-L have 4 and IsUnstruct 3 differences). FD1(d_o,T) from DisProt and FD1(d_o,T) predicted by IsUnstruct are very close to zero but with opposite signs, and such variations are expected. A little bit larger incompatibilities are for AAs Metionine (M) and Arginine (R) that are also close to the transition from positive to negative. It must be noticed that FD values for current version of DisProt differs from version 3.4 (shown of Figure 5) both in order of AAs related to FD and in orientation. For example, AAs M and R are disorder oriented in version 3.4, but order oriented in DisProt version 7.03. The reason can be the different percentage of AAs in ordered/disordered regions in later added proteins to DisProt database. Also the results from predictors better fits to FD in DisProt 3.6 than in 7.03. The probable reason for

that can be that the set of proteins used for predictors training which is more similar to set in version 3.6 than in 7.03.

From previous figures it can be seen that all three predictors produce similar results. For this reason, even though the calculation was done for all three predictors, in the further text the most of the results will be illustrated only for IsUnstruct predictor, while the results or other two will be presented only if there are large differences in results.

Number of n-grams that have positive FD(d_o) is shown on Table 7. Observing that only fractional differences can not give precise characterization of disordered or ordered regions because numbers are too high. Even if consider only those n-grams that occur only in disordered regions their number remains too high. Appendix table A4 contains list of some n-grams that occur only in disordered regions predicted by IsUnstruct predictor.

Table 7. Number of n-grams with positive FD(d_o) regions and their percentage in the sample

N-gram length	Number of grams with positive FD(d_o)			Total number of n-grams	Percent of positive n-grams		
	VSL2b	IsUnstruct	IUPred-L		VSL2b	IsUnstruct	IUPred-L
1	12	11	10	20	60.00%	55.00%	50.00%
2	160	156	154	400	40.00%	39.00%	38.50%
3	2,702	2,643	2,482	8,000	33.77%	33.03%	31.02%
4	51,234	48,278	45,588	159,955	32.03%	30.18%	28.50%
5	900,484	795,388	637,986	2,606,403	34.54%	30.51%	24.47%
6	3,064,465	2,427,476	1,436,112	8,565,717	35.77%	28.33%	16.76%
7	2,286,561	1,623,033	1,023,642	6,696,891	34.14%	24.23%	15.28%
8	2,076,342	1,416,707	948,566	6,003,445	34.58%	23.59%	15.80%
9	2,082,655	1,393,438	960,963	5,825,621	35.74%	23.91%	16.49%
10	2,111,453	1,391,255	982,448	5,702,909	37.02%	24.39%	17.22%

One criterion for selecting "better" n-grams (the ones that not only appear in disordered regions, but also have positive fractional difference in disordered regions), can be the position of such n-grams in the list of n-grams ordered according to their mole fractions in descending order. For each n-gram length, first 100 n-grams with the highest mole fractions are shown in Appendix Table A5. It is interesting that among the n-grams that occur only in predicted disordered regions the most of the n-grams include some kind of homorepeats [30], either partial or full (for example EEEEG, PPPSPPPS,

SSSSSSS, etc), or a repeat structure (for example PAPAPA). These repeat structures also can be found in many of the n-grams that prefer disordered regions (for example EEE, DDD, PPPSP, etc., see Table A5) where such n-grams are included in longer n-grams that appear only in disordered regions. Similar tables are presented for characteristic n-grams in ordered regions (Appendix tables A6 and A7) and for borders between ordered and disordered regions (Appendix tables A8 an A9). Characteristic n-gram of smaller length are combined with order promoted AAs (for example WIC, CYW, LCYL, VLYV, etc.) or rudimentary (homo)repeats (like YYVV or ILILL) but for longer n-grams no clear pattern can be observed except that trigram LLL appears as a part of various longer n-grams. Although there are a lot of n-grams in Tables A8 and A9, no clear pattern for border n-grams can be observed. Because of the huge number of n-grams in previously described sets (for example, set of n-grams of any length that appear only in disordered regions have cardinality of 3.5M, while set of n-grams that have positive disorder fractional difference and appear not only in disordered regions have cardinality of 2.7M), additional restriction can be provided by increasing threshold for eliminating n-grams accepting the rule that n-grams with very small mole fractions will be removed from sets. This procedure will be used for sets produced as combination of different approaches.

As a verification of the method of using mole fraction and fractional differences for determining characteristic n-grams, a comparison with fractional differences of identical n-grams available from DisProt proteins was performed. The comparison results are presented in Table 8. Percentage of identical n-grams that belong to the same type of region grows up to 99.38% as n-gram length increase. Additional information about widespread of n-grams over proteins shows that the most of the n-grams appear in different proteins and in proteomes of different phyla and classes of viruses. different classes of viruses. For example, n-gram GGGGGGG belongs to 450 different proteins in material (from 3 phyla and 12 classes), and to 5 proteins from DisProt, n-gram GGGSGGG to 70 proteins (3 phyla, 9 classes) in material and 4 proteins from DisProt, etc.

4.2.2 Fractional differences of nucleotide n-grams

Nucleotide n-grams can also be ordered by the concentration in disordered or ordered regions. Fractional differences of nucleotide monograms divided into three sets

according to their starting positions ("ORF" on figure) are shown on Figure 14. Concentration is almost uniform regardless of ORF-s: nucleotide T has larger concentration in ordered regions while other nucleotides have large concentration in disordered regions with exception of C in ORF3.

Table 8. Number of matched regions according to fractional difference of AA n-grams that appear in predicted regions and regions from DisProt database.
Number of equal - number of n-grams available both in materials used in research and in DisProt
Number of matched - number of n-grams that belongs to the same type of region (comparing FD related to predicted regions and FD related to regions from DisProt)
Number of non-matched - number of n-grams that belongs to the opposite type of region (comparing FD related to predicted regions and FD related to regions from DisProt)
Matched/number of equal - percent of n-grams with matched FD related to total number of n-grams

N-gram length	Number of equal	Number of matched	Number of non-matched	matched/number of equal
1	20	17	3	85.00%
2	400	344	56	86.00%
3	7,213	5,276	1,937	73.14%
4	45,224	24,118	21,106	53.33%
5	61,346	42,455	18,891	69.20%
6	20,705	19,273	1,432	93.08%
7	2,950	2,910	40	98.64%
8	1,140	1,131	9	99.21%
9	799	794	5	99.37%
10	648	644	4	99.38%

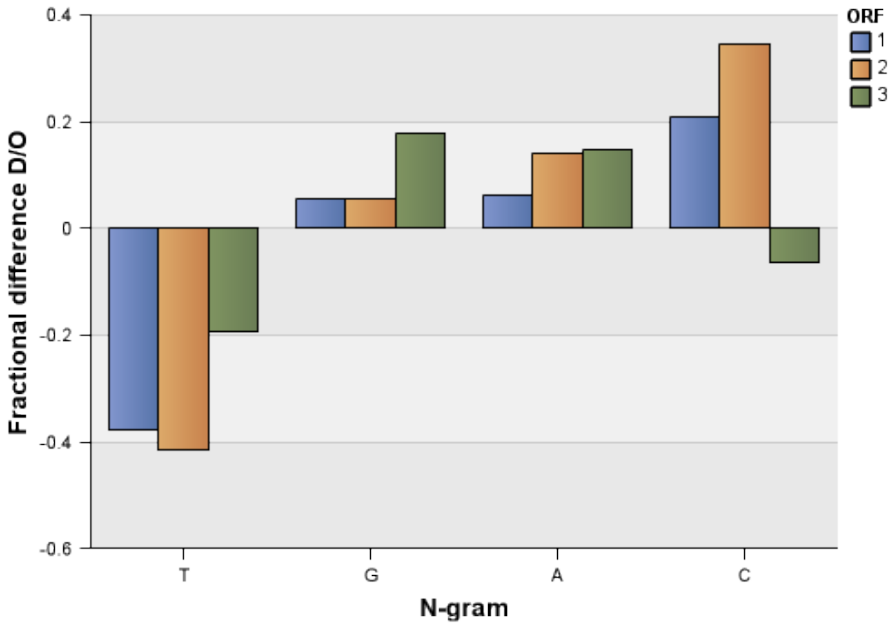


Figure 14. Fractional differences of nucleotide monograms grouped according to their starting positions. Ordered/disordered regions are predicted using IsUnstruct predictor.

Fractional differences of nucleotide trigrams can be used for determining if there are any difference related to ordered/disordered regions and AA codon usage. Figure 15 presents fractional differences of nucleotide trigrams ordered according AA codon usage (related to translation table 11). The most interesting are values for ORF1. Such differences exist for some AAs but these results are predictor depending and can not be generalized without further verification. For example, depending on predictor used order/disorder codon dependencies are

- VSL2b: amino acid A:
  - GCT-order, all other codons - disorder
- IsUnstruct (shown on Figure 15):
  - amino acid G: GGG-order, all other codons - disorder
  - amino acid N: AAC-disorder AAT-order
  - amino acid T: ACT-order, all other codons - disorder
- IUPred-L:
  - amino acid H: CAC-disorder CAT-order
  - amino acid K: AAA-order, AAG-disorder
  - amino acid N: AAC-disorder AAT-order

4.3 Z-score

Z-score value is used as an additional confirmation if specific n-gram characterize ordered or disordered region. Z-score is calculated only for n-grams in disordered or ordered regions. It is not possible to calculate it for n-grams in N (border) regions because there is not guarantee for n-gram in border region that all its sub-n-grams also belong to the border region (which is necessary for z-score calculation). Also, some n-grams and their sub-n-grams can occur many times in proteins in the same type of region, but only once in any of (individual) proteins. Such n-grams have z-score equal to zero and do not satisfy any confidence level, despite they are potential markers for some type of region.

The most restrictive criterion for z-values is chosen by selecting n-grams with confidence level 99% (see Table 1). The selection algorithm for n-gram *N* and region with type *R* (ordered, disordered) can be illustrated with the following pseudocode:

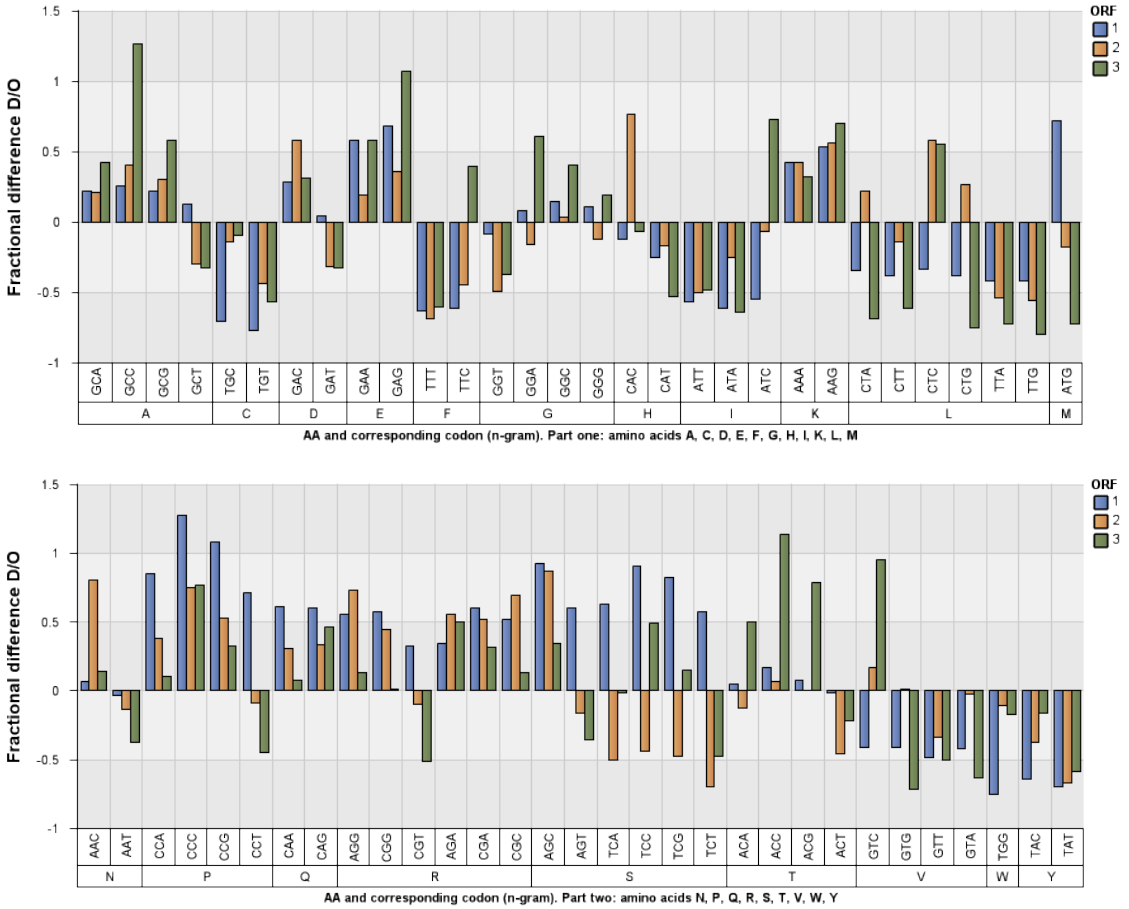


Figure 15. Fractional differences of nucleotide trigrams. Part one: amino acids A, C, D, E, F, G, H, I, K, L, M. Pat two: amino acids N, P, Q, R, S, T, V, W, Y. Ordered/disordered regions are predicted with IsUnstruct predictor.

```
for each n-gram N an region type R
    if exist z-score for N in type R regions
       then if abs(z-score)>2.58
               then if exist z-score for N in opposite-type regions
                       then if abs(z-score_opposite-type)<1.65
                               then N characterize R type regions
                               else without-characterization
                       else without-characterization
               else N characterize R type regions (exclusive)
       else without-characterization
```

Applying the previous algorithm, n-grams that characterize only one type of region (requirement abs(z-score)>2.58) but not the opposite (requirement abs(z-score)<1.65) were selected. N-grams that characterize specific regions are shown in

Appendix tables A10 (ordered regions) and A11 (disordered regions). In both tables n-gram patterns with similar structures (homorepeats and repeats) can be seen for disordered and ordered regions. As in the previous methods, for both ordered and disordered regions, numbers of selected n-grams have peak for length 6 and 7, and decrease as n-gram length increase or decrease (Table 9).

Table 9. Number of selected characteristic n-grams based on z-score values

N-gram		
length	number /disordered regions	number /ordered regions
3	568	1532
4	12789	31699
5	126827	330136
6	641019	2215554
7	406744	3121903
8	61952	567521
9	8638	54022
10	2374	13331

4.4 Combination of fractional difference, z-score and mole fractions

4.4.1 Combination of Fractional difference and Mole fractions for AA n-grams

Numbers of significant n-grams decrease in a very small percent (about 2.65%) if z-score method combined with method based on fractional difference. More significant reduction is obtained if combination includes fractional difference, z-score and n-grams with mole fractions larger than specific value (which is increasing threshold level, see section 3.1). Percentages of decreasing n-gram numbers depending of mole fractions are shown in Table 10.

Number of n-grams that characterize ordered regions is reduced much faster than number of n-grams that characterize disordered ones. This is caused by average number and standard deviation of n-gram occurrences which is both between 2 and 3 for n-gram length>5, but with lower average number and higher standard deviation of n-gram occurrences in disordered compared to ordered regions, which is especially emphasized

for n-gram lengths 6 and 7. Related to number of n-grams in Table 10 middle (but satisfactory) level of reducing is obtained by taking condition "*mole fraction>1E-6*". N-grams that satisfy this condition are shown in Appendix Tables A12 (ordered regions) and A13 (disordered regions). Among the n-grams that characterize ordered regions, the same pattern as in the previous tables is observed (for example 'LLL'), but for disordered regions patterns are more uniform than in previous cases. On the top of the list for all n-gram lengths are homorepeats ('QQQ', 'SSSS', 'GGGG', 'PPPPPP', 'EEEEEEE', etc.), tandem repeats ('APAP', 'SRSRSR', 'PEPEPE', 'AATTTAATTT', etc.) or palindromes ('APAPA', 'SDSDSDS', 'PKPAPKP', 'DEDDEDDED', etc.) or their shorter versions of disorder promoting AAs (see Figure 13) combined with some other AAs.

Table 10. Number of n-grams and percentage of initial n-grams for different mole fractions used

Initial n-gram number - n-grams that satisfy fractional difference and z-score conditions

Mole fractions >			5E-6		1E-6		5E-7		1E-7	
Type	N-gram length	Initial n-gram number	n-gram number	Percent of initial	n-gram number	Percent of initial	n-gram number	Percent of initial	n-gram number	Percent of initial
Disorder	3	242	242	100.00	242	100.00	242	100.00	242	100.00
	4	7176	5678	79.1248	7102	98.9687	7163	99.8188	7176	100.00
	5	108560	4471	4.1184	55374	51.0077	79551	73.2783	108560	100.00
	6	635966	813	0.1278	73993	11.6347	216916	34.1081	635966	100.00
	7	406634	423	0.1040	63583	15.6364	149604	36.7908	406634	100.00
	8	61939	166	0.2680	20641	33.3247	36584	59.0645	61939	100.00
	9	8633	89	1.0309	3590	41.5846	5822	67.4388	8633	100.00
	10	2371	46	1.9401	1147	48.3762	1741	73.4289	2371	100.00
Order	3	1218	1216	99.8357	1218	100.00	1218	100.00	1218	100.00
	4	24662	10579	42.8959	22100	89.6115	23789	96.4601	24620	99.8296
	5	281140	333	0.1184	59787	21.2659	140114	49.8378	265399	94.4010
	6	2144995	126	0.0058	7860	0.3664	78970	3.6815	1205359	56.1940
	7	3081651	104	0.0033	4096	0.1329	45270	1.4690	1114921	36.1793
	8	563499	33	0.0058	1484	0.2633	23994	4.2580	333600	59.2015
	9	53794	12	0.0223	289	0.5372	3600	6.6921	40082	74.5101
	10	13283	2	0.0150	103	0.7754	992	7.4681	9507	71.5726

N-grams determined under these conditions can be compared with n-grams generated from DisProt database proteins. The percents of agreement of predicted characteristic n-grams with corresponding n-grams in disordered and ordered regions are shown on Figure 16. Shorter n-grams (length<5) more precisely characterize ordered regions than disordered. For disordered regions, longer n-grams agreed with n-grams from DisProt

database in high (9-grams) or very high (other n-grams, n>4, n≠9) percent; for ordered regions pentagrams agreed in high percent while longer n-grams agreed in very high percent.

Based on these results, it is expected that also the data mining analysis will confirm that regions are more precisely characterized with longer n-grams. This expectation is in compliance with the results presented in the next chapter. Also, because using z-score values excludes set of n-grams that characterize border regions, for border regions the final results will be produced by intersecting sets obtained with fractional difference and mole fractions methods with set of n-grams produced with data mining.

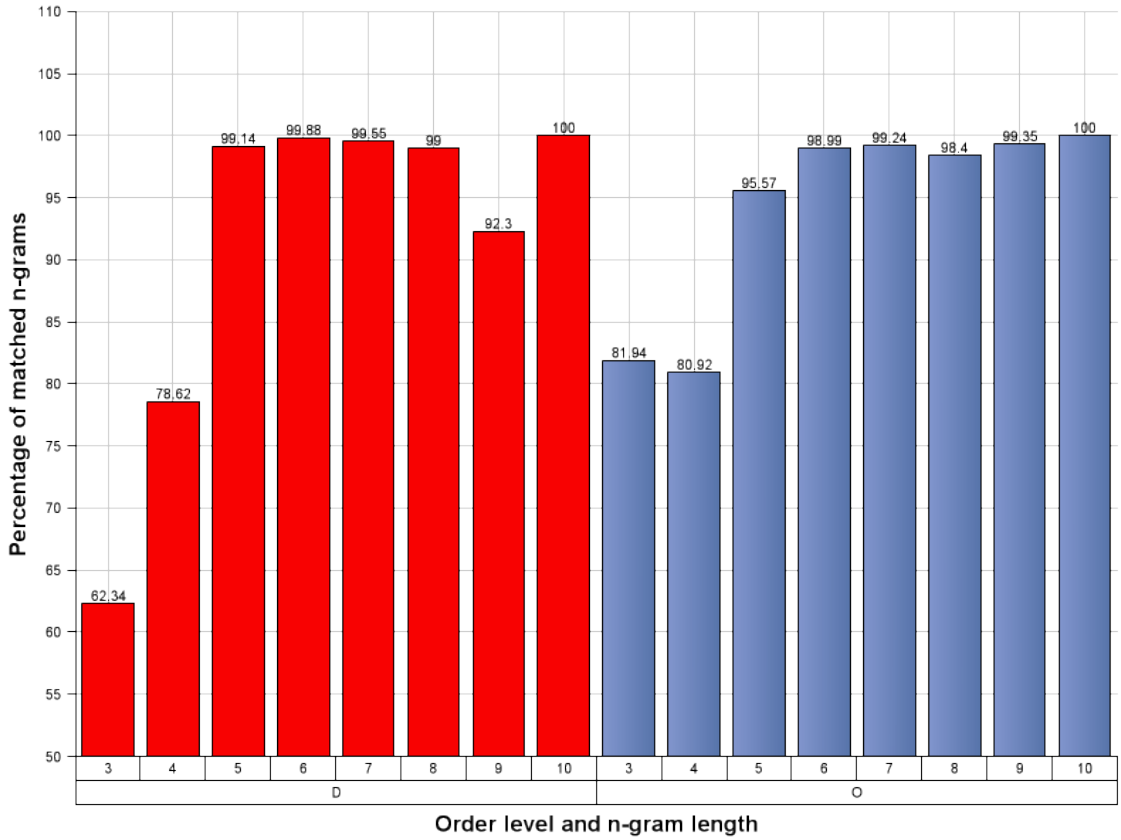


Figure 16. Agreement in characterization regions with identical n-grams from used material and DisProt database. D - disordered regions; O - ordered regions. Ordered/disordered regions are predicted with IsUnstruct predictor.

4.4.2 Combination of Fractional difference and Mole fractions for nucleotide n-grams

Z-score values were not calculated for nucleotide n-grams. In addition to the previously mentioned disadvantage, it is not possible to calculate z-scores for nucleotide n-grams divided into ORF groups because sub-n-grams (necessary for z-score calculation) belong to different ORF group. If methods that include combination of fractional difference and mole fractions are applied on nucleotide n-grams, some interesting facts can be observed:

- Percentages of retain (initial) n-grams are comparable for all ORF-s; also number of n-grams have the same order of magnitude in all ORF-s for the same mole fractions restriction level. Table 11 presents how the number of n-grams is related to increasing mole fractions for ORF=1[6].
- Number of order related n-grams decrease more rapidly compared to disorder ones, as mole fraction increase, regardless their number significantly exceed number of disordered related n-grams. This leads to conclusion that longer order related n-grams have a smaller cardinality of occurrences than longer disorder related n-grams, i.e. for all n-grams lengths exists some disordered related n-grams with sufficient number of occurrences that with high probability can be considered as markers for disordered regions.
- For each n-gram lengths some significant nucleotide n-grams exist in different ORF-s. Some of these n-grams (with length divided by 3) corresponds to AAs n-grams that are also among significant ones (for example, `'GGTCAGCACATTTCCATCCGA'` with corresponding AA n-gram `'GQHISIR'`, `'AATCCAGCTCCGACGTCAAGTCCT'` which correspond `'NPAPTSSP'`, etc).

[6] Table not include some n-grams lengths necessary to demonstrate the trend of decreasing percents of retained material. Maximum n-gram length is equal 30 whih corresponds to AA n-grams with length 10.

Table 11. Number of nucleotide n-grams and percent of retain initial n-grams for different mole fractions used. N-grams belong to ORF=1 i.e. start on position correspond to AAs n-grams

Mole fractions >			5E-6		1E-6		5E-7		1E-7	
Type	N-gram length	Initial n-gram number	n-gram number	Percent of initial	n-gram number	Percent of initial	n-gram number	Percent of initial	n-gram number	Percent of initial
Disorder	1	3	3	100.00	3	100.00	3	100.00	3	100.00
	2	10	10	100.00	10	100.00	10	100.0	10	100.00
	3	37	36	97,2972	36	97,2972	37	100.00	37	100.0
	7	6.556	6.533	99,6491	6.533	99,6491	6.533	99,6491	6.553	99,9542
	8	24.842	23.968	96,4817	24.805	99,851	24.806	99,855	24.836	99,9758
	9	99.374	49.350	49,6608	96.335	96,9418	98.963	99,5864	99.361	99,9869
	13	3.913.917	662	0,0169	88.176	2,2528	340.739	8,7058	3.455.361	88,2839
	14	3.196.310	402	0,0125	70.169	2,1953	270.353	8,4582	2.721.386	85,1414
	15	2.320.596	254	0,0109	60.434	2,6042	229.094	9,8722	1.913.555	82,4596
	16	1.831.777	192	0,0104	53.916	2,9433	200.902	10,9676	1.377.924	75,2233
	17	1.645.754	169	0,0102	52.923	3,2157	195.910	11,9039	1.208.295	73,4189
	18	1.560.665	127	0,0081	63.902	4,0945	186.066	11,9222	1.130.913	72,4635
	19	1.545.559	106	0,0068	59.336	3,8391	172.069	11,1331	1.043.707	67,5294
	20	1.535.807	105	0,0068	58.830	3,8305	170.658	11,1119	1.034.296	67,3454
	21	1.513.175	97	0,0064	56.494	3,7334	164.093	10,8442	1.015.957	67,1407
	24	1.500.663	79	0,0052	50.556	3,3689	146.869	9,7869	941.656	62,7493
	27	1.492.550	182	0,0121	60.052	4,0234	240.903	16,1403	877.990	58,8248
	30	1.485.934	171	0,0115	54.535	3,67	219.899	14,7987	821.506	55,2854
Order	1	1	1	100.00	1	100.00	1	100.00	1	100.00
	2	6	6	100.00	6	100.00	6	100.00	6	100.00
	3	26	26	100.00	26	100.00	26	100.00	26	100.00
	7	8.597	8.351	97,1385	8.351	97,1385	8.352	97,1501	8.420	97,9411
	8	35.159	31.692	90,139	34.728	98,7741	34.730	98,7798	34.771	98,8964
	9	128.503	44.411	34,5602	120.629	93,8725	126.937	98,7813	127.696	99,3719
	13	10.246.752	4	0	4.171	0,0407	60.288	0,5883	1.880.747	18,3545
	14	8.724.060	4	0	1.097	0,0125	28.198	0,3232	1.330.687	15,253
	15	6.289.576	1	0	539	0,0085	18.155	0,2886	1.020.736	16,229
	16	4.930.514	1	0	401	0,0081	14.973	0,3036	900.736	18,2686
	17	4.444.240	1	0	362	0,0081	14.253	0,3207	868.765	19,5481
	18	4.169.242	--	--	238	0,0057	12.408	0,2976	808.236	19,3856
	19	4.042.036	--	--	200	0,0049	11.778	0,2913	785.001	19,4209
	20	4.012.001	--	--	182	0,0045	11.640	0,2901	779.692	19,4339
	21	3.906.581	--	--	147	0,0037	18.179	0,4653	739.870	18,939
	24	3.742.557	--	--	87	0,0023	16.069	0,4293	688.272	18,3904
	27	3.600.261	--	--	49	0,0013	14.361	0,3988	644.961	17,9142
	30	3.472.107	--	--	39	0,0011	12.943	0,3727	607.449	17,4951

4.5 Data mining

Previously described sets of n-grams and repeats were used as the input to Data Mining process. Two different data mining techniques were applied: association rules and classification. In process of determining association rules, the complete set of n-grams (repeats) is used as input. In classification process, data were divided into two subsets: model and test (see section 3.2). Classification models were built using model subset as input and verified on test subset. For both techniques results were obtained using IBM Intelligent miner [31].

4.5.1 Association rules

Association rules were obtained using SIDE (Simultaneous Depth-first Expansion) algorithm [32] with the following parameters: confidence>=51%, support>=0.0001 and lift >=1.05 or lift <=0.95. Association rules were obtained for each n-gram or repeat length from 2 to 10. Typical result produced by Intelligent miner is shown on Figure 17 and includes association rules, rule support, confidence, lift, absolute support (number of n-grams that satisfy rule), rule body, rule head, number of items in rule body and rule head, group (rules having head 'ORDER_LEVEL_IU='D' belong to group 2, while rules indicating order level 'O' belong to group 1), and weight mean (here empty). More information about meaning of each field can be found in [33].

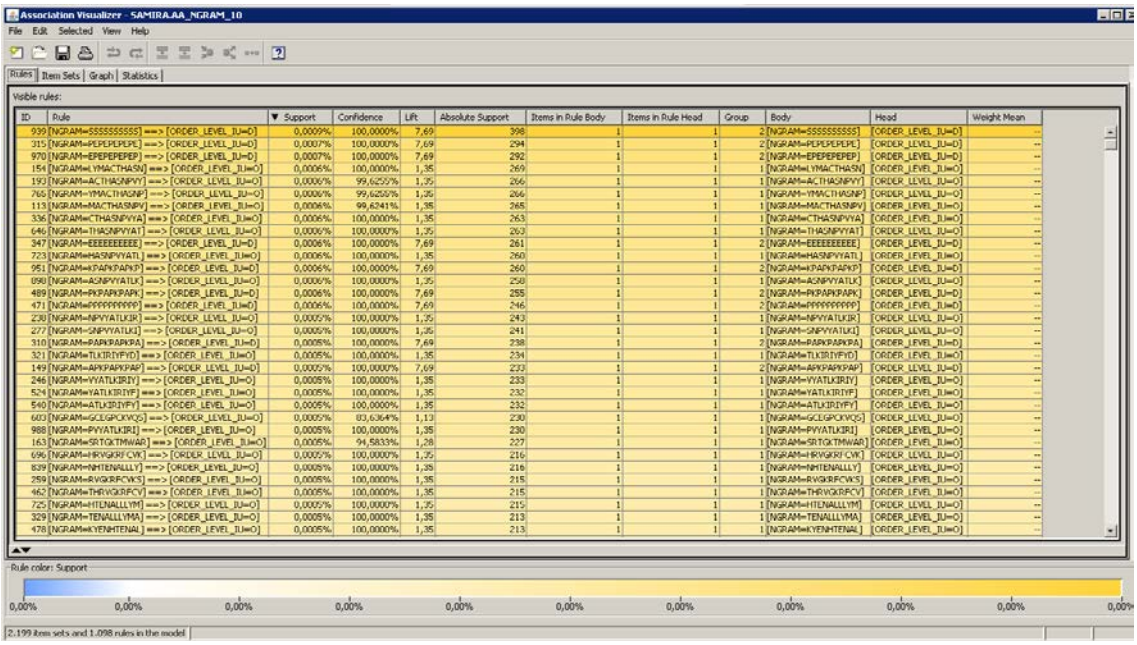


Figure 17. Association rules for n-grams with length=10 produced by IBM Intelligent miner. Information about each rule includes rule and related support, confidence, lift, absolute support, number of items in rule body and rule head, group, rule body, rule head and weight mean.

Association rules can also be represented graphically. If number of rules is large, presenting all rules on a single picture would make the picture cumbersome and ambiguous. For this reason on Figure 18, for example, only the rules related to disordered regions are shown. Rule head is in the middle of the figure while n-grams

that belong to rule bodies are on the circle. Two of three measure parameters (support, confidence, lift) can be (arbitrary) selected for presentation on the figure:

- by line colour; confidence level is presented on the figure by colour spectrum from highest (ocher in tone) to lowest (blue).
- by line width; support is presented on the figure by line width - n-grams with higher support are connected to rule head with wider line.
- by numbers; numerical values of the parameters presented by colour and width are shown on corresponding line.

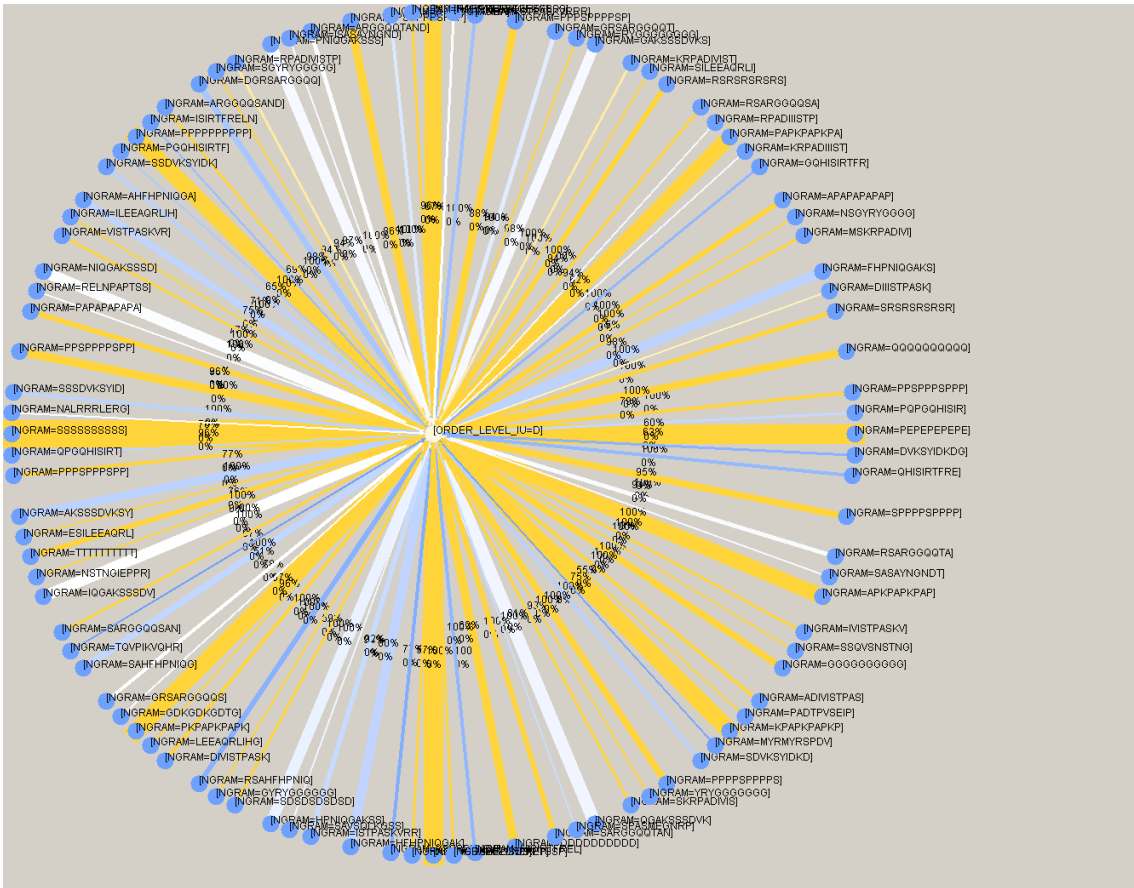


Figure 18. Graphical presentation of association rules

4.5.1.1 Association rules of AA n-grams

Total numbers of discovered rules per n-gram lengths are shown in Table 12. Each rule includes only one n-gram. Although rules for ordered regions are more numerous and have larger support, they have significantly lower average value of lift, and uniform but

small standard deviation of lift. As higher lift means, by default, more interesting rule, the conclusion that can be derived from Table 12 is that rules for disordered regions are, in general, more significant, and that n-grams much better characterize disordered than ordered regions. Appendix tables A14, A15 and A16 contains first 100 rules for each n-gram length that characterize all three types of regions.

Parameters used in association rules results in significantly lower number of rules (i.e. n-grams) compared with corresponding number of n-grams for z-score values (Table 9) or fractional difference (Table 7). The reasons are:

- different meaning of rules compared to classification (for example confidence>=51% implies that majority of n-grams appears in specific region)
- more restrictive support level than in mole fractions or fractional difference method (support>=0.0001 can be considered as mole fractions threshold equal to 1E-6 on global level, not on the level of individual order level as mole fractions are)[7]
- additional filtering with lift interval which discards rules with low level of interestingness (i.e. rule that are expected to occur).

Combining the results obtained from the association rules, mole fractions, fractional difference and z-score methods produce smaller set of n-grams that characterize regions from different points of view and very high confidence. Numbers of n-grams in the intersection set are shown in Table 13. Numbers of n-grams in this table are relatively small because of the different characteristics of methods. For example, because of confidence≥51% for association rules, first condition that some n-gram can be marked as characteristic one, for some region type, is that more than half occurrences of that n-gram are found in the regions of such type. On the other side, fractional difference or z-score values can be higher for n-gram in such region if majority of occurrences of this n-gram belongs to region with different type. Also, some n-grams have standard deviation equal to zero and hence their z-score can not be calculated (this is especially expressed for n-grams with length 8, 9 and 10). Percentages of order levels agreement

[7] Support level and previously used threshold guarantee that no n-gram with small (e.g. statistically non-significant) number of occurrences will appear in results. For example, if number of n-grams with specific length is 1.500.000 than n-gram of such length which occurs less than 150 times will not be taken into account.

between methods are shown on Figure 19 (A: for n-grams in disordered; B: for n-grams in ordered regions).

Table 12. Association rules characteristics for disordered and ordered regions.
Parameters used for discovering rules are: confidence>=51%, support >=0.0001 and lift >=1.05 or lift <=0.95

Rules for disordered regions									
N-gram length	Number of rules	Lift				Support		Confidence	
		Average	Standard deviation	Min	Max	Average	Standard deviation	Average	Standard deviation
2	1	2,88	0	2,88	2,88	0,1479133	0	58,17	0
3	150	2,95	0,33	2,63	4,45	0,0098004	0,0059749	56,20	6,30
4	5.119	3,29	0,44	2,79	5,58	0,0005698	0,0005616	59,05	7,89
5	6.778	4,28	0,77	2,96	5,92	0,0001690	0,0001894	72,30	13,09
6	781	5,72	0,74	3,13	6,26	0,0002048	0,0002725	91,48	11,86
7	339	6,23	0,72	3,33	6,61	0,0002100	0,0002423	94,26	11,00
8	187	6,51	0,83	3,51	6,96	0,0002178	0,0002045	93,50	12,04
9	135	6,81	0,90	3,70	7,32	0,0002103	0,0001729	93,03	12,36
10	97	7,00	1,03	3,93	7,68	0,0002160	0,0001553	91,13	13,43

Rules for ordered regions									
N-gram length	Number of rules	Lift				Support		Confidence	
		Average	Standard deviation	Min	Max	Average	Standard deviation	Average	Standard deviation
2	329	1,03	0,16	0,64	1,26	0,1874637	0,1163087	80,87	12,96
3	6.533	1,07	0,17	0,64	1,29	0,0093590	0,0076481	82,93	13,25
4	109.586	1,08	0,17	0,64	1,29	0,0005483	0,0004824	83,81	13,70
5	99.911	1,14	0,14	0,65	1,30	0,0001429	0,0000515	87,60	11,33
6	3.167	1,27	0,09	0,65	1,31	0,0001556	0,0000874	97,13	6,93
7	1.811	1,30	0,07	0,66	1,32	0,0001674	0,0000963	98,41	5,42
8	1.363	1,31	0,07	0,66	1,33	0,0001718	0,0000977	98,38	5,56
9	1.146	1,32	0,07	0,69	1,34	0,0001712	0,0000977	98,36	5,76
10	948	1,33	0,07	0,70	1,35	0,0001734	0,0000980	98,44	5,46

Rules for border regions									
N-gram length	Number of rules	Lift				Support		Confidence	
		Average	Standard deviation	Min	Max	Average	Standard deviation	Average	Standard deviation
2									
3									
4									
5	55	10,13	1,58	7,64	13,89	0,0001489	0,0000630	67,66	10,60
6	57	9,71	2,00	6,16	12,32	0,0001545	0,0000557	78,82	16,23
7	58	8,56	1,63	5,28	10,56	0,0001545	0,0000534	81,04	15,46
8	57	7,53	1,45	4,65	9,31	0,0001553	0,0000512	80,86	15,63
9	56	6,83	1,19	4,30	8,37	0,0001539	0,0000503	81,58	14,23
10	53	6,23	1,14	3,90	7,64	0,0001521	0,0000521	81,55	14,95

Table 13. Numbers of n-grams in intersection set of fractional difference, z-score and association rules methods depending on fractional difference.

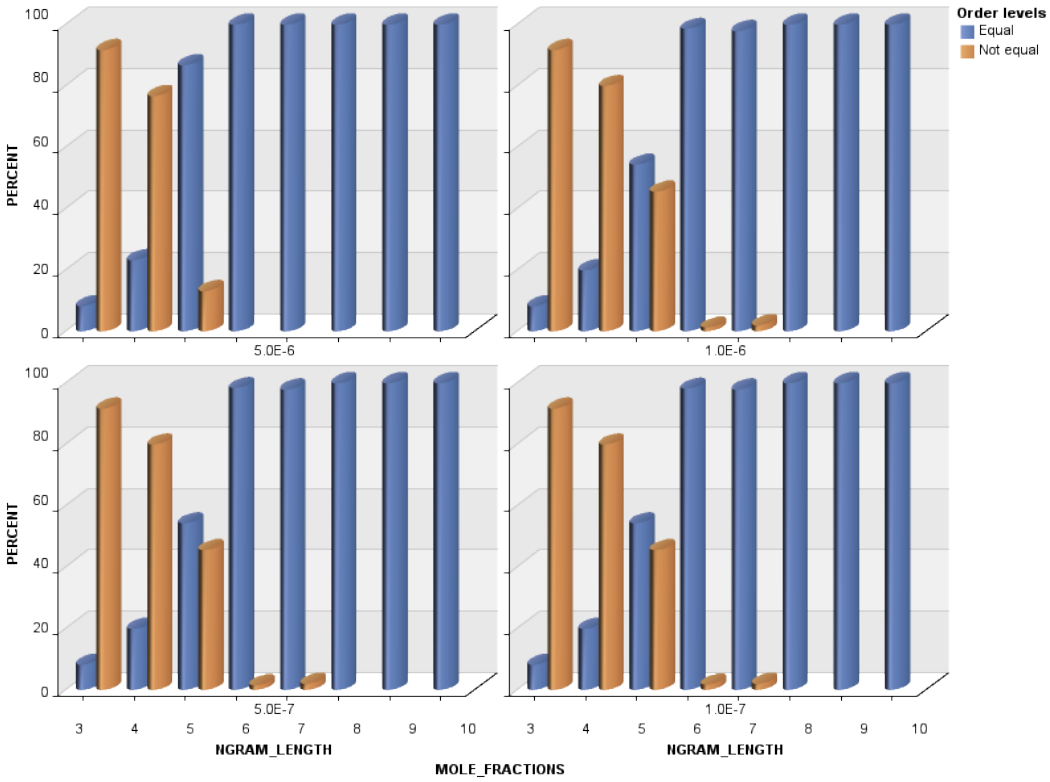
Order level: FD/z-score - order level of n-gram in combination of fractional difference and z-score methods; association rules - order level of n-gram according found association rule. Blue cells: numbers of n-grams with identical order level in all methods; yellow cells: numbers of n-grams with different order level in FD/z-score and association rules methods.

N-gram length	Order level		Minimal value of n-gram mole fraction			
	FD/z-score	association rules	5.0E-6	1.0E-6	5.0E-7	1.0E-7
3	D	D	15	15	15	15
		O	166	166	166	166
	O	O	1.087	1.089	1.089	1.089
4	D	D	1.046	1.046	1.046	1.046
		O	3.446	4.180	4.180	4.180
	O	O	9.143	18.881	18.881	18.881
5	D	D	2.458	2.458	2.458	2.458
		O	375	2.062	2.063	2.063
	O	O	322	33.231	33.231	33.231
		D	0	2	5	8
6	D	D	436	436	436	436
		O	0	6	7	8
	O	O	126	2.746	2.746	2.746
		D	0	5	5	6
7	D	D	190	190	190	190
		O	0	4	4	4
	O	O	104	1.553	1.553	1.553
		D	0	5	6	6
8	D	D	60	60	60	60
	O	O	33	493	493	493
		D	0	2	3	5
9	D	D	30	30	30	30
	O	O	12	125	125	125
10	D	D	10	10	10	10
	O	O	2	41	41	41

It is interesting that number of n-grams with identical order levels in all methods does not dramatically change for various mole fractions smaller than 5E-6. Also, as length of n-grams increase, numbers of order levels differences decrease, and for lengths 9 and 10 there are no differences in order levels for the identical n-grams. These trends remain the same if percentages are considered instead of n-grams numbers.

N-grams that belong to resulting set, without restriction related to mole fractions, are listed in Appendix tables A17 (disordered regions) and A18 (ordered regions). The minimal n-gram length is 3 because no z-score exists for shorter n-grams. Tables includes up to 100 n-grams (if there are so many characteristic n-grams for appropriate length) ordered according lift, confidence, and support, all in descending order.

A



B

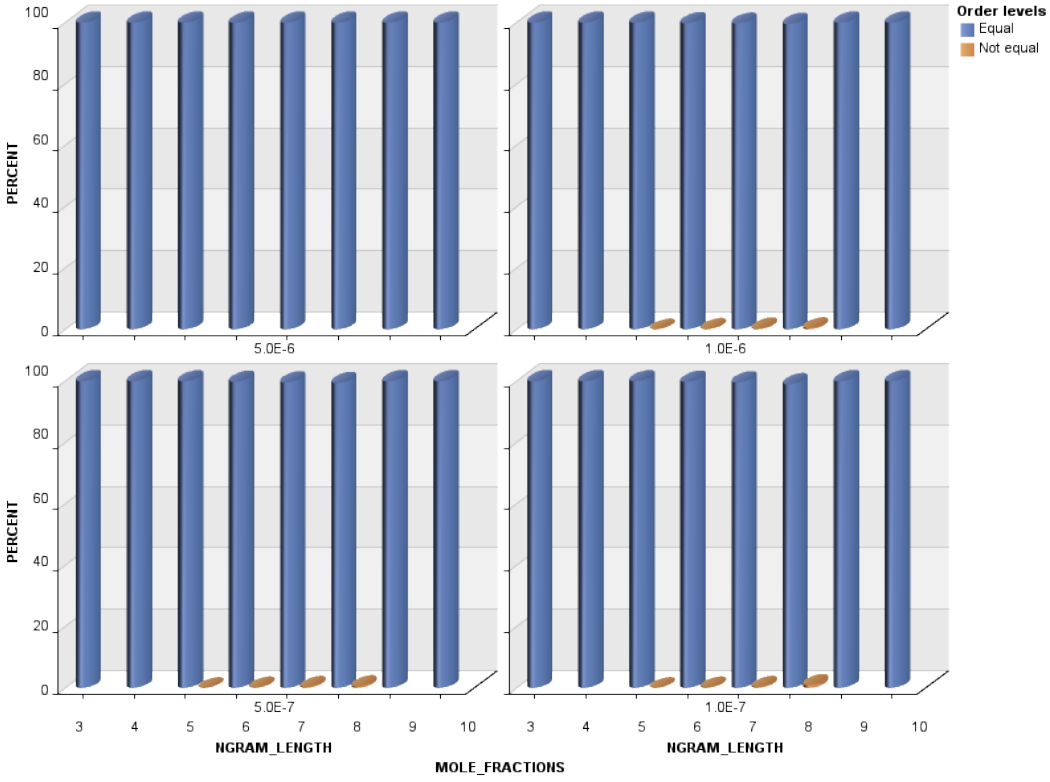


Figure 19. Percentage of order levels agreement between FD/z-score and association rules methods. A: n-grams in disordered regions; B: n-grams in ordered regions

Final set of n-grams that characterize disordered regions includes:

1) homorepeat n-grams of various lengths (like HHHH, KKKKK, GGGGGG, NNNNNN, PPPPPPP, TTTTTTT, EEEEEEEE, DDDDDDDDD, QQQQQQQQQ, SSSSSSSSSS, etc) of AAs that are disorder promoting (G, K, E, D, Q, S, P, E) or border promoting (N, T, H). Homorepeats of A amino acids are found in association rules but are not member of final set because they do not satisfy z-score condition - either have large z-score in both ordered and disordered regions or have smaller absolute value of z-score than necessary for confidence level of 99% (±2.58).
2) their combinations with some AA (like PPPA, REEEE, TGGGGG, GAGGGGGS, RYGGGGGGG, etc)
3) tandem repeats like n-grams of disorder promoting AAs (for example, KPAPKPAP, PSPPPPSPPP, PEPEPEPE, GGEGGEGG, etc)
4) palindromes of disorder promoting AAs (for example, QPQPQ, DEEEED, PAPAPAPAP, etc.)

Final set of n-grams that characterize ordered regions includes:

1) n-grams that include bigrams or trigrams of order promoting AAs (bigrams: VV, FF, WW, YY; trigram LLL)
2) almost all n-grams that include bigram CC or II. More than 99.5% of n-grams that include bigram II are classified as order or border characteristics, with exception n-grams where II is surrounded by disorder promoting AAs. N-grams RPADIII, IISTPA, ADIIIST, PADIII, ADIIIS, IIISTPAS, PADII are marked as disorder promoting while n-gram MKKII is marked as border promoting. Also, about 90.5% n-grams that include bigram CC characterize order region, while others characterize border regions.

Only a few of the patterns can be observed in the final set of n-grams that characterize border regions:

1) n-grams that contain HCP, PLLN, YFYDS characterize border regions only
2) n-grams that contain QID, TRS, FQI, TEG, and YFY prefer border regions but also characterize order regions. Some of them (like PLL or YFY) are sub-n-grams of n-grams that characterize border regions only

4.5.1.2 Comparison with data from DisProt database

Previous results can be compared with corresponding data from DisProt database. The same methods (mole fractions, fractional difference, z-score and association rules) were applied on data available from DisProt database. Due to the initial smaller number of n-grams, the final set of DisProt n-grams also have low cardinality and the intersection between this set and set of obtained results is too small. That's why results of comparison will be shown in three figures: Figure 20 includes results of comparison order levels of identical n-grams from final (intersected) sets; Figure 21 includes results of comparison of order levels generated by association rules, and Figure 22 includes results of comparison of order levels generated by fractional difference and z-score. In all three figures numbers and percentages of identical n-grams with equal/not equal order levels in DisProt and used material are presented.

Also, as explained in the beginning of the Chapter 3, results of comparison where n-gram is predicted to be disorder related, but in DisProt it is order related, should be taken with reserve. These numbers in the corresponding tables on figures 20-22 are marked yellow, while results of comparison where n-gram is in disordered region in DisProt and ordered region in material are marked red.

Order levels	N-gram length	Order level in DisProt	
		D	O
Equal	3		45
	4	34	398
	5	20	148
	6	1	2
	8	1	
Not equal	3	3	
	4	34	27
	5	13	44
	6		1
	7		3

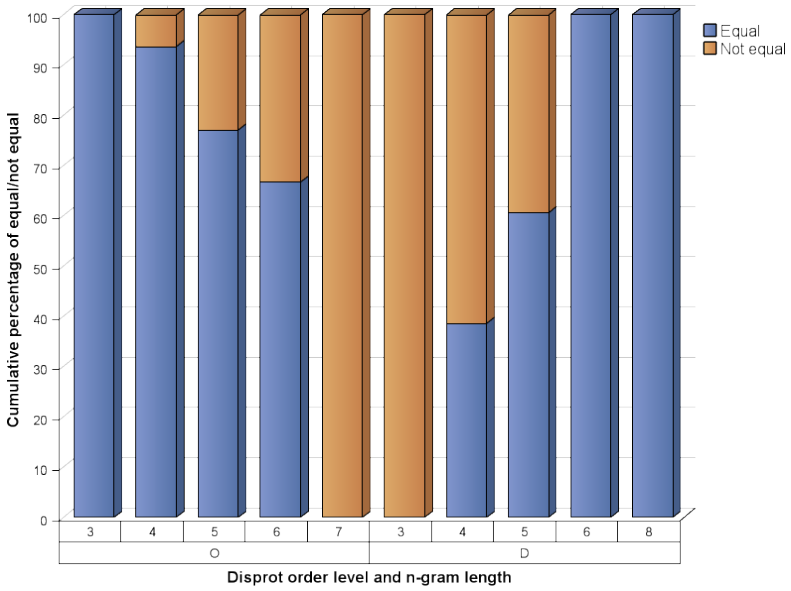


Figure 20. Number and percentage of equal/not equal order levels related to identical n-grams obtained from intersection of result sets of n-grams from DisProt and used material. There are no identical n-grams with length 7 in DisProt and used material in intersection of sets.

Order levels	N-gram length	Order level in DisProt	
		D	O
Equal	2		235
	3	7	4.782
	4	595	23.944
	5	282	1.567
	6	40	8
	7	15	1
	8	5	
	9	4	
	10	4	
Not equal	2		1
	3	42	122
	4	1.515	1.910
	5	133	433
	6	1	35
	7		9
	8		6
	9		6
	10		4

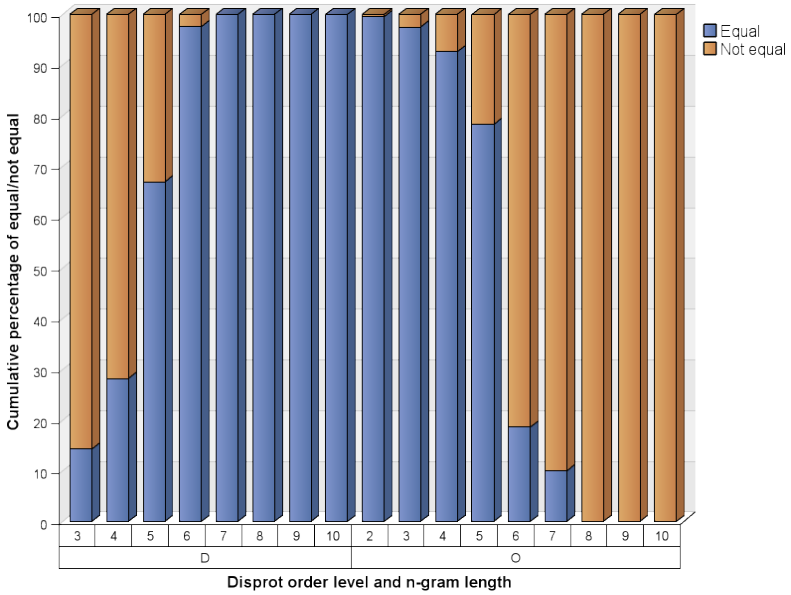


Figure 21. Number and percentage of equal/not equal order levels related to identical n-grams from association rules generated on n-grams from DisProt and used material

Order levels	N-gram length	Order level in DisProt	
		D	O
Equal	3	14	50
	4	238	693
	5	1.183	7.234
	6	120	2.321
	7	8	9
	8	1	
	9	1	
Not equal	3	37	3
	4	704	160
	5	1.548	2.885
	6	89	1.302
	7		19
	8		2
	10		1

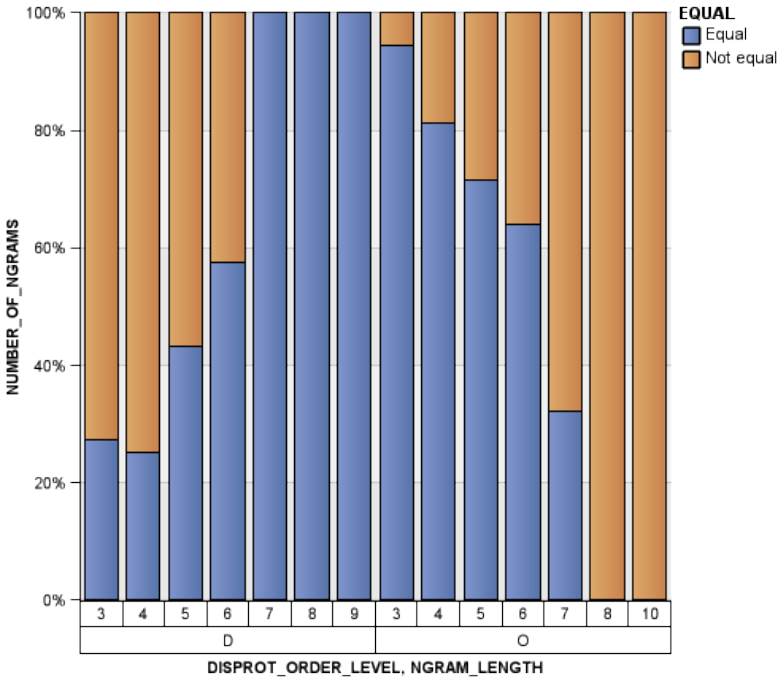


Figure 22. Number and percentage of equal/not equal order levels related to identical n-grams from fractional difference/z-score results based on n-grams from DisProt and used material

Although the numbers of n-grams with equal/not order level are not too high (which in some cases seems that they are not enough representative) the common trend can be observed in all three figures:

1) For the n-grams length 3 and 4 accuracy of characterization is significantly lower than 50% for disordered regions and significantly higher for ordered regions
2) N-gram length 5 is a crossing point - accuracy moved to disorder side. Further increasing n-grams length increase the accuracy of characterization for disordered regions and decrease accuracy of characterization for ordered regions.
3) It is possible that accuracy of prediction for longer n-grams is higher than presented. As previously noticed current version of DisProt database include precise information only about protein regions that are experimentally proved as disorder. Consequence is that set of disorder related n-grams selected from DisProt is not complete. It is possible that longer n-grams currently not recognized as disorder related in DisProt are actually disorder-characterize ones.

For example, differences between order levels in association rules based on DisProt and used material are related to the following n-grams which consist of disorder promoting AAs only.

```
Length  N-grams
------------------------------------------------------------------------------
 7      DSDSDSD,GGGGSGG,GGGSGGG,GPPGPPG,PEPEPEP,QQQQQQQ,SDSDSDS,SGGGGGG,SSSSSSS
 8      DSDSDSDS,EEEEEEEE,QQQQQQQQ,SDSDSDSD,SGGGGGGG,SSSSSSSS
 9      DSDSDSDSD,EEEEEEEEE,QQQQQQQQQ,SDSDSDSDS,SGGGGGGGG,SSSSSSSSS
10      DSDSDSDSDS,QQQQQQQQQQ,SDSDSDSDSD,SSSSSSSSSS
```

4.5.1.2.1 Finding patterns in characteristic n-grams

Additional research was done to discover patterns related to characteristic n-grams. All substrings of n-grams with lengths great than or equal 3 was considered as potential pattern. Such substring is marked as characteristics for order (disorder) region if it is not part of any n-gram that characterize disorder (order) region. Surprisingly, number of such patterns is not too low; number of patterns for material used in this research (download from NCBI), material from DisProt database and intersection of these two sets are shown in of Table 14.

Table 14. Number of sequences (sub-n-grams) that belong to n-grams and characterize some region type.

Order level	Pattern length	NCBI material	DisProt material	Intersection
D	3	100	184	4
	4	2463	1268	131
	5	2641	811	27
	6	470	188	3
	7	206	34	2
	8	77	17	1
	9	36	7	--
	10	10	2	--
N	3	1	--	--
	4	37	9	--
	5	83	9	--
	6	70	9	--
	7	73	9	--
	8	70	9	--
	9	63	10	--
	10	53	10	--
O	3	5559	2819	2060
	4	38010	6038	2201
	5	33946	4201	171
	6	2873	616	3
	7	1610	42	--
	8	581	6	--
	9	157	2	--
	10	41	1	--

For example, patterns of length 6 that belong to both sets and characterize ordered regions are `GGLEGL`, `GSGKST`, `TGSGKS`, while patterns of the same length that characterize disordered regions are `APAPAP`, `GGGGGG`, `SGSSSS`. It is interesting that no intersection between sets exists for sequences that characterize borderline region. Also, it is interesting that, if hydrophobicity (according Kyte-Doolittle scale, further KD scale) of amino acids in patterns are considered then patterns that characterize disordered regions are much hydrophilic than patterns related to ordered regions. Hydrophobicity of patterns is calculated on two ways: as majority of hydrophobic/hydrophilic AA (in this case 'neutral' means that numbers of hydrophilic and hydrophobic AAs are equal), and as a sum of hydrophobic/hydrophilic values according to KD scale (see Table 15). If sum is negative than the pattern is marked as hydrophilic; if sum is positive than the pattern is marked as hydrophobic, and otherwise it is marked as neutral. It can be concluded that pattern in intersection set that characterize disordered regions and can be considered as 'proved disordered' are almost completely hydrophilic. Due to the previously mentioned reasons patterns in DisProt material (and consequently in the intersection) can not be considered as 'proved order' and not commented here.

Table 15. Hydrophobicity of n-gram patterns that characterize regions.

Majority of hydrophobic/hydrophilic AA - majority of AAs in pattern are hydrophobic or hydrophilic
Neutral - pattern consists of equal number of hydrophilic and hydrophobic AAs
Hydrophobic/hydrophilic value - sum of hydro-values of AAs from pattern denotes hydrophilic/hydrophobic object
Neutral value - sum of hydro-values of AAs from pattern is equal to 0
All values are according Kyte-Doolittle scale of AAs hydrophobicity

Source	Pattern length	Disordered regions						Borderline regions						Ordered regions					
		Percentage of pattern with																	
		majority of hydrophilic AAs	hydrophilic value	majority of hydrophobic AAs	hydrophobic value	equal number of hydrophilic and hydrophobic AAs	neutral value	majority of hydrophilic AAs	hydrophilic value	majority of hydrophobic AAs	hydrophobic value	equal number of hydrophilic and hydrophobic AAs	neutral value	majority of hydrophilic AAs	hydrophilic value	majority of hydrophobic AAs	hydrophobic value	equal number of hydrophilic and hydrophobic AAs	neutral value
DisProt material	3	86,41	77,71	13,58	22,28	0,00	0,00							68,96	57,21	31,03	42,42	0,00	0,35
	4	74,05	79,33	4,25	20,34	21,68	0,31	22,22	33,33	33,33	66,66	44,44	0,00	54,38	60,69	13,36	38,65	32,24	0,64
	5	88,03	77,80	11,96	21,82	0,00	0,36	33,33	33,33	66,66	66,66	0,00	0,00	74,45	58,98	25,54	40,89	0,00	0,11
	6	88,82	85,10	2,65	13,82	8,51	1,06	33,33	33,33	22,22	66,66	44,44	0,00	67,85	64,12	10,55	35,55	21,59	0,32
	7	97,05	88,23	2,94	11,76	0,00	0,00	66,66	55,55	33,33	44,44	0,00	0,00	92,85	90,47	7,14	9,52	0,00	0,00
	8	94,11	94,11	0,00	5,88	5,88	0,00	66,66	55,55	11,11	44,44	22,22	0,00	100,00	100,00	0,00	0,00	0,00	0,00
	9	100,00	85,71	0,00	14,28	0,00	0,00	90,00	50,00	10,00	50,00	0,00	0,00	100,00	100,00	0,00	0,00	0,00	0,00
	10	100,00	100,00	0,00	0,00	0,00	0,00	60,00	50,00	0,00	50,00	40,00	0,00	100,00	100,00	0,00	0,00	0,00	0,00
Intersection	3	100,00	75,00	0,00	25,00	0,00	0,00							63,39	51,01	36,60	48,68	0,00	0,29
	4	81,67	92,36	0,76	7,63	17,55	0,00							36,48	41,52	21,26	57,74	42,25	0,72
	5	92,59	85,18	7,40	14,81	0,00	0,00							50,87	27,48	49,12	72,51	0,00	0,00
	6	66,66	66,66	0,00	33,33	33,33	0,00							100,00	66,66	0,00	33,33	0,00	0,00
	7	50,00	50,00	50,00	50,00	0,00	0,00												
	8	0,00	0,00	0,00	100,00	100,00	0,00												
NCBI material	3	97,00	95,00	3,00	5,00	0,00	0,00	100,00	100,00	0,00	0,00	0,00	0,00	65,49	54,65	34,50	45,04	0,00	0,30
	4	88,63	93,78	0,85	6,21	10,51	0,00	75,67	83,78	0,00	16,21	24,32	0,00	41,88	46,50	18,74	52,85	39,36	0,64
	5	95,11	89,85	4,88	10,03	0,00	0,11	86,74	73,49	13,25	26,50	0,00	0,00	61,68	38,89	38,31	60,69	0,00	0,41
	6	87,65	84,04	4,68	15,95	7,65	0,00	78,57	71,42	2,85	28,57	18,57	0,00	55,72	54,12	15,94	45,59	28,33	0,27
	7	91,26	86,40	8,73	13,59	0,00	0,00	91,78	83,56	8,21	15,06	0,00	1,36	76,83	58,38	23,16	41,55	0,00	0,06
	8	87,01	84,41	6,49	15,58	6,49	0,00	88,57	84,28	0,00	15,71	11,42	0,00	65,92	58,86	12,04	40,96	22,03	0,17
	9	91,66	86,11	8,33	13,88	0,00	0,00	98,41	90,47	1,58	9,52	0,00	0,00	77,07	56,68	22,92	42,67	0,00	0,63
	10	100,00	90,00	0,00	10,00	0,00	0,00	98,11	84,90	0,00	15,09	1,88	0,00	73,17	63,41	14,63	36,58	12,19	0,00

4.5.1.3 Association rules of nucleotide n-grams

Discovering association rules for complete set of n-grams exceeds computational capability of computer system used for this research. Due to a huge number of n-grams (ranging from 42M to 140M) association rules were discovered on smaller subsets of direct non-complementary nucleotide repeats (n-grams) only, but not on other sort of nucleotide repeats (direct complementary, inverse complementary and inverse non-complementary).

For each n-gram length, set of n-grams is divided in three parts, according to their corresponding ORF-s. Number of discovered rules rapidly decrease as n-gram length increase, and is smaller than number of rules of AAs of corresponding length because of different codon usage tables used for translating AAs. Number of association rules for nucleotide n-gram lengths 15, 18, 21, 24, 27 and 30 is shown in Table 16, and results of the comparison of their translation (using translation table 11) to corresponding AA n-grams is shown in Table 17. It can be seen that longer n-grams are mostly related to disordered regions regardless of ORF.

Table 16. Number of discovered association rules for nucleotide n-grams

N-gram length	ORF					
	1		2		3	
	D	O	D	O	D	O
15	16	15	13	15	14	14
18	11	12	11	9	9	8
21	6	7	6	3	4	3
24	4	2	4	1	4	1
27	2	--	3	--	2	--
30	1	--	2	--	1	--

In all three ORFs nucleotide n-grams behave regularly as well as the corresponding AA n-grams. Longer nucleotide n-grams more precisely characterize both types of regions. Additionally, n-grams with length 27 and 30 characterize only disordered regions, as in the case of similar (with lengths 9 and 10) AA n-grams. Also, some of the n-grams that have different order level than corresponding AA n-grams (equivalent to their translation), are homorepeats of disorder promoting AAs (as S or Q) which are possible disorder related, as previously mentioned.

Table 17. Number and percentage of equal/not equal order levels related to translations of nucleotide n-grams and identical AA n-grams

Order levels	N-gram length	ORF1				ORF2				ORF3			
		D		O		D		O		D		O	
		num	perc	num	perc	num	perc	num	perc	num	perc	num	perc
Equal	15	23	92.0	15	100.0	12	80.00			10	71.42		
	18	14	93.33	12	100.0	8	100.0			5	83.33		
	21	8	100.0	7	100.0	4	100.0						
	24	5	100.0	2	100.0	3	100.0						
	27	2	100.0			2	100.0						
	30	2	100.0			1	100.0						
Not equal	15	2	8.00			3	20.00	3	100.0	4	28.58		
	18	1	6.67					2	100.0	1	16.67		
	21												
	24												
	27												
	30												

There are n-grams that occur in association rules related to all three ORF-s. These n-grams have maximal length 9, and as this correspond to AA n-grams of length 3 or shorter which, as shown earlier, do not have high precision in regions characterization (especially not satisfactory level of characterization for disordered regions) so no such n-grams are considered.

4.5.1.4 Association rules of inverse non complementary AA repeats

Inverse non-complementary repeat (in further text *IN repeats*) represents palindrome with a gap of arbitrary (≥1) length between left and right components of the repeat. Left and right component can belong to different types of regions, so 9 different "double order levels" exist: DD, DO, DN, OD, OO, ON, ND, NO, NN[8]. Repeats characterize region type 'X' if both components (left and right) fall into regions of such types, so high accuracy is reached in the research only for DD, OO and NN combinations. Also, in process of determination of association rules only left component of repeat (on Figure 23 "REPEAT_LEFT") and double order level combinations are considered because right component is unambiguously determined by the left one.

Association rules are determined for both sets of all IN repeats and statistically significant IN repeats. The similar parameters were used as in determining association

[8] "Double order level" DO is different from OD because DO determines that left component of repeat is related to disordered while right component is related to ordered region, and vice-versa for OD.

rules for n-grams: confidence>=51%, support>=0.0005 and lift >=1.05 or lift <=0.95. Support threshold for association rules is increased to 0.0005 because, as the number of repeats is significantly lower than number of n-grams with the same length, using support equal to 0.0001 as in association rules for n-grams lead to plenty of association rules for small repeat lengths. The results obtained have similar form as in the case of ordinary n-grams, as illustrated on Figure 23[9].

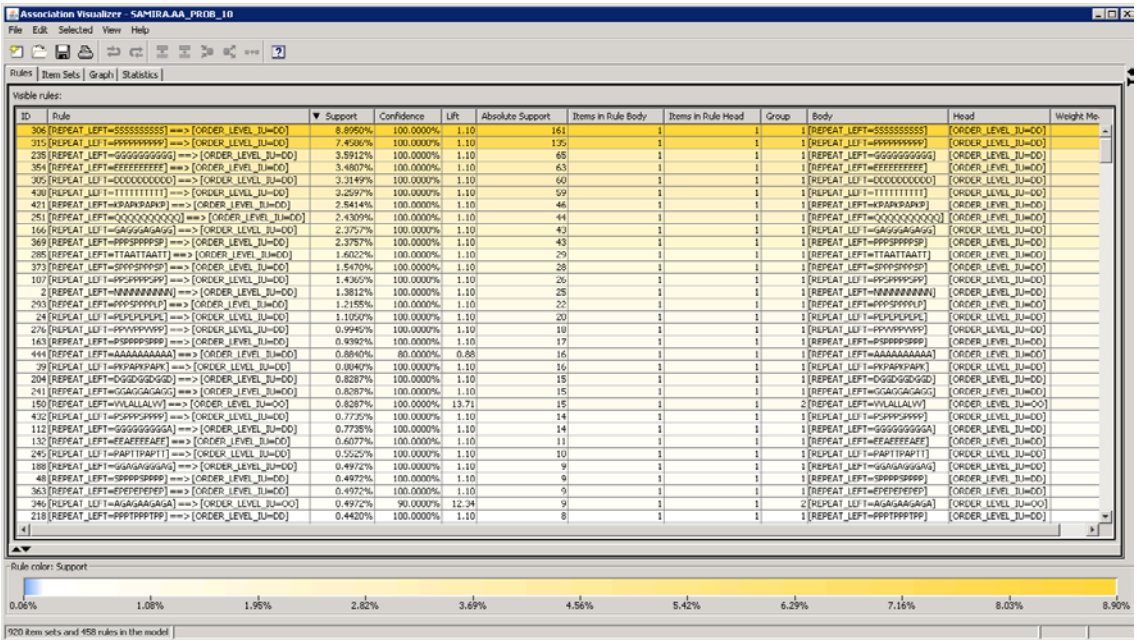


Figure 23. Association rules for IN repeats with length=10 produced by IBM Intelligent miner. Information about each rule includes rule and related support, confidence, lift, absolute support, number of items in rule body and rule head, group, rule body, rule head and weight mean.

Association rules are determined for all repeat lengths from 3 to 10 where term "repeat length" is related to length of either left or right component of repeat. Association rules are also determined for DisProt repeats. Number of rules found for different repeat lengths are shown in Table 18. There is one anomaly in the table: regardless the significantly higher number (about 20 times) of all repeats with length 3 in the used material (from NCBI) than in DisProt material, the number of association rules for this category of repeats is higher for DisProt material. The reason is very small absolute support (two) for DisProt repeats which corresponds to support 0.0005. For the same reason sets with smaller number of repeats (repeat length>5 in NCBI material and

[9] Support 0.0005 additionally filters input dataset, so palindromes with small number (<5) of occurrences in used material were eliminated and not appear in association rules. Due to smaller number of data, for DisProt data threshold for elimination palindromes is less than 2 occurrences.

repeat length>3 for DisProt material) produce majority of rules that occurs only once in the complete results (see Table 18).

Table 18. Number of discovered association rules for inverse non-complementary repeats where source material is originated from NCBI and DisProt. Results are shown for sets of all and statistically significant repeats. Also, for each set, number of association rules where repeat included in the body of the rule occurs more than 1 in complete material (column "absolute support>1") is shown. Repeat length is related to length of either left or right component of repeat.

Repeat length	Repeats from NCBI material				Repeats from DisProt			
	All repeats		Statistically significant		All repeats		Statistically significant	
	All rules	Abs. support>1	All rules	Abs. support>1	All rules	Abs. support>1	All rules	Abs. support>1
3	4186	4186	3566	3566	4645	4645	2925	2925
4	17175	17175	15130	15130	7342	2738	5360	2616
5	8691	8691	8021	8021	2421	582	2234	538
6	5057	2539	5011	2520	363	73	361	73
7	9369	3091	9328	3076	353	49	353	49
8	1471	521	1471	521	70	16	70	16
9	1902	584	1902	584	82	16	82	16
10	457	152	457	152	19	9	19	9

Analysis of rules obtained for all and statistically significant repeats produce the following results:

1. For smaller repeat length all rules have absolute support>1
2. If number of rules is equal for all repeats and statistically significant repeats than rules are identical
3. For larger repeat length the set of repeats have smaller cardinality and predefined support 0.0005 is equivalent to absolute support 1 with consequence that all repeats are taken into consideration and produce some rules. There are no guarantees that such rules with minimal possible absolute support are valid in general. Because those rules can not produce highly accurate results, they will not be taken into consideration.
4. If it is assumed that the probability of appearance each individual AAs is equal, the following filter can be applied on rules based on smaller repeat length: if support for rule is smaller than probability for repeat occurring (for trigrams 0.0125, for tetragrams 0.000625) than this rule is ignored. Although this presumption does not hold in real life (because frequency of occurring is not the same for different AAs and depends on content of material) proposed filter is useful for decreasing number of rules with low probability. Rule is not

applicable on repeats with length longer than 4 for NCBI based material and longer than 3 for Disprot based material because probability of occurring specific repeats is lower than predefined support for association rules. When this filter is applied on repeats from NCBI based material number of rules for statistically significant repeats becomes larger than number of rules for all repeats [10] (827 for all repeats and 869 for statistically significant repeats of length 3, and 12691 and 15130 for length 4). For DisProt material number of repeats decrease to 1245 (all repeats) and 1100 (statistically significant repeats).

5. An additional reduction in the number of rules that are considered in further analysis is achieved by using only those rules with double order level 'OO', 'DD' or 'NN', which are useful for region characterization. Percentage of these rules in total number of rules before and after applying filters (including probability filter and absolute support>1) are

	Material from NCBI		Material from DisProt	
	Before	After	Before	After
All repeats	95.24%	88.29%	93.22%	71.36%
Stat. signif. rep.	95.44%	89.44%	91.80%	76.71%

Rules with longer repeats do not contain other order levels than 'OO','DD' or 'NN'. Numbers of rules after applying previous filters are shown in Table 19.

6. In general, rules for repeats with length>3 that do not belong to the set of statistically significant repeats have the following characteristics:
    a. Higher support corresponds to lower confidence. Majority of such rules have double order level 'OO', confidence between 0.51 and 0.65 and lift near 0.95 and 1.05. As this value of lift indicates that the rule body and the rule head appear almost as often together as expected, means that the occurrence of the rule body has almost no effect on the occurrence of the rule head, these rules will not be taken into account for determining characterization strings.

[10] These numbers may look like an error because set of statistically significant repeats is subset of set of all repeats. But, because of larger number of repeats in set of all repeats large number of rules have minimal support which didn't passed filter. This is evident if compare average support per rule for all repeats and statistically significant repeats: 0.0128/0.0150 for and 0.0035/0.0040 for repeats with length 3 and 4 repsectively.

b. Majority of rules with high confidence have small support and very low absolute support (2 or 3).

7. Rules based on repeats from NCBI with length=3 that do not belong to the set of statistically significant repeats have the following characteristics:
    a. All rules have double order level 'OO'
    b. There are no rules with confidence 100%. About 65% of rules have lift smaller than 0.95 and confidence below 61.5%.
    c. About 30% rules have confidence>70% and lift>1.08. Repeats in these rules are potential characteristic sequences (for ordered regions). Majority of these repeats have palindromes as left and right components. Left components of these repeats are: AVI, CDC, CEC, CKC, CRC, CSC, CTC, ELG, FCF, FHF, FMF, FWF, HAH, HDH, HEH, HFH, HIH, HKH, HNH, HVH, HYH, IDA, IED, IWI, KVI, MFM, MIM, MVM, NWN, PCP, PWP, RTL, WLW, YCY, YHY, YMY
8. Rules based on repeats from DisProt with length=3 and support>0.0125 that do not belong to the set of statistically significant repeats, have the following characteristics:
    a. All rules have double order level 'OO'
    b. About 30% of rules include repeats that have palindromes as left and right components.
    c. There are ≈10% rules with confidence 100%, ≈17% of rules have lift smaller than 0.95 and confidence below 61.5%.
    d. About 78% rules have confidence>70% and lift>1.08. As previously mentioned, there is no guarantee that protein regions in DisProt database that are not marked as disordered are ordered. Based on this premise, there is no guarantee that repeats in such rules can be used as strings that characterize ordered regions, and hence such repeats were not listed.

Based on the results of this analysis, the set of rules based on statistically significant repeats from NCBI material with previously described filters applied was used as a base for determining repeats that characterize protein regions. For verification of obtained results set of rules based on statistically significant repeats from DisProt material with applied same filters was used. Numbers of rules obtained after applying filers are shown in Table 19.

Table 19. Number of association rules based on repeats after applying filters

Repeat length	Repeat from NCBI material				Repeats from DisProt material			
	Order level				Order level			
	All	DD	OO	NN	All	DD	OO	NN
3	869	78	791	--	1032	31	1001	--
4	13872	1898	11969	5	2375	218	2157	--
5	7589	2230	5311	48	507	147	360	--
6	2463	991	1427	45	68	28	40	--
7	3066	1475	1454	137	49	28	21	--
8	517	371	124	22	16	8	8	--
9	584	442	124	18	16	10	6	--
10	152	134	15	3	9	5	4	--

Appendix Tables A20 -- A22 include left components of repeats that characterize disordered, ordered and borderline regions from NCBI and Tables A23-A24 include left components of repeats that characterize disordered and ordered regions from DisProt material respectively. Tables include first 100 repeats (if exists), ordered by confidence, lift and support, all in descending order. Although it seems that if some n-gram 'X' characterize some region type 'Y' that repeat with left or right component equal to 'X' characterize region type 'Y' (i.e. 'YY') this is not always true. For example, repeat with left/right components ATTTAA/AATTAA have order level 'OO' while both n-grams ATTTAA and AATTTA have order level 'D' in association rules. Of course, if left and right components of repeat in association rule related to n-grams have confidence 100% than both rules types characterize the same order level.

Results of comparison of order levels in association rules based on material from NCBI and DisProt are shown in Table 20. As in previous cases, results of comparison where repeats are predicted to be disorder related, but in DisProt they are order related, should be taken with reserve. These numbers in the Table 20 are marked yellow, while results of comparison where repeats are in disordered region in DisProt and ordered region in material from NCBI are marked red. As in previous comparison with DisProt, there are no disagree in order levels for longer repeats when order level in DisProt is equal to 'DD', i.e. method provide high accuracy for repeat length ≥7. Again, as in previous cases, as left components of repeats that are not equal when order level in DisProt is equal to 'OO' for length ≥7 are

```
Len.  Repeats
 7    AEATAEA,DSDSDSD,GGGGGGG,GGGGSGG,GGGSGGG,GGRGRGG,GPPGPPG,HHHHHHH,PEPEPEP,PEPSPEP,QQQAQQQ,QQQQQQQ
 8    GGGGGGGG, GPPGPPGP, HHHHHHHH, PGPPGPPG, SSSSSSSS
 9    DSDSDSDSD, EEEEEEEEE, GGGGSGGGG, PGPPGPPGP, QQQQQQQQQ, SSSSSSSSS
10    EEEEEEEEEE, QQQQQQQQQQ, SSSSSSSSSS
```

which include only disorder promoting AAs, it can be supposed, with a high probability, that the characterization of the disorder regions is one hundred percent correct for repeats with length≥7.

Table 20. Numbers of equal/not equal order levels related to identical repeats in association rules. Source: materials from DisProt an NCBI.

Order levels	Repeat length	Order level in association rules based on DisProt repeats	
		DD	OO
Equal	3	7	357
	4	75	624
	5	72	109
	6	14	5
	7	16	--
	8	3	--
	9	5	--
	10	2	--
Not equal	3	6	41
	4	29	212
	5	15	22
	6	5	17
	7		11
	8		5
	10		6
			3

Some general characteristics related to repeats (material from NCBI) that characterize regions are:

1) Homorepeats of all amino acids except Y characterize some type of region. In general, homorepeats of disorder promoting AAs characterize disordered regions and homorepeats of order promoting AAs characterize ordered regions. Exceptions are M, which characterizes ordered regions, and H and N, which characterize disordered regions. There is no overlapping or duplicate characterization – not even one amino acid characterizes different region type for different homorepeat length. Only homorepeats of amino acid A have lift smaller than 1 (more precisely smaller than 0.878). Characterizations of region types by homorepeats are very accurate. As illustration, found homorepeats, their lengths, lift and confidence of corresponding association rule are shown in Appendix table A25.

2) All rules with repeats whose length is 10 have confidence 100% and lift 1.0963, regardless support which varies between 8.895 and 0.110. The only exception is rule with repeat AAAAAAAAAA with confidence 80%, support 0.884 and lift 0.877, from which can be concluded that amino acid A (which is small and hydrophobic) behaves little different than other disorder promoting AAs.
3) Majority of left and right components of repeats are palindromes itself (see Figure 24).

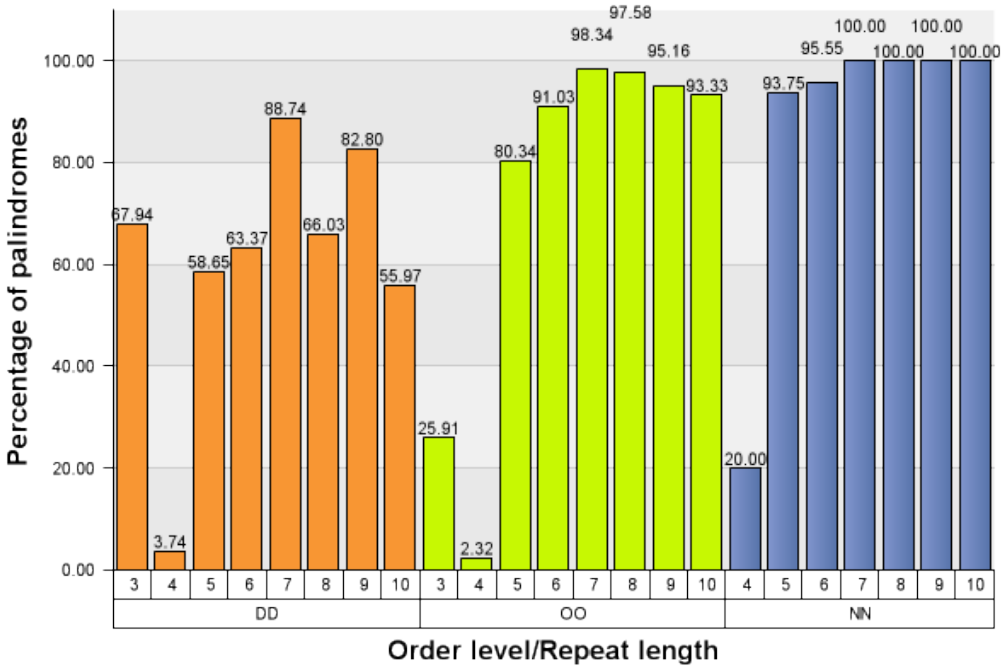


Figure 24. Percentage of palindromic left/right components of repeats that characterize regions

Left and right components that are not palindromes are

- tandem repeats, or
- combinations of smaller homorepeats or palindrome with some AAs

4) Tandem repeats are highly represented in repeats which characterize some region (see Table 21). It is interesting that almost all longer repeats that are not palindromes itself are tandem repeats (see Table 20), while shorter repeats that are neither palindromes nor tandem repeats also includes some sub-palindrome (length ≥3) combined with other AAs.[11]

[11] Tandem repeats are defined as pair of identical sequences with minimal sequence length 2. According this definition minimal repeat length that can include tandem repeat is 4, so percentage calculation is not applicable on repeats with length 3.

Table 21. Percentage of tandem repeats in set of all repeats and in non-palindrome repeats

Repeat length	Tandem repeat percentage			Non-palindrome Tandem repeat percentage		
	DD	NN	OO	DD	NN	OO
4	7,00	0,00	0,65	4,05	0,00	0,61
5	22,15	8,33	7,26	24,51	0,00	4,98
6	57,31	31,11	19,90	75,75	100,00	27,34
7	76,54	36,49	42,09	95,18	0,00	37,50
8	93,26	50,00	50,80	96,03	0,00	100,00
9	95,24	72,22	70,96	100,00	0,00	100,00
10	99,25	100,00	100,00	100,00	0,00	100,00

5) If repeat includes only order promoting AAs, it does not characterize disordered region, with exception of only 20 repeats:

- 7 homorepeats of AA Asparagine (length from 4 to 10)
- 7 homorepeats of AA Histidine (length from 4 to 10)
- repeats with very small support/absolute support:
    - NNNYNNN (abs. support=4),
    - LHHHHL, HHNHH, INNNNN, HHYHH, HHLHH (abs. support=2)

4.5.2 Classification

Another method for discover characteristic n-grams can be applying tree classification method on available set of repeats to predict order/disorder class. Although the obtained model has very limited capabilities [12] for correct prediction on previously unseen material it can be used for discovering n-gram sequences that characterize order/disorder regions (class in model). Due to a large number of n-grams/palindromes the model could not be constructed based on complete sets of n-grams/palindromes. Instead of that, the initial sets are divided by the association of the phyla. For each phylum, sets of n-grams/palindromes are divided into two parts, as described in chapter 3.2. Classification models are constructed using tree based algorithms SPRINT (Scalable PaRallelizable INduction of decision Trees) [32] for each phylum and

[12] Low capability is consequence of using repeat sequences only in model construction. N-grams have categorical type with possible (depends on their length) very large number of values. As dataset used for model construction does not include all possible n-gram values, class for previously unseen value can not be predicted correctly.

checked on corresponding test sets. Quality of each of classification models were between 82% and 96%, while quality of applying constructed models on test data were between 68% and 85%. Sets of n-grams and palindromes produced in models as characteristics of regions confirm previously obtained results from association rules mining. An example of characteristic n-grams obtained with classification is shown on figure 23.

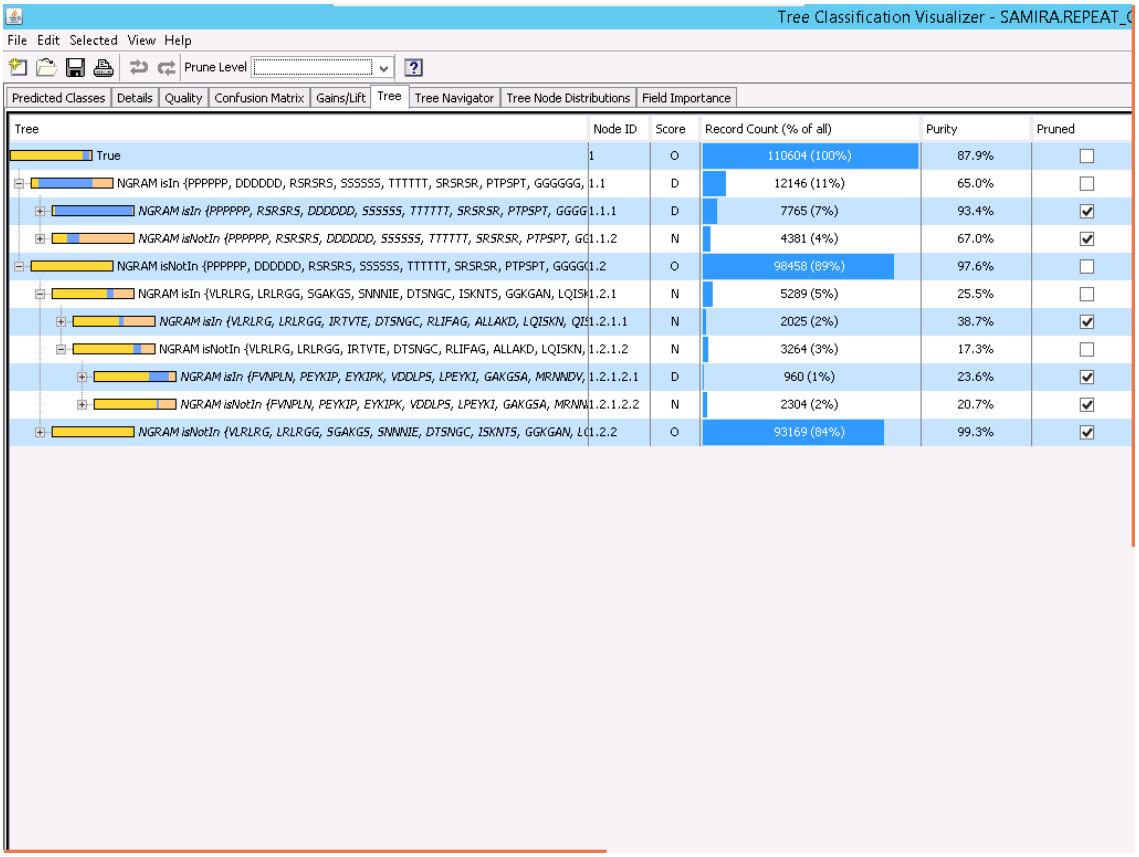


Figure 23. N-grams from classification model for Anelloviridae phylum that characterize specific type of regions

5 Conclusion

Discovering characteristic sequences for ordered/disordered regions in proteins is very important. Intrinsic disorder of proteins are implicated in most important cellular processes such as: cell signaling, transcription and chromatin remodeling functions. On the other side, they are involved in a number of diseases, such as neurological, cardiovascular and malignant pathological states. Taking this in mind, studying structural and dynamical properties of intrinsically disordered proteins is of great importance for better understanding of their actions and developing new medicaments.

In this thesis a new method for determining sequences that characterize ordered/disordered regions with very high confidence is presented. Proposed method establish correspondence with amino acid n-grams to specific region type using n-gram (repeat) characteristics (mole fraction, fractional difference, z-score) and data mining techniques (association rules and classification) applied on both repeats and palindromes. Each of these characteristics/techniques produces n-grams sequences that characterize regions with very high percent of confidence. Sets of sequences produced with various techniques intersect in a very large degree and can be used as characterization sequences for specific region types. General principles that can be observed from the results are:

- type of characterized region depends on sequence (either repeat or palindrome) length
    - shorter n-grams (length up to 6) more precisely characterize ordered regions
    - longer n-grams (length 6 or longer) more precisely characterize disordered regions
- sequences that appear in intersection of results obtained by different methods (fractional characteristics, z-score, association rules) have almost 95% confidence for characterization
- ordered regions are characterized with
    - AAs patterns (VV, FF, WW, YY, LLL)
    - almost all n-grams with patterns CC and II
    - homorepeats of order/border promoting AAs with exception H and N

    - tandem repeats of order promoting AAs
- disordered regions are characterized with
    - homorepeats of various lengths of disorder/border promoting AAs with exception M, and their combination with some AA
    - tandem repeats of disorder promoting AAs
    - palindromes of disorder promoting AAs
    - combinations of homorepeats of disorder/border promoting AAs and some (disorder/border promoting) AA (like PPPA, REEEE, TGGGGG, GAGGGGGS, RYGGGGGGG, etc.)
    - border regions are characterized with some specific n-grams (HCP, ...) or pattern (PLL or YFY)

The proposed method is verified by compared obtained results with results obtained with applying identical methods on material from DisProt database. Results of this thesis show that exists significant correlation between ordered/disordered regions and specific n-grams which can be used for improvement of disorder prediction.

References

[1] A. M. Lesk: *Introduction to Bioinformatics*, 3rd ed. Oxford University Press, 2008

[2] G.N.Ramachandran, C. Ramakrishnan, V. Sasisekharan: *Stereochemistry of polypeptide chain configurations*, Journal of Molecular Biology. **7**: 95–9. (1963)

[3] G. H. Reginald, C. M. Grisham: *Biochemistry*, fourth Edition, Belmont, CA: Brooks/Cole, 2013

[4] A. J. Cozzone: *Proteins: Fundamental Chemical Properties*, Institute of Biology and Chemistry of Proteins, CNRS, Lyon, France, 2002.

[5] P. Tompa, A. Fersht: *Structure and Function of Intrinsically Disordered Proteins.* Boca Raton: Chapman and Hall/CRC Taylor and Francis Group; 2010.

[6] DisProt Database - Database of protein disorder http://www.disprot.org/

[7] V. N. Uversky, A. K. Dunker: *Understanding protein non-folding*, Biochim Biophys Acta - Proteins & Proteomics 2010, 1804(6):1231-1264.

[8] D. Eliezer: *Biophysical characterization of intrinsically disordered proteins*, Current Opinion in Structural Biology 2009, 19:23-30

[9] M. Punta, I. Simon, Z. Dosztanyi: *Prediction and Analysis of Intrinsically Disordered Proteins***,** In Owens J R (ed.), *Structural proteomics: High-Troughput Methods*, Methods in Molecular Biology, vol. 1261, SpringerScience+Business Media New York, 2015, pp. 35-59.

[10] Z. Dosztányi, B. Mészáros, I. Simon: *Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins*, Briefings In Bioinformatics, Vol 11. No 2, 225-243, 2009.

[11] B. Xue, R. L. Dunbrack, R. W. Williams, A. K. Dunker and V. N. Uversky: *PONDR-FIT: A Meta-Predictor of Intrinsically Disordered Amino Acids*. Biochim

Biophys Acta 1804(4):996-1010, 2010.

[12] M. Lobanov and O. Galzitskaya. *The Ising model for prediction of disordered residues from protein sequence alone. Phys. Biol*. 8 (2011) 035004 (9pp).

[13] P. Romero, Z. Obradovic, C. Kissinger, J. E. Villafranca, and A. K. Dunker. *Identifying Disordered Regions in Proteins from Amino Acid Sequence*. Proceedings of the 1997 IEEE International Conference on Neural Networks. Part 4, pp90-95 (1997).

[14] J. Flint, V. R. Racaniello, G. F. Rall, AM Skalka, L. W. Enquis: *Principles of Virology*, Garland science, Taylor & Francis Group, USA, (2015)

[15] N. Tokuriki, C. J. Oldfield, V. N. Uversky, I. N. Berezovsky, D. S. Tawfik: *Do viral proteins possess unique biophysical features?*, Trends in Biochemical Sciences, 34, 53-59, (2008))

[16] B. Xue, A. K. Dunker, V. N. Uversky: *Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life*, Journal of Biomolecular Structure and Dynamics, 30, 137–149, (2012).

[17] D. Tauritz: *Application of n-Grams*, Department of Computer Science University of Missouri-Rolla; 2002.

[18] A. Jelović, N. Mitić, S. Eshafah, M. Beljanski: *Finding statistically significant repeats in nucleic acids and proteins*, Journal of Computational Biology,
DOI: 10.1089/cmb.2017.0046

[19] P. Woolf, C. Burge, A. Keating, M. Yaffe: *Statistics and Probability Primer for Computational Biologists*, Massachusetts Institute of Technology, 2004

[20] PN. Tan, M. Steinbach, V. Kumar: *Introduction to Data Mining, Pearson Education*, 2006

[21] M. Kantardzic: *Data mining : concepts, models, methods, and algorithms*, John Wiley & Sons, 2011

[22] IBM SPSS Modeler 18.0 *Algorithms Guide*, IBM Corporation 2016.

[23] Z. Dosztányi, V. Csizmok, P. Tompa, I. Simon: *IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content*, Bioinformatics 21:3433-3434, 2005.

[24] K. Peng, P. Radivojac, S. Vučetić, AK . Dunker, Z. Obradović: *Length-dependent prediction of protein intrinsic disorder*, BMC Bioinformatics 7:208, 1-17, 2006.

[25] M. Yu Lobanov, I. V. Sokolovskiy, O. V. Galzitskaya: *IsUnstruct: prediction of the residue status to be ordered or disordered in the protein chain by a method based on the Ising model*, Journal of Biomolecular Structure and Dynamics 2013, 31(10), pp. 1034-1043

[26] M. Ganapathiraju, D. Weisser, J. Klein-Seetharaman, R. Rosenfeld, J. Carbonell, R. Reddy: *Comparative n-gram analysis of whole-genome sequences*. HLT'02: Human Language Technologies Conference: 2002 San Diego; 2002.

[27] H. U. Osmanbeyoglu, M. K. Ganapathiraju: *N-gram analysis of 970 microbial organisms reveals presence of biological language models***,** BMC Bioinformatics 2011**,** 12:12**.**

[28] M. Ganapathiraju, A. Mitchell, M. Thahir, K. Motwani, S. Ananthasubramanian: *Suite of Tools for Statistical N-gram language modeling for pattern mining in whole genome sequences***,** Journal of Bioinformatics and Computational Biology, Dec;10(6) 2012.

[29] G. Pavlovic-Lazetic, N. Mitic, M. Beljanski: *n-Gram characterization of genomic islands in bacterial genomes*, Computer Methods and Programs in Biomedicine, (2009), vol. 93 No. 3, pp. 241-256

[30] M. Yu. Lobanov, O. V. Galzitskaya: *Occurrence of disordered patterns and homorepeats in eukaryotic and bacterial proteomes*, Mol. BioSyst., 2012,8, 327–337.

[31] IBM corporation: Intelligent miner https://www.ibm.com/support/knowledgecenter/SSEPGG_10.5.0/com.ibm.im.overview.doc/c_im_benefits.html

[32] Dynamic Warehousing: Data Mining Made Easy, SG24-7418-00, IBM corporation, 2007, http://www.redbooks.ibm.com/redbooks/pdfs/sg247418.pdf

[33] IBM InfoSphere Warehouse: Visualizing mining models, IBM Corporation, 2008, SH12-6840-03

Appendix

Table A1. Amino acid codes

Amino acid names	One letter code	Three letter code*
Alanine	A	Ala
Asparagine or aspartic acid	B	Asx
Cysteine	C	Cys
Aspartic acid	D	Asp
Glutamic acid	E	Glu
Phenylalanine	F	Phe
Glycine	G	Gly
Histidine	H	His
Isoleucine	I	Ile
Leucine or Isoleucine	J	Xle
Lysine	K	Lys
Leucine	L	Leu
Methionine	M	Met
Asparagine	N	Asn
Pyrrolysine	O	Pyl
Proline	P	Pro
Glutamine	Q	Gln
Arginine	R	Arg
Serine	S	Ser
Threonine	T	Thr
Selenocysteine	U	Sec
Valine	V	Val
Tryptophan	W	Trp
Unspecified or unknown	X	Xaa
Tyrosine	Y	Tyr
Glutamine or glutamic acid	Z	Glx
N-Formylmethionine		fMet

* N-Formylmethionine has only four-letter code

Table A2: Summary of disorder-prediction methods

Xue, B., R. L. DunBrack, R.W. Williams, A.K. Dunker, and V. N. Uversky (2010) "PONDR-Fit: A meta-predictor of intrinsically disordered amino acids." Biochim. Biophys. Acta 1804(4):996-1010, PMID: 20100603	PONDR-FIT™
Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB. "Protein disorder prediction: implications for structural proteomics." Structure. 2003;11(11):1453-9, PMID: 14604535	DisEMBL™
Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. "Prediction and functional analysis of native disorder in proteins from the three kingdoms of life." J Mol Biol. 2004;337(3):635-45, PMID: 15019783	DISOPRED2
MacCallum B. "Order/Disorder Prediction With Self Organising Maps." CASP 6 meeting, Online paper	DRIPPRED
Cheng J, Sweredoski M, Baldi P. "Accurate Prediction of Protein Disordered Regions by Mining Protein Structure Data" Data Mining and Knowledge Discovery. 2005; 11(3):213-222, Online Paper	DISpro
Prilusky J, Felder CE, Zeev-Ben-Mordehai T, Rydberg EH, Man O, Beckmann JS, Silman I, Sussman JL. "FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded." Bioinformatics. 2005;21(16):3435-8, PMID: 15955783	FoldIndex©
Linding R, Russell RB, Neduva V, Gibson TJ. "GlobPlot: Exploring protein sequences for globularity and disorder." Nucleic Acids Res. 2003;31(13):3701-8, PMID: 12824398	GlobPlot 2
Dosztanyi Z, Csizmok V, Tompa P, Simon I. "IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content." Bioinformatics. 2005;21(16):3433-4, PMID: 15955779	IUPred
Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, Dunker AK. "Sequence complexity of disordered protein." Proteins. 2001;42(1):38-48, PMID: 11093259	PONDR®
Coeytaux K, Poupon A. "Prediction of unfolded segments in a protein sequence based on amino acid composition." Bioinformatics. 2005;21(9):1891-900, PMID: 15657106	PreLink
Yang ZR, Thomson R, McNeil P, Esnouf RM. "RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins." Bioinformatics. 2005;21(16):3369-76, PMID: 15947016	RONN

Vullo A, Bortolami O, Pollastri G, Tosatto S. "Spritz: a server for the prediction of intrinsically disordered regions in protein sequences using kernel machines" Nucleic Acids Res. 2006;34(Webserver Issue):W164-W168, PMID: 16844983	SPRITZ
Garbuzynskiy SO, Lobanov MY, Galzitskaya OV. "To be folded or to be unfolded?" Protein Sci. 2004;13(11):2871-77., PMID 15498936 Galzitskaya OV, Garbuzynskiy SO, Lobanov MY. "Prediction of natively unfolded regions in protein chain." Mol Biol (Mosk). 2006;40(2):341-8., PMID 16637275	FoldUnfold
Vucetic S, Brown CJ, Dunker AK, Obradovic Z. "Flavors of protein disorder." Proteins. 2003 Sep 1;52(4):573-84, PMID: 12910457	VL2
Obradovic Z, Peng K, Vucetic S, Radivojac P, Brown CJ, Dunker AK. "Predicting intrinsic disorder from amino acid sequence." Proteins. 2003;53 Suppl 6:566-72, PMID: 14579347	VL3, VL3H, VL3E
Obradovic Z, Peng K, Vucetic S, Radivojac P, Dunker AK. "Exploiting heterogeneous sequence properties improves prediction of protein disorder." Proteins. 2005;61 Suppl 7:176-82, PMID: 16187360	VSL2
M. Lobanov and O. Galzitskaya. " The Ising model for prediction of disordered residues from protein sequence alone". *Phys. Biol*. 8 (2011) 035004 (9pp).	IsUnstruct
Walsh,I., Martin,A.J., Di Domenico,T., and Tosatto, S.C. (2012) ESpritz: accurate and fast prediction of protein disorder. Bioinformatics., 28(4), 503-509.	ESpritz

(partially reproduced from http://disorder.compbio.iupui.edu/predictors.php).

Table A3: Distribution of proteins over phyla and classes

Phylum	Class	Number of viruses	Number of proteins	Length of proteins (AA)
Retro-transcribing viruses	Hepadnaviridae	12	53	20,809
	Caulimoviridae	57	257	128,320
	Retroviridae	61	230	119,194
Satellites	Satellite_Nucleic_Acids	156	168	30,943
dsDNA viruses, no RNA stage	Adenoviridae	51	1,630	548,121
	Papillomaviridae	125	871	299,201
	Phycodnaviridae	16	7,288	1,582,026
	Baculoviridae	60	8,302	2,321,064
	Caudovirales	1,269	135,640	27,989,577
	Herpesvirales	72	7,025	3,267,167
	Iridoviridae	16	2,780	767,150
	Polyomaviridae	65	332	116,481
	Poxviridae	35	7,154	2,129,256
	unclassified_dsDNA_phages	21	1,227	255,521
	unclassified_dsDNA_viruses	24	6,479	1,635,691
	Fuselloviridae	10	328	53,646
	Ligamenvirales	12	635	115,366
dsRNA viruses	Partitiviridae	37	78	41,932
	Totiviridae	42	81	64,531
	Reoviridae	58	624	401,085
ssDNA viruses	Anelloviridae	45	139	44,616
	Inoviridae	34	371	68,102
	Microviridae	18	181	33,826
	unclassified_ssDNA_viruses	63	231	56,241
	Circoviridae	56	140	35,069
	Geminiviridae	361	2,197	435,961
	Parvoviridae	76	284	144,333
ssRNA viruses	ssRNA_positive-strand_viruses_no_DNA_stage	919	3,154	2,438,730
	ssRNA_negative-strand_viruses	254	1,478	999,093
unclassified phages	Undefined	22	1,202	250,274
unclassified viruses	unclassified_Gemycircularvirus	29	67	20,312
Summary		**4,076**	**190,626**	**46,413,638**

Table A4. N-grams that occur only in disordered regions

For each length first 100 n-grams that appear only in disordered regions sorted according their mole fractions in descending order are presented, except for length 4 where only four such n-grams exist.

N-gram length						
4	5	6	7	8	9	10
HHHH	GGGGG	GGGGGG	GGGGGGG	SSSSSSSS	SSSSSSSSS	SSSSSSSSSS
SNAM	PPPPP	PPPPPP	PPPPPPP	GGGGGGGG	PPPPPPPPP	PEPEPEPEPE
GHMA	APAPA	TTTTTT	EEEEEEE	PPPPPPPP	PEPEPEPEP	EPEPEPEPEP
GSHM	PSPPP	PEPEPE	DDDDDDD	EEEEEEEE	EPEPEPEPE	EEEEEEEEEE
	NNNNN	EPEPEP	PEPEPEP	PEPEPEPE	EEEEEEEEE	KPAPKPAPKP
	EEEED	GGGGGA	EPEPEPE	EPEPEPEP	PKPAPKPAP	PKPAPKPAPK
	PPAPP	PKPAPK	PKPAPKP	DDDDDDDD	PAPKPAPKP	PPPPPPPPPP
	SSTSS	KPAPKP	TTTTTTT	PKPAPKPA	KPAPKPAPK	PAPKPAPKPA
	KKKKK	AGGGGG	PPPSPPP	KPAPKPAP	GGGGGGGGG	APKPAPKPAP
	DEEDE	PPSPPP	KPAPKPA	APKPAPKP	APKPAPKPA	DDDDDDDDDD
	PAPPP	APKPAP	PAPKPAP	PAPKPAPK	DDDDDDDDD	PSPPPPSPPP
	KKSKK	PSPPPP	APKPAPK	TTTTTTTT	QQQQQQQQQ	PPSPPPPSPP
	EEDDD	GGGGGS	QQQQQQQ	QGAKSSSD	PPPPSPPPP	PPPSPPPPSP
	NSSSS	APAPAP	GGGGGGA	QQQQQQQQ	PSPPPPSPP	QQQQQQQQQQ
	PPAAP	SGGGGG	PPSPPPP	PPPSPPPP	TTTTTTTTT	SPPPPSPPPP
	KKEKK	PAPAPA	PPPPSPP	PPPPSPPP	PPSPPPPSP	GGGGGGGGGG
	SKKKK	PPPPSP	MDSRTGE	RSRSRSRS	PPPSPPPPS	PPPPSPPPPS
	ESSSS	NNNNNN	PAPAPAP	GGGGGGGA	SPPPPSPPP	RSRSRSRSRS
	RRRGR	SRSRSR	GAKSSSD	PSPPPPSP	RSRSRSRSR	APAPAPAPAP
	AAPPA	QGAKSS	QGAKSSS	PPSPPPPS	PAPAPAPAP	TTTTTTTTTT
	GGGDD	SPPPPS	RSRSRSR	SPPPPSPP	SRSRSRSRS	PAPAPAPAPA
	EEEEG	SDSDSD	AGGGGGG	AGGGGGGG	APAPAPAPA	SARGGQQTAN
	HHHHH	PQGPQG	GAGGGGG	APAPAPAP	KGDKGDKGD	SRSRSRSRSR
	PPPPQ	DSDSDS	SRSRSRS	SRSRSRSR	NNNNNNNNN	PPSPPPSPPP
	STTST	GGGGGR	GPQGPQG	GPQGPKGD	SARGGQQTA	NNNNNNNNNN
	DSDEE	TGGGGG	PSPPPPS	PAPAPAPA	PSPPPSPPP	TTAATTTAAT
	DEDSD	RGGGGG	NNNNNNN	NNNNNNNN	GGGGGGGGA	TAATTTAATT
	DDDDK	GPQGPK	APAPAPA	KGDKGDKG	DSDSDSDSD	PPPSPPPSPP
	DDDKD	PQGPKG	GGGGGGS	KGDKGDTG	PPPSPPPSP	DIVISTPASK
	REEEE	SPPPSP	SPPPPSP	GAGGGGGG	PPSPPPSPP	AATTTAATTT
	SPSPG	PSPPPS	SGGGGGG	SARGGQQT	RSARGGQQS	ADIVISTPAS
	EKKKS	EEEEED	GPQGPKG	SDSDSDSD	TTAATTTAA	SARGGQQSAN
	GGSRS	AAPAPA	GGGGGSG	SGGGGGGG	AATTTAATT	VISTPASKVR
	SSSVD	PPPPPS	GGGGGAG	GGAGGGGG	RGGQQSAND	ARGGQQSAND
	EAEED	APPPPP	GGGGAGG	SPPPSPPP	GAGGGGGGG	IVISTPASKV
	PSPEP	SSSSSD	DSDSDSD	PPPSPPPS	ATTTAATTT	RSARGGQQSA
	NTERH	TTAATT	PQGPKGD	PSPPPSPP	DIVISTPAS	TTTAATTTAA
	PQQQP	SPPPPP	PSPPPSP	DSDSDSDS	SDSDSDSDS	ATTTAATTTA
	KKKAA	GGGGGY	SDSDSDS	PPSPPPSP	IVISTPASK	MSKRPADIVI
	PAATS	SSSTSS	MSKRPAD	GGGGGGAG	ADIVISTPA	SDSDSDSDSD
	TPEPP	QGPQGP	KGDKGDK	TNGIEPPR	SQLKGSSST	SKRPADIVIS
	QQEEE	PPPPPA	DKGDKGD	GGGGGGGS	ARGGQQSAN	DSDSDSDSDS
	KKTSS	MSKRPA	KGDKGDT	GGGGGAGG	TTTAATTTA	PADIVISTPA
	PKPRP	SKRPAD	PPSPPPS	GPQGPQGP	SARGGQQSA	ELNPAPTSSP
	RGEET	EEEEDE	DEDEDED	QGPKGDTG	VISTPASKV	RYGGGGGGGG
	KPTPP	PKGDTG	TNGIEPP	TTAATTTA	AGGGGGGGG	NSTNGIEPPR
	KRPPP	TGPQGP	PTPSPTP	TTTAATTT	SKRPADIVI	ISLGSGLSMS
	APEDP	LPPPPP	GPAGPQG	ATTTAATT	MSKRPADIV	PADTPVSEIP
	MEEEE	TPPPTP	SPPPSPP	SARGGQQS	ALRRRLERG	SSRASSRASS
	THMPR	YGGGGG	GGGGSGG	GGQQSAND	MPKRDAPWR	SQLKGSSSTS
	KKGKS	PPTPPP	GGSGGGG	RGGQQSAN	GGGGGGGAG	SILEEAQRLI
	KSASS	SSSSGS	YGGGGGG	VISTPASK	LNPAPTSSP	ESILEEAQRL
	QQPPQ	EDEEEE	SSSSSSD	IVISTPAS	SGGGGGGGG	ILEEAQRLIH
	DSPPS	APAPAA	PTPPPTP	DIVISTPA	YGGGGGGGG	PPGPEEGEGP
	PEPPS	GGGGGD	GGGGGGR	ARGGQQSA	SRASSRASS	NSGYRYGGGG
	SSEKP	PGGGGG	TTAATTT	SPASMEGN	TGPQGPKGD	LEEAQRLIHG
	DSPPP	PPAPPA	NGIEPPR	QLKGSSST	DKGDKGDTG	SGYRYGGGGG
	NKGPE	NGIEPP	GSGGGGG	GGGGGGSG	EEQKQLTLF	SSQVSNSTNG
	AQAQE	SPSPPP	GPEGPEG	EDEDEDED	QGPKGDKGD	GPPGPEEGEG
	GPSSG	GGGGGV	TTTAATT	GPEGPEGP	SSRASSRAS	YRYGGGGGGG
	SPEPP	PAGPQG	PQGPQGP	EGPEGPEG	STNGIEPPR	GYRYGGGGGG
	SQPEE	SPSPSP	PPPPLPP	YGGGGGGG	GGAGGGGGS	DISLGSGLSM
	LMPCE	GQQTAN	PPPPPPS	AGTSKVSR	APAAPAAPA	PCESSSQVSN
	GPLGS	PPPTPP	SPPPPPP	SKRPADIV	TNGIEPPRG	KGDKGDKGDT
	KRPGP	EDEEDE	DDEDDED	SSSSSSSD	GDGDGDGDG	GDGDGDGDGD
	PKRPR	GGGGGN	GGGGGGY	QGPKGDKG	RYGGGGGGG	SSSQVSNSTN
	VASMQ	PTPPPT	QPEESVG	DEDEDEDE	NSTNGIEPP	QLKGSSSTSS
	KGPPY	PPPPAP	SSRASSR	AGGGGGSG	PAPVPKPAP	MPCESSSQVS
	VKGPP	QGIQGP	PPPTPPP	ASSRASSR	ILEEAQRLI	SRASSRASSR
	QQPQA	EDEDEE	QGPKGDT	GAGGGGGS	GAGGGGGSG	STNGIEPPRG
	GVPRG	QPEESV	ASSSSSS	MPKRDAPW	ADTPVSEIP	CESSSQVSNS
	QPRRR	EGPEGP	SSSSSSA	RASSRASS	RASSRASSR	GEGGEGGEGG
	AHSTQ	SSSSDS	PSSSSSS	TSSSSSSS	PADTPVSEI	DGDGDGDGDG
	EPRHH	PAPPPP	QPQPEES	GPTGPTGP	QLKGSSSTS	SNSTNGIEPP
	ESPPP	APTSSP	TGGGGGG	LRRRLERG	ISLGSGLSM	ESSSQVSNST

PKPPE	PPPPPL	RGGGGGG	NPAPTSSP	DGDGDGDGD	GGEGGEGGEG
QQTQQ	SRASSR	ARGGQQS	ALRRRLER	DISLGSGLS	PSPPPSPPPS
DDQAS	PEGPEG	DDDDDDE	SRASSRAS	SLGSGLSMS	TDISLGSGLS
EPEEM	VGGGGG	GGQQSAN	PKRDAPWR	PGPEEGEGP	QTANDAAAEA
PQSPS	GPEGPE	GQQSAND	GPQGIQGP	SSQVSNSTN	GGAGGGGGSG
QPPRR	SSSSSE	RGGQQSA	GGGGGGGY	SILEEAQRL	AGGGGGSGRR
SQPSQ	TPPPPP	EDDEDDE	PAPVPKPA	GGGGGSGRR	SQVSNSTNGI
EMNRQ	NGGGGG	AGGGGGS	GDGDGDGD	GYRYGGGGG	GAGGGGGSGR
EQKES	RRSPSP	GPQGPAG	SSRASSRA	GPPGPEEGE	GGAGAGGGAG
GHMAS	GGGGGL	VISTPAS	DKGDKGDT	EEAQRLIHG	QVSNSTNGIE
MEGRE	AGPQGP	PVPKPAP	VQPQPEES	SQVSNSTNG	QSGTSARRAE
PASQP	DEEEED	LKGSSST	DEDDEDDE	SGYRYGGGG	EGGEGGEGGE
PSRPR	PPPPPT	PSPPPPP	PPPLPPPP	GAGAGGGAG	LMPCESSSQV
QPPEE	QQQQQP	IVISTPA	EEQKQLTL	ESILEEAQR	RHKLAEKRAR
SPPQP	PSPSPS	SPASMEG	TGPQGPKG	GEGGEGGEG	ATDISLGSGL
AQQQT	GGGGGT	PASMEGN	APTSSPTS	PPGPEEGEG	ALRRRLERGE
EPKKP	RSPSPR	PGGGGGG	APVPKPAP	NSGYRYGGG	PAAPAAPAAP
KGPEQ	DEEDEE	RRRSSGG	STNGIEPP	LEEAQRLIH	VSNSTNGIEP
QQQAS	EEDEDE	ENTERHT	GPAGPQGP	YRYGGGGGG	ASSRASSRAS
QREQM	GGGGGP	KRDAPWR	NGIEPPRG	NATNGIEPP	GNEMVLPAET
RYCRK	QPQPEE	KRPADIV	RYGGGGGG	CESSSQVSN	MVLPAETRPG
SPEPA	PPPPTP	PPLPPPP	PSSSSSSS	SNSTNGIEP	QATEFDSPFA
AGHQQ	TTTPTT	EGPEGPE	NSTNGIEP	TANDAAAEA	VLPAETRPGA
RQQQE	KKKKKK	GTSKVSR	ISLGSGLS	SSSQVSNST	NGAAAREQAT
RRHHH	EEEEDD	EEEEEED	GGEGGEGG	PCESSSQVS	TEFDSPFADR

Table A5. N-grams with positive disorder fractional difference

Table includes for each length first 100n-grams occurring both in disordered and ordered regions with positive disorder/order fractional difference sorted according mole fractions in descending orders, except for length one where 11 monograms exists.

N-gram length									
1	2	3	4	5	6	7	8	9	10
S	SS	SSS	GGGG	SSSSS	SSSSSS	AKSSSDV	AKSSSDVK	HPNIQGAKS	FHPNIQGAKS
E	EE	GGG	SSSS	EEEEE	DDDDDD	KSSSDVK	PNIQGAKS	FHPNIQGAK	AHFHPNIQGA
A	AA	PPP	PPPP	QQQQQ	EEEEEE	NIQGAKS	HPNIQGAK	HFHPNIQGA	SAHFHPNIQG
K	KK	EEE	DDDD	PEPEP	PAPKPA	PNIQGAK	RSARGGQQ	GRSARGGQQ	DGRSARGGQQ
R	GG	RRR	EEEE	PPPSP	KGDKGD	HPNIQGA	FHPNIQGA	AHFHPNIQG	IDGRSARGGQ
P	AS	DDD	APAP	PAPKP	RSRSRS	SARGGQQ	HFHPNIQG	DGRSARGGQ	AVSQLKGSSS
D	RR	PAP	PAPA	PKPAP	AKSSSD	RSARGGQ	GRSARGGQ	TPASKVRRR	TPASKVRRRL
G	SA	APA	QQQQ	PPSPP	MDSRTG	GPKGDKG	SQLKGSSS	TAATTTAAT	SAVSQLKGSS
T	EA	KKK	PPPS	SPPPP	KSSSDV	ISTPASK	TPASKVRR	VSQLKGSSS	ISASAYNGND
Q	AE	SPS	PPAP	PPPPS	DSRTGE	GDKGDTG	PASKVRRR	AVSQLKGSS	SASAYNGNDT
M	KE	PSP	PSPP	KPAPK	GPQGPQ	TPASKVR	AATTTAAT	PASKVRRRL	PASKVRRRLN
	SG	PSS	PEPE	SRSRS	IQGAKS	QLKGSSS	TAATTTAA	SAVSQLKGS	ISIRTFRELN
	PP	PTP	PPSP	QGPQG	NIQGAK	SQLKGSS	VSQLKGSS	IIISTPASK	ASKVRRRLNF
	SE	EED	SPPP	APKPA	PNIQGA	EDEDEDE	AVSQLKGS	ISASAYNGN	IEQSVISASA
	PS	SSP	GPQG	PPPPA	DEDEDE	PASKVRR	KSYIDKDG	SASAYNGND	EQSVISASAY
	SK	PPA	APPP	PPPTP	EDEDED	APAAPAA	ASKVRRRL	ASKVRRRLN	QSVISASAYN
	EK	PPS	PAPP	RRRSS	SARGGQ	AAPAAPA	SAVSQLKG	ASAYNGNDT	SVISASAYNG
	ES	APP	PSSS	PSPSP	ARGGQQ	GPQGIQG	GGAGAGGG	SIRTFRELN	VISASAYNGN
	DE	KRK	EEED	QGPKG	TSSSSS	SDWSFLK	IIISTPAS	IEQSVISAS	RPMNRKPRMY
	GS	SRS	EPEP	RSPSP	DEEEEE	SDVKSYI	IISTPASK	SVISASAYN	PMNRKPRMYR
	DD	SES	PTPP	PPPPT	PAAPAA	GPTGPTG	GAGAGGGA	QSVISASAY	LSAVSQLKGS
	SD	PEP	RRRS	QGAKS	ASSSSS	ASKVRRR	IEQSVISA	KSYIDKDGD	NLSAVSQLKG
	ST	TPP	PPPA	MDSRT	PKGDKG	ATTTAAT	EQSVISAS	EQSVISASA	STHFHPNIQG
	AP	SSE	SPSP	SPSPS	DDEDDE	KSYIDKD	VISASAYN	VISASAYNG	SSTWYPQPGQ
	KS	RRS	PPTP	PQGPQ	PSSSSS	ASSRASS	SVISASAY	MNRKPRMYR	LNERTATETR
	PA	APS	SSSP	SPSPP	DDEDED	GGRGGGG	QSVISASA	RPMNRKPRM	KLNERTATET
	KA	EES	PAPK	PAPEP	DKGDKG	NDDDDDD	NRKPRMYR	PMNRKPRMY	EDIKGYKPHT
	AK	ESS	SSPS	PPPPR	GDKGDK	AVSQLKG	TGPTGPTG	LSAVSQLKG	IEDIKGYKPH
	AR	RKR	APKP	MSKRP	SSSSSA	GGGAGAG	PMNRKPRM	NLSAVSQLK	ANLSAVSQLK
	SP	RSS	PSPS	PQGPK	GPKGDK	ALRRRLE	TTFKDSTG	LTASDWSFL	NGNIHVSKLP
	ED	QQQ	KPAP	RRSSS	STPASK	SAVSQLK	MNRKPRMY	THFHPNIQG	LIAARGYVYT
	TS	RSR	PKPA	PPRPP	ISTPAS	AGAGGGA	TEFDSPFA	KLNERTATE	EFGFDGGDSE
	RA	SPP	PPPT	PQPQP	DKGDTG	DDDDDDN	LSAVSQLK	DIKGYKPHT	AARGYVYTAA
	SR	PKP	PRRR	QQQQP	GDKGDT	SSSSSAS	GDKGDTGA	NERTATETR	ENGNIHVSKL
	KR	EPE	PAPS	PPQPP	PLPPPP	ESILEEA	AGGAGAGG	LNERTATET	AVLIAARGYV
	ER	KPK	PSSP	QPQPQ	TPSPTP	IISTPAS	TTGLSKAK	SSTWYPQPG	VLIAARGYVY
	RS	RRK	PPGP	PPPEP	PTPSPT	IIISTPA	GFDGGDSE	EDIKGYKPH	IAARGYVYTA
	PE	PRP	PPSS	PQQQQ	LKGSSS	GDDDDDD	SSTWYPQP	DAEQRELLD	TVTITADVRD

	TP	PPT	PQGP	SPPSP	QLKGSS	LEEAQRL	LNERTATE	IEDIKGYKP	ESGHIQEFDD
	DS	MSK	PPPR	QQQQR	RRRSSG	DDDDDDG	NLSAVSQL	NGNIHVSKL	NRPMNRKPRM
	RK	PRR	SPTP	QPQQQ	PASKVR	IEQSVIS	LTASDWSF	ARGYVYTAA	SGLLDDGANY
	RE	EEQ	PRPP	GPPPP	SSRASS	GGGSGRR	IKSTDSTI	VLIAARGYV	MSGLLDDGAN
	GE	PPR	SPPS	PPPQP	GGQQTA	QSVISAS	KLNERTAT	FGFDGGDSE	RPGESWASRS
	PT	EQE	PEPP	PEPPK	GSSSSS	GPTGPAG	DIKGYKPH	PEFGFDGGD	EAVTDALSPA
	TE	RRP	SSPP	ESSEE	SGSSSS	EQSVISA	IKGYKPHT	ENGNIHVSK	AEAVTDALSP
	ET	PKK	PKPK	PSPPS	ASKVRR	SVISASA	ERTATETR	AARGYVYTA	AELEAEAVTD
	SN	PGP	PPEP	RRRSP	GGGGSS	GTSARRA	NERTATET	ANLSAVSQL	LAELEAEAVT
	EQ	MKK	PQPP	PPAPR	SSSSSC	VISASAY	IEDIKGYK	IAARGYVYT	GGGDPEDIER
	MS	RPR	PPQP	PPTPS	EEEDEE	APAAAPA	PPLGLTDP	LIAARGYVY	EAEAVTDALS
	KP	QEE	RPPP	EGEGP	EEEDED	FNVPQKH	DAEQRELL	AVLIAARGY	TLAELEAEAV
	DP	RPS	PPRR	PPEPP	STSSSS	AESKEEA	PSDWSFLK	EFGFDGGDS	HISIQTFREL
	PK	RPP	GPPP	PRPPP	RRRSRS	TFKDSTG	EDIKGYKP	TLAELEAEA	PSRSAHFHPN
	RP	EEP	PPPK	PEGPE	SRTGEL	TTFKDST	LDPGDSAS	TVTITADVR	MIEDIKGYKP
	DK	PDP	QPQP	PPPRR	SEEEEE	QYLVTTF	KSPLFQDN	PLAESARAV	LEAEAVTDAL
	QA	PPE	PPPQ	PQPEE	PAPAAP	TGPTGPT	AARGYVYT	GDFTPKPGA	ELEAEAVTDA
	PR	RSP	AQQQ	ESSSQ	DEDDDD	NRKPRMY	ANLSAVSQ	ESGHIQEFD	AVTDALSPAD
	TK	MSS	QPPP	RPPSP	SSSSAS	MNRKPRM	VLIAARGY	GLLDDGANY	SRSAHFHPNI
	EP	PQP	QGPQ	EPKEE	SSSSTS	RPGAVKG	LIAARGYV	RPGESWASR	KEGIPPDQQR
	RG	KKP	PKPP	QQPPP	DWSFLK	TTGLSKA	RGYVYTAA	SGLLDDGAN	QVPIKVQHRL
	PG	QSS	QQQP	EPEEP	GGFGST	TEFDSPF	ARGYVYTA	MSGLLDDGA	GLLDDGANYE
	AQ	PSR	PQQQ	KPKPT	EEEDDE	REQATEF	NGNIHVSK	KEGIPPDQQ	ATETRRGVAE
	NS	MSD	PGPP	PPSYE	IVISTP	EFDSPFA	IAARGYVY	LPEFGFDGG	DTRTRDTHRH
	PD	SPR	KRPR	PPEPE	PEESVG	LSAVSQL	FGFDGGDS	GGGDPEDIE	TRTRDTHRHL
	SQ	SRP	PEEP	PEPDP	GPTGPT	DKGDTGA	LAESARAV	SRSAHFHPN	RTRDTHRHLD
	QS	SSQ	EPPP	PPQPQ	SASSSS	NREQIEQ	LAELEAEA	HISIQTFRE	KLANLSAVSQ
	QE	PPK	RRPP	EPPPP	DEDDDE	NFFEKRV	TVTITADV	LEAEAVTDA	TRDTHRHLDG
	KQ	EPP	PKPS	SPPPT	GPQGPA	GFDGGDS	TLAELEAE	GGDPEDIER	LANLSAVSQL
	EN	MSE	PQPQ	PPQQP	SSSSSP	LNKMLKG	DFTPKPGA	TQAVSQRLL	SIRTFRELNQ
	MA	EPS	PEPS	PVGPQ	PTSSPT	TGLSKAK	PLAESARA	EAEAVTDAL	KKLNERTATE
	QQ	SEP	QQPQ	RQQQQ	VPKPAP	GPQGPRG	GDFTPKPG	EAVTDALSP	NLPGIREVLK
	NE	MTT	QPQQ	APSKP	NPAPTS	INALRRR	SGLLDDGA	AEAVTDALS	VAARDGDDAI
	GP	KRP	QQQR	PEGPQ	PTPPPP	PASAEAI	ESGHIQEF	AELEAEAVT	LNLPGIREVL
	NP	GPP	EPSP	PPKPK	TGPAGP	FDGGDSE	HRSVSHQL	MIEDIKGYK	NTHDTNMRDD
	QR	EQQ	DPPP	MEGNR	EKEKEK	GGGGASS	RPGESWAS	ELEAEAVTD	ETRRGVAEIA
	MK	EPK	PQQP	PPPKR	GGRRRS	PAPTPAP	DAYAAALN	LRYPGGKSR	QHISIRTYRE
	QK	MSN	PSQP	QQQRP	PAPAPP	LNERTAT	KEGIPPDQ	AVTDALSPA	KDTRTRDTHR
	RQ	QQR	QGPK	PRPGE	GEGPGG	LAFLKSI	GLLDDGAN	LAELEAEAV	TRRGVAEIAT
	MT	PPQ	QQEE	PSRSP	KGSSST	RLEENDK	LLDDGANY	PSRSAHFHP	RRGVAEIATQ
	QP	QPP	QPEE	QQEQQ	DSDDSD	NERTATE	EGIPPDQQ	AQRRGRVGR	THDTNMRDDL
	PQ	QPS	PNPP	PPPPD	APAKKA	DEDDDED	MSGLLDDG	VTDALSPAD	IRTFRELNQA
	TQ	QEQ	QPSP	RPPPR	PAGPAG	LTASDWS	RYPGGKSR	GTSSTTACV	ISLEEVKDQN
	QT	SQP	RSPP	PQQPP	EDEDDD	LSSSSSS	TFRELNQA	EGIPPDQQR	HDTNMRDDLD
	QG	QQE	PPQQ	RQQRE	LNPAPT	LSKAKRR	GGGDPEDI	SAGEHFNPT	TETRRGVAEI
	PN	QPA	SPQP	EPEAP	KPKPKP	NLSAVSQ	GGDPEDIE	TETRRGVAE	ISIQTFRELN
	ME	PSQ	KPPP	PPPKK	EGGEGG	LEENDKT	GGRGGRGG	ATETRRGVA	CTQVPIKVQH
	QD	SPQ	KQQQ	RSPVR	DIIIST	EDVNSLV	PSHTAGGT	RLVKRAERR	GQHISIRTYR
	MN	KPP	QQPP	QPPQP	PAPKPK	GRGGRGG	QMLSSLLV	LLDDGANYE	GISLEEVKDQ
	NQ	QQS	QSQP	DPEPE	RASSRA	GSSGGGG	QAVSQRLL	DTRTRDTHR	DVAARDGDDA
	DQ	QQP	QPPQ	PRGPP	DDSDDD	KLNERTA	SVLQQIGS	TRDTHRHLD	LDVAARDGDD
	QN	GPQ	QPEP	PPAKR	GPTGPQ	DPGDSAS	PSRSAHFH	QTKERLTSP	DTSRVTVRRL
	MR	PQQ	QPRP	PEPQP	SASSAS	TFRELNP	VTDALSPA	TRTRDTHRH	RRLNFDSPYS
	MD	QGP	PKPQ	PRQPR	GPPGPP	KSTDSTI	EAVTDALS	HGGLRADSD	DKNNVPYKKE
	MP	SQQ	EEPQ	EPDPE	EDEEED	SALSSLA	LRYPGGKS	LANLSAVSQ	ETESGHIQEF
	EM	QSP	MWDP	TTHMP	PPGPPG	KGYKPHT	TQAVSQRL	KLANLSAVS	GKFGATASDI
	SM	QPQ	HPPP	PKPKA	ERERER	DDDDDND	SRSAHFHP	RDTHRHLDG	TLDKNNVPYK
	MQ	QRQ	PSQQ	QQQAP	SDSDDD	IKGYKPH	ELEAEAVT	RTRDTHRHL	DTNMRDDLDE
	HP	MSQ	PQRP	KREMK	RRRRRS	SPLFQDN	AELEAEAV	DKNNVPYKK	ERYRLVHPTG
	PM	QPR	QEEP	QQQQH	SSPSSS	DIKGYKP	GDPEDIER	GVSAGRGFG	LGISLEEVKD
	QM	EPQ	KGPP	SPPQA	NDAAAE	SPSVKSV	MIEDIKGY	PGQHISIQT	YTLDKNNVPY
	MM	MPP	MMPP	SSRPP	PSLDDI	PTPPPSP	GDTGATGP	NLPGIREVL	METESGHIQE

Table A6. N-grams that appear only in ordered regions

For each length first 100 n-grams that appear only in ordered regions sorted according their mole fractions in descending order are presented, except for length 3 where only 10 such n-grams exist.

N-gram length							
3	4	5	6	7	8	9	10
WIC	IFII	YNVID	IKGGIP	LYMACTH	LYMACTHA	THASNPVYA	LYMACTHASN
YCW	FINY	IKGGI	NVIDDV	YMACTHA	HASNPVYA	YMACTHASN	THASNPVYAT
WCY	FVFL	VGKRF	YNVIDD	QIKGGIP	MACTHASN	HASNPVYAT	CTHASNPVYA
WYW	ILYV	ATLKI	YMACTH	ASNPVYA	NPVYATLK	SNPVYATLK	HASNPVYATL
CWF	FIII	ACTHA	LYMACT	ACTHASN	SNPVYATL	ASNPVYATL	ASNPVYATLK
FWW	LLLW	GKRFC	QIKGGI	SNPVYAT	ASNPVYAT	THRVGKRFC	NPVYATLKIR

CWY	YILV	LYMAC	SNPVYA	NPVYATL	THRVGKRF	NPVYATLKI	SNPVYATLKI
CYW	LYVY	MACTH	MACTHA	PVYATLK	HRVGKRFC	PVYATLKIR	TLKIRIYFYD
HWW	VLAC	YMACT	NPVYAT	YNVIDDV	VYATLKIR	LKIRIYFYD	VYATLKIRIY
CWW	YYVL	ALLLY	CTHASN	NHTENAL	PVYATLKI	TLKIRIYFY	ATLKIRIYFY
	VLLC	KYENH	PVYATL	THRVGKR	KIRIYFYD	YATLKIRIY	YATLKIRIYF
	IFLC	GPHNY	NHTENA	HRVGKRF	LKIRIYFY	VYATLKIRI	PVYATLKIRI
	CVLV	RFFDL	HTENAL	WMDENIK	YATLKIRI	ATLKIRIYF	NHTENALLLY
	FVIF	PHNYL	THRVGK	RVGKRFC	ATLKIRIY	YENHTENAL	HRVGKRFCVK
	FIVF	IYFYD	SDVTRG	LGPHNYL	TLKIRIYF	KYENHTENA	THRVGKRFCV
	TMWA	KIRIY	HRVGKR	YATLKIR	GKRFCVKS	NHTENALLL	RVGKRFCVKS
	YAYI	TMWAR	RVGKRF	VYATLKI	YENHTENA	HRVGKRFCV	HTENALLLYM
	VFYL	RIYFY	GPHNYL	IRIYFYD	ENHTENAL	HTENALLLY	KYENHTENAL
	VVIF	LLYMA	VGKRFC	KIRIYFY	ENALLLYM	RVGKRFCVK	TENALLLYMA
	IVIF	PLYFK	LGPHNY	LKIRIYF	KYENHTEN	TENALLLYM	ENALLLYMAC
	LFIY	LYFKI	WMDENI	TLKIRIY	TENALLLY	VGKRFCVKS	ALLLYMACTH
	VLYY	RFCVK	MDENIK	ATLKIRI	NHTENALL	ENALLLYMA	NALLLYMACT
	YAIY	KIWMD	YATLKI	KRFCVKS	HTENALLL	LLYMACTHA	LLLYMACTHA
	LYYY	GKIWM	ATLKIR	GKRFCVK	RVGKRFCV	NALLLYMAC	GKIWMDENIK
	YYVV	LLLYM	RIYFYD	VTGGQYA	VGKRFCVK	LLLYMACTH	HNYLCGHLDL
	FAII	IWMDE	IRIYFY	YENHTEN	LLYMACTH	ALLLYMACT	GGIPTIFLCN
	LICL	NNVIR	KIRIYF	ENHTENA	NALLLYMA	GKIWMDENI	KGGIPTIFLC
	YYIL	NYLCG	LKIRIY	ENALLLY	GKIWMDEN	KIWMDENIK	KHFKEFMGAQ
	VAYY	HNYLC	NPLYFK	KYENHTE	LLLYMACT	HNYLCGHLD	IKGGIPTIFL
	WYVD	IFLCN	RFCVKS	HTENALL	ALLLYMAC	NYLCGHLDL	LKHFKEFMGA
	LLWL	LCGHL	GKRFCV	NALLLYM	KIWMDENI	GGIPTIFLC	NYLCGHLDLS
	YINF	TIFLC	KYENHT	TENALLL	IWMDENIK	GIPTIFLCN	GIPTIFLCNP
	IYIV	YLCGH	KRFCVK	VGKRFCV	HNYLCGHL	KGGIPTIFL	WARSLGPHNY
	YLVF	AQRDW	VTGGQY	GKIWMDE	NYLCGHLD	LKHFKEFMG	ARSLGPHNYL
	YIYT	KNYFL	GKIWMD	LLYMACT	YLCGHLDL	KHFKEFMGA	LGPHNYLCGH
	YLYF	YLKHF	NALLLY	ALLLYMA	GIPTIFLC	HFKEFMGAQ	YLCGHLDLSP
	YFTF	HFKEF	ENALLL	KIWMDEN	GGIPTIFL	IKGGIPTIF	GPHNYLCGHL
	YIGF	NYFLT	ALLLYM	LLLYMAC	IPTIFLCN	YLCGHLDLS	YGKPVQIKGG
	IAWL	FMGAQ	YENHTE	IWMDENI	WARSLGPH	WARSLGPHN	KYGKPVQIKG
	LVFY	FLCNP	ENHTEN	HNYLCGH	KHFKEFMG	GPHNYLCGH	PHNYLCGHLD
	FVFI	FFDLV	TENALL	NYLCGHL	KGGIPTIF	ARSLGPHNY	KPVQIKGGIP
	YGIF	CRELH	KIWMDE	YLCGHLD	LKHFKEFM	IPTIFLCNP	GKPVQIKGGI
	YIAI	YFLTY	LLYMAC	NPLYFKI	IKGGIPTI	RSLGPHNYL	GKRFCVKSVY
	YYVY	NHNLR	LLLYMA	LCGHLDL	FKEFMGAQ	LCGHLDLSP	QIKGGIPTIF
	VWLA	FGQVF	IWMDEN	PTIFLCN	HFKEFMGA	LGPHNYLCG	LGKIWMDENI
	LVCV	LGKIW	HNYLCG	GGIPTIF	LCGHLDLS	GKPVQIKGG	SLGPHNYLCG
	YFVI	LLLLV	NYLCGH	IPTIFLC	VTGGQYAS	YGKPVQIKG	RSLGPHNYLC
	LLWF	WYNVI	LKHFKE	GIPTIFL	GPHNYLCG	PHNYLCGHL	YENHTENALL
	VCVL	VVYNH	PLYFKI	LKHFKEF	ARSLGPHN	KYGKPVQIK	ENHTENALLL
	YYYL	FNHNL	LCGHLD	KGGIPTI	PHNYLCGH	KPVQIKGGI	WYNVIDDVDP
	IYFV	VISIN	YLCGHL	KHFKEFM	RSLGPHNY	GKRFCVKSV	LCGHLDLSPK
	LHYY	AWYNV	CGHLDL	ARSLGPH	PTIFLCNP	PVQIKGGIP	YNVIDDVDPH
	RIYY	IQFEG	IFLCNP	KEFMGAQ	CGHLDLSP	LGKIWMDEN	YLKHFKEFMG
	CIAL	LILLL	GAQRDW	WARSLGP	WYNVIDDV	AWYNVIDDV	DDVDPHYLKH
	FIAY	QVFNM	KHFKEF	HFKEFMG	SLGPHNYL	QIKGGIPTI	VGKRFCVKSV
	LIFY	VYNHQ	GGIPTI	IKGGIPT	NVIDDVDP	KRFCVKSVY	IDDVDPHYLK
	CLAI	DPHYL	PTIFLC	FKEFMGA	LGPHNYLC	NVIDDVDPH	AWYNVIDDVD
	ICAL	FLRVF	TIFLCN	CGHLDLS	KPVQIKGG	YNVIDDVDP	HFKEFMGAQR
	LTWL	HVLIQ	EFMGAQ	TGGQYAS	KYGKPVQI	SLGPHNYLC	KEFMGAQRDW
	LYYF	LHVLI	KGGIPT	IDDVDPH	GKPVQIKG	ENHTENALL	FKEFMGAQRD
	DIIC	VLIQF	GIPTIF	PHNYLCG	VQIKGGIP	WYNVIDDVD	DVDPHYLKHF
	FYVI	ICREL	HFKEFM	LGKIWMD	YGKPVQIK	CGHLDLSPK	VDPHYLKHFK
	TFIY	AGKYE	RSLGPH	RSLGPHN	KRFCVKSV	YLKHFKEFM	PHYLKHFKEF
	YIPI	VVVVV	KEFMGA	GPHNYLC	LGKIWMDE	DDVDPHYLK	GKTMWARSLG
	LAWI	RCMLA	GGQYAS	GHLDLSP	PVQIKGGI	DVDPHYLKH	HYLKHFKEFM
	FIYF	GFRCM	FKEFMG	TIFLCNP	RFCVKSVY	KEFMGAQRD	VIDDVDPHYL
	LCVI	HTNSV	GHLDLS	SLGPHNY	QIKGGIPT	EFMGAQRDW	DPHYLKHFKE
	LFCL	FRCML	DDVDPH	WYNVIDD	AWYNVIDD	IDDVDPHYL	NVIDDVDPHY
	YLFV	LIIGL	YNHQEA	KYGKPVQ	VIDDVDPH	FKEFMGAQR	CGHLDLSPKV
	FAVY	CGCSY	VQIKGG	PVQIKGG	YNVIDDVD	GKTMWARSL	GHLDLSPKVY
	YYEI	ILSLI	KPVQIK	NVIDDVD	KNYFLTYP	VDPHYLKHF	TMWARSLGPH
	VWVV	LLVVL	GKPVQI	KPVQIKG	YLKHFKEF	DPHYLKHFK	MWARSLGPHN
	LYML	GCGKT	PHNYLC	NYFLTYP	GHLDLSPK	PHYLKHFKE	KTMWARSLGP
	IGYF	DAWYN	SLGPHN	YGKPVQI	DVDPHYLK	HYLKHFKEF	LHVLIQFEGK
	CFAL	IIILL	HLDLSP	VIDDVDP	KEFMGAQR	VIDDVDPHY	GKYENHTENA
	FTYI	YVLGK	WYNVID	RFCVKSV	DDVDPHYL	KTMWARSLG	GFTHRGTHHC
	FYLY	GITHR	LGKIWM	VQIKGGI	VDPHYLKH	GHLDLSPKV	ITHRVGKRFC
	YFLF	NDAWY	KNYFLT	GKPVQIK	EFMGAQRD	QEAGKYENH	GITHRVGKRF
	FYIV	VHGFR	YGKPVQ	FCVKSVY	IDDVDPHY	HLDLSPKVY	QEAGKYENHT
	IWEI	FLLLL	VIDDVD	AWYNVID	FMGAQRDW	HVLIQFEGK	AGKYENHTEN
	YLCD	HGFRC	YFLTYP	YLKHFKE	GKTMWARS	TMWARSLGP	EAGKYENHTE
	CLGI	ILLVL	PVQIKG	KNYFLTY	KTMWARSL	MWARSLGPH	DAWYNVIDDV
	LFYY	QIRFN	NLDRIF	LIQFEGK	PHYLKHFK	LHVLIQFEG	LDLSPKVYSN
	LYLC	VLLLV	NYFLTY	HLDLSPK	DPHYLKHF	GKYENHTEN	HLDLSPKVYS
	IIFF	ILALL	FNHNLR	FGQVFNM	TMWARSLG	GFTHRGTHH	NDAWYNVIDD
	LFLC	VVLAL	YLKHFK	EFMGAQR	HYLKHFKE	FTHRGTHHC	GLTHRVGKRF
	FTFY	ALGIH	AWYNVI	DDVDPHY	HLDLSPKV	GITHRVGKR	LTHRVGKRFC
	YIAF	GDLIY	CVKSVY	DVDPHYL	LDLSPKVY	ITHRVGKRF	SNDAWYNVID
	CGLI	ILILL	FGQVFN	VDPHYLK	VLIQFEGK	EAGKYENHT	YSNDAWYNVI
	CVIA	LAVLG	GQVFNM	MGAQRDW	EAGKYENH	AGKYENHTE	HGFTHRGTHH
	LCYL	GNIIG	LIQFEG	DPHYLKH	QEAGKYEN	LDLSPKVYS	ENIKTKNHTN
	LIGW	NYVVY	IQFEGK	FMGAQRD	HVLIQFEG	DLSPKVYSN	KRFCVKSVYI
	VVYC	VLLLA	VYNHQE	VVYNHQE	VVYNHQEA	NDAWYNVID	LVRDRRPYGT

	WLAI	ALVIL	DVDPHY	KTMWARS	MWARSLGP	DAWYNVIDD	DENIKTKNHT
	LKCF	GKVMC	VDPHYL	TMWARSL	LHVLIQFE	LVRDRRPYG	WQSNCKYGKP
	FLCA	IILLL	NNVIRA	PHYLKHF	GKYENHTE	TVKNDLRDR	WMDENIKTKN
	VMFF	LLLLG	FMGAQR	HYLKHFK	ITHRVGKR	KNHTNSVMF	MDENIKTKNH
	YFKY	AVLLV	PHYLKH	MWARSLG	FTHRGTHH	LTHRVGKRF	RGNGITHRVG
	IDWI	ISLLL	DPHYLK	LDLSPKV	GITHRVGK	FWLVRDRRP	NGITHRVGKR
	CFLT	LGVVA	MGAQRD	QEAGKYE	THRGTHHC	GLTHRVGKR	GNGITHRVGK

Table A7. N-grams with positive order fractional difference

Table includes for each length first 100 n-grams occurring both in disordered and ordered regions with positive order/disorder fractional difference sorted according their mole fractions in descending orders, except for length one where 9 monograms exists.

N-gram length									
1	2	3	4	5	6	7	8	9	10
L	LL	YII	LIVL	LLLLL	SRTGKT	GPCKVQS	CEGPCKVQ	CEGPCKVQS	GCEGPCKVQS
V	VL	YIY	LLLF	NVIDD	GPCKVQ	CEGPCKV	EGPCKVQS	GCEGPCKVQ	DNEPSTATVK
I	LV	YFY	DIIL	KYGKP	VYATLK	EGPCKVQ	GCEGPCKV	QSNTKYGKP	VPRGCEGPCK
N	LI	CVI	LVIL	ASNPV	PCKVQS	GCEGPCK	EGDSRTGK	FDNEPSTAT	RGCEGPCKVQ
F	VV	IIC	ILVL	VYATL	YGDTDS	GDSRTGK	SNTKYGKP	RGCEGPCKV	CEGPCKVQSY
Y	IL	YCL	YILN	PVYAT	WARSLG	EGDSRTG	QSNTKYGK	EGPCKVQSY	MFDNEPSTAT
H	VI	CYL	NLIV	YATLK	VRDRRP	SNTKYGK	FDNEPSTA	MFDNEPSTA	PRGCEGPCKV
C	IV	FCV	VYVL	THASN	GFTHRG	NTKYGKP	MFDNEPST	GPCKVQSYE	EGPCKVQSYE
W	II	CVF	IGII	DVTRG	TGGQYA	QSNTKYG	NTKYGKPV	DVPRGCEGP	DVPRGCEGPC
	IG	WLY	IIFL	NPLYF	DSRTGK	HRGTHHC	KEEALSQL	CKVQSYEQR	FDNEPSTATV
	LF	LIW	VLAY	HTENA	GDSRTG	FDNEPST	QSYEQRHD	PCKVQSYEQ	PCKVQSYEQR
	FL	YVC	VFID	WMDEN	KYGKPV	MFDNEPS	VQSYEQRH	SNTKYGKPV	GPCKVQSYEQ
	YL	FWL	INFI	THRVG	IDDVDP	TKEEALS	RGTHHCSS	KVQSYEQRH	PDVPRGCEGP
	LY	WIV	YKII	LGVIS	ARSLGP	RELHEDG	HRGTHHCS	VQSYEQRHD	QSNTKYGKPV
	FV	CVY	IILV	MDENI	FCVKSV	TKYGKPV	KFLNQVNA	HRGTHHCSS	CKVQSYEQRH
	VF	VWI	LYYL	LGPHN	GAGKST	QSYEQRH	DEYQLSHD	DVPKGCEGP	KVQSYEQRHD
	FI	CFV	IVLF	KNHTN	TKYGKP	SYEQRHD	VPKGCEGP	RKFLNQVNA	PDVPKGCEGP
	VY	YIC	IIIA	FCVKS	NTKYGK	GTHHCSS	RKFLNQVN	KEEALSQLQ	VGSGKSTGLP
	YV	IYC	KYII	FTHRG	SNTKYG	RGTHHCS	SSYKEFLD	VEGDSRTGK	ESRRKFLNQV
	YI	ICF	LYIL	INNVI	LDLSPK	LGIHSPS	VELEGVNG	ESRRKFLNQ	RRKFLNQVNA
	GY	WFL	LFIL	QFEGK	QSNTKY	NGAGKST	ESRRKFLN	RDEYQLSHD	SRRKFLNQVN
	FT	WLF	VVVF	YGKPV	NDLRDR	YGDTDSV	KVDGRTMK	LVAEVERLR	ELVAEVERLR
	YA	LFW	GIII	LKHFK	RGTHHC	KVDGRTM	SRRKFLNQ	ELVAEVERL	HRDEYQLSHD
	AY	CYI	NYIL	ENHTE	HRGTHH	SYKEFLD	SYKEFLDE	RRKFLNQVN	ESHRDEYQLS
	YG	YCI	LIFL	ILLLL	SIVIEG	SSYKEFL	ELVAEVER	SRRKFLNQV	RHPNISQLST
	IF	VCY	LIVF	IVIEG	SLTKEE	SSYKEYL	LVAEVERL	SSYKEFLDE	SHRDEYQLSH
	IY	CYV	YFYD	GAQRD	ATVTGG	FLNQVNA	RHPNISQL	IGVVKPLAI	IESHRDEYQL
	YT	WII	YALV	KEFMG	VKNDLR	KFLNQVN	VAEVERLR	ESHRDEYQL	IVEGDSRTGK
	YN	YVW	VAII	LLLLI	GLTHRV	ELVAEVE	RRKFLNQV	HRDEYQLSH	NYIESHRDEY
	TY	CFI	AYVV	CVKSV	ELHEDG	GAGFGAG	IGVVKPLA	RHPNISQLS	YIESHRDEYQ
	NY	IIW	LLIF	HLDLS	VIEGDS	PNSSYKE	LSSFTNVP	SHRDEYQLS	IGVVKPLAIT
	FF	WFI	NILY	LALLL	VYSNDA	RKFLNQV	GVVKPLAI	YIESHRDEY	PMYRKPRMYR
	YF	VWY	VLLY	TVVDN	SNLDRI	ELEGVNG	ESHRDEYQ	FVKTLTGKT	MSAEVLDRTK
	YY	FWT	VLYA	IRDLI	DENIKT	PTSSYKE	HRDEYQLS	IESHRDEYQ	RPMYRKPRMY
	FY	WVF	LYIT	SNTKY	STVVDN	RRKFLNQ	SHRDEYQL	IVEGDSRTG	SAEVLDRTKQ
	LC	YFC	YIIE	LLAVL	AKFKGK	TKNTFSL	YIESHRDE	NYIESHRDE	GIGVVKPLAI
	WL	VYW	GYIL	LLVEL	LGRVGR	VDGRTMK	YNNRWVKD	GVVKPLAIT	MYRKPRMYRM
	HI	IWI	IFLK	IIIDE	GKLKLS	VELEGVN	IESHRDEY	KTLTGKTIT	PNSSYKEFLD
	CL	YCF	VIGF	IGAGI	SYEQRH	ESRRKFL	VKTLTGKT	MYRKPRMYR	SSYKEFLDEE
	IH	FWV	IRIY	HYLKH	HQEAAK	KGTVKIE	NYIESHRD	PMYRKPRMY	GVVKPLAITN
	VC	FYC	TFVL	ALVLA	YEQRHD	SRRKFLN	VVKPLAIT	AEVLDRTKQ	YKEFLDEEKN
	AW	YWV	LIYL	MGAQR	ASNEQA	LVAEVER	KTLTGKTI	MSAEVLDRT	DNEPSTATIK
	GW	FWI	TIVF	LKLST	NGVTLD	YKEFLDE	MYRKPRMY	SAEVLDRTK	EPSTATIKND
	CV	FCY	LVLY	ALLLT	THHCSS	VAEVERL	TLTGKTIT	SYKEFLDEE	NEPSTATIKN
	CG	WIY	YFLT	VTLAL	GTHHCS	KTLTGKT	YKEFLDEE	GIGVVKPLA	SYKEFLDEEK
	LW	CFY	LYLF	HRGTH	LPATAD	LSSFTNV	AEVLDRTK	PNSSYKEFL	NSSYKEFLDE
	CA	YIW	NVII	INNIK	LRVLAA	SSFTNVP	EALSQLQN	NEPSTATIK	SGIGVVKPLA
	WA	CYF	RIYF	GTHHC	RKALGI	QSGLDFK	EVLDRTKQ	NSSYKEFLD	SSFTNVPDEM
	AC	WYI	RILI	AAVLL	SHVGKV	RHPNISQ	MSAEVLDR	KEFLDEEKN	LSSFTNVPDE
	DW	DWW	FVVV	LVALT	GVSSRG	GAGLGAG	SAEVLDRT	VVKPLAITN	NSGIGVVKPL
	WV	FIW	LYMA	ALVAG	LKDPIP	NNRWVKD	EFLDEEKN	YKEFLDEEK	RKPRIYRTLR
	VW	FWF	YIVD	LLALI	HENGEP	GVVKPLA	EPSTATIK	AKEAFHPMY	FVKTLTGKTI
	IC	IWY	IILF	VNGVL	IQIKGG	IGVVKPL	GIGVVKPL	DNEPSTATI	VKTLTGKTIT
	CD	VWC	IINY	LGLLL	LKAELR	AGKSLIQ	NSSYKEFL	EEALSQLQN	EEALSQLQNL
	GC	WYF	IIFN	NHTNS	PAGTGK	DDIDDID	PNSSYKEF	EPSTATIKN	EGGQHLNVNV
	HF	FFW	YLVN	INNII	STAKHS	ESHRDEY	ALSQLQNL	PSTATIKND	FLEKISIPRG

	WT	YFW	LFYL	VLGKI	RIQRLG	EVILPRG	NEPSTATI	EALSQLQNL	GGQHLNVNVL
	HY	FWY	YLLV	LVAAI	SYKEFL	EVYVSPF	DIDGIREP	SFTNVPDEM	MQEWADDYFG
	WD	CLW	FIIL	ILAIL	GLADAL	RNALDGN	KEFLDEEK	SSFTNVPDE	LLDSIQGRAP
	CK	YWF	VFIL	LLGLA	NISPET	VVKPLAI	VKPLAITN	KPRIYRTLR	SFTNVPDEMQ
	FM	CIC	IDFV	VYSND	NKKFIK	HRDEYQL	GGPGDFRV	SGIGVVKPL	FNYIESHRDE
	CI	CWL	YNIV	LILDE	AARGGH	SGQPSTV	KEAFHPMY	GLPNLKRAN	FTNVPDEMQE
	TW	FYW	YLIL	NNYVV	LEELLK	IESHRDE	LLDSIQGR	GQHLNVNVL	NIRAGKYRGS
	RW	CCI	LFIG	LLQLL	AIAEEL	SHRDEYQ	PSTATIKN	LDSVQGRGP	LNKVVSHLPG
	DC	WWL	FAIL	LNAIL	LAAALG	YIESHRD	SFTNVPDE	LSSFTNVPD	NKVVSHLPGV
	KC	WWA	IIFI	QFHNL	VDPLTG	YNNRWVK	STATIKND	EGGQHLNVN	QSLLDSIQGR
	YH	LWW	TLVY	LFLLL	AVSQDQ	AGSGKST	FTNVPDEM	MQEWADDYF	KTGVSKKTGK
	YM	WVC	IYIL	LIDAG	SYKEYL	ERKQIRL	QEWADDYF	QEWADDYFG	QVFSQTTGAE
	WI	WVW	FIIN	VLLDE	YKEFLD	ESGDFAR	SSFTNVPD	VKTLTGKTI	TRKDRLGNTL
	TC	WAW	VIAF	AALAI	AGTGKS	LKVSATP	DYNLNSPL	FLEKISIPR	YKTGVSKKTG
	FH	ICC	VGVY	ALVVL	GAGFGA	KNRVVYD	LDQFPLGR	GGQHLNVNV	YQVFSQTTGA
	WG	WWV	VIGY	GILGG	VDGRTM	LDSVQGR	LPNLKRAN	LDSIQGRAP	DAARLELERD
	CT	CYC	YIVT	LISLI	DFASLY	ALSQLQN	PRIYRTLR	LEKISIPRG	DDAARLELER
	KW	WWT	GVIY	LVDLA	ENGTSP	GVRRSAR	DSVQGRGP	FTNVPDEMQ	KDRLGNTLVG
	FC	WCI	VLFF	VAGVV	SSYKEF	NYIESHR	EGESRTGK	LLDSIQGRA	KYQSLLDSIQ
	CN	TWW	VYTV	VLAIV	GAVGSG	FTLDEEF	EKISIPRG	RNAHNFPLD	MEQMWPKVED
	WN	WHY	FLLI	AGIAL	LPPLGG	LTGKTIT	GKLRAKGH	FNYIESHRD	PIKVQHRIAK
	NC	WTC	FLGY	ALAVG	ALVKKF	VKPLAIT	GLPNLKRA	IRAGKYRGS	RKDRLGNTLV
	IW	VWW	YKYI	EGDLI	DLPPLG	KEFLDEE	GQHLNVNV	NIRAGKYRG	SLLDSIQGRA
	YC	CWV	VLYI	IILLI	DTDSLF	TLTGKTI	KVQHRIAK	IKVQHRIAK	VFSQTTGAED
	NW	CVW	YIID	LIGAL	IGAGIA	AEVLDRT	LDSVQGRG	KVVSHLPGV	YQSLLDSIQG
	CF	ICW	IALY	ALLLV	SKEQAL	DTAEELE	LTARLSRS	LNKVVSHLP	DALVAAKIKP
	WF	HFW	EYII	DILKL	PFLRPE	DYNLNSP	PELVAEVE	NKVVSHLPG	DQVPEELEEW
	CY	CWT	IVFI	FNQPI	GVGKTT	EVLDRTK	QHLNVNVL	QSLLDSIQG	EEFNETIKSR
	WY	WWI	LIFN	VIIDE	LVAEVE	FLDEEKN	EGGQHLNV	SLLDSIQGR	EFNETIKSRG
	FW	WMF	LYVI	ALAIL	EATDTS	KPLAITN	EWADDYFG	KTGVSKKTG	EWQVFQQSSP
	QW	WFW	YLVA	ALVGL	KFLNQV	LDSIQGR	FLEKISIP	QVFSQTTGA	FNETIKSRGR
	WQ	WQW	IVGF	AVGVL	KVSATP	MSAEVLD	GGQHLNVN	RKDRLGNTL	GSVEWQVFQQ
	YW	WIW	FINI	FHNLN	RMTDNE	PNLKRAN	GSTALNGA	TGVSKKTGK	IDALVAAKIK
	QC	FMW	FIIK	GLGIG	FLNQVN	SAEVLDR	KERKQIRL	TRKDRLGNT	MATEVDHVRY
	HC	NWC	IVYG	GVVAL	HWKELI	VLDRTKQ	KYQSLLDS	VFSQTTGAE	MVRRSMEAID
	CH	YWC	YVSF	IALAG	KGWGKD	AAAGGHL	MQEWADDY	YKTGVSKKT	QIDALVAAKI
	CC	CFW	LAYI	ILLLI	LLNEFP	EAFHPMY	RLGVIALA	YQVFSQTTG	QVFQQSSPLY
	WH	WWF	ILYI	LIALL	LNQVNA	EFLDEEK	TITAGTGL	AARLELERD	SVEWQVFQQS
	WM	CWN	IYLI	NILKY	NSSYKE	EPSTATI	DSIQGRAP	DAARLELER	TESRRKFLNQ
	HW	QCW	IVFV	IIKLL	AKVTGG	GESRTGK	EEFNETIK	KHKFNRSGL	VEWQVFQQSS
	CM	WQC	VIFL	ILLII	GRVLRK	LSQLQNL	LDSIQGRA	KYQSLLDSI	VFQQSSPLYW
	WC	IWW	YILA	ILSLL	LEGVNG	NSSYKEF	LEKISIPR	MEQMWPKVE	WQVFQQSSPL
	WW	WCC	NIVY	LLGKV	PNSSYK	PSTATIK	RNAHNFPL	PIKVQHRIA	PAFGSVEWQV
	CW	HWC	LGIY	LVGLL	VKAIAE	WADDYFG	IDPSGRGK	YQSLLDSIQ	VRAQIDALVA

Table A8. N-grams that appear only in border between disordered and ordered regions

For each length first 100 n-grams that appear only on border between ordered and disordered regions sorted according their mole fractions in descending order are presented, except for length 4 where only 1 such n-gram exists.

N-gram length						
4	5	6	7	8	9	10
MWCW	IWRFP	FYDSVT	YFYDSVT	YFYDSVTN	MEGNRPTFV	ASMEGNRPTF
	GPAWY	NRPTFV	FYDSVTN	EGNRPTFV	SMEGNRPTF	SMEGNRPTFV
	PAWYW	VVYKYE	EGNRPTF	MEGNRPTF	LYDALEAPA	LYDALEAPAD
	HAWMP	VYKYEE	GNRPTFV	QVVYKYEE	GQVVYKYEE	IRIYFYDSIT
	HQQLW	CHLKNP	QVVYKYE	LYDALEAP	IYFYDSITN	KNYGHPRENF
	HWMEI	FYDSIT	VVYKYEE	YFYDSITN	KNYGHPREN	NKNYGHPREN
	WMEIP	GHPREN	FYDSITN	YGHPRENF	NYGHPRENF	RIYFYDSITN
	SWWRH	HPRENF	GHPRENF	GQVVYKYE	RIYFYDSIT	EGNRPTFVVQ
	WADHG	ICHLKN	YFYDSIT	IYFYDSIT	EGNRPTFVV	GNRPTFVVQN
	HGMPD	DLDYVG	YGHPREN	NYGHPREN	GNRPTFVVQ	IRIYFYDSVT
	MVVFK	MKKIIL	ICHLKNP	VICHLKNP	NRPTFVVQN	MEGNRPTFVV
	AEKTH	PTFVVQ	VICHLKN	GNRPTFVV	PTFVVQNET	NRPTFVVQNE
	IYWGM	HNTRDG	NRPTFVV	IYFYDSVT	RIYFYDSVT	PTFVVQNETQ
	MDEDH	WVTLGG	PTFVVQN	NRPTFVVQ	RPTFVVQNE	RPTFVVQNET
	MDINW	YDSVQN	RPTFVVQ	PTFVVQNE	IYFYDSVTN	RIYFYDSVTN
	MPFRD	HPNLRM	VYKYEEE	RPTFVVQN	QVVYKYEEE	WVTLGGAGGG
	NCYDR	LYIPEQ	EFAPDAP	VVYKYEEE	KEFAPDAPL	DGKRVSPPRE
	PWNIQ	PNLRML	FAPDAPL	EFAPDAPL	WVTLGGAGG	DREPDLYIPE

	QQHFN	PPREVR	FYDSVQN	KEFAPDAP	DREPDLYIP	DTLVELEGVN
	QWHAR	SPPREV	WVTLGGA	WVTLGGAG	DTLVELEGV	GKRVSPPREV
	SHTWG	HLKNPE	YHNTRDG	YFYDSVQN	EPDLYIPEQ	HPNLRMLDDD
	WNIQH	IFNAFM	YYHNTRD	YYHNTRDG	GKRVSPPRE	IPPHPNLRML
	YWGMR	KYEEEQ	DLYIPEQ	DTLVELEG	HPNLRMLDD	KRVSPPREVR
	ECEFR	MPKEKY	DTLVELE	EPDLYIPE	IPPHPNLRM	PHPNLRMLDD
	FWTQM	PRENFA	EPDLYIP	HPNLRMLD	KRVSPPREV	PNLRMLDDDA
	GAHNI	FPSVEP	FDLDYVG	KRVSPPRE	PHPNLRMLD	PPHPNLRMLD
	HTWAV	IWRFPS	HPNLRML	PDLYIPEQ	PNLRMLDDD	PPREVRIVQV
	IQMST	RLGIRP	PDLYIPE	PHPNLRML	PPHPNLRML	PREVRIVQVV
	LDWWE	VGHTSE	PHPNLRM	PNLRMLDD	PPREVRIVQ	PRSNMIRHYL
	QMQDS	WRFPSV	PNLRMLD	PPHPNLRM	PREVRIVQV	REPDLYIPEQ
	STWSR	GLRGYN	PPREVRI	PPREVRIV	PRSNMIRHY	REVRIVQVVL
	CYGCQ	IPEQTV	PREVRIV	PREVRIVQ	REPDLYIPE	RVSPPREVRI
	DMQRW	QATIFD	REVRIVQ	REPDLYIP	REVRIVQVV	SPPREVRIVQ
	FVMKR	TDAEQR	RVSPPRE	REVRIVQV	RSNMIRHYL	VSPPREVRIV
	HCMRN	GPAWYW	SPPREVR	RSNMIRHY	RVSPPREVR	YDREPDLYIP
	KAWVF	IERRDA	VSPPREV	RVSPPREV	SPPREVRIV	GQVVYKYEEE
	KPAWW	LGIRPP	SFDLDYV	SPPREVRI	VSPPREVRI	RPRSNMIRHY
	MADFC	LKNAEN	CHLKNPE	VSPPREVR	VSFDLDYVG	CHLKNPEKGK
	MFYQP	LRGYNV	HLKNPEK	SFDLDYVG	CHLKNPEKG	ICHLKNPEKG
	MIDWW	PAWYWT	LKNPEKG	VSFDLDYV	DDQIFNAFM	VICHLKNPEK
	PHNHE	ADNNSG	QIFNAFM	CHLKNPEK	HLKNPEKGK	GHPRENFADI
	PPTWG	CGGGRH	YKYEEEQ	DQIFNAFM	ICHLKNPEK	NYGHPRENFA
	PQNAW	DHLIPS	HPRENFA	HLKNPEKG	VICHLKNPE	QVVYKYEEEQ
	PTHTW	ERGNFD	MPKEKYY	ICHLKNPE	GHPRENFAD	YGHPRENFAD
	QHCMR	ESIEYG	PRENFAD	LKNPEKGK	HPRENFADI	EKYYLYREDG
	RFVPW	FGGQTA	IWRFPSV	GHPRENFA	IYFYDSVQN	KEKYYLYRED
	RPRYQ	GEGKIV	KEKYYLY	HPRENFAD	VVYKYEEEQ	MPKEKYYLYR
	SYDKC	GHQQLW	KKVEYKG	PRENFADI	YGHPRENFA	PKEKYYLYRE
	THDPQ	GHWLGI	KYEEEQE	VYKYEEEQ	EKYYLYRED	RIYFYDSVQN
	TQHCM	GIQNNK	KYNAKKV	EKYYLYRE	IWRFPSVEP	VDERLNKMLK
	WTFHK	GIRPPK	KYYLYRE	IWRFPSVE	KEKYYLYRE	VVYKYEEEQE
	WWCIR	GKLPIV	RFPSVEP	KEKYYLYR	KYYLYREDG	AKKVEYKGIV
	YDPHQ	GNGGHW	VDERLNK	KYYLYRED	MPKEKYYLY	DLYIPEQTVK
	YTQHC	GRVMVK	WRFPSVE	MPKEKYYL	PKEKYYLYR	ELGLQATIFD
	YWHEK	GSPIWE	AKKVEYK	PKEKYYLY	VDERLNKML	EPDLYIPEQT
	DATWH	HAWMPP	IPEQTVK	VDERLNKM	VVICHLKNP	IPEQTVKDRD
	EPFWH	HLIPSC	IRLGIRP	VVICHLKN	VYKYEEEQE	KKVEYKGIVF
	FGHVQ	HQQLWD	KVDERLN	WRFPSVEP	AKKVEYKGI	KYNAKKVEYK
	FHMTS	HWLGIY	LQATIFD	YKYEEEQE	DLYIPEQTV	KYYLYREDGT
	FHRQA	IGPAWY	LYIPEQT	AKKVEYKG	GLIRLGIRP	LGLQATIFDI
	HNADE	IPSCAG	MSEEIKV	DLYIPEQT	GLQATIFDI	LYIPEQTVKD
	KGCAM	IQNNKK	PSGLRGY	GLQATIFD	IPEQTVKDR	MSKYNAKKVE
	KPPYW	IRPPKN	QATIFDI	IPEQTVKD	KKVEYKGIV	NAKKVEYKGI
	MCWMA	KSERGN	RKLAAEK	KKVEYKGI	KYNAKKVEY	PDLYIPEQTV
	PYWPC	KWLAAE	SGLRGYN	KYNAKKVE	LGLQATIFD	SKYNAKKVEY
	WHRGW	KYLPTK	SKYNAKK	LIRLGIRP	LYIPEQTVK	YIPEQTVKDR
	WMGRV	LIPSCA	YIPEQTV	LQATIFDI	MPSGLRGYN	YNAKKVEYKG
	WTQQS	MFEITS	AWSRPWG	LYIPEQTV	MSKYNAKKV	ARKLAAEKAA
	YYMPD	MPPTKD	DLDEDED	MPSGLRGY	NAKKVEYKG	GPAWYWTVAR
	ACEMQ	MSGEWF	EEYAAAQ	MSKYNAKK	PDLYIPEQT	IKQHGLEYEE
	CEMQW	NNSGDK	GGRLEAA	NAKKVEYK	SKYNAKKVE	KLAAEKAAET
	CPFQI	NSGDKP	GLRGYNV	PSGLRGYN	YIPEQTVKD	KVDERLNKML
	DFWVH	PDDSHW	GPAWYWT	SKYNAKKV	YNAKKVEYK	LAAEKAAETK
	DWMKF	PESIEY	PAWYWTV	YIPEQTVK	ARKLAAEKA	LARKLAAEKA
	EMQWK	PGNLLD	ALTDAEQ	YNAKKVEY	AWSRPWGLE	MPSGLRGYNV
	GIQCK	PGSSEK	AWMPPTK	ARKLAAEK	GPAWYWTVA	NQMLSSLLVS
	HSLVW	PIWERM	ERQLESC	AWSRPWGL	IKQHGLEYE	PRAWSRPWGL
	IWDPM	QKWLAA	ESGSPIW	GGRLEAAT	KLAAEKAAE	RAWSRPWGLE
	NDPWN	QLWDTV	GHTSEDD	GPAWYWTV	KVDERLNKM	RKLAAEKAAE
	QQMWE	QNNKKP	HAWMPPT	KVDERLNK	LAAEKAAET	YIKQHGLEYE
	RQQMW	QQLWDT	IPSCAGS	LAAEKAAE	LARKLAAEK	DEQDRLINLV
	RRYWY	RLITMF	IQKWLAA	PAWYWTVA	PAWYWTVAR	DYKLDLDEDE
	TFAMC	RRHVTP	LGIRPPK	RAWSRPWG	PRAWSRPWG	FGGQTAISTG
	VNDPW	RSPLHW	LVGHTSE	RKLAAEKA	PSGLRGYNV	GPGNLLDVEL
	WAVCH	RVYQLR	LWDTVMK	SGLRGYNV	RAWSRPWGL	HLIPSCAGSG
	YMPNE	SGDKPI	LYAVSNS	EGKIVTTL	RKLAAEKAA	IAYYGRVMVK
	CRITP	SGSPIW	MQQQAYI	GIGPAWYW	ALPGEVVGD	KKKLREVETE
	CWDES	SNSDDP	MVKEETP	GLTDPHPI	DPSLGTRRI	LIPSCAGSGA
	FQNHS	SPIWER	NNSGDKP	LPGEVVGD	EGKIVTTLK	LPVPFRSLTK
	KVPIW	SPLHWP	NVVVDSD	LPVPFRSL	EQDRLINLV	LSALTDAEQR
	MTWRV	SRSPLH	PEEYAAA	LWDTVMKR	GHQQLWDTV	LYAVSNSDDP
	PDHKM	TGGSFE	PLHWPHE	NGGHWLGI	KWAGIGPAW	PVPFRSLTKQ
	QVAWE	TRRIRM	QLWDTVM	NVAPGEGK	LGIQNNKKP	QQLWDTVMKR
	SEMHS	VYQLRA	RIRMPGQ	PEEYAAAQ	LSALTDAEQ	RIRMPGQIGG
	WQWGC	WDTVMK	RVYQLRA	QIQKWLAA	LYKSERGNF	RKSRSPLHWP
	YKNQR	WHAWMP	SESGSPI	RPGPGNLL	PIWERMNSV	TADNNSGDKP
	YQFYG	WLEETQ	SNLGQLG	RSPLHWPH	VMVKEETPE	TDPHPIVRDL
	CSMTP	WMPPTK	TAISTGA	VAPGEGKI	VSNSDDPTN	TSPERQLESC
	FCKMC	WTLPGS	VKEETPE	VGHTSEDD	VWHAWMPPT	VSNSDDPTNN
	FMQTM	YAVSNS	WDTVMKR	WDTVMKRR	WHAWMPPTK	WHAWMPPTKD

Table A9. N-grams with positive fractional difference on border between disordered and ordered regions

Table includes for each length first 100 n-grams that appear on border between ordered and disordered regions, and in ordered or disordered regions or both, but prefer border region, sorted according their mole fractions in descending orders, except for length two where 78 bigrams exists.

N-gram length								
2	3	4	5	6	7	8	9	10
EL	MIY	DWWE	YFYDS	YFYDSV	RIYFYDS	IRIYFYDS	KIRIYFYDS	LKIRIYFYDS
LS	NIY	KFQI	FYDSI	IYFYDS	IYFYDSV	RIYFYDSV	IRIYFYDSV	KIRIYFYDSV
SL	WVS	GHWL	RPTFV	YFYDSI	LVSPTRS	DLVSPTRS	FDLVSPTRS	KIRIYFYDSI
LE	MFY	PAWY	LYIPE	LVSPTR	IYFYDSI	RIYFYDSI	IRIYFYDSI	VSPTRSAHFH
LK	MYY	GGHW	VYKYE	YDSVTN	TGELITA	RTGELITA	DSRTGELIT	FFDLVSPTRS
KL	WVK	AWYW	YKYEE	IKFNLY	VSPTRSA	SRTGELIT	DLVSPTRSA	MDSRTGELIT
IE	WEV	HWLG	HLKNP	NDTEGL	GQPSTVV	LVSPTRSA	LVSPTRSAH	FDLVSPTRSA
KI	WIE	DSHW	YDSIT	DTEGLL	SRTGELI	VSPTRSAH	VSPTRSAHF	DLVSPTRSAH
IK	HII	FVLQ	DLDYV	TEGLLK	DLVSPTR	SGQPSTVV	SRTGELITA	LVSPTRSAHF
SV	KIC	YVVI	WDPLV	YDSITN	NSGQPST	DSRTGELI	GQPSTVVDN	DSRTGELITA
EI	VWE	FHEM	GQLGI	MWDPLV	YDALEAP	GQPSTVVD	MDSRTGELI	SGQPSTVVDN
VS	MWL	HWTF	PTFVV	PSTVVD	DALEAPA	FDLVSPTR	SGQPSTVVD	GNNSGQPSTV
VE	WIK	TIYI	DLYIP	RGPAGW	QPSTVVD	NNSGQPST	GNNSGQPST	NNSGQPSTVV
DL	WFK	WNLH	GHPRE	NMFDNE	DTEGLLK	NSGQPSTV	NNSGQPSTV	KSYIDKDGDT
RL	WVD	WRHK	WVTLG	RPWGLE	GNDTEGL	YDALEAPA	NSGQPSTVV	NSGQPSTVVD
SI	QLW	CWGL	QAYIN	APDAPL	NDTEGLL	DALEAPAD	YDALEAPAD	YDALEAPADT
LR	AVW	TKMW	YDSVQ	SDAIDL	NGNDTEG	PLLNEFPE	DALEAPADT	DALEAPADTP
EV	YWE	KSCY	YHNTR	YDNEPS	NMFDNEP	ALEAPADT	DPLLNEFPE	WDPLLNEFPE
KV	CYD	WMEI	YIPEQ	GQLGIL	SSDVKSY	QPSTVVDN	ALEAPADTP	CCCPHCPRHK
IS	WKV	KMWQ	EPEFI	RPTFVV	TEGLLKE	SSDVKSYI	CCPHCPRHK	KSSSDVKSYI
VK	WIP	KQFW	INLVM	WDPLVN	MEGNRPT	LEAPADTP	SSSDVKSYI	RFFDLVSPTR
LP	MIW	SFWV	MFKKW	YKYEEE	PSTVVDN	DTEGLLKE	FFDLVSPTR	AYNGNDTEGL
PL	CIN	WWRH	RHLID	AISIRK	SRGPAGW	GNDTEGLL	AYNGNDTEG	GNDTEGLLKE
QL	HTW	FFEL	EKYYL	DTLVEL	FNMFDNE	NDTEGLLK	GNDTEGLLK	GQPSTVVDNT
RI	WID	GMWM	FVRPP	DAIDLI	TLSDAID	NGNDTEGL	NDTEGLLKE	NGNDTEGLLK
PV	WIR	MISW	HCTQV	FAPDAP	YDNEPST	YNGNDTEG	NGNDTEGLL	SAYNGNDTEG
VR	IWS	WNIQ	KYYLY	YGHPRE	MWDPLVN	FNMFDNEP	YNGNDTEGL	YNGNDTEGLL
RV	CEF	AWFA	LINLV	YHNTRD	SDAIDLI	NMFDNEPS	VFNMFDNEP	DPLLNEFPET
FS	KWI	CMVK	WRFPS	DLYIPE	DAISIRK	SSSDVKSY	FNMFDNEPS	FNMFDNEPST
LQ	WYE	ELWG	ASYAF	FDLDYV	DAIDLIN	SMEGNRPT	NMFDNEPST	NMFDNEPSTA
SF	CIR	FYYK	ATIFD	FYDSVQ	EAPADTP	VFNMFDNE	PLLNEFPET	VFNMFDNEPS
IR	MCI	HILA	GSPIW	GGRVKP	LSDAIDL	MTLSDAID	KSSSDVKSY	PLLNEFPETV
VP	MIC	NPQW	LIPSC	PDLYIP	GGRVKPL	YDNEPSTA	ASMEGNRPT	AKSSSDVKSY
IP	WVN	QVHC	LRGYN	PREVRI	NYGHPRE	SDAIDLIN	LRAVLTEAL	HFHPNIQGAK
FK	CYE	WWGG	MFEIT	REVRIV	RFDSQTK	LSDAIDLI	QPSTVVDNT	IDKDGDTLEW
EF	WWR	WYLS	MPPTK	VSPPRE	YFYDSVQ	TLGGAGGG	YDNEPSTAT	SYIDKDGDTL
FE	EWY	YWPA	NLIEL	YYHNTR	DKPIPLS	TLSDAIDL	LEAPADTPV	ALEAPADTPV
KY	ICN	FCLH	RRIRM	DKPIPL	GDKPIPL	GGRVKPLP	DTEGLLKEI	ELRAVLTEAL
QI	WWE	IFGV	VYVKG	GDKPIP	NQMFKKW	KNYGHPRE	LSDAIDLIN	KGNNSGQPST
YS	WIN	KWWQ	YLPTK	LINLVM	RSNMIRH	LGGAGGGG	MTLSDAIDL	PASMEGNRPT
PI	KMW	VVFM	AWYWT	QAYINA	HCTQVPI	LRAVLTEA	TLGGAGGGG	QPCCCPHCPR
KF	WTM	CEMQ	DVKTF	QMFKKW	KPIPLSG	RFDSQTKE	TLSDAIDLI	YDNEPSTATV
MI	GHW	EPFW	GHWLG	KPIPLS	PRSNMIR	DKPIPLSG	LGGAGGGGG	LEAPADTPVS
SY	YWG	FIHR	KLPIV	PRSNMI	RAIRRRR	GDKPIPLS	NKNYGHPRE	VKSYIDKDGD
EY	HWL	GWMK	KLYAN	RAIRRR	WRDPSTP	PRSNMIRH	RFDSQTKER	DTEGLLKEIE
YK	SWY	IWNG	LDYIG	RSNMIR	WSRPWGL	TEGLLKEI	VTLGGAGGG	FDSQTKERLT
QV	CHN	LMVI	MFKVY	DPLVNE	KNPEKGK	YDAISIRK	DKPIPLSGI	MTLSDAIDLI
YE	MCY	MCGI	NVVVD	HCTQVP	RHLIDTS	YNQMFKKW	ELRAVLTEA	NDTEGLLKEI
VQ	DFC	THAW	SHGIA	PLYSGS	RRHLIDT	HCTQVPIK	GDKPIPLSG	TLGGAGGGGG
IQ	CFR	TKCF	SKLYA	RHLIDT	RRRRVDL	KPIPLSGI	TEGLLKEIE	TLSDAIDLIN
ML	LHW	VFIG	SRDPY	WVSGWS	RWVSGWS	WRDPSTPT	HCTQVPIKV	LGGAGGGGGS
PF	AWY	VQYI	WKDGE	APGEGK	WVSGWSE	EAPADTPV	KPIPLSGIK	RFDSQTKERL
FR	IWK	VTWL	WLEET	ARKEYL	DPLVNEF	GRRHLIDT	TYNQMFKKW	RNKNYGHPRE
FP	MCW	YYHK	ADLKW	ARRFYD	GRRHLID	RNSTLSAL	EAPADTPVS	VTLGGAGGGG
PY	WGY	LFYF	ESQNY	IEIKPK	PLVNEFP	RRHLIDTS	GRRHLIDTS	DKPIPLSGIK
QF	MVW	NCYD	EYFYE	PKEKYY	RLINLVM	RWVSGWSE	RNSTLSALM	GDKPIPLSGI
MF	WWQ	PKHW	GGHWL	PLVNEF	RPRSNMI	WSRPWGLE	RPRSNMIRH	TEGLLKEIED
YP	MFW	WTQM	HWLGI	RETRNS	TANDDVE	DPLVNEFP	MWDPLVNEF	KPIPLSGIKG
YQ	MMW	GLFM	IGPAW	RRHLID	WDPLVNE	DRLINLVM	QPCCCPHCP	QPSTVVDNTL
QY	CYP	HHCG	IPSCA	SRPWGL	EKYYLYR	MWDPLVNE	RGSWQKKKL	EAPADTPVSE
MV	DWW	IWNA	IQNNK	YEEEQE	ITGEKYP	RPRSNMIR	RWVSGWSEA	HCTQVPIKVQ
FQ	FFC	MYYY	KWLAA	EKYYLY	PKEKYYL	VWRDPSTP	TVWRDPSTP	RNSTLSALMP
KH	MWW	TWGW	KYLPT	GEKYPE	TGEKYPE	WDPLVNEF	VWRDPSTPT	ETVWRDPSTP
DM	WYK	WIKT	LWDTV	ITGEKY	DERLNKM	WVSGWSEA	WDPLVNEFP	MWDPLVNEFP
HE	YWP	WVTL	MSGEW	KEKYYL	KEFAPDA	ITGEKYPE	IRNSTLSAL	RGSWQKKKLR
WE	IWQ	WWKN	NGGHW	KKVEYK	LEGPLYS	IYFYDSVQ	MMTANDDVE	STPASKVRRR
WS	WWG	YCNS	PDINE	KYNAKK	LIRLGIR	MTANDDVE	VKSYIDKDG	TVWRDPSTPT
MY	HLC	YWGM	PHPIV	KYYLYR	NAKKVEY	DERLNKML	DERLNKMLK	VRGSWQKKKL
HK	NFW	AGWW	PIWER	RFPSVE	PRHMEVF	GLIRLGIR	KSVGITGQL	DVKSYIDKDG
SW	WWK	AVWA	PLHWP	TGEKYP	SRPWGLE	NPASAEAI	RIYFYDSVQ	GIRNSTLSAL
WK	MHW	AYLI	QLWDT	VMVKEE	VSGWSEA	SPRHMEVF	SVGITGQLT	IMMTANDDVE
WR	MCC	CHCS	QQAYI	VSGWSE	YNAKKVE	SVGITGQL	GLEYEEQKQ	IRNSTLSALM

CP	MYW	CYGC	RIRMP	YNAKKV	AWYWTVA	AWYWTVAR	HGLEYEEQK	DERLNKMLKG
PW	CFQ	CYKE	RLAFV	AKKVEY	GLEYEEQ	GLEYEEQK	KQHGLEYEE	IRIYFYDSVQ
WP	LCM	DYHH	RLITM	ASYAFG	GRLEAAT	HGLEYEEQ	LEYEEQKQL	KSVGITGQLT
MW	WMM	FIWE	RRFYD	ATIFDI	GVRQNTS	KLAAEKAA	QHGLEYEEQ	SVGITGQLTG
WQ	KWC	GCYE	RRHVT	EFAPDA	HGLEYEE	KQHGLEYE	ADLKWAGIG	VKSVGITGQL
MC	PWW	HEFF	SIEYG	IRLGIR	KLAAEKA	LEYEEQKQ	AWYWTVARP	GLEYEEQKQL
	WCS	IWDP	SPLHW	NAKKVE	LEYEEQK	QHGLEYEE	DPHPIVRDL	HGLEYEEQKQ
	YWH	LCVK	SQTLI	PIEIRP	QHGLEYE	RETRNSSF	GRIEAATSS	KQHGLEYEEQ
	WVH	MCVL	TIFGI	RHMEVF	RETRNSS	TLEGPLYS	LADLKWAGI	LEYEEQKQLT
	CCK	MEWV	VIGMQ	SEEIKV	TFDNSPG	ADLKWAGI	NQMLSSLLV	QHGLEYEEQK
	DWC	MNFW	VYQLR	SGLRGY	AAKVIAD	DLKWAGIG	NSTLSALMP	DPHPIVRDLY
	FWQ	PPYW	WDTVM	YIPEQT	ADLKWAG	DPHPIVRD	PHPIVRDLY	LADLKWAGIG
	HNW	RCHC	WLAAE	YLPTKR	DLKWAGI	GRIEAATS	PLGLTDPHP	NSTLSALMPC
	SWC	VYFK	WMPPT	AWYWTV	DRLINLV	LADLKWAG	PPLGLTDPH	PPLGLTDPHP
	CWE	WCIR	YFYEE	DKNEVI	EEALAWA	LARKLAAE	RAFESGDFA	RAFESGDFAR
	FWP	WITL	ADDQW	EEYAAA	GGSFELA	LGLTDPHP	STLSALMPC	STLSALMPCE
	MWC	WVTS	DQWVP	GNLLDV	GLTDPHP	NQMLSSLL	WQKKKLREV	SWQKKKLREV
	WGC	YCSQ	DSHWT	GVRQNT	GRIEAAT	NSTLSALM	AATASPASM	AWYWTVARPD
	WWS	CLMK	GASCA	ISIRKP	LADLKWA	PHPIVRDL	AGEDGLTYR	DDSHWTFSSD
	CHC	LCFQ	HRYQI	KNAENF	LLYAVSN	PLGLTDPH	ARWVSGWSE	DENGNIHVSK
	WCE	LWNQ	LSCEY	LEYEEQ	LRRLADE	PRAWSRPW	DENGNIHVS	DSHWTFSSDL
	CWI	LWYI	MKAIC	LTSLGG	NQMLSSL	PSTVVDNT	GALPGEVVG	KVVSHLPGVV
	HCM	MCYG	PEFIT	MQQQAY	NSNLGQL	QKKKLREV	GFSGCEHRS	LADENGNIHV
	WCK	NIIY	QWVPD	PEEALA	PHPIVRD	RAFESGDF	GLNKVVSHL	LPGVVHEMRS
	CWD	QVYI	RKAFL	VDRVER	PRAWSRP	RIEAATSS	PGVVHEMRS	LSRVTDATTS
	WKW	SWIL	SCEYS	WKDGEL	QQQAYIN	RSRSYIKL	SHLPGVVHE	RLSRVTDATT
	CWM	WDRN	SKVIL	WRDPST	RAWSRPW	STLSALMP	SRVTDATTS	RSGLNKVVSH
	CWH	WLIG	YQIKD	YKLDLD	RIEAATS	TRAFESGD	VVHEMRSEA	YYGRSGLNKV

Table A10. Characteristic n-grams in ordered regions by z-score values

N-grams presented in the table have abs(z-score)>2.58 in ordered and abs(z-score)<1.65 in ordered regions. Table includes for each length first 100 n-grams sorted according z-score values in descending order.

N-gram length							
3	4	5	6	7	8	9	10
INN	GCPW	CRELH	NALLLY	KGSGKSM	INVIDDVE	SHASNPVYQ	KVTGGQYASN
NII	YMAC	HASNP	HLDLSP	DGDTDSY	HNVIDDVR	THASNPVYA	TVTGGQYASK
WAR	CEGP	HNYLC	RTGKTM	FSGKSTE	VNVIDDVI	WYNVIDDVD	LHLHVLIQFK
DDV	CPWD	HRGTH	ENALLL	EASNPVG	VGPCKVQD	GPHNYLCGH	YWLVRDRRPN
DTI	EGPC	TIFLC	CVAEAW	QASNPVH	GGKTMWAV	SLGPHNYLS	TYDLIRDLIA
AMA	WMDE	FLTYP	CNIDLH	FASNPVW	EYATLKID	ALGPHNYLS	LWARSLGPHI
NIG	WQSN	HGFTH	YGKPVQ	DGTGKSG	LRTGKTMN	SLGPHNYLC	STHFAKFKGR
FIN	MACT	MKIDH	ATLKIR	HASNPVY	SKYENHTF	ALGPHNYLC	MSTAKHSVDV
IFN	CKVQ	HTENA	NLDRIF	RLCNPGW	ATGGQYAQ	EYNVIDDVA	FDRINVRRLF
NYI	GNHD	YLCGH	YATLKI	CLCNPGF	DNALLLYN	VATNIIENG	MRADVKEFEQ
VYI	PHNY	HRVGK	HDDLVM	QAAAVAI	RNALLLYF	GAQRDWQSN	FRADVKEFEA
VNG	KTMW	MACTH	ALLLYM	GHASNPQ	YALLLYMF	PVQIKGGIP	NQVPINATGH
ITV	CTHA	DVDPH	MDENIK	TVRDRRF	YPHLHVLF	TSLYPSIIR	ISDVTRGNGI
FTL	GPCK	KYENH	CAIIAW	RVGKRFC	EKYENHTV	ASLYPSIIR	GSSYKEFLDK
DPR	ACTH	HGFRC	RLEAIC	DGPCKVP	NLKHFKEN	YLRVLAALK	ISDVTRGNGL
KKF	DECH	RDRRP	SGHLDF	DWARSLR	AKYGKPVE	YFWRPEEVS	VLPTSAGKSA
ITI	IWMD	QIKGG	CISDVT	GCKVQSK	GIPTIFLC	LENALMLYS	NLPTSAGKSL
GEV	DRRP	NVIDD	HRVGKR	TCKVQSN	KVQIKGGF	SSLYPSIII	LGGDFLTSLV
FGP	NHQE	PVYAT	CVKSVY	RDVTRGR	CKPVQIKR	EGTGKTTLS	RCVSDVTRGS
QQL	TMWA	KGGIP	WKELIG	HVGKRFG	GGKPVQIA	FVGSGKSTY	CGYSQGAIVC
GDV	PHLH	CTHAS	HGSTIM	QGTGKTW	NVIDDVDP	FGLMVWCII	FLVRDRRPVD
ISS	YNHQ	QFEGK	HGSTVM	HLAAAGW	QGTGKTTY	IGGDFLTSF	DYSPDTLGYE
LDD	GHLD	FMGAQ	NVIDDV	RIDDVDF	MAGTGKTV	FGLPATADL	NEQALVKRFW
SLI	YYCW	FKEFM	GNGITH	VYNVIDW	NKNDLRDG	WLVRDRRPY	GDPFWYEDDV
DDL	TPLH	FTHRG	IKGGIP	PVYATLK	EGPCKVQS	ERIDANLLN	YDFASLYPSN
IYN	CKIC	HVLIQ	QIKGGI	YTHRVGE	DFFDLVSA	KRIDANLLD	VSDVTRGNGI
DNI	CRTC	QSNTK	QLRRAW	QIYFYDG	ENALLLYM	ASLYPSIIQ	LWARSLGPHN
FAP	CIPC	GKRFC	KNDLRD	YMDENIR	FAAAAAAH	MSNLCTEIS	YRFFDLVSPS
VDG	MWAR	KIRIY	WTKTVW	TENALLL	RPSTATVG	NNKKFIKIL	IPRRHGKTWI
VVG	HTEN	NHTEN	QIGRVP	TPVYATN	TENALLLY	TAGFGAGFT	VPRRHGKTWF
IHS	CDKC	GKTMW	ELIGAQ	NPVYATL	NALLLYMA	CPGSGKSTW	VPRRHGKTWI
WRL	THAS	QVFNM	FDLVSP	YATLKIR	GNLRKALY	IYDKYNDVY	VSDVTRGNGL
NIL	LYMA	YATLK	MLAIKY	IVTGGQT	LRGCEGPA	QYDKYNDVN	FGPAGTGKTS
RRL	CAVC	MDENI	YGALGN	PVTGGQQ	WEPSTATK	GPTSAGKST	KQAIELLPDF

VFI	YCWM	LGKIW	WALKNA	QVYATLE	YNVIRAVY	TIHSRSYTH	VTDIAGYAGV
CPW	DWQS	SLGPH	CGVAAC	FLKIRIR	GKPVQIKG	VLTEGDSAS	RVGIAVDTGT
KMI	MVWC	THASN	HITNAH	SKRFCVV	NKYGKPVR	ANPFLRPEF	LGKTTVVAIF
NTG	HASN	SNPVY	CGHLDL	LRTGKTG	AGSGKSTC	ANPFLRPEL	LAGLPATADK
HNL	NHTN	NHNLR	SDVTRG	DRIYFYQ	IHTNSVMR	FDAIVQALK	KSCSQGGIRG
LNF	DEAH	CNPGP	WKALSH	VKYGKPT	FGTGKTTF	SLSICNAHV	VEGPIDSLFL
IKF	GKIW	YNVID	VGKRFC	PFCVKSK	YGKPVQIK	IGFKTRYGM	FLPEKTLGWQ
RVW	MDEN	QRDWQ	HAIAQC	HGKTTLF	QLRKALGH	IFLTYPQCD	DKQGARWTGR
IMV	RDWQ	IRIYF	QEAGKY	MGKTTLY	AQEAGKYL	GIELLPDFY	KVSDAAPYIF
WGP	CAYC	GDTDS	FPETVH	EINNVIH	KIRIYFYD	WDTETTGLP	QVSDAAPYII
KPA	KRFC	KYGKP	IKLKNH	GENALLG	PSTATVKN	KGARWAGEA	DYETAVREFI
WNG	RFCV	SDVTR	HFKEFM	FYATLKK	VCISDVTF	GEMTVAGKK	ASLYPSIIRA
AVP	CGCS	IFLCN	GIPTIF	FKHFKEE	IVRDRRPT	RGARWAGES	NEMDAGIYYA
KCA	PMYR	CSLTK	WARVAT	CRFCVKR	GPSTATVD	YFLTYPQCS	DEIIDNSVDE
NGV	DPPY	RIYFY	YNVIDD	PLGVISW	VGYSQGAE	FNKITKGGL	LPTLYFSADM
VFD	RPMY	SNDAW	CSLTKE	PYNHQEP	LTRGNGIL	MIFLAMLVI	RFLRGQLALV
IHD	HLHV	RVGKR	CVSDVT	QAAAIGY	TGKTMWAR	GTLFLTEGK	KSPKWLNDLI
WRP	YRKP	HDISH	NIKTKN	HRVGKRF	LKLKRLRG	MIHSRSYTY	SAGLPATADF
WLR	RIYF	ENHTE	HIGDLM	AENHTET	ERGNGITL	MIHSRSYTH	FDLNSLYPHL
WPI	PAGT	LLLYM	ASNPVY	YGTGKTN	NIIENGVT	IAEVERLRS	NKPGDDFQLG
ITS	LCNP	LYMAC	WARALG	RTGKTMW	LGQVFNMV	GFGAGFGAG	MIDLPPLGGT
CTH	CAIC	HDKRM	HTNSVM	GGIPTIF	IEALSQLG	LRVLAALSR	WLVRDRRPVT
LNS	YFYD	WYNVI	QINVAH	RLLLYMH	TLKNHTNL	RRVLAALSR	TVSGAVPGQM
KRY	CVVC	TLKIR	ELLPDF	CDSRTGF	ASINNVIY	KKALGIHKA	SVSGAVPGQI
RGW	LLYM	RGTHH	YLKLKH	YNLDRIV	SRVLAALD	KSIELAQDS	GGGDIYHNTT
LYN	NCKY	PVQIK	HVTGGH	DGTGKTD	VIENGVTD	RKALGIHKC	EKQGARWTGM
PGW	WTFP	EGDSR	YVSFAC	DNYLCGT	RIRFYVST	YSIELAQDL	PHLHVLIQFE
AIN	HRVG	NDAWY	WDIEIC	GKRFCVK	ASTATVKS	SPTGSGKSL	ACNLGHINLS
FSN	NCRC	PHYLK	WDLGGM	IRDRRPN	IKLKNHTN	VIGLHHVTG	EIARMYGVTR
LLS	CNLC	KNHTN	SRTGKT	DKYGKPF	YKLKRLRS	IIGLHHVTA	ATGNAAIEEA
LVG	DTDS	THRVG	CALINM	FGTGKSE	NENGVTLI	EVNRFIIYA	SFDKQGARWT
HGC	GFTH	FLCNP	RVGKRF	ACTHASN	FAEVERLA	TENALLLYM	DIARMYGVTP
YAS	KNHT	TGGQY	CVIGLH	WDDVDPE	GKVMCISD	SGMYASALN	PCNLGHINLA
RAL	WPWP	RSLGP	NIVAAH	VLCNPGE	GQYASKEQ	STPNGLNHY	THVVYNHQEQ
FTN	PVQI	MFFLV	WLVGEH	WPVQIKS	FNGAGKSF	GDFLTSLIN	LADRIADRIA
IVW	NHTE	CAYFW	CSLAAD	WMDENIK	RSLGPHNY	ITLFKEIRR	WYDPLAQSFI
GTD	WGHP	YLKHF	PIDSLF	CAARAAH	HGLPATAE	PTIGIGHLI	VFCIMLGTGM
ISI	SNPV	VVYNH	TLKIRI	FAGKSTE	QSYEQRHD	WNISPETII	IFCIMLGTGT
IYL	WYMW	GFRCM	MIIATY	GCVKSVS	FGSLKAEG	AVAIFLAHY	SSHQYGGTTL
GFN	FRCM	PLYFK	KAELRP	ECVKSVE	QRLGRVGR	VVAIFLAHF	EPIAYNATPN
RRS	CAFC	EQRHD	EAGKYE	PTIFLCN	HAAAAVAM	DAELNAILA	NFLRGQLALI
YGP	NYLC	HFKEF	WNGSLK	RALAAGM	ALRDRYQN	AFKTRYGIC	TEATDTSFVL
LVA	HKCF	AQRDW	FFLAAW	SEFMGAM	AGAGKSTS	WDPLAQSFL	LKPGDDFQLA
DRF	VFNM	SPKVY	FNIASY	AVYATLP	GETVHGFS	RDLECGCSA	REKIHGTNFS
NVG	WYVD	IYFYD	ENIKTK	VIDDVDP	FAKFKGKL	RALDNLLDY	QRLRDHGEYM
YSN	PLPW	CGHLD	MLAVKY	VGDSRTF	SNTKYGKP	KAAELRNFA	LLAHVGYPRL
LHV	FLTY	IKGGI	SNPVYA	FLCNPGP	KATNIIEW	SKEQALVKR	WVVEFDPNIP
EYV	FWKH	CMKID	PIAGLE	YQRDWQL	VATNIIEN	LEINREVVD	PAGTGKTTLT
NNL	GDIV	PTIFL	HLEAIF	WYLKHFY	KSVYVLGK	LSGIKGQIG	EDLNSLYPHI
KIA	ENHT	YFLTY	MVTAPC	DPHYLKH	IQFDSSLY	AGTGKTTLT	ARIFGGAWEQ
IIK	GTHH	NPVYA	YPAGTW	NFMGAQI	CVSDVTRG	RIHSPSRVA	GRIFGGAWER
GYN	CWEC	PHNYL	WLAGGW	FEQALVA	VDLIRDLQ	CNPFLRPEL	IADRIADRIT
SAI	YPQC	GKIWM	SDRRPQ	ARSLGPH	IKLKRLRF	IFLTYPQCS	WVGIAVDTGN
GND	LWFM	YNHQE	WDLTNC	GRELHEF	DVDPHYLK	ILTEGDSAA	LHGEDPHPFE
YNN	DPHY	WARSL	CGTALC	HNYLCGH	FSLKDPIP	HTKQAIELS	AFIQDIYDKI
IIS	EPWH	DRKPH	HVLIQF	NKRFCVG	KVCVDDFN	DDIDDIDDI	GGIRGGSATC
QTI	CECG	LCNPG	WDLDKD	CGHLDLS	NIDLHYFS	RTKQAIELL	HMQATLPGGT
NYG	FNMF	NMFDN	HASNPV	GHRVGKS	SLKDPIPW	PGAGKSTMM	FHGEDPHPFA
ESF	GARW	FCVKS	IRIYFY	TLVRDRH	PIPWKLYY	AQFDSSLTG	NGIRGGSATV
GHL	HQEA	DVTRG	YGDTDS	HFKEFMG	IVRGLLCT	SWWRNYAHA	SVNRFIIYSE
YIV	VWCI	VGKRF	HTENAL	QWMDENS	TPTRQFSS	VLFGKPFRS	VGSGKSTGLP
IKQ	AWYN	YMACT	LVRDRR	TLKIRIY	LVRDRRPT	FWTAKKRYA	KPKSIGVATT
IFV	CQIC	VTGGQ	PVYATL	KGGIPTI	LFRAPTVD	SQFDSSLTP	EVVFKHDYEE
IYK	PLYF	DWQSN	CTHASN	YTGGQYV	EKTLTGKV	IKGGIPTIF	ANTDCDGDKK
DYV	PTIF	GPHNY	THRVGK	KHGFTHA	VYATLKIR	VQIKGGIPT	AICNAHIPGN
YIK	IYFY	LGPHN	NPVYAT	NYFLTYP	AAKFKGKK	QVFNMFDNE	HPWMSPAGYR

Table A11. Characteristic n-grams in disordered regions by z-score values

N-grams presented in the table have abs(z-score)>2.58 in disordered and abs(z-score)<1.65 in disordered regions. Table includes for each length first 100 n-grams sorted according z-score values in descending order.

N-gram length							
3	4	5	6	7	8	9	10
DDD	SSSS	WALKC	KSSSDV	HAPAPAH	YSSSSSSI	WGGGGGGGF	PAGGGGGGGR
QQQ	WSFL	CRKRW	WKKKGW	EPPPSPF	IPKPAPKA	AAPKPAPKK	FSSSSSSSSY
DED	GIQG	SSSSS	ISASAY	PGDKGDM	WRRRRRRW	GPEPEPEPH	RAAAAAAAAI
HHH	PSPP	PKPAP	AKSSSD	FGGGAGH	FAGGGGGE	QGGGGGGSC	FEEEEEEEES
MSP	PQGP	CDGSC	SPPPPS	WDEDEDW	CPAPAPAC	VKPAPKPAV	FGGGGGGGAS
PPE	NNNN	IQGAK	YPLSPY	MGGGGSQ	GKPAPKPI	NSSSSSSSN	SGGGGGGGAL
PDP	MWDP	WEREW	MPAPPC	IAKSSSI	SQGPKGDV	DAPAPAPAD	PTTAATTTAV
EAE	EPEP	EEEEE	HGGRGM	KSARGGF	TGGGGGRI	NGGGGGGAD	VPPPPSPPPL
MNI	QGPQ	HLVEF	CASGAC	LGPEGPF	LGGGGGGK	VPSPPPSPH	YGGGGGGGGA
SPS	MDSR	FTKRH	CESSSQ	FGGSSSP	RPEPEPEQ	AEPEPEPET	ASSSSSSSDE
KEK	DWSF	SSDVK	YAELKF	HGGGSGL	QDEDEDDY	KGGGGGGGW	LPPSPPPSPL
WYC	GEQG	RPADI	WSGGSI	VAPKPAR	NPSPPPPY	VAPAPAPAK	SDEDEDEDER
SES	HHHH	GPTGP	IGGGAE	KGAGAGM	KPEPEPES	KPKPAPKPK	RAGGGGGSGV
EME	DKGD	KGPPY	CRERAN	MSSDVKV	PPEPEPEK	DRSRSRSRD	VGGGGGGGAG
GGP	APKP	WSPPF	MKGDKP	MPAPAAD	KGGGGGYR	SPSPPPPSP	EPAAPAAPAP
PNP	WLNC	CGPEF	PRELNF	TQGPQGF	HGGGGGYQ	AGGGGGSGR	STNGIEPPRG
MGG	ISMC	QSAND	SSSSSS	HPAAAPD	LGGGSGGP	PGGGGGGGK	AGAGGGGGGR
DSE	PTGP	LVTTF	ISTPAS	IEDEEDV	SAAPAPAQ	DSRSRSRST	TEEEEEEEEI
ESE	DDED	GAKSS	HPKGDH	QGSGGGK	PSDSDSDF	GSRSRSRSQ	PSPPPPSPPT
MAN	GDQG	WAVQW	QGAKSS	HGGGGGY	QSGGSGGP	NSGGGGGGR	HGPAGPQGPR
MKP	SDSD	SARGG	YSLEEF	VSGSGSQ	ANNNNNNL	TDSDSDSDT	QSSSSSSSCT
QKQ	PKPA	GGGGG	CEDDDK	HGGNGGK	VSDDDDDF	PPSPPPPSP	AGGGGGSGGV
PHP	GERG	CGDDC	CKRLRC	TAKSSSK	GQLKGSSQ	VDDEDDEDQ	GSDSDSDSDG
MGL	AHFH	TTDPW	YAARAC	NAGGGAQ	SQLKGSSS	KPSPPPSPK	VGPKGDTGAD
MKG	WCCW	HEQDW	NPASAE	KEDDEDH	TTGGGGGP	EAATTTAAL	MAAAAAAAAE
PQL	LMPC	FASFH	CKEKVH	KPAPKPA	DIQGAKSA	AQQQQQQQM	VSSSSSSSSW
MDG	GLQG	SDVKS	CAPLPM	WSSSGGW	GPAAPAAS	EPTPPPTPE	AEEEEEEEER
MNP	WYPQ	TTTTT	GEEQKF	HEPEPET	HAAPAAAC	VGGGGGGAQ	SPPPSPPPPT
NPA	PALP	AKSSS	FDSDDM	IGPEGPL	NPAAPAAG	PPAPAPAPP	QQSANDAYAE
DDP	GSTG	PSPPP	HAGTPN	GAKSSSD	RSSSSSCN	SSRSRSRSA	RPMNRKPRMY
GQG	GKDG	PPPSP	MDEEEF	YGGAGGD	VPSPPPPY	AGGGGGGRV	NAAPAPAAPE
RMR	EDDE	PPSPP	WPRPAM	KDDDGDT	QSSSSDSM	GGAGGGGGS	ATNGIEPPRG
PPK	VKSY	CTGKW	DPKGDF	GAPKPAT	PGIEPPRE	KTTTTTTTK	AGGSGGSGGA
EYE	IWDQ	FMKKW	MEEKKF	GEEAEEM	DSPPPSPR	MGGGAGGAV	IGGGGGGSGH
SKS	PSSP	YSGKW	NDAAAE	RGGAGAN	FGGGGGYG	AGGRGGGGK	APAAPAAPAA
LAP	DSDD	VMGGH	HGGGDN	IRGGQQP	YGGGGGYA	GPKPAPKPS	RYGGGGGGGG
MPP	GHMA	CAPGF	FEAAAD	TEEDDDM	VPPTPPPV	TPSPPPSPI	YGGGGGGSRF
DTD	QHIS	YASDC	CSSTSC	NHPNIQP	QGGGGGGQ	YSDDDDDDS	KPGGGGNGGH
ENE	WAPW	QTAND	MGTGGQ	PPKPAPQ	ERGGGGGY	RYGGGGGGG	TSSSSSSSDG
FGF	YWFW	MVASM	YEEVEH	QAEAKAH	RGSGGGGN	ASSSSSSCV	CQSANDAYAE
TTG	DNDD	YQRVC	EPKGDE	EGGTGGP	LDDDEDDN	LSSSSGSSL	STTTPTTTTA
MGP	MKTY	APAPA	MRSSSP	GPQGPQG	TGGGGGAV	KAAPAAPAK	YAGGGGGGGL
MKS	PLFQ	FTALM	VELADH	PSRSRSH	TDDEDDDL	PAPAAPAPK	RGGGGGAGGA
PQI	PPGP	TERHT	RGDKGF	REEDDDN	PDEDEDEY	GGAGAGGGA	SEGDRRRVRI
MAD	FYHY	GPVGP	MPPLPK	TRRARRN	EPPPPTPD	QSDSDSDSE	TGGKGGNGGS
NQN	GTGG	ANDAY	GSGLSM	DGPEGPD	KPPPPTPN	SPSPPPPSL	VDGKDGKDGV
ESG	MGNL	WSFLK	FAATPC	QSSGGGY	KGGGGGPF	PPPPSPPPP	GDRRRVRIEV
MEA	NYGH	IVIST	MDSRTG	HTTAATL	TGGGGGPM	PDDDDDEDP	QGPQGPKGDG
GSL	GAKS	YIVKY	YPLPAM	FAGAGAN	RAAAPAAN	AGGSGGGGK	VSEGDRRRVR
GMI	DEED	YGLGW	YAREQT	EAAAAGR	FGGGGSSC	SGGGGGGAV	GAPAPAPAPS
TGP	GAPG	PAAAP	FIEKLI	MRRRRGE	DAAPAPAS	APAPAPAPA	KRRQKREDER
PPF	QGPK	YRQEW	STPASK	QGNGGGC	SGGAGAGL	PGRGGGGGC	RELLDLARQQ
EMQ	KGDT	WARAF	HPPPER	GSRTGEG	ADDGDDDW	PDSDSDSDV	TLTQQEQQAQ
PVE	PTPP	YGADY	WASTGH	SSRTGEK	AAAAEALR	ESSSSSSSA	EGPQGPQGPE
MNS	GSSG	NGIEP	SKRPAD	VPAPKPR	AAGAGGGM	KYGGGGGGS	ASSSSSSSSD
MAG	PVGP	GPPGP	NSPTPY	DAAAEAL	TAGGGGGR	ISGGGGGGP	PGEDVNSLVI
VMP	HLMC	QLKGS	EGDTGW	NSSSSTP	LAGGAGGK	RPAPAAPAR	SGGKGGNGGA
MDF	VMKW	FVENM	FLDELW	SPLPPPW	KPPAPPAE	FGGGGGGGA	AGATGPQGPM
MTN	YLVT	YLPFW	FSKSPW	DAKSSSA	TTPPTPPE	NESWASRSE	GDKGDKGDKG
GGR	SIRT	CDSSC	GERLEV	YDGDGDL	PGGSSGGA	VPGGGGNGA	NLAAARASTQ
TMA	DGDD	CKITC	EPEPEP	EPPAPPE	YFGGGGGW	SAGGGAGAD	QEAPEWAPPK
MEP	SSTW	SQLKG	YGGAGT	EPKPAPM	QGAKSSSD	GGAGAGGGG	KGGNGGSSPS
SLS	GEEG	WVRPI	FSSGSV	RATTTAR	DAAKKAAN	SAPAAPAAK	VRDALAGKRA
RYH	KKKK	YTAQF	MSGTTL	NSPSPSN	IGGGGSGE	VGGGGGGGA	NGGNGGSSPT
WFC	DGGD	QGPKG	NTASDF	QSSSDSY	TAEKAAEN	SLGSGLSMS	QGLGTEAPSN

TES	PNSP	CNTAH	YADADM	WEPEPER	QSANDAYA	AGPQGPKGE	DPAPAPAPPK
MPK	QHFA	FVSDC	FAPRTW	KDGRSAI	KAAPAPAK	KGKDGKDGC	EGPGGPPGPE
NGR	QPQQ	YGVLC	YLPSLW	PAGGGGK	PAATTTAI	KPSPSPSPK	HKSGKNKGQP
YPY	PSEP	WDDIF	YPRRRY	LGPTGPE	DAAAEALN	DNPASAEAI	LMPCESSSQV
DMA	PEVP	HGPRN	CNRRRI	KRRAARI	DGGAGGGD	GEGEGPGGE	GEEEEEEEEG
TPL	PGAP	CQQQC	YPARPF	NPPTPPM	QSSSSGSP	APAPAPAPP	FMPCESSSQI
DPQ	DSDS	MVKPW	QKLKKY	GDEEDEL	STGGGGGR	APKPAPKPA	AATRAVTAAG
PIG	QDVQ	HGFQM	FTPSSH	FAERAAS	VAAAAEAN	VGPVGPQGS	PGGGSGGGGA
EHE	YHAY	WNESH	CTAPAY	PPPSPPP	FAPAAPAS	ERTATETRR	AKGDTGAQGE
APY	IVIS	SSSDV	GPPQPM	MPPPPTA	QPAPPPPV	PAPAAPAPA	AGTPLRRYPL
HYP	PAIP	NNNNN	WTSKPH	PKPAPKP	HRSPSPRK	PPPPSPPPS	GGVSGNPRAN
QLR	QPQP	FNVPQ	GPNIQC	KAKSSST	KDEEEEEK	AAPAPAPAA	MSGLLDDGAN
IKG	MPCE	HALDH	LPNIQN	YAAATAE	SSGGGGGL	AGGGGGGGN	RGVSGNPRAD
GTS	DDSD	MSKRP	KGGGGK	PEPEPEP	TGGSGGSA	TGGGGGGVD	SGTPLRRYPM
SHS	DDGD	NTERH	QESDDH	GPQGPKG	NSRSRSRG	KGPQGPQGQ	TAAAPAPSKG
RHH	DDDP	FGFGV	NERLEI	PAPAPAP	SNPAPTSE	EEGEGPGGP	FKGAKGDKGE
GEL	NECY	FPDFH	LPGPGC	PAGGAGN	APPLPPPA	DGAGAGGAD	PAAAPAPSKP
TPP	SRYC	QGIQG	CTTTAL	NDEDEEY	YSGGGGGK	NGRGGGGGV	EPAPKPKPAA
RPQ	PAVP	HDTNM	SPDPDT	YLSSSSE	TRSSSPSV	LRGGGGGGK	PGPEGPQGPA
GYE	PSKP	VRGSW	YESLPC	GPSPSPG	RGGGGGYV	ATDLRGSGG	APGGGGNGGD
KTT	PVEP	YSDQM	YLERQH	NPPQPPG	RAPAPAAR	YGGGGGGGG	GGGGGGGGAG
IAG	TTST	PAAQW	CPSGSH	SAAAKAV	FRSSSSSP	KGAKGDKGE	AGGGGGGGRR
MGD	EEDE	MGSLI	FVESEW	AEDEDEV	MSSASSAT	KGAKGDKGD	DSSSSSSSGE
IRA	RLIH	FDEPH	GAKSSS	LSSSSEN	VRSSSSSC	TYGGGGGGA	AGGGGGGGVR
DYE	PSLP	HMSHH	MAAGPW	LGGGSGP	ELNPAPTS	YYGGGGGGA	QENTERHTAG
MRA	QMIA	QINGW	HSPSPG	CGGGAGS	LRRRLERG	APKPAPKPK	EEEEEEEEEE
SLG	GEPG	FGPHM	FQAPAR	GPPPAPR	CESSSQVS	SGPAGPQGA	VKGDKGDPGN
TAI	APAP	HSTQV	PKEQEP	IKSSSDM	ALRRRLER	SDPREEQVS	LSDEQLEALL
PEE	PLQP	YDESY	ISPADY	MGRRRSH	DLSDEELR	KPAAAPAPS	GAGGGGGSGR
LNP	WDPL	FQMRF	FELQEP	IEDDDDM	PAAPAPAA	SDDEDEDEI	TGPKGDKGDN
DID	PTEP	RSPSP	NATAAM	VRRGRRE	IIISTPAS	PGGGGNGGH	GEGEEGGGEG
GDP	RGGQ	YLVTT	KPAPKP	TSTPASD	TTTAATTT	VGPTGPTGD	KDLTESQKEK
IQG	LFQD	YVRVH	PKPAPK	NAELEAR	ILEEAQRL	AGGGGASSG	NSKFSEKKKS
RTA	KVFI	TASDW	RSRSRS	MTGGSGP	ESSSQVSN	GDAAAAAAP	ESSYLDARHK
ATR	GVQG	PPPPP	PEPEPE	GAEAERY	SSQVSNST	DARAAAAAP	SKVGRFTVMT

Table A12. Characteristic n-grams in ordered regions produced by combination of z-score, fractional difference and mole fractions

N-grams presented in the table have mole fractions>1E-6, abs(z-score)>2.58 in ordered and abs(z-score)<1.65 in disordered regions. Table includes for each length first 100 n-grams sorted according fractional difference in ordered regions in descending order.

N-gram length							
3	4	5	6	7	8	9	10
SAL	DALA	AALAR	PCKVQS	EGPCKVQ	VPRGCEGP	DNEPSTATV	SFDQVPEELE
LDD	ARAL	ALAAA	ARSLGP	QSNTKYG	PRGCEGPC	EGPCKVQSY	VGSGKSTGLP
STL	LSLS	LAALA	SRTGKT	HASNPVY	CKVQSYEQ	RGCEGPCKV	RKPRIYRTLR
VDE	LADA	AALAG	YGDTDS	QIKGGIP	EGPCKVQS	LVAEVERLR	NEPSTATIKN
LSN	LDAA	GAVAA	VRDRRP	THASNPV	EGDSRTGK	SRRKFLNQV	NYIESHRDEY
VKA	LLEK	LAALS	IKGGIP	ACTHASN	QSYEQRHD	SSYKEFLDE	FDNEPSTATV
NLK	ELLD	AAGLA	NVIDDV	ASNPVYA	VQSYEQRH	THASNPVYA	EGDSRTGKTM
EID	RALA	AALGG	YNVIDD	MACTHAS	SNTKYGKP	VYATLKIRI	FKEFMGAQRD
TLK	VDAA	AGAAV	HASNPV	SNPVYAT	THASNPVY	TENALLLYM	HYLKHFKEFM
LNS	ELLA	LRKAL	ASNPVY	NPVYATL	CTHASNPV	EGDSRTGKT	VIDDVDPHYL
RAL	RRLL	AAALL	YMACTH	CTHASNP	NPVYATLK	GAQRDWQSN	MWARSLGPHN
DDV	LLKE	VVAAA	LYMACT	PVYATLK	VYATLKIR	IKGGIPTIF	PHLHVLIQFE
ALD	EALL	AVAAG	QIKGGI	YNVIDDV	KIRIYFYD	GPHNYLCGH	DFGQVFNMFD
TTV	VEAL	GTGKS	SNPVYA	NHTENAL	TGKTMWAR	PVQIKGGIP	KVTGGQYASN
TIE	ALLS	AAVGA	MACTHA	HRVGKRF	LKIRIYFY	SLGPHNYLC	MDFGQVFNMF
DIS	LGAA	ALALA	THASNP	THRVGKR	TLKIRIYF	ENHTENALL	QSNCKYGKPV
DIE	GAAV	LLAAL	ACTHAS	RVGKRFC	YATLKIRI	WYNVIDDVD	TVTGGQYASK
VAR	TLTA	ALAVA	NPVYAT	WMDENIK	SRTGKTMW	CGHLDLSPK	ERIQRLGRVG
GEV	ADAV	AALAV	CTHASN	LGPHNYL	ENALLLYM	MGAQRDWQS	VKSVYILGKI
RLG	RALG	ARSLG	PVYATL	YATLKIR	TENALLLY	HLHVLIQFE	NHVVYNHQEA
EIA	AALG	LAAGL	NHTENA	VYATLKI	HTENALLL	EGPCKVQSF	FDRINVRRLF
TIK	ALGG	RDRRP	HTENAL	LVRDRRP	VGKRFCVK	STATVKNDL	KLKNHTNSVM
LAA	VATA	AGTGK	THRVGK	GKTMWAR	NALLLYMA	MFFLVRDRR	LSTAKHSVDI

LNA	VGAA	AAVLA	HRVGKR	IRIYFYD	GIPTIFLC	VATNIIENG	TKYGKPIQIK
GLD	SLGL	GDTDS	SDVTRG	TGKTMWA	KGGIPTIF	QVFNMFDNE	NLNSNLDRIF
AAV	KELI	AVALA	GKTMWA	KIRIYFY	LKHFKEFM	VKSVYILGK	ISDVTRGNGI
VEG	AGAL	SDVTR	GPHNYL	RTGKTMW	DSRTGKTM	WLVRDRRPY	FRCMLAIKYL
LNN	KLIE	GALAL	RVGKRF	LKIRIYF	VTGGQYAS	KVTGGQYAS	QIKGGIPSIV
LTA	VSAL	VTGGQ	LGPHNY	TLKIRIY	CGHLDLSP	VQIKGGIPT	SETIHSRSYT
ITS	LEAI	RVGKR	VGKRFC	ATLKIRI	RSLGPHNY	IWMDENIKT	ALEAIRFYVS
TLD	GALA	AILAA	WMDENI	GKRFCVK	WYNVIDDV	GVISINNVI	IRDLISVIRA
NNL	VAAG	TENAL	MDENIK	KRFCVKS	NVIDDVDP	NTKYGKPVQ	NLPTSAGKSL
SAI	GLGA	LVRDR	YATLKI	SRTGKTM	GKPVQIKG	VTRGNGITH	RVNNYVVYNQ
AGL	DELV	QIKGG	ATLKIR	VTGGQYA	YGKPVQIK	TVTGGQYAS	LKRLRFKGTV
NIK	RLLA	TLKIR	LVRDRR	YENHTEN	PVQIKGGI	HVVYNHQEA	FRCMLAVKYL
VTA	LDAV	TGGQY	KTMWAR	ENALLLY	YNVIDDVD	RDRYQVLRK	ERIVSILEWD
LGD	AVGA	ENALL	RIYFYD	HTENALL	YLKHFKEF	SCMKIDHCV	AVGSGKSTGL
LTN	AALL	WARSL	TGKTMW	NALLLYM	DVDPHYLK	YGTPMDFGQ	TYSPDTLGYD
TVA	LLGG	GGQYA	IRIYFY	TENALLL	DFGQVFNM	ERIQRLGRV	KQAIELLPDF
VNA	SLLT	KHFKE	KIRIYF	GKIWMDE	HYLKHFKE	IQIKGGIPT	KQLSFFWRPE
TLN	AGVA	LLLLL	LKIRIY	VGKRFCV	QEAGKYEN	SVYVLGKIW	LGKTTVVAIF
SGI	ALTG	NVIDD	RTGKTM	LLYMACT	VVYNHQEA	KLSTAKHSV	NLSRQLGKTT
DTV	GTLA	PVYAT	TLKIRI	ALLLYMA	PHLHVLIQ	DRINVRRLF	FLVRDRRPVD
GDV	DLIK	KYGKP	NPLYFK	KIWMDEN	MWARSLGP	KGKLKLSTA	KSCSQGGIRG
PLL	LLAA	VYATL	RFCVKS	DSRTGKT	GFTHRGTH	FKGKLKLST	TSLYPSIIRQ
AGV	ADLV	YATLK	GKRFCV	LLLYMAC	GKYENHTE	PETVHGFRC	LEGLRQKGWS
VSV	ELLL	THASN	KRFCVK	IWMDENI	LVRDRRPY	IPFRAPTVK	ALTPLRGSDP
AID	LAAV	DVTRG	KYENHT	HNYLCGH	RFFDLVSP	QELRVLAAL	FRADVKEFEA
NAV	DDVL	HTENA	VTGGQY	NYLCGHL	NHTNSVMF	NHGFTHRGT	GYDLIRDLIS
LLS	VLDA	THRVG	GKIWMD	NPLYFKI	ATVKNDLR	YFLTYPQCS	ILADGDDAGM
VVK	IAAG	LGVIS	NALLLY	YLCGHLD	VHGFRCML	KNDLRDRFQ	KQIKSRYGDK
VVS	LALA	MDENI	ALLLYM	LCGHLDL	HGFRCMLA	VQIKGGIPS	QQVPINATGS
LLQ	ALAL	LGPHN	ENALLL	IFLCNPG	TATVKNDL	APTVKILSK	SDPKNFQPVM
VDG	LVAA	KNHTN	YENHTE	GGIPTIF	CSLTKEEA	NVIRAVRFA	TSGSGMGKST
GVT	VELL	FCVKS	ENHTEN	GIPTIFL	SLTKEEAL	FASLYPSII	NEMDAGIYYA
VGT	LSLL	FTHRG	TENALL	IPTIFLC	CISDVTRG	TIHSRSYTH	DEIIDNSVDE
VLS	AVAL	INNVI	KIWMDE	LKHFKEF	CVSDVTRG	YVVYNHQEA	IHSRSYTHIM
LVS	DALV	QFEGK	LLYMAC	PTIFLCN	FGQVFNMF	APKDFVLQF	PGPNSSYKEF
SVL	LLSL	YGKPV	LLLYMA	QRDWQSN	FFLVRDRR	CMLAVKYLQ	MGGDFLTSLI
VLE	VLAG	YNVID	IWMDEN	ARSLGPH	IKTKNHTN	HFIVATNII	SEKGVSWAAE
IDT	LVSL	VIDDV	HNYLCG	KEFMGAQ	IVIEGDSR	VLCNPGEGA	VPRRHGKTWF
NIT	NALL	IKGGI	NYLCGH	KGGIPTI	LTKEEALS	VKNDLRDRF	WADNAVSFTA
DNI	LLLK	KGGIP	PHLHVL	KHFKEFM	ISINNVIR	GEMTVAGKK	GDFARPNLFE
VNG	ALGI	VGKRF	LKHFKE	WARSLGP	DENIKTKN	IRAVRFATD	LLVLKNNKGV
DTI	ALGV	FLTYP	LCGHLD	GAQRDWQ	MDENIKTK	GAGFGAGFG	NEQALVKRFW
INN	TLAL	LCNPG	PLYFKI	HFKEFMG	TRGNGITH	LPTSAGKSL	SLPIAGLEDI
RLV	GDVV	SNPVY	YLCGHL	AQRDWQS	KSVYILGK	EGRGQDYHA	IGKVMCISDV
NGV	AGVL	HASNP	CGHLDL	IKGGIPT	GNGITHRV	VNNYVVYNQ	LSLPIAGLED
GTI	VLGA	ATLKI	IFLCNP	CGHLDLS	IKLKNHTN	GFGAGFGAG	LSSSFDQVPE
VLR	GVVA	NPVYA	FLCNPG	FKEFMGA	ATNIIENG	YFLTYPKCS	REKIHGTNFS
TGI	LRLL	ACTHA	GAQRDW	TGGQYAS	NIIENGVT	ELRVLAALS	VPTLYFSADS
LFS	LSVL	GKRFC	KHFKEF	IDDVDPH	VATNIIEN	ERIVSILEW	YLDNLGVISI
ALL	LVLS	LYMAC	GGIPTI	PHNYLCG	VLQFHNLN	GDFLTSLIN	ISKRAGIGIN
SVI	VVAG	DWQSN	QRDWQS	GHLDLSP	LGVISINN	LNFQVWTTS	TPYLRLPIHD
VIE	VAGV	MACTH	IPTIFL	GPHNYLC	PKVYSNDA	RGARWAGES	FVLQFHNLNS
SLI	AGIA	YMACT	EFMGAQ	LGKIWMD	TNIIENGV	RKALGIHKC	IHAELNAILF
TVV	IAGL	CTHAS	GIPTIF	RSLGPHN	VISINNVI	YSIELAQDL	INESGLYSLI
ISI	LLAL	ALLLY	KGGIPT	WYNVIDD	GKVMCISD	INSLYGALG	LLAHVGYPRL
IVS	ALVG	KYENH	PTIFLC	KYGKPVQ	NIKLKNHT	NGLMVWCIE	RVTAEEIRYV
ITV	LVGA	HRVGK	RDWQSN	KPVQIKG	VYNHQEAG	VAFDMRGQQ	VSDVTRGNGL
VVA	LLTL	NHTEN	TIFLCN	NVIDDVD	GQYASKEQ	ALGPHNYLS	YGLNLHYIPP
VVN	VVVD	GHLDL	AQRDWQ	NYFLTYP	QYASKEQA	FLGLPFNIA	YGVFSTGISV
GVV	LTLL	GPHNY	GGQYAS	PVQIKGG	IIENGVTL	KICRELHED	LQTIGRVLRK
LIR	VALL	RFFDL	HFKEFM	VIDDVDP	KLKNHTNS	KICRELHEN	MGFKTRYGIG
LVA	GDIV	GKTMW	KEFMGA	YGKPVQI	VTRGNGIT	LPFNIASYA	PVSPMGCRSF
NVL	LVAL	NALLL	RSLGPH	RFCVKSV	QRLGRVGR	RALDNLLDY	QLIMKSKLPY
SII	LIAL	PHNYL	FKEFMG	FCVKSVY	INNVIRAV	SKEQALVKK	WKHFQTAVKS
VGV	LVLD	IYFYD	GHLDLS	HGFTHRG	YVLGKIWM	TSAGKSLIQ	WVVEFDPNIP
IIS	LVGL	KIRIY	DDVDPH	AWYNVID	MCISDVTR	TTLFLTEGD	ANTDCDGDKK
VVG	AVLV	KTMWA	YNHQEA	YLKHFKE	YDLIRDLI	WLAIQPVIS	DIARMYGVTP
LIA	IALL	IRIYF	VQIKGG	KNYFLTY	DLRDRYQV	AGFGAGFGA	EDLLIRVNEY
LIN	LAIL	LKIRI	KPVQIK	LIQFEGK	ENIKLKNH	AIELLPDFL	EGMATSIAEL
IIK	LVLL	KRFCV	GKPVQI	WQSNTKY	RRPYGTPM	LSGIKGQIG	GAKEAFHPMY
LVG	LLVL	TMWAR	PHNYLC	FGQVFNM	DENIKLKN	LYQSCHILQ	GPAGTGKTTL
NIL	VLLL	RIYFY	SLGPHN	FLCNPGP	NDLRDRYQ	NIFLAMLVN	IGRTWIQITW
VIA	LLLV	LLYMA	HLDLSP	HLDLSPK	SFFSLKDP	SKRYLYQDN	NGPAGTGKTT
GVI	LLIL	GIPTI	LGKIWM	DDVDPHY	YLSGHLDF	CGMYASALT	PAGTGKTTLT
GIV	LLLI	PLYFK	WYNVID	DVDPHYL	FFSLKDPI	TSLYPSIIR	PCNLGHINLA
IVG	LILL	YENHT	KNYFLT	EFMGAQR	NIDLHYFS	VLQFHNLNA	RVAHIHVVNG
NII	ILLL	RFCVK	YGKPVQ	VDPHYLK	SLKDPIPW	VNNYVVYNH	SINNVIRAVD

Table A13. Characteristic n-grams in disordered regions produced by combination of z-score, fractional difference and mole fractions

N-grams presented in the table have mole fractions>1E-6, abs(z-score)>2.58 in disordered and abs(z-score)<1.65 in ordered regions. Table includes for each length first 100 n-grams sorted according fractional difference in disordered regions in descending order.

N-gram length							
3	4	5	6	7	8	9	10
QQQ	PSPP	GGGGG	GGGGGG	SSSSSSS	GGGGGGGG	PEPEPEPEP	SSSSSSSSSS
PPR	SSSS	PPPPP	PPPPPP	PPPPPPP	PPPPPPPP	EPEPEPEPE	EEEEEEEEEE
PPQ	EPEP	APAPA	TTTTTT	EEEEEEE	EEEEEEEE	EEEEEEEEE	HPNIQGAKSS
SPS	PQGP	PSPPP	PEPEPE	DDDDDDD	PEPEPEPE	PKPAPKPAP	PSPPPPSPPP
TPP	PTPP	NNNNN	EPEPEP	PEPEPEP	EPEPEPEP	PAPKPAPKP	SPPPPSPPPP
PPK	PAPP	EEEED	QQQQQQ	EPEPEPE	KPAPKPAP	KPAPKPAPK	GGGGGGGGGG
DDD	APAP	PPAPP	GGGGGA	PKPAPKP	APKPAPKP	APKPAPKPA	PPSPPPSPPP
PPE	APKP	SSTSS	PKPAPK	TTTTTTT	PAPKPAPK	DDDDDDDDD	AATTTAATTT
SES	PKPA	KKKKK	KPAPKP	PPPSPPP	QGAKSSSD	QQQQQQQQQ	TTTAATTTAA
PDP	PPPA	DEEDE	AGGGGG	KPAPKPA	GAKSSSDV	PPPPSPPPP	VISTPASKVR
QKQ	KPAP	PAPPP	PPSPPP	PAPKPAP	PPPSPPPP	PSPPPPSPP	ATTTAATTTA
ESE	PPPR	KKSKK	PPPSPP	APKPAPK	QQQQQQQQ	PPSPPPPSP	RYGGGGGGGG
PEE	APPP	EEDDD	APKPAP	QQQQQQQ	NIQGAKSS	PAPAPAPAP	GDGDGDGDGD
RQQ	PPGP	PSPSP	PSPPPP	GGGGGGA	KSSSDVKS	APAPAPAPA	STNGIEPPRG
PNP	SPPS	PKPAP	APAPAP	PPSPPPP	GGGGGGGA	PSPPPSPPP	APAAPAAPAA
AQQ	PSSP	SPPPP	PAPAPA	PPPPSPP	PSPPPPSP	DSDSDSDSD	PADIIISTPA
PSA	PAPA	RSPSP	PPPPSP	MDSRTGE	PPSPPPPS	PPSPPPSPP	SLGSGLSMSG
SST	PPPT	PPPSP	NNNNNN	PAPAPAP	SPPPPSPP	TTAATTTAA	TDISLGSGLS
MEE	KKKK	QGPQG	SRSRSR	GAKSSSD	AGGGGGGG	ADIVISTPA	AGGGGGSGRR
EAE	SPSS	QGPKG	QGAKSS	QGAKSSS	APAPAPAP	ATTTAATTT	GAGGGGGSGR
SKS	EEDE	SPSPS	QGPKGD	RSRSRSR	ISTPASKV	TTTAATTTA	LMPCESSSQV
KEK	DEEE	PPSPP	GAKSSS	SRSRSRS	MSKRPADI	SKRPADIVI	PAAPAAPAAP
DED	SDSD	PQGPQ	SPPPPS	GPQGPQG	RGGQQTAN	YGGGGGGGG	AGGGGGGGGG
SPT	PPPL	EEEEE	SDSDSD	IQGAKSS	SPPPSPPP	NDAAAEALN	AAPAAPAAPA
RGP	KSSS	SSSSS	PQGPQG	PSPPPPS	PSPPPSPP	TGPQGPKGD	TSSSSSSSSS
REQ	EDDE	RRRSS	SSDVKS	SSSDVKS	TTAATTTA	GDGDGDGDG	GAGGGAGAGG
PTS	DDED	PPPPT	DSDSDS	NNNNNNN	TTTAATTT	GGAGGGGGS	APAAPAPAAP
DSE	RKRK	PPPAP	TGGGGG	APAPAPA	ATTTAATT	RYGGGGGGG	GGGGGGGGGY
MAD	SSST	PSPTP	GPQGPK	SPPPPSP	PAAPAAPA	STNGIEPPR	PAAPAPAAPA
GGS	EERK	GPPGP	PQGPKG	GPQGPKG	APAAPAAP	GAGGGGGSG	IISTPASKVR
AKK	DSDS	GAKSS	SPPPSP	SSDVKSY	EDEDEDED	DGDGDGDGD	GDRRRVRIEV
PGS	PSTS	DEEEE	PSPPPS	GGGGAGG	ELNPAPTS	DISLGSGLS	KPGGGGNGGH
NSS	SDSS	SDSDS	DEDDED	DSDSDSD	GGGGGGSG	SLGSGLSMS	KRRQKREDER
ASQ	KRKR	AKSSS	PPPPPS	PSPPPSP	GPEGPEGP	GAGAGGGAG	RELLDLARQQ
GGP	DEED	SSSPS	PPPLPP	SDSDSDS	AGGGGGSG	MPCESSSQV	SEGDRRRVRI
NPS	TPTP	PLPPP	TTAATT	GDKGDKG	ASSRASSR	PAAPAAPAA	TLTQQEQQAQ
ASR	SSES	GPTGP	PPPPLP	MSKRPAD	DGDGDGDG	SPPPSPPPS	VSEGDRRRVR
RPT	EEDD	TPPPP	SPPPPP	GPKGDTG	GAGGGGGS	AAPAAPAAP	AINALRRRLE
TES	SDSE	PTPSP	GGGGGY	DKGDKGD	ALRRRLER	ADIIISTPA	GDKGDKGDKG
TTP	TSSS	PAPTP	SSSTSS	STPASKV	LNPAPTSS	GGAGAGGGA	LTPSDWSFLK
PPL	DDDE	TGPQG	QGPQGP	SKRPADI	LRRRLERG	AGGGGGSGR	NLAAARASTQ
SQT	KKSK	PPPLP	PAAPAP	DEDEDED	NPAPTSSP	ATDISLGSG	ESARAVREGQ
DDP	ASSS	DSSSS	PPPPPA	TNGIEPP	DAAAEALN	GGGGGGGGY	QEAPEWAPPK
SET	SSEE	EDEDE	MSKRPA	PTPSPTP	AAPAAPAA	SSSTPPSIK	EGPGGPPGPE
NSP	APTP	EEDEE	EEEEDE	GPAGPQG	GGGGGGGY	SSSSTPPSI	FTSSDLAFLK
TGP	SSTP	EDDED	GPKGDT	GGQQTAN	NALRRRLE	TATDISLGS	IPKEQARIDL
ENE	SSTS	EEEDE	SKRPAD	RGGQQTA	NDAAAEAL	DEDDEDDED	QGLGTEAPSN
GRS	EKKE	SSSSP	PKGDTG	GGSGGGG	PAPVPKPA	GAGGGAGAG	RLNKMLKGEK
AKE	KKEE	RRRSR	KRPADI	YGGGGGG	SSRASSRA	NSSSSTPPS	FQTTGLSKAK
GES	DSDD	RSSSS	LPPPPP	GQQTAND	APTSSPTS	SPPPSPPPP	KGGISQQPDI
GDP	DDSD	APPAP	EDDDDD	PTPPPTP	DEDDEDDE	AAPAPAAPA	PEESVGDTQM
QSG	KKRK	EDDDE	TPPPTP	TTAATTT	DKGDKGDT	APAAPAPAA	PKPAPVPKPA
RGG	SAPS	GPVGP	YGGGGG	GSGGGGG	EEQKQLTL	GAGGAGAGG	VRDALAGKRA
ADP	SSGS	DDEEE	EDEEEE	NGIEPPR	PPPLPPPP	HSTQVPIKV	ANLPTTHMPR
KRN	ESEE	STSSS	SSSSGS	GPEGPEG	NSTNGIEP	PAAPAPAAP	GGGGGGGGGA
SKT	KKEK	DDEDD	SSSSSS	PAAPAAP	RYGGGGGG	PQPQPQPQP	YEKKPRSVSQ
KST	SGSS	SSSDV	EEEEEE	PPPPLPP	DISLGSGL	AGAGGGAGA	YGGGGGGGGA
ESG	SSGG	EDDEE	PAPKPA	PQGPQGP	DTPVSEIP	AGGGAGAGG	PPPRHPGRRS
GDS	KSKK	TTTTT	RSRSRS	DDEDDED	GGEGGEGG	APAPAAPAP	AKGDKGEPGQ
DAE	ERRR	QGIQG	AKSSSD	QPEESVG	GGGGGSGR	DNPASAEAI	FKGAKGDKGE
GGR	EAEE	DEDDE	MDSRTG	SSRASSR	ILEEAQRL	GGGGGGSGG	GEGEEGGGEG
GAS	EKEK	EDEDD	GPQGPQ	ADIVIST	ISLGSGLS	QPQPQPQPQ	GGGGGGGAGY
ERG	RRGR	SEEEE	IQGAKS	PPPTPPP	SSSSSSSC	RTATETRRG	HKSGKNKGQP
NNS	EAER	APAAP	NIQGAK	DEDDEDD	GEGGEGGE	GGGGGGAGG	KDLTESQKEK
NNP	SAPA	DDDED	EDEDED	NPASAEA	SLGSGLSM	ERTATETRR	NSKFSEKKKS
TKA	SSDD	SSSAS	ARGGQQ	RELNPAP	AGAGGGAG	GGGAGAGGG	PELPSLDDID
SAG	NNNN	SSSST	DDEDDE	QPQPEES	APAAPAPA	GGSGGGGGG	PSDWSFLKGI
STG	GGSS	DSDDD	PSSSSS	TSSSSSS	EEAQRLIH	LNKMLKGEK	ELRTERLERI
SGK	SSNS	SSDVK	DKGDKG	GGQQSAN	EGGEGGEG	LPTTHMPRQ	ESSYLDARHK
ANP	SGGG	TSSSS	GDKGDK	GQQSAND	GGGGSGRR	APPAPAPAP	GAGAGGGGGG

NPT	RRRA	PAAAP	SSSSSA	RGGQQSA	GPEEGEGP	DEEYYEEDR	LNENANKDSR
TAE	RRAR	ASSSS	STPASK	AGGGGGS	GPPGPEEG	DRAKANLAA	NGGNGGSSPT
NPA	EKAE	DEDDD	ISTPAS	EDDEDDE	PPPPLPPP	DRRRVRIEV	RGVSGNPRAD
GTS	EEGE	DDEDE	DKGDTG	GPQGPAG	QVSNSTNG	EDDEDDEDD	SGTPLRRYPM
ATR	PGGG	SSSSG	KSSSDV	IVISTPA	SGYRYGGG	ITPSGAVDD	AKFHSPKSPM
SLS	GSSG	PPAPA	PLPPPP	LKGSSST	SSQVSNST	KPGGGGNGG	ENDKTMFEKF
TDD	EARE	DDSDD	TPSPTP	PSPPPPP	TANDAAAE	KPLTQEHAD	LSDEQLEALL
SLP	ASTS	PAPVP	DDDDDD	PVPKPAP	TPSPTPSP	MGLIPTAPL	QGPQGIQGPQ
TAQ	STTS	IQGAK	PTPSPT	VKSYIDK	CESSSQVS	MPSESSSVV	SGAPEMSPAS
ATT	DDGD	GGSSS	PNIQGA	PASMEGN	ESSSQVSN	PGGGGNGGH	ADGGGDPEDI
RTA	STST	SGGGG	DEDEDE	PPPLPPP	KGSSSTSS	PTPPPTPPP	AGGAGAGGAG
ELR	DGDD	DGDDD	SARGGQ	RRRSSGG	QTANDAAA	QQEQQAQLD	LKQIQFKRSK
AGR	GASS	SSGGG	PAAPAA	SPASMEG	SSSQVSNS	QRELLDLAR	PAAAPAPSKP
QLR	RGGG	SDVKS	GDKGDT	ASMEGNR	GSGLSMSG	SDPREEQVS	SSSSSSSSGS
KTT	AKAK	GGGNG	EDDEDD	TPASKVR	GTSARRAE	ATNGIEPPR	EKAEKAAEKK
DEN	KAKA	GGGGN	DDDEDD	APAAPAA	YEEQKQLT	SRLIKASTS	GPQGSPGLNG
DTD	LSSS	SARGG	APAAPA	AAPAAPA	PAAPAPAA	VPEVPEVPE	SGGAGGTTSI
PVE	EARA	GSSGG	GFGSTG	RSARGGQ	PELPSLDD	KRPPPRHPG	TGGKGGNGGS
NGS	TSSA	GSGGG	SSSASS	SDWSFLK	QSGTSARR	KTLAELEAE	TTNNNSTNND
VPE	TTST	AASSS	HPNIQG	ASKVRRR	SGTSARRA	LSTPSLPPA	VGGGGGGGAG
DLE	STTT	GGRGG	PAAAPA	ATTTAAT	APAPAAPA	PEVPEVPEV	AAAPAAVAAD
DTT	VSSS	DDDGD	GSGGGG	SDVKSYI	SQLKGSSS	PKPKPAPKP	RPMNRKPRMY
ETG	REAA	NGGGG	AAPAAP	TAATTTA	GRSARGGQ	QRQAPQGAQ	KLNERTATET
TGR	EVEE	LSSSS	GGGGAG	GGSGGSG	AATTTAAT	TLAELEAEA	MSGLLDDGAN
ELA	GSTG	SSLSS	GGSGGS	GAGAGGG	TAATTTAA	ASAYNGNDT	KEGIPPDQQR
ALE	GSAA	AAAEA	SGGSGG	GGAGAGG	IIISTPAS	EGDRRRVRI	QVPIKVQHRL
TTG	GTGG	GGTGG	GGGAGG	AATTTAA	GAGAGGGA	AGSAAGSAA	GQHISIRTFR
AQL	GAAA	AGAGG	AGAGGG	SIRTFRE	QSANDAYA	GESWASRST	CQSANDAYAE
LAP	AAAV	AKAAA	GGGAGA	GGAGGAG	MSDVVERA	SANDAYAEA	KVRRRLNFDS
DVE	AAAL	AGAAA	GGAGAG	GQHISIR	RPMNRKPR	VSEGDRRRV	VRRRLNFDSP

Table A14. Characteristic n-grams in disordered regions produced by association rules

N-grams presented in the table belong to the body of association rules with head *ORDER_LEVEL='D'*. Parameters used in mining are confidence>=51%, support>=0.0001 and lift >=1.05 or lift <=0.95. Except for n-gram with length two where only one rule exists, table includes for each length first 100 n-grams, sorted according lift and confidence, both in descending order.

N-gram length								
2	3	4	5	6	7	8	9	10
PP	PPP	GHMA	AAPPA	AAPAPA	AAPAPAA	ADTPVSEI	ADIVISTPA	ADIVISTPAS
	QQQ	GSHM	APAPA	AAPPPP	ADTPVSE	AGGGGGGG	ADTPVSEIP	APAPAPAPAP
	PSP	HHHH	APEDP	AGPQGP	AGAGGGG	AGGGGGSG	AGGGGGGGG	APKPAPKPAP
	SPP	SNAM	DDDDK	APKPAP	AGGGGGG	AGTSKVSR	ALRRRLERG	ARGGQQSAND
	PAP	PPPP	DDDKD	APPAPP	AGGGGGS	ALRRRLER	APAAPAAPA	DDDDDDDDDD
	SSS	PSPP	DEDSD	APPPPP	AGTSKVS	APAPAPAP	APAPAPAPA	DIVISTPASK
	PQP	QQQQ	DEEDE	APTPPP	ANDAAAE	APKPAPKP	APKPAPKPA	DSDSDSDSDS
	PKP	QPQP	DSDEE	AQQQQQ	APKPAPK	APTSSPTS	ARGGQQSAN	EEEEEEEEEE
	PPS	EPEP	DSPPS	AQRLIH	APPPPPP	ARGGQQSA	DDDDDDDDD	ELNPAPTSSP
	QQP	SSSS	EAEED	ASMEGN	APTSSPT	DDDDDDDD	DGDGDGDGD	EPEPEPEPEP
	PEP	PPPS	EEDDD	CESSSQ	ASGGGGG	DEDDEDDE	DIVISTPAS	ESILEEAQRL
	QPP	EEEE	EEEED	DDDDDS	ASSSSSS	DEDEDEDE	DKGDKGDTG	GGGGGGGGGG
	RPP	PPSP	EEEEG	DDDDGD	DDDDDDD	DIVISTPA	DSDSDSDSD	GYRYGGGGGG
	PPR	SPSP	EKKKS	DEEDEK	DDEDEDD	DKGDKGDT	EEAQRLIHG	ILEEAQRLIH
	EEE	PQPQ	ESSSS	DEEEED	DSSSSSS	DSDSDSDS	EEEEEEEEE	IVISTPASKV
	EPP	QGPQ	GGGDD	DSDSDS	DTPVSEI	DTPVSEIP	EEQKQLTLF	KPAPKPAPKP
	QPQ	PPQQ	GGGGG	DTGPQG	EAQRLIH	EAQRLIHG	EPEPEPEPE	LEEAQRLIHG
	PPQ	SPPP	GGSRS	EDEDEE	EDDEDDE	EDEDEDED	ESILEEAQR	MSKRPADIVI
	PRP	PQGP	HHHHH	EDEEDE	EDEEDEE	EEEEEEEE	GAGAGGGAG	NSGYRYGGGG
	PQQ	QPQQ	KKEKK	EEEEDD	EDEEEEE	EEQKQLTL	GAGGGGGGG	NSTNGIEPPR
	PPA	PPPQ	KKGKS	EESVGD	EEEEEEE	EGPEGPEG	GAGGGGGSG	PADIVISTPA
	GPP	PQPP	KKKAA	ENTERH	EEQKQLT	EPEPEPEP	GDGDGDGDG	PADTPVSEIP
	SSP	PEPP	KKKKK	GGGGGG	EGPEGPE	GAGGGGGG	GEGGEGGEG	PAPAPAPAPA
	APP	PPQP	KKSKK	GGGGGL	ENTERHT	GAGGGGGS	GGAGGGGGS	PAPKPAPKPA
	SPS	RPPP	KKTSS	GGGGGN	EPEPEPE	GDGDGDGD	GGGGGGGAG	PEPEPEPEPE
	MSK	PSPS	KPTPP	GGGGGQ	GDTGPQG	GGAGGGGG	GGGGGGGGA	PKPAPKPAPK
	PTP	PTPP	KRPPP	GGGGGS	GGGGGAS	GGEGGEGG	GGGGGGGGG	PPPPPPPPPP
	PGP	QQQP	KSASS	GGGGGV	GGGGGGA	GGGGGAGG	GGGGGSGRR	PPPPSPPPPS
	KPP	GPQG	MEEEE	GPEGPE	GGGGGGG	GGGGGGAG	GYRYGGGGG	PPPSPPPPSP
	TPP	QPPQ	NNNNN	GPPGPE	GGGGGGR	GGGGGGGA	ILEEAQRLI	PPPSPPPSPP
	EPE	PAPK	NSSSS	GPVGPQ	GGGGGGV	GGGGGGGG	IVISTPASK	PPSPPPPSPP
	PSS	PQQQ	NTERH	GQQSAN	GGGGGGY	GGGGGGGS	KGDKGDKGD	PPSPPPSPPP
	RRR	PAPP	PAATS	GRRRSS	GGGGGSG	GGGGGGGY	KPAPKPAPK	PSPPPPSPPP
	PPK	QQPP	PAPPP	GYRYGG	GGGGNGG	GGGGGGSG	LEEAQRLIH	QQQQQQQQQQ
	EEP	QQEE	PEPPS	KPAPAP	GGGGSGG	GGGGGSGG	LNPAPTSSP	RSARGGQQSA
	PRR	QPPP	PKPRP	KQLTLF	GIEPPRG	GGGGGSGR	MPKRDAPWR	RSRSRSRSRS

	QGP	QQPQ	PPAAP	KRDAPW	GPAGPQG	GGQQSAND	MSKRPADIV	RYGGGGGGGG
	RPR	PPAP	PPAPP	LPPPPP	GPEGPQG	GPAGPQGP	NATNGIEPP	SARGGQQSAN
	MSS	PPPK	PPPPP	MNETEL	GPQGLQG	GPEGPEGP	NSGYRYGGG	SARGGQQTAN
	MKK	PKPP	PPPPQ	MRSSSP	GPQGPKG	GPQGPKGD	NSTNGIEPP	SDSDSDSDSD
	PPT	QSQP	PQQQP	MSKRPA	GPQGPQG	GPQGPQGP	PADTPVSEI	SGYRYGGGGG
	QQE	GGGG	PSPEP	NATNGI	GPTGPQG	ILEEAQRL	PAPAPAPAP	SILEEAQRLI
	QEE	MWDP	PSPPP	NGGGGG	GPVGPQG	IVISTPAS	PAPKPAPKP	SKRPADIVIS
	SQP	APAP	QQEEE	NKNYGH	GQQSAND	KGDKGDKG	PEPEPEPEP	SPPPPSPPPP
	QQR	KPAP	QQPPQ	NNNNNN	GSGGGGG	KGDKGDTG	PKPAPKPAP	SRSRSRSRSR
	DDD	PGPP	REEEE	NPASAE	GSSSSSS	KPAPKPAP	PPPPPPPPP	SSQVSNSTNG
	MSE	PEPE	RGEET	NTERHT	GTSKVSR	LEEAQRLI	PPPPSPPPP	SSSSSSSSSS
	GPQ	GPPP	RRRGR	PASMEG	KGDKGDK	LRRRLERG	PPPSPPPPS	TTTTTTTTTT
	RRS	DDDD	SKKKK	PEGPQG	KGDKGDT	MPKRDAPW	PPPSPPPSP	VISTPASKVR
	QQS	PRPP	SPSPG	PERGSG	KPAPKPA	NATNGIEP	PPSPPPPSP	YRYGGGGGGG
	QEQ	PPPA	SSEKP	PKGDTG	KRDAPWR	NGIEPPRG	PPSPPPSPP	KRPADIVIST
	KPK	PKPA	SSSVD	PKPAPK	KRPADIV	NPAPTSSP	PSPPPPSPP	RPADIVISTP
	SEP	PSQP	SSTSS	PKPKPA	LRRRLER	NSTNGIEP	PSPPPSPPP	DIIISTPASK
	QSP	PQQP	STTST	PPAPPP	MDSRTGE	PADTPVSE	QGPKGDKGD	GRSARGGQQS
	EQE	PPPR	THMPR	PPPAAP	MPKRDAP	PAPAPAPA	QQQQQQQQQ	IQGAKSSSDV
	RRP	PPGP	TPEPP	PPPPPQ	NGIEPPR	PAPKPAPK	RGGQQSAND	ARGGQQTAND
	PKK	SSSP	PSPSP	PPPPPR	NSTNGIE	PEPEPEPE	RSARGGQQS	AVSQLKGSSS
	GGG	APKP	PQPQP	PPPPPV	PAAPAPA	PKPAPKPA	RSRSRSRSR	NALRRRLERG
	KRP	QGPK	PKPAP	PPPPSP	PAGPQGP	PKRDAPWR	RYGGGGGGG	NIQGAKSSSD
	EPS	PPRR	QGPQG	PPQPQP	PAPTSSP	PPPLPPPP	SARGGQQSA	RELNPAPTSS
	MSN	APPP	PEPEP	PQGPAG	PEPEPEP	PPPPPPPP	SARGGQQTA	GDKGDKGDTG
	MSD	QPEE	RPPPP	PQGPQG	PPPPLPP	PPPPPPPS	SDSDSDSDS	RSARGGQQTA
	RPS	PAPA	KPAPK	PSPPPP	PPPPPPA	PPPPSPPP	SGGGGGGGG	GAKSSSDVKS
	RSP	PPSS	PPPSP	PSPSPS	PPPTPPP	PPPSPPPP	SGYRYGGGG	ISASAYNGND
	EPQ	SPQP	QPQQQ	PSPTPP	PPSPPPP	PPPSPPPS	SILEEAQRL	SASAYNGNDT
	EQQ	PKPK	QQPQQ	PSPTPS	PQGIQGP	PPSPPPPS	SKRPADIVI	PNIQGAKSSS
	KKK	RSPP	PEPPK	PTPPPT	PSPPPPP	PPSPPPSP	SPPPPSPPP	KRPADIIIST
	EEQ	PSSP	TERHT	QGAKSS	PSPPPSP	PSPPPPSP	SQVSNSTNG	RPADIIISTP
	SPQ	SPPS	PQGPQ	QGIQGP	PSPSPSP	PSPPPSPP	SRSRSRSRS	QGAKSSSDVK
	SQQ	PPTP	SPPPP	QGPQGP	PSPTPSP	PSSSSSSS	SSQVSNSTN	SAVSQLKGSS
	PPE	KRPR	PPPPS	QKQLTL	PVPKPAP	QGAKSSSD	SSSSSSSSS	HPNIQGAKSS
	SRS	SSPP	QPQPE	QPEESV	QGAKSSS	QGPKGDKG	STNGIEPPR	GRSARGGQQT
	QPS	PPEP	SPSPP	QPQPEE	QGPKGDT	QGPKGDTG	TGPQGPKGD	RGGQQTANDA
	MTT	PPPT	RPPSP	QQQQQP	QPQPEES	QGPQGPQG	TNGIEPPRG	SPASMEGNRP
	QSS	RRRS	MSKRP	RDAPWR	QQQQQQQ	QQQQQQQQ	TTTTTTTTT	ISTPASKVRR
	RKR	PNPP	PPPPR	RGGGGG	RGGGGGG	RGGQQSAN	VISTPASKV	PQPGQHISIR
	MPP	QQQR	PQGPK	RGRGRG	RGGQQSA	RNKNYGHP	YGGGGGGGG	SAHFHPNIQG
	SES	PKPS	NPPPP	RNKNYG	RNKNYGH	RSRSRSRS	YRYGGGGGG	AHFHPNIQGA
	EPK	SPTP	RRSSS	RPADIV	SDDDDDD	RYGGGGGG	ARGGQQTAN	QPGQHISIRT
	MSQ	EPPP	QQQRQ	RPGRPR	SGGGGGG	SARGGQQS	KRPADIVIS	HFHPNIQGAK
	EES	PEEP	PPEPE	RSSSPS	SPASMEG	SARGGQQT	PADIVISTP	AKSSSDVKSY
	SPR	RRPP	QQQQP	SDSDSD	SPPPPPP	SDSDSDSD	RPADIVIST	SSSDVKSYID
	ESS	EEED	PPQPQ	SNSSSS	SPPPSPP	SGGGGGGG	IQGAKSSSD	FHPNIQGAKS
	PSR	KPPP	EEEEE	SPRRRR	SPSPPPP	SKRPADIV	ELNPAPTSS	SDVKSYIDKD
	APA	QPAP	QPQPQ	SPSPPP	SSGGGGG	SNSTNGIE	DIIISTPAS	SSDVKSYIDK
	EED	HPPP	RQQQQ	SPSPSP	SSSSGSS	SPASMEGN	IIISTPASK	PGQHISIRTF
	RSR	PKPQ	PPPEP	SPTPPP	SSSSSDS	SPPPPSPP	QGAKSSSDV	DGRSARGGQQ
	KRK	SPPQ	PPKPK	SPTPSP	SSSSSSA	SPPPSPPP	VSQLKGSSS	STPASKVRRR
	QPR	SSPS	PQQQQ	SSSQVS	SSSSSSD	SRSRSRSR	SPASMEGNR	ISIRTFRELN
	APS	PPKK	QQEQQ	SSSSSE	STNGIEP	SSSSSSSC	RELNPAPTS	QHISIRTFRE
	QPA	QPRP	RPPPR	SSTPPS	STSSSSS	SSSSSSSD	AVSQLKGSS	HISIRTFREL
	KKP	QRQQ	PQQPP	STPPSI	TGGGGGG	SSSSSSSS	NIQGAKSSS	GQHISIRTFR
	SSE	QPPR	QQQQR	TPAPAP	TGPQGPK	STNGIEPP	GDKGDKGDT	DVKSYIDKDG
	RSS	AQQQ	SSSSS	TPPPPP	TNGIEPP	TGPQGPKG	RSARGGQQT	RSAHFHPNIQ
	PDP	PEPS	PAPEP	TPPSIK	TPSPTPS	TNGIEPPR	RGGQQTAND	KSSSDVKSYI
	SSQ	PSSS	RQQRE	TSETNA	TPVSEIP	TSSSSSSS	AKSSSDVKS	MYRMYRSPDV
	QRQ	QPQA	PPPTP	TSSPTS	TSSPTST	TTTTTTTT	ISASAYNGN	TQVPIKVQHR
	SRP	QEEP	EPEAP	TTAATT	TTTTTTT	VISTPASK	SASAYNGND	ADIVISTPAS
	PSQ	PRRR	QQQQQ	TTTTTT	VISTPAS	VQPQPEES	ASAYNGNDT	APAPAPAPAP
	KRR	PSQQ	QPPQP	YGGGGG	VQPQPEE	YGGGGGGG	RPADIIIST	APKPAPKPAP

Table A15. Characteristic n-grams in ordered regions produced by association rules

N-grams presented in the table belong to the body of association rules with head *ORDER_LEVEL='O'*. Parameters used in mining are confidence>=51%, support>=0.0001 and lift >=1.05 or lift <=0.95. Except for n-gram with length two where only one rule exists, table includes for each length first 100 n-grams, sorted according lift and confidence, both in descending order.

N-gram length								
2	3	4	5	6	7	8	9	10
WW	CWF	ACDW	AFGVL	AAKYEN	AAELRNF	ADGSQFDS	AFDICGVQP	AGPSKHFKSN
WC	CWW	ADWW	AFVGL	ADGSQF	ADSDAFT	AGKYENHT	ALEAIRFYV	AKYENHTENA
CW	CWY	AWCV	AGKYE	AEDPYI	ALDNLLD	ALDNLLDY	AVKSCSQGG	AQDLRAVHGM
CF	CYW	AWNY	AIFLA	AMSRRY	ALLFTWR	ARVATGRE	CAITHIDYG	AWCLMLISRG
WI	FWW	AWVC	AIIGI	ASYALL	ALLLYMA	ASGLADAL	CMLAIKYLQ	CGHLDLSPKV
YW	HWW	AYAC	AQAFF	CAWCLM	ARVATGR	ASSPDAVR	CMLAVKYLQ	CRELHENGEP
IC	WCY	CAVY	ATAYL	CDADGS	ASYGVFS	AVSQDQTK	CNIDLHYFS	CVSDVTRGNG
CI	WIC	CCEY	AVGYV	CLVWDI	ATNIIEN	AWYNVIDD	CSTLKDLIE	DADGSQFDSS
CY	WYW	CCIY	AVYEV	CNIDLH	AVEDLVN	CAWCLMLI	CVIEYRQQV	DAWYNVIDDV
FW	YCW	CDHI	AYTVL	CVIEYR	CRELHED	CMLAIKYL	CVKSVYVLG	DFASLYPSII
IW	YWI	CELF	DIAVG	DAAPYI	CVIEYRQ	DFASLYPS	DADGSQFDS	DLHYFSSSFF
WY	IWY	CFDI	DLLYI	DLECGC	DGSQFDS	DHCVIEYR	DFLQPGIVE	DRRPYGTPMD
YC	CLW	CFVL	DNWID	DLTSLY	DKRMTDN	DIPFRAPT	DFVLQFHNL	EMTVAGKKFF
WF	FYW	CGHI	DYGLY	DNLLDY	DNSLFEI	DKYNDVNR	DLRDRYQVM	FASLYPSIIQ
FC	CFW	CICD	EIKDY	DSIAWL	DSDAFTQ	DLIRDLIS	EAAKYENHT	FCVKSVYILG
VW	WWF	CLMT	EYPLK	EEYTRL	DVVVDFG	DLPCGCSY	ELFGARIHS	FDELFGARIH
LW	WWY	CNAF	FDINN	EFMGAQ	EATDTSF	EAGKYENH	ELPRILVDH	FDRINVRRLF
CC	WVC	CNCF	FFLAL	EGKYQC	ECLPNVC	EFLRETWT	ETSLWTLPD	FLVRDRRPVD
WV	FCW	CYYT	FGLIA	EGVFSL	ERIQRLG	EFMGAQRD	ETYCAITHI	FQPVMGFKTR
VC	YFW	DYCV	FLLGV	FGARIH	EWRVYLG	ELPRILVD	FASLYPSII	FWLVRDRRPY
WL	CWV	EWCG	FTDFP	FHNLNS	FDEILEG	FCVKSVYI	FHNLNANLD	GAIDLPPLGG
YF	CVW	FFMF	FTLAV	FIENET	FQVWTTS	FFSLKDPI	FLGLPFNIA	GARIHSHGNL
YI	WWI	FFWG	FYGLR	FIVATN	FYAKVTG	FGARIHSH	FRCMLAVKY	GFRCMLAVKY
CV	WMF	FHLC	FYSGL	FTLEKS	GCLVWDI	FIIASRNV	GFRCMLAVK	GGDFLTSLIN
LC	CYL	FIDC	GAYYG	GKTTVV	GCSTLKD	FLRETWTR	GGQYASKEQ	GKEFLRETWT
CL	LFW	FIHC	GDIVY	GNGLTH	GDFARPN	FRCMLAVK	GLPFNIASY	GKIWMDENIK
YY	WIY	FIMV	GFPAV	GRGQDY	GDTDSVF	FSLKDPIP	GPVAFSHFD	GKTMWARALG
FY	CYI	FIWL	GIVNV	HFKEFM	GKYENHT	FSSSFFSL	GQIYKHACA	HDKRMTDNES
FI	WII	FKLC	GQLIA	HGMDAD	GPLCKGD	FVHPVGFG	GRETCAWCL	HNLNSNLDRI
IY	WFW	FTIC	GRVTL	HLDFNS	GPPDTGK	GFRCMLAV	GSQFDSSLT	HTNSVMFWLV
IF	WQW	FWLF	GSLLI	HNLRKA	GPVAFSH	GHLDLSPK	GVISINNVI	HYLKHFKEFM
VY	CYC	FYDC	GYAVI	HTFDEL	GQVFNMF	GKEFLRET	GYSQGAIVT	ICRELHENGE
YV	YWF	FYGC	HLLAF	HTNTVM	GQVFNMY	GKSLGLCS	HTNSVMFFL	IIENGVTLDI
II	WCV	FYVC	IAAVI	IARGDS	HGEMTVA	GKTTVVAI	ILVYVASYN	IKICRELHED
FV	WFI	FYVM	IAMAL	IGIGHL	IENGTSP	GLPFNIAS	IVYFAETYC	IPSIVLCNPG
FF	WMC	GCYL	IECNG	ILEWDR	IKGGIPS	GVGPLCKG	KFKGKLKLS	IVKPFPFLAD
YL	YWC	GIWF	IFINY	IPFRAP	IKGLGSL	HACATGSG	KIWMDENIK	IYKHACATGS
LY	YVW	HLCI	IFNNG	ISKNAL	ILGKIWM	HDDLVMSL	KNDLRDRFQ	KALGIHKCFL
VF	CCW	HLWA	IIASR	IVHFKE	IPFRAPT	HNYLCGHL	KNHTNTVMF	KNFQPVMGFK
HW	IIC	HWLL	IIDIS	KELAPK	IQFEGKF	INAKNYFL	KPVQIKGGI	KNHTNSVMFW
CM	YFC	HWVL	IITSL	KGKLKL	IQFEGKY	INVRRLFN	KRYLYQDNE	KSVYILGKIW
WH	IIW	HYYF	ILGYA	KIWMDE	IQRLGRV	IPSIVLCN	KSCSQGGIR	KTKNHTNTVM
FL	VYW	ICTI	ILIGI	KTLITG	KILSKQF	IRCNIDLH	LADALVILA	KVQSFESRHD
IV	VWI	ILWY	INIFL	LAADIA	KIWMDEN	ITHRVGKR	LCGHLDLSP	LARYAFDFYE
VI	FYC	IWFV	IVCLL	LGGVYS	KPVQIKG	IVSILEWD	LCGPVAFSH	LGYDLIRDLI
CH	CMW	IWIL	IVTLV	LGIILL	KQLSFFW	KEFLRETW	LEAIRFYVS	LHENGEPHLH
LI	FWC	IYWE	KHNLV	LGNDLR	KSVYILG	KFIKICRE	LFVNILRLE	LINSLYGALG
LF	ICW	KCYG	KIAYT	LGQQLS	LCGPVAF	KGKLKLST	LGPHNYLSG	LKHWKELIGA
IL	VWC	LCHF	KPDFV	LGVVPV	LCSLAAD	KIRIYFYD	LKLSTAKHS	LLLYMACTHA
AW	CFI	LFCL	LAIFL	LGYTDA	LEGVNGE	LAELCGPV	LNSFTLEKS	LVYVASYNEV
HC	WYF	LQWW	LGLVN	LHVLIQ	LETSLWT	LGKTTVVA	MDENIKTKN	LWTLPDNPLD
WM	HFW	LWCA	LIRLF	LIAAAP	LKIRIYF	LNSNLDRI	MTDNESLQA	MTDNESLQAS
WT	WAW	LWFM	LITTM	LIQFEG	LKNHTNS	LPTPIMAG	NAKNYFLTY	MWARSLGPHN
TW	VFW	LWHT	LLIDI	LKTGMY	LLIRVNE	LRDRYQVM	NALEAIRFY	NALLLYMACT
CT	WLY	LWIY	LLLDI	LLALLA	LLTVGHP	LRRLGAPI	NFQPVMGFK	NCKYGKPVQI
WA	YIC	LWVV	LLLGI	LLVGYG	LLVYYCW	LSGIKGQI	NNYVVYNQQ	NDLRDRYQVM
CN	YCF	LYIC	LLLIV	LMLISR	LMLISRG	LTYPKCSL	NPGEGASYK	NIKLKNHTNS
WN	CCI	LYNC	LLTTF	LPCGCS	LNSNLDR	LVRDRRPY	NVIRAVRFA	NIKTKNHTNS
NW	CYF	NFTC	LRLIV	LPIHDE	LPRILVD	MKIDHCVI	PETVHGFRC	NLPTSAGKSL
GW	IFW	NWYI	LSVII	LSFDVT	LRDRFQV	NDAWYNVI	QEAGKYENH	NLSRQLGKTT
VL	IWW	NYLW	LTLVG	LSHDLT	LSGIKGQ	NLNANLDR	QPGIVEWNK	NNVIRAVRFA
AC	YCL	PQLW	LVFLA	LYNGGP	LSNALYG	NYKYQYDK	RFWKVNNHV	PFRAPTVKIL
LV	FIW	PWKF	LVNRA	MALPPC	MCISDVT	PFLADNSP	RGNGITHRV	PHNYLCGHLD
CA	CVI	QVGC	LVYAL	MLYMAC	MEGGGVG	PGIVEWNK	RIQRLGRVG	PKDYVLQFHN
YM	WLC	RWAW	NLIYQ	NDVNRW	MKIDHCV	PKNFQPVM	RLMEGGGVG	PKVYSNDAWY
GC	WCI	RWFP	NLMIL	NHNLRK	MTDNESL	PNIMMNNN	RPYGTPMDF	PVQIKGGIPT
NC	WHY	RWGW	QIVAL	NLHYIP	NDAWYNV	PTIFLCNP	SAFDRINVR	PYVALTPLRG
HY	YCI	TVCH	REGWA	NLPRIA	NHQEAAK	PVQIKGGI	SLQASWTFP	QIKGGIPTIF
WG	CYV	TWAI	RIVAI	NPVYAT	NIFLAML	QFEGKFQC	SNPVYATLK	QLGKTTVVAI

YH	FFW	VCVL	RLVRV	PKCSLT	NISPETI	QKDWQSNC	SVYVLGKIW	QQEAGKYENH
CG	YIW	VFWD	RVGIV	QIMAHF	NLDRIFT	RAPTVKIL	TGGQYASNE	RLETSLWTLP
TC	YWV	VHCL	RVLAW	QRTHFA	NSNLDRI	RCMLAVKY	THFAKFKGK	RQQVPINATG
VV	WIW	VLWH	RVNNY	QWAYDN	NSVMFFL	RCNIDLHY	TKYGKPIQI	RRIGVGHLGV
HF	WCW	VSMC	SFPNI	RDSIAW	NWRNIVK	RDDIVYFA	TTTPNGLNH	SLQASWTFPI
CD	WCC	VTMC	TDTSF	RFKGTV	PIPWKLY	REWRVYLG	TVVDNTLMV	SPKVYSNDAW
FM	CWC	VYIM	TGKIY	RHGKTW	PSIVLCN	RGQQKRFA	TYPQCSLTK	TAVKSCSQGG
FH	WIV	VYWG	VAIVV	RTHFAK	PVISSGR	RIGVGHLG	VAGKKFFLC	TGRETCAWCL
CK	IYW	WACL	VALLY	RVFKTQ	PWKLYYR	RINVRRLF	VEIHDKRMT	TKNHTNSVMF
QW	CIY	WFFN	VDLFY	SLGYIG	PYVALTP	RLETSLWT	VHGFRCMLA	TKNHTNTVMF
YT	ICC	WICK	VIIII	SNPVYA	QFPSTAS	RVNNYVVY	VKNDLRDRF	TMWARALGPH
YA	CFY	WISI	VITLG	SRSCMK	RETCAWC	RYQVLRKW	VLGKIWMDE	TVAGKKFFLC
LL	WHC	WIVY	VLAFL	SVIVEI	RGTGLTH	SLARYAFD	VLIQFEGKY	TVKNDLRDRY
YG	FWV	WKAC	VLTCL	TFVSFD	RKALGIH	SLPTPIMA	VMCISDVTR	VDIPFRAPTV
IH	VWY	WLWR	VTRDI	VGHPYF	RTFTAAP	SNDAWYNV	VSDVTRGNG	VEIHDKRMTD
GY	YLW	WMIN	VVCTN	VIDRNE	SFFHGEM	SYSLKEKE	VWGPSAPDA	VIDDVDPHYL
WD	VCY	WQQC	VVDGI	VISSGR	SHQYGGT	TLGYDLIR	VYKKAQAFD	VISINNVIRA
KC	WLW	WSFI	VVGLL	VLCNPG	SQFDSSL	TRGNGLTH	VYVASYNEV	VKNDLRDRFQ
KW	YYW	WWYA	WLQRA	VSEGIH	SRDEGLH	TYPKCSLT	VYVLGKIWM	VNNHVVYNHQ
YN	VIW	WYKF	WSGKE	VVVVDR	SSAVEDL	VGIAVDTG	WCIENGTSP	VNNYVVYNQQ
TY	VYC	WYLF	WYQRS	WIVIHA	TAGYTPF	VGKSLGLC	WMDENIKTK	VQIKGGIPTI
IG	ICY	YCFG	YDMYR	WSGKEF	TDNESLQ	VKRFWKVN	WSGKEFLRE	VYNHQEAGKY
QC	YLC	YCTF	YDVIK	YCAITH	TFVSFDL	VKSVYILG	YCDADGSQF	WARVATGRET
HI	FWF	YFAI	YFIRL	YDLIRD	VGKVMCI	VLCNPGEG	YFLTYPQCS	WTFPIRCNID
DC	WLF	YFKW	YIKKY	YGNDLR	VKILSKQ	WARVATGR	YGKPVQIKG	WYNVIDDVDP
RW	FWI	YINW	YISDI	YHAKRF	VKSCSQG	YCAITHID	YKHACATGS	YASKEQALVK
DW	WTC	YLCF	YITDI	YIDQYA	VKSVYIL	YGTPMDFG	YLCGHLDLS	YCAITHIDYG
AY	LWW	YMYY	YIVEL	YIKICR	WLAIQPV	YIKGLGSL	YMACTHASN	YCDADGSQFD
FT	WFY	YNWA	YKHAC	YPTASA	WMDENIK	YPQCSLTK	YNHQEAGKY	YFLTYPKCSL
NY	WIF	YRWD	YLDFA	YVDSRI	YAKVTGG	YQSCHILQ	YQDNERVAH	YGTPMDFGQV
IM	WDC	YWIT	YNVLR	YYIDLE	YTLGQQL	YRQQVPIN	YYFHGHIVP	YQYDKYNDVN

Table A16. Characteristic n-grams in border regions produced by association rules

N-grams presented in the table belong to the body of discovered association rules with head *ORDER_LEVEL='N'*. Parameters used in mining are confidence>=51%, support>=0.0001 and lift >=1.05 or lift <=0.95. N-grams in table are sorted according lift and confidence, both in descending order.

N-gram length					
5	6	7	8	9	10
VYKYE	CHLKNP	EGNRPTF	EGNRPTFV	GQVVYKYEE	ASMEGNRPTF
WDPLV	FYDSIT	FYDSITN	GQVVYKYE	IYFYDSITN	IRIYFYDSIT
NRPTF	FYDSVT	FYDSVTN	IYFYDSIT	KNYGHPREN	KNYGHPRENF
CHLKN	GHPREN	GHPRENF	LYDALEAP	LYDALEAPA	LYDALEAPAD
YKYEE	HPRENF	GNRPTFV	MEGNRPTF	MEGNRPTFV	NKNYGHPREN
FYDSQ	ICHLKN	ICHLKNP	NYGHPREN	NYGHPRENF	RIYFYDSITN
HPREN	NRPTFV	QVVYKYE	QVVYKYEE	RIYFYDSIT	SMEGNRPTFV
PRENF	VVYKYE	VICHLKN	VICHLKNP	SMEGNRPTF	KIRIYFYDSV
HLKNP	VYKYEE	VVYKYEE	YFYDSITN	IRIYFYDSV	AYNGNDTEGL
WDPLL	YDSVTN	YFYDSIT	YFYDSVTN	AYNGNDTEG	GNDTEGLLKE
FYDSV	IKFNLY	YFYDSVT	YGHPRENF	GNDTEGLLK	NGNDTEGLLK
DPLLN	YDSITN	YGHPREN	RIYFYDSV	NDTEGLLKE	SAYNGNDTEG
MKKII	GNRPTF	IYFYDSV	DTEGLLKE	NGNDTEGLL	YNGNDTEGLL
RPTFV	YFYDSV	DTEGLLK	GNDTEGLL	YNGNDTEGL	VSPTRSAHFH
LDYVG	NDTEGL	GNDTEGL	NDTEGLLK	MWDPLLNEF	MWDPLLNEFP
YDSIT	MWDPLV	NDTEGLL	NGNDTEGL	WDPLLNEFP	WDPLLNEFPE
FYDSI	SRGPAG	NGNDTEG	YNGNDTEG	DPLLNEFPE	KIRIYFYDSI
KFNLY	PLLNEF	GSKSEAL	DPLLNEFP	IRIYFYDSI	DPLLNEFPET
DLDYV	WDPLLN	MWDPLLN	MWDPLLNE	PLLNEFPET	PLLNEFPETV
IKFNL	TEGLLK	DPLLNEF	WDPLLNEF	YDALEAPAD	YDALEAPADT
CGGGR	YFYDSI	PLLNEFP	PLLNEFPE	WGEFQIDGR	WGEFQIDGRS
MKKLL	IKFNIY	WDPLLNE	RIYFYDSI	DALEAPADT	GEFQIDGRSA
DSVTN	DPLLNE	IYFYDSI	YDALEAPA	GEFQIDGRS	DALEAPADTP
PLYSG	DALEAP	TEGLLKE	DALEAPAD	YIDKDGDTL	LEWGEFQIDG
SRGPA	MWDPLL	YDALEAP	GEFQIDGR	EFQIDGRSA	SPTRSAHFHP
KFNIY	SKSEAL	DALEAPA	ALEAPADT	SPTRSAHFH	EFQIDGRSAR
LYIPE	DTEGLL	ALEAPAD	EFQIDGRS	ALEAPADTP	EWGEFQIDGR
YFYDS	ALEAPA	EFQIDGR	FQIDGRSA	CCPHCPRHK	CCCPHCPRHK
YDSVT	EFQIDG	FQIDGRS	WGEFQIDG	FQIDGRSAR	FQIDGRSARG
NDTEG	GELITA	GEFQIDG	RTGELITA	DKDGDTLEW	DKDGDTLEWG
KNPEK	FQIDGR	TGELITA	KDGDTLEW	LEWGEFQID	KSYIDKDGDT
EFQID	AFNYIE	DGDTLEW	PTRSAHFH	EWGEFQIDG	KDGDTLEWGE
QQRLI	TGELIT	DKDGDTL	DGDTLEWG	DGDTLEWGE	PTRSAHFHPN

DTEGL	VSPTRS	KDGDTLE	SRTGELIT	PTRSAHFHP	DGDTLEWGEF
LKNPE	KDGDTL	RTGELIT	CCPHCPRH	KDGDTLEWG	DSRTGELITA
MKKLI	GEFQID	AFNYIES	EWGEFQID	SYIDKDGDT	MDSRTGELIT
CPRHK	GNDTEG	LVSPTRS	LVSPTRSA	SRTGELITA	TRSAHFHPNI
DSITN	RTGELI	VSPTRSA	VSPTRSAH	DSRTGELIT	DLVSPTRSAH
DPLVN	EGNRPT	WGEFQID	DLVSPTRS	CCCPHCPRH	FFDLVSPTRS
MWDPL	DGDTLE	CCPHCPR	TRSAHFHP	LVSPTRSAH	LVSPTRSAHF
AFNYI	QIDGRS	MEGNRPT	CCCPHCPR	TRSAHFHPN	FDLVSPTRSA
ALEAP	LGGAGG	TRSAHFH	LEAPADTP	DLVSPTRSA	PCCCPHCPRH
QIDGR	LEAPAD	SRTGELI	CPHCPRHK	VSPTRSAHF	QIDGRSARGG
FQIDG	SPTRSA	LEAPADT	DSRTGELI	FDLVSPTRS	LKIRIYFYDS
GPLYS	CCPHCP	QIDGRSA	QIDGRSAR	PCCCPHCPR	IDGRSARGGQ
PHCPR	CPHCPR	SPTRSAH	SPTRSAHF	MDSRTGELI	ASKVRRRLNF
HCPRH	DKDGDT	IDKDGDT	YIDKDGDT	QIDGRSARG	GNNSGQPSTV
PLLNE	IYFYDS	PHCPRHK	PCCCPHCP	KIRIYFYDS	SGQPSTVVDN
SPTRS	HCPRHK	CPHCPRH	IRIYFYDS	IDGRSARGG	NNSGQPSTVV
IEAAT	PHCPRH	CCCPHCP	IDGRSARG	GNNSGQPST	NSGQPSTVVD
FDSQT	IDGRSA	RIYFYDS	SGQPSTVV	SGQPSTVVD	PASKVRRRLN
SYIEK	SGQPST	PTRSAHF	NNSGQPST	NNSGQPSTV	TPASKVRRRL
CPHCP	PTRSAH	IDGRSAR	NSGQPSTV	NSGQPSTVV	GQPSTVVDNT
AVSNS	GQPSTV	GQPSTVV	GQPSTVVD	GQPSTVVDN	
GNRPT	TRSAHF	SGQPSTV	ASKVRRRL	ASKVRRRLN	
	LVSPTR	NSGQPST	FDLVSPTR	PASKVRRRL	
	QPSTVV	DLVSPTR	QPSTVVDN		
		QPSTVVD			

Table A17. Characteristic n-grams in disordered regions produced by combination of z-score, fractional difference, mole fractions and association rules

N-grams presented in the table characterize disordered regions by association rules, and have abs(z-score)>2.58 in disordered and abs(z-score)<1.65 in ordered regions, mole fractions>1E-7 and positive fractional difference in disordered regions. Table includes, for each n-gram length, (at most) first 100 n-grams sorted according lift, confidence, and support, all in descending order.

N-gram length							
3	4	5	6	7	8	9	10
QQQ	HHHH	GGGGG	GGGGGG	PPPPPPP	GGGGGGGG	PEPEPEPEP	SSSSSSSSSS
PPR	SNAM	PPPPP	PPPPPP	EEEEEEE	PPPPPPPP	EPEPEPEPE	EEEEEEEEEE
PPQ	GHMA	APAPA	TTTTTT	DDDDDDD	EEEEEEEE	EEEEEEEEE	PSPPPPSPPP
SPS	GSHM	PSPPP	PEPEPE	PEPEPEP	PEPEPEPE	PKPAPKPAP	SPPPPSPPPP
TPP	PSPP	NNNNN	EPEPEP	EPEPEPE	EPEPEPEP	PAPKPAPKP	GGGGGGGGGG
PPK	QPQP	EEEED	GGGGGA	PKPAPKP	KPAPKPAP	KPAPKPAPK	PPSPPPSPPP
DDD	EPEP	PPAPP	PKPAPK	TTTTTTT	APKPAPKP	APKPAPKPA	VISTPASKVR
PPE	SSSS	SSTSS	KPAPKP	PPPSPPP	PAPKPAPK	DDDDDDDDD	RYGGGGGGGG
MPP	PQPQ	KKKKK	AGGGGG	KPAPKPA	QGAKSSSD	QQQQQQQQQ	HPNIQGAKSS
SES	QGPQ	DEEDE	PPSPPP	PAPKPAP	QQQQQQQQ	PPPPSPPPP	GQHISIRTFR
PDP	PQGP	PAPPP	APKPAP	APKPAPK	PPPSPPPP	PSPPPPSPP	
HHH	QPQQ	KKSKK	PSPPPP	QQQQQQQ	GGGGGGGA	PPSPPPPSP	
QKQ	PQPP	EEDDD	APAPAP	GGGGGGA	PSPPPPSP	PAPAPAPAP	
ESE	PPQP	PPAAP	PAPAPA	PPSPPPP	PPSPPPPS	APAPAPAPA	
PEE	PTPP	KKEKK	PPPPSP	PPPPSPP	SPPPPSPP	PSPPPSPPP	
	QPPQ	SKKKK	NNNNNN	MDSRTGE	AGGGGGGG	DSDSDSDSD	
	PQQQ	ESSSS	SRSRSR	PAPAPAP	APAPAPAP	PPSPPPSPP	
	PAPP	AAPPA	QGAKSS	GAKSSSD	SPPPSPPP	ADIVISTPA	
	QPPP	RRRGR	SPPPPS	QGAKSSS	PSPPPSPP	SKRPADIVI	
	MWDP	GGGDD	SDSDSD	RSRSRSR	GPEGPEGP	YGGGGGGGG	
	APAP	HHHHH	PQGPQG	SRSRSRS	GGGGGGSG	TGPQGPKGD	
	KPAP	PPPPQ	DSDSDS	GPQGPQG	EDEDEDED	GDGDGDGDG	
	PGPP	STTST	TGGGGG	PSPPPPS	AGGGGGSG	GGAGGGGGS	
	PPPA	DEDSD	GPQGPK	APAPAPA	GAGGGGGS	STNGIEPPR	
	PKPA	DDDKD	PQGPKG	SPPPPSP	ALRRRLER	RYGGGGGGG	
	PPPR	REEEE	SPPPSP	GPQGPKG	NPAPTSSP	GAGGGGGSG	
	PPGP	EKKKS	PSPPPS	GGGGAGG	LRRRLERG	DGDGDGDGD	
	APKP	GGSRS	PPPPPS	DSDSDSD	GGGGGGGY	GAGAGGGAG	
	QGPK	SSSVD	TTAATT	PSPPPSP	DKGDKGDT	ASAYNGNDT	
	APPP	EAEED	SPPPPP	SDSDSDS	APTSSPTS	NDAAAEALN	
	PAPA	PSPEP	GGGGGY	MSKRPAD	DEDDEDDE		
	PSSP	NTERH	SSSTSS	DKGDKGD	EEQKQLTL		

	SPPS	PQQQP	QGPQGP	DEDEDED	PPPLPPPP		
	PPPT	QQEEE	PPPPPA	TNGIEPP	NSTNGIEP		
	QRQQ	KKTSS	MSKRPA	PTPSPTP	RYGGGGGG		
	QQQE	PKPRP	EEEEDE	GPAGPQG	GGGGGSGR		
	QQQS	RGEET	SKRPAD	GGSGGGG	GGEGGEGG		
	QQEQ	KPTPP	PKGDTG	YGGGGGG	ILEEAQRL		
	QQQH	KRPPP	LPPPPP	PTPPPTP	DTPVSEIP		
	PPPE	APEDP	TPPPTP	TTAATTT	SSSSSSSC		
	MPKR	THMPR	YGGGGG	GSGGGGG	SQLKGSSS		
	EEPE	QQPPQ	EDEEEE	NGIEPPR	PAAPAAPA		
	PSPQ	DSPPS	SSSSGS	GPEGPEG	ELNPAPTS		
	EQEQ	SSEKP	PGGGGG	PPPPLPP	ISTPASKV		
	QQSQ	PSPSP	PPAPPA	PQGPQGP	DGDGDGDG		
	APQP	PQPQP	NGIEPP	DDEDDED	LNPAPTSS		
	PPPG	PKPAP	PAGPQG	QPEESVG	IIISTPAS		
	PSKP	QGPQG	SPSPSP	PPPTPPP	GEGGEGGE		
	PGPS	RPPPP	GQQTAN	QPQPEES	GAGAGGGA		
	PAPQ	PPPSP	EDEEDE	GGQQSAN	RGGQQTAN		
	QEQQ	QPQQQ	GGGGGN	GQQSAND	GAKSSSDV		
	PHPP	QQPQQ	PTPPPT	RGGQQSA	APAAPAAP		
	QEPQ	TERHT	EDEDEE	AGGGGGS	NIQGAKSS		
	QKPP	PQGPQ	QGIQGP	GPQGPAG	AAPAAPAA		
	QGPP	SPPPP	QPEESV	EDDEDDE	KSSSDVKS		
	MSEQ	QPQPE	APTSSP	IVISTPA	MSKRPADI		
	EEKP	MSKRP	SSSSDS	PSPPPPP	DAAAEALN		
	RKRP	PPPPR	GPEGPE	PVPKPAP	NDAAAEAL		
	MSKR	PQGPK	PEGPEG	PASMEGN	GRSARGGQ		
	PQSP	NPPPP	SSSSSE	SPASMEG	QSANDAYA		
	PAPR	QQQRQ	NGGGGG	EGPEGPE			
	PPPD	PPQPQ	AGPQGP	ENTERHT			
	RHHH	EEEEE	DEEEED	PPLPPPP			
	KKKK	QPQPQ	GGGGGL	EEEEEED			
	MAPP	PPPEP	RRSPSP	GTSKVSR			
	APQQ	QQEQQ	PPPPPT	GPEGPQG			
	MDSR	RPPPR	PSPSPS	GPVGPQG			
	EEPP	PQQPP	RSPSPR	PAAPAPA			
	PSEP	SSSSS	QPQPEE	GPAGPAG			
	PSRP	RQQRE	TTTPTT	MRSSSPS			
	PDPP	PSPPS	GPAGPA	AGTSKVS			
	PEPA	RRSPP	GRRRSS	GGGGNGG			
	SDSD	PRGPP	NPASAE	PAPTSSP			
	AGPP	PPEPP	RGGQQS	PTSSPTS			
	DEEE	QQQAP	RGRGRG	VGGGGGG			
	PKPT	PEPPQ	GGGGGE	APAPAAP			
	RQQR	PSPQP	PQPEES	GPTGPQG			
	PRSP	SPPQA	AGPAGP	DDEDEDD			
	SPSS	PPPPE	GGQQSA	SSSSGSS			
	EEDE	PKEEP	GQQSAN	SSSSSSG			
	PTGP	PPSPP	QQSAND	PAPKPKP			
	PPMP	RPSSP	GEGEGE	SSSSSDS			
	SRSP	QPPQQ	PQGPAG	MPKRDAP			
	QQAQ	PAQQP	RSSSPS	VQPQPEE			
	EDDE	PQGEQ	TSSPTS	DEEEEEE			
	PPDP	QEESS	GPPGPE	PSPTPSP			
	PKKK	QQPQP	MRSSSP	AAPAPAA			
	SKRP	PQPEE	RSSSSS	LRRRLER			
	PESP	PPQPP	ASMEGN	GPPGPPG			
	SEPP	QGPKG	ENTERH	APTSSPT			
	GQQP	PSPGP	EDDEEE	APPPPPP			
	MPCE	RSPSP	PQPQPQ	PPPPPSP			
	KSSS	PEPAP	SSGSSS	RRRLERG			
	PRPA	QQPPP	TSKVSR	NSTNGIE			
	KKPK	PPPQP	KRDAPW	PQGIQGP			
	PPPL	SPSPS	NTERHT	QKQLTLF			
	EQQE	EPEEP	PASMEG	PCESSSQ			
	PRKP	RRRSS	PKPKPK	PPPPPPA			
	SQQQ	GPPGP	PSPTPS	STPPSIK			
	DDED	PSSPP	SPASME	ANDAAAE			

Table A18. Characteristic n-grams in ordered regions produced by combination of z-score, fractional difference, mole fractions and association rules

N-grams presented in the table characterize ordered regions by association rules, and have abs(z-score)>2.58 in ordered and abs(z-score)<1.65 in disordered regions, mole fractions>1E-7 and positive fractional difference in ordered regions. Table includes, for each n-gram length, (at most) first 100 n-grams sorted according lift, confidence, and support, all in descending order.

N-gram length							
3	4	5	6	7	8	9	10
WIC	IFII	YNVID	IKGGIP	QIKGGIP	NPVYATLK	THASNPVYA	FKEFMGAQRD
YCW	LLLW	IKGGI	NVIDDV	ACTHASN	VYATLKIR	VYATLKIRI	HYLKHFKEFM
WYW	VLAC	VGKRF	YNVIDD	ASNPVYA	KIRIYFYD	TENALLLYM	VIDDVDPHYL
CLW	IFLC	ATLKI	YMACTH	SNPVYAT	LKIRIYFY	IKGGIPTIF	MWARSLGPHN
WWF	CVLV	ACTHA	LYMACT	NPVYATL	TLKIRIYF	GPHNYLCGH	KVTGGQYASN
WWY	FVIF	GKRFC	QIKGGI	PVYATLK	YATLKIRI	PVQIKGGIP	QSNCKYGKPV
FCW	FIVF	LYMAC	SNPVYA	YNVIDDV	ENALLLYM	SLGPHNYLC	TVTGGQYASK
YFW	TMWA	MACTH	MACTHA	NHTENAL	TENALLLY	ENHTENALL	ERIQRLGRVG
CWV	VVIF	YMACT	NPVYAT	HRVGKRF	HTENALLL	WYNVIDDVD	VKSVYILGKI
CVW	YAIY	ALLLY	CTHASN	THRVGKR	VGKRFCVK	CGHLDLSPK	NHVVYNHQEA
WWI	LICL	KYENH	PVYATL	RVGKRFC	NALLLYMA	VATNIIENG	FDRINVRRLF
WIY	VAYY	GPHNY	NHTENA	WMDENIK	GIPTIFLC	VKSVYILGK	LSTAKHSVDI
WFW	WYVD	RFFDL	HTENAL	LGPHNYL	KGGIPTIF	WLVRDRRPY	KLKNHTNSVM
CYC	YLYF	PHNYL	THRVGK	YATLKIR	LKHFKEFM	VQIKGGIPT	TKYGKPIQIK
WMC	YFTF	IYFYD	HRVGKR	VYATLKI	VTGGQYAS	KVTGGQYAS	NLNSNLDRIF
VWI	IAWL	KIRIY	SDVTRG	IRIYFYD	CGHLDLSP	IWMDENIKT	ISDVTRGNGI
FWC	YIAI	TMWAR	GPHNYL	KIRIYFY	RSLGPHNY	GVISINNVI	FRCMLAIKYL
VWC	YYVY	RIYFY	RVGKRF	LKIRIYF	WYNVIDDV	VTRGNGITH	QIKGGIPSIV
ICW	LLWF	LLYMA	LGPHNY	TLKIRIY	NVIDDVDP	NTKYGKPVQ	SETIHSRSYT
CFI	LHYY	PLYFK	VGKRFC	ATLKIRI	GKPVQIKG	TVTGGQYAS	ALEAIRFYVS
WYF	CIAL	LYFKI	WMDENI	GKRFCVK	YGKPVQIK	HVVYNHQEA	IRDLISVIRA
YIC	CLAI	RFCVK	MDENIK	KRFCVKS	PVQIKGGI	SCMKIDHCV	NLPTSAGKSL
IWW	LTWL	KIWMD	YATLKI	VTGGQYA	YNVIDDVD	YGTPMDFGQ	RVNNYVVYNQ
FIW	DIIC	GKIWM	ATLKIR	YENHTEN	YLKHFKEF	IQIKGGIPT	LKRLRFKGTV
IYW	YIPI	LLLYM	RIYFYD	ENALLLY	DVDPHYLK	ERIQRLGRV	FRCMLAVKYL
ICC	FIYF	IWMDE	IRIYFY	HTENALL	HYLKHFKE	SVYVLGKIW	ERIVSILEWD
WHC	YLFV	NNVIR	KIRIYF	NALLLYM	QEAGKYEN	KLSTAKHSV	TYSPDTLGYD
VWY	YYEI	NYLCG	LKIRIY	TENALLL	VVYNHQEA	DRINVRRLF	KQLSFFWRPE
YLW	VWVV	HNYLC	NPLYFK	GKIWMDE	MWARSLGP	KGKLKLSTA	LGKTTVVAIF
YLC	IGYF	IFLCN	RFCVKS	VGKRFCV	GKYENHTE	FKGKLKLST	NLSRQLGKTT
LWW	CFAL	LCGHL	GKRFCV	LLYMACT	LVRDRRPY	PETVHGFRC	FLVRDRRPVD
WIF	IWEI	TIFLC	KRFCVK	ALLLYMA	NHTNSVMF	QELRVLAAL	TSLYPSIIRQ
WFV	YLCD	YLCGH	KYENHT	KIWMDEN	VHGFRCML	IPFRAPTVK	KSCSQGGIRG
WVY	CLGI	AQRDW	VTGGQY	LLLYMAC	HGFRCMLA	NHGFTHRGT	MDFGQVFNMF
WVW	VVYC	KNYFL	GKIWMD	IWMDENI	CISDVTRG	YFLTYPQCS	PHLHVLIQFE
WCL	LKCF	YLKHF	NALLLY	HNYLCGH	CVSDVTRG	VQIKGGIPS	DFGQVFNMFD
WWH	VMFF	HFKEF	ALLLYM	NYLCGHL	FGQVFNMF	KNDLRDRFQ	KQAIELLPDF
LWY	CFLT	NYFLT	ENALLL	NPLYFKI	IKTKNHTN	NVIRAVRFA	EGDSRTGKTM
CCY	CLYL	FMGAQ	YENHTE	YLCGHLD	ISINNVIR	APTVKILSK	NYIESHRDEY
QWW	GLCF	FLCNP	ENHTEN	LCGHLDL	DENIKTKN	YVVYNHQEA	AVGSGKSTGL
FLW	CVVC	FFDLV	TENALL	GGIPTIF	MDENIKTK	FASLYPSII	VGSGKSTGLP
CQW	TIWN	CRELH	KIWMDE	GIPTIFL	TRGNGITH	TIHSRSYTH	
LWF	VCVC	YFLTY	LLYMAC	IPTIFLC	KSVYILGK	APKDFVLQF	
GCW	IFYV	NHNLR	LLLYMA	LKHFKEF	GNGITHRV	CMLAVKYLQ	
YWW	CYLN	FGQVF	IWMDEN	PTIFLCN	IKLKNHTN	HFIVATNII	
WFM	DCII	LGKIW	HNYLCG	ARSLGPH	ATNIIENG	VLCNPGEGA	
CCV	IYNM	LLLLV	NYLCGH	KEFMGAQ	NIIENGVT	VKNDLRDRF	
VWW	YLFQ	WYNVI	LKHFKE	KGGIPTI	VATNIIEN	GEMTVAGKK	
IWL	FFIF	VVYNH	LCGHLD	KHFKEFM	VLQFHNLN	IRAVRFATD	
LCY	FWLV	FNHNL	PLYFKI	WARSLGP	LGVISINN	LPTSAGKSL	
WWV	WAVL	VISIN	YLCGHL	HFKEFMG	PKVYSNDA	EGRGQDYHA	
IWI	CYNL	AWYNV	CGHLDL	IKGGIPT	TNIIENGV	VNNYVVYNQ	
CCF	VIYF	IQFEG	IFLCNP	CGHLDLS	VISINNVI	YFLTYPKCS	
YCY	WLYN	LILLL	GAQRDW	FKEFMGA	GKVMCISD	ELRVLAALS	
YYF	ILWL	QVFNM	KHFKEF	TGGQYAS	NIKLKNHT	ERIVSILEW	
YYC	YLPY	VYNHQ	GGIPTI	IDDVDPH	VYNHQEAG	RKALGIHKC	
WYV	AWGY	DPHYL	EFMGAQ	PHNYLCG	GQYASKEQ	YSIELAQDL	
YCC	CFTV	FLRVF	GIPTIF	GHLDLSP	QYASKEQA	GDFLTSLIN	
CFL	CYGL	HVLIQ	KGGIPT	GPHNYLC	KLKNHTNS	LNFQVWTTS	
IFC	CLYA	LHVLI	PTIFLC	LGKIWMD	VTRGNGIT	NGLMVWCIE	
VCF	GFFY	VLIQF	TIFLCN	RSLGPHN	IIENGVTL	INSLYGALG	
FCL	VIMV	ICREL	GGQYAS	WYNVIDD	QRLGRVGR	VAFDMRGQQ	
HLW	WALF	AGKYE	HFKEFM	KYGKPVQ	INNVIRAV	FLGLPFNIA	
QWC	VYIW	VVVVV	KEFMGA	KPVQIKG	YVLGKIWM	ALGPHNYLS	
CLI	FALC	RCMLA	RSLGPH	NVIDDVD	MCISDVTR	WLAIQPVIS	
IVW	FLIC	GFRCM	FKEFMG	NYFLTYP	YDLIRDLI	TSAGKSLIQ	
VWH	FRYY	HTNSV	GHLDLS	PVQIKGG	ENIKLKNH	KICRELHEN	
WCT	LVCF	FRCML	DDVDPH	VIDDVDP	RRPYGTPM	SKEQALVKK	

VWV	VYIY	LIIGL	YNHQEA	YGKPVQI	DENIKLKN	KICRELHED	
FYY	IFQY	CGCSY	VQIKGG	RFCVKSV	SFFSLKDP	RALDNLLDY	
WCM	LFYC	ILSLI	KPVQIK	FCVKSVY	YLSGHLDF	LYQSCHILQ	
LIC	VYVC	LLVVL	GKPVQI	AWYNVID	FFSLKDPI	AIELLPDFL	
WCQ	CILV	GCGKT	PHNYLC	YLKHFKE	FSLKDPIP	SKRYLYQDN	
YIF	LWFL	DAWYN	SLGPHN	KNYFLTY	IENGVTLD	NIFLAMLVN	
VWL	CVKI	IIILL	HLDLSP	LIQFEGK	KVCVDDFN	LSGIKGQIG	
CLV	IDCV	YVLGK	LGKIWM	FGQVFNM	MDENIKLK	QYDKYNDVN	
IYY	RVWL	GITHR	WYNVID	HLDLSPK	NIDLHYFS	CGMYASALT	
NCY	VTCK	NDAWY	KNYFLT	DDVDPHY	SLKDPIPW	VNNYVVYNH	
FVF	WIVA	VHGFR	YGKPVQ	DVDPHYL	KSVYVLGK	VLQFHNLNA	
IWT	WLVV	FLLLL	VIDDVD	EFMGAQR	LRKALGIH	TSLYPSIIR	
HVW	LFRM	HGFRC	NLDRIF	VDPHYLK	PIPWKLYY	IDLHYFSSS	
TWC	CHIL	ILLVL	NYFLTY	DPHYLKH	CVIEYRQQ	LGKTTVVAI	
VCL	IGWI	QIRFN	PVQIKG	MGAQRDW	DRYQVLRK	TKNHTNTVM	
IFI	RFCI	VLLLV	YFLTYP	VVYNHQE	NLRKALGI	SRQLGKTTV	
ICL	WNLV	VVLAL	FNHNLR	TMWARSL	NRFFDLVS	RQLGKTTVV	
FYF	CVYV	ALGIH	AWYNVI	PHYLKHF	PIQIKGGI	DCSSAVEDL	
YLF	IGDW	GDLIY	CVKSVY	HYLKHFK	IEYRQQVP	HNLNANLDR	
LYY	YVFI	ILILL	YLKHFK	LDLSPKV	IQRLGRVG	YKKAQAFDE	
CIH	HHII	GNIIG	FGQVFN	DLSPKVY	KPIQIKGG	ICFAGDDMC	
WRW	CNIT	NYVVY	GQVFNM	QEAGKYE	RIQRLGRV	GVSEGIHPI	
FWM	CNLC	ALVIL	LIQFEG	VLIQFEG	SCMKIDHC	EIHAELNAI	
ICA	LAFW	GKVMC	IQFEGK	AGKYENH	YGKPIQIK	ASLYPSIIQ	
VVW	LLVW	IILLL	VYNHQE	EAGKYEN	AKYENHTE	VVAIFLAHF	
YNW	CSIY	LLLLG	DVDPHY	HVLIQFE	FAKFKGKL	TDIAGYAGC	
FLY	ICII	AVLLV	VDPHYL	LHVLIQF	IRCNIDLH	VHGMDADAE	
FIY	LFMI	ISLLL	FMGAQR	KNHTNSV	STAKHSVD	NLGVISINN	
CNW	VLWY	LGVVA	NNVIRA	VRDRRPY	YGTPMDFG	VDLPCGCSY	
LCL	YDIC	AVIRF	PHYLKH	ICRELHE	LKLKHWKE	NGVGPLCKG	
ACF	IIYM	ILGAV	DPHYLK	DVTRGNG	DGSQFDSS	SQFDSSLTP	
LCV	CAFI	IINIL	MGAQRD	THRGTHH	WTFPIRCN	FLVRDRRPV	

Table A19. Characteristic n-grams in bordered regions produced by combination of fractional difference, mole fractions and association rules

N-grams presented in the table characterize bordered regions by association rules, and have mole fractions>1E-7 and positive fractional difference in bordered regions. Table includes n-grams sorted according lift, confidence, and support, all in descending order.

N-gram length					
5	6	7	8	9	10
VYKYE	FYDSVT	YFYDSVT	YFYDSVTN	MEGNRPTFV	ASMEGNRPTF
WDPLV	NRPTFV	FYDSVTN	EGNRPTFV	SMEGNRPTF	SMEGNRPTFV
NRPTF	VVYKYE	GNRPTFV	MEGNRPTF	LYDALEAPA	LYDALEAPAD
CHLKN	VYKYEE	EGNRPTF	QVVYKYEE	NYGHPRENF	IRIYFYDSIT
YKYEE	GHPREN	VVYKYEE	YFYDSITN	IYFYDSITN	RIYFYDSITN
FYDSQ	FYDSIT	QVVYKYE	LYDALEAP	RIYFYDSIT	KNYGHPRENF
HPREN	CHLKNP	YGHPREN	YGHPRENF	GQVVYKYEE	NKNYGHPREN
PRENF	HPRENF	YFYDSIT	GQVVYKYE	KNYGHPREN	KIRIYFYDSV
HLKNP	ICHLKN	FYDSITN	NYGHPREN	IRIYFYDSV	AYNGNDTEGL
WDPLL	YDSVTN	GHPRENF	IYFYDSIT	AYNGNDTEG	YNGNDTEGLL
FYDSV	IKFNLY	ICHLKNP	VICHLKNP	NDTEGLLKE	GNDTEGLLKE
DPLLN	YDSITN	VICHLKN	RIYFYDSV	NGNDTEGLL	NGNDTEGLLK
MKKII	GNRPTF	IYFYDSV	DTEGLLKE	YNGNDTEGL	SAYNGNDTEG
RPTFV	YFYDSV	GNDTEGL	YNGNDTEG	GNDTEGLLK	VSPTRSAHFH
LDYVG	NDTEGL	NDTEGLL	NGNDTEGL	MWDPLLNEF	MWDPLLNEFP
YDSIT	MWDPLV	DTEGLLK	NDTEGLLK	WDPLLNEFP	WDPLLNEFPE
FYDSI	SRGPAG	NGNDTEG	GNDTEGLL	DPLLNEFPE	KIRIYFYDSI
KFNLY	PLLNEF	GSKSEAL	DPLLNEFP	IRIYFYDSI	DPLLNEFPET
DLDYV	WDPLLN	MWDPLLN	MWDPLLNE	PLLNEFPET	PLLNEFPETV
IKFNL	TEGLLK	DPLLNEF	WDPLLNEF	YDALEAPAD	YDALEAPADT
CGGGR	YFYDSI	PLLNEFP	PLLNEFPE	WGEFQIDGR	WGEFQIDGRS
MKKLL	IKFNIY	WDPLLNE	RIYFYDSI	DALEAPADT	GEFQIDGRSA
DSVTN	DPLLNE	IYFYDSI	YDALEAPA	GEFQIDGRS	DALEAPADTP
PLYSG	DALEAP	TEGLLKE	DALEAPAD	YIDKDGDTL	LEWGEFQIDG
SRGPA	MWDPLL	YDALEAP	GEFQIDGR	EFQIDGRSA	SPTRSAHFHP
KFNIY	SKSEAL	DALEAPA	ALEAPADT	SPTRSAHFH	EFQIDGRSAR
LYIPE	DTEGLL	ALEAPAD	EFQIDGRS	ALEAPADTP	EWGEFQIDGR
YFYDS	ALEAPA	EFQIDGR	FQIDGRSA	CCPHCPRHK	CCCPHCPRHK
YDSVT	EFQIDG	FQIDGRS	WGEFQIDG	FQIDGRSAR	FQIDGRSARG
NDTEG	GELITA	GEFQIDG	RTGELITA	DKDGDTLEW	DKDGDTLEWG
KNPEK	FQIDGR	TGELITA	KDGDTLEW	LEWGEFQID	KSYIDKDGDT
EFQID	AFNYIE	DGDTLEW	PTRSAHFH	EWGEFQIDG	KDGDTLEWGE

QQRLI	TGELIT	DKDGDTL	DGDTLEWG	DGDTLEWGE	PTRSAHFHPN
DTEGL	VSPTRS	KDGDTLE	SRTGELIT	PTRSAHFHP	DGDTLEWGEF
LKNPE	KDGDTL	RTGELIT	CCPHCPRH	KDGDTLEWG	DSRTGELITA
MKKLI	GEFQID	AFNYIES	EWGEFQID	SYIDKDGDT	MDSRTGELIT
CPRHK	GNDTEG	LVSPTRS	LVSPTRSA	SRTGELITA	TRSAHFHPNI
DSITN	RTGELI	VSPTRSA	VSPTRSAH	DSRTGELIT	DLVSPTRSAH
DPLVN	EGNRPT	WGEFQID	DLVSPTRS	CCCPHCPRH	FFDLVSPTRS
MWDPL	DGDTLE	CCPHCPR	TRSAHFHP	LVSPTRSAH	LVSPTRSAHF
AFNYI	QIDGRS	MEGNRPT	CCCPHCPR	TRSAHFHPN	FDLVSPTRSA
ALEAP	LGGAGG	TRSAHFH	LEAPADTP	DLVSPTRSA	PCCCPHCPRH
QIDGR	LEAPAD	SRTGELI	CPHCPRHK	VSPTRSAHF	QIDGRSARGG
FQIDG	SPTRSA	LEAPADT	DSRTGELI	FDLVSPTRS	LKIRIYFYDS
GPLYS	CCPHCP	QIDGRSA	QIDGRSAR	PCCCPHCPR	IDGRSARGGQ
PHCPR	CPHCPR	SPTRSAH	SPTRSAHF	MDSRTGELI	GNNSGQPSTV
HCPRH	DKDGDT	IDKDGDT	YIDKDGDT	QIDGRSARG	ASKVRRRLNF
PLLNE	IYFYDS	PHCPRHK	PCCCPHCP	KIRIYFYDS	SGQPSTVVDN
SPTRS	HCPRHK	CPHCPRH	IRIYFYDS	IDGRSARGG	NNSGQPSTVV
IEAAT	PHCPRH	CCCPHCP	IDGRSARG	GNNSGQPST	NSGQPSTVVD
FDSQT	IDGRSA	RIYFYDS	SGQPSTVV	SGQPSTVVD	PASKVRRRLN
SYIEK	SGQPST	PTRSAHF	NSGQPSTV	NSGQPSTVV	TPASKVRRRL
CPHCP	PTRSAH	IDGRSAR	NNSGQPST	NNSGQPSTV	GQPSTVVDNT
AVSNS	GQPSTV	GQPSTVV	GQPSTVVD	GQPSTVVDN	
GNRPT	TRSAHF	SGQPSTV	ASKVRRRL	ASKVRRRLN	
	LVSPTR	NSGQPST	FDLVSPTR	PASKVRRRL	
	QPSTVV	DLVSPTR	QPSTVVDN		
		QPSTVVD			

Table A20. Left components of characteristic inverse non-complementary repeats (material downloaded from NCBI) related to disordered regions

Table includes, for each repeat length, (at most) first 100 n-grams sorted according confidence, lift, and support, all in descending order.

Repeat length							
3	4	5	6	7	8	9	10
PPP	QQQE	GGGGG	GGGGGG	PPPSPPP	PPPSPPPP	PPPPSPPPP	SSSSSSSSSS
PPS	HHHH	PPPPP	PPPPPP	GGGGGGG	GGGGGGGG	SSSSSSSSS	PPPPPPPPPP
QQQ	PEPE	PSPPP	PSPPPP	PPPPPPP	SSSSSSSS	PPPPPPPPP	GGGGGGGGGG
SPP	EQEQ	PPPSP	TTTTTT	PKPAPKP	PPPPSPPP	GGGGGGGGG	EEEEEEEEEE
PGP	GEGP	EEGEG	PPPPSP	DDDDDDD	PPPPPPPP	PPPPLPPPP	DDDDDDDDDD
PSP	KPAP	PPPPS	PPSPPP	TTTTTTT	EEEEEEEE	QQQQQQQQQ	TTTTTTTTTT
PAP	QPQQ	PKPAP	SPPPPS	EEEEEEE	DDDDDDDD	EPEPEPEPE	KPAPKPAPKP
PKP	DGKP	SPPPP	PPPSPP	PPSPPPP	TTTTTTTT	DDDDDDDDD	QQQQQQQQQQ
PQP	QPPP	APAPA	PPPPLP	PAPAPAP	QQQQQQQQ	PSPPPSPPP	GAGGGAGAGG
GPR	APRP	SSTSS	TTAATT	PSPPPPS	PPPPLPPP	EEEEEEEEE	PPPSPPPPSP
QPQ	QQPQ	NNNNN	NNNNNN	PTPSPTP	SPPPSPPP	TTTTTTTTT	TTAATTAATT
PEP	QQAQ	PPAPP	PLPPPP	EPEPEPE	ATTAATTA	PPPPTPPPP	SPPPSPPPSP
RPP	EQEK	KKKKK	EDEEDE	QQQQQQQ	PSPPPPSP	PEPEPEPEP	PPSPPPPSPP
RPG	PPQP	PSPGP	AGGGGG	PPPPSPP	APKPAPKP	RSRSRSRSR	NNNNNNNNNN
GPP	PQPQ	PAPKP	GGGGGA	RSRSRSR	NNNNNNNN	SRSRSRSRS	PPPSPPPPLP
PTP	PQQQ	PSPSP	PRPPRP	PSPGPSP	CSSSSSSS	PPPSPPPSP	PEPEPEPEPE
PRP	QQVP	KKSKK	PKPAPK	SPPPPSP	PPPTPPPP	PPPPAPPPP	PPVVPPVVPP
SSS	GPGP	TETTN	DEEEED	PEPEPEP	EEAEEAEE	GAGGAGAGG	PSPPPPSPPP
EEE	QEQE	PEPGP	PEPEPE	SRSRSRS	EEDEEDEE	DSDSDSDSD	PKPAPKPAPK
PPA	PSKP	KKEKK	QAQQAQ	PTPPPTP	DEEYYEED	NNNNNNNNN	DGGDGGDGGD
PPG	ISPQ	HHHHH	KPAPKP	NNNNNNN	TTAATTTA	SPPPSPPPP	GGAGGAGAGG
SPS	RPPP	PPEPP	KKKKKK	GEDEGED	TTAATTAA	GAGGAGGAG	GGGGGGGGGA
PPR	SPSR	PSKSP	HHHHHH	DEDEDED	PAPAAPAP	PAPAPAPAP	PSPPPSPPPP
EQE	KPEE	PTPSP	SPSPPP	PAPKPAP	SPSPPPPS	GGAGAGGAG	EEAEEEEAEE
EDD	PPSS	PSPTP	EDEGED	TTAATTA	PEPEPEPE	GEDEGEDEG	PAPTTPAPTT
EPE	AKRR	VPEPA	SGSGSG	APAPAPA	GGSGGSGG	APKPAPKPA	EPEPEPEPEP
DEG	SDSE	TTTTP	EEEAEE	DSDSDSD	PPPLPPPP	DEGEDEGED	GGAGAGGGAG
DDD	APSP	PGPPG	PPSSPP	GPPGPPG	GPSPGPSP	PPSPPPPSP	SPPPPSPPPP
APP	AQTQ	PQQQP	EEAEEE	PSPPPSP	PPSPPPPS	AEKAKAKEA	EDEDEEDEDE
DEE	GPQG	PPPLP	PAPKPA	PPPTPPP	SPPPPSPS	KPAPKPAPK	GPPGPPGPPG
QEQ	KPPP	GPPGP	CSSSSS	SDSDSDS	GGGGGGAG	CSSSSSSSS	PLPPPPSPPP
SSP	TTET	PPPAP	PTPSPT	DEGEDEG	PPPPAPPP	DEDDEDDED	PPPTPPPTPP
RRR	DGED	PTPKP	GGDGGD	KPAPKPA	ATTTAATT	PTPSPTPTP	APAPAAPAPA
EED	PAQQ	LPPPP	PTTTTT	DDEDEDD	GAGAGGAG	PTPTPSPTP	GGRGGGGRGG
DDE	PQPP	PKPTP	PPPSPS	GGAGGAG	APTGGTPA	PPPPNPPPP	PAPKPAPKPA

PSS	TPSP	PTPPP	EPEPEP	TTAATTT	DDEDDEDD	APAPAPAPA	PPPPSPPPSP
KPK	VEED	PKRKP	PPAAPP	PTPTPTP	HHHHHHHH	DEDEDEDED	PPPSPPPPPP
PPT	RSPS	PPDPP	PSPPSP	GGGGGAG	PAPTTPAP	GDGGDGGDG	QAQQAQAQQA
APA	TDGK	QPAPQ	EEQEQE	AGGGGGA	PPTPPTPP	PKPAPKPAP	APKPAPKPAP
TPP	VPQQ	REQER	RSRSRS	SPSPSPS	SALSSLAS	GGGGGGGAG	EEEAEEAEEE
RPR	GNMN	KRPRK	AAPPAA	TTTPTTT	TSALSSLA	PEEVVVEEP	PPPPPSPPPP
GEG	QQPP	SHTHS	GEEGEG	AGGGGGG	DGEDEGED	SSSSGSSSS	PPPPSPPPPS
RRS	RMAE	PPPTP	SSPPSS	GRGRGRG	SSASSASS	PPPSPPPPS	QAQAQQAQAQ
EDE	RPGE	CSSSS	PAPAPA	PPPAPPP	SSSSSSSC	TTTTPTTTT	SCSSSSSSSS
SRS	SPTS	QAQQA	TPTTTT	CSSSSSS	AGGGGGGA	PSPTPSPTP	SSSASSASSS
RSR	AEGP	RSPSK	PPPTPP	GEGEGEG	EPEPEPEP	EDEDEDEDE	GGDGGDGGDG
DED	ERKR	PGPEP	SSGGSS	DDDGDDD	GDGGDGGD	EEAEEEAEE	GGGGYYGGGG
PDP	NGPQ	QGNGQ	AASAAS	PAPEPAP	RKSKKSKR	GEGEGEGEG	NPPPPSPPPP
RKR	NNGP	KEAER	QPQQPQ	PSPTPSP	SDSDDSDS	PAPKPAPKP	PAAPAAPAAP
RQR	PGMG	QQTQQ	SRSRSR	APASAPA	PPAPPAPP	QAQQAQQAQ	PSPTPPPTPS
KKK	PSPT	PPSPS	AEKKEA	APKPAPK	PPTPSPTP	AGSTATSGA	RSRSRSRSRS
QRQ	QAID	EAEEE	PPRRPP	ATTAATT	GRGRGRGR	EKQASAQKE	SRSRSRSRSR
PMP	SHEA	EEEDE	PSSSSP	DGGDGGD	PAPKPAPK	PGPPGPPGP	AAPAAAAPAA
EGE	SQGG	PVPKP	RGGGGG	EDDEDDE	PKPAPKPA	PPAPAPAPP	AQQAQAQQAQ
GGG	TGPD	QEREQ	SKKKKS	TETKTET	PPGAAGPP	AGGGGGGGA	DEDDEDDEDD
KRK	DEAP	GDGGD	SKRRKS	HHHHHHH	PPPAPPPP	AKKAPAKKA	DKEDKKDEKD
SES	GDCG	PQSQP	TTTTPT	KKKKKKK	GGRGGRGG	EEEAEEAEE	EDDEDDEDDE
PHP	KGGD	PTPIP	AAAAQG	PAPPPAP	KKKKKKKK	EETSESTEE	EEDEEDEEEE
QAQ	RAQA	EEEED	GGGGGS	LPPAPPL	PPPPPSPP	PTPKPTPKP	EPKPEEPKPE
QKQ	RSPP	QQHQQ	DAKKAD	APATAPA	SPSPGPSP	PTPTPTPTP	GAGGAGGAGG
SDS	QQQR	AEKAK	PQQQQP	KEEAEEK	AGGGGGGG	APAAPAAPA	GGAGAGGAGG
EEK	SPKP	AQQAQ	PTDDTP	PVPKPAP	KKSKKSKK	GGGGSGGGG	HHHHHHHHHH
PVP	APPR	PEEPK	SDDDDS	EEDEDEE	PPSPPPSP	GRGRGRGRG	KKKAEEAKKK
EEA	AQPQ	DDDED	SSRRSS	EEDEEDE	RRRSSRRR	HHHHHHHHH	PTPTPPTPTP
EAE	EAER	DEDDE	AKAKEA	SSSTSSS	RSRSRSRS	PPPPVPPPP	PVRRRRRRVP
GPG	EQTP	EDEEE	EEPKPE	EDEEEDE	APAKKAPA	RRSPSPSRR	RRSPSPSRRS
ESE	KKQP	EPQPE	EESSEE	NKKSKKN	DGGDGGDG	SDSDSDSDS	SSSSAASSSS
ERE	QRQQ	GCQCG	GSGGGG	PKPTPKP	DSSSSSSD	SESESESES	TPSPTPSPTP
DSD	KDEK	PDPGP	KPAPAP	PSPSPSP	EDEDEDEE	ERERERERE	AGGGGGGGGA
SQS	KEDK	QQQVP	PPPPAP	RGRGRGR	EDEDEEDE	GDGDGDGDG	EEDEEEEDEE
PNP	PGSP	RRRSS	PPQQPP	SSSASSS	EDEGEDEG	GGGGAGGGG	EEEEDDEEEE
QSQ	PPDP	TPAAP	STSSTS	APPAPPA	EEDEDEDE	PKPAPKPKP	EEEEIIEEEE
PGA	PTSP	APPAP	APKPAP	APPPPPA	GEDEGEDE	PPPSPPPPP	ELDALLADLE
APS	QEAK	APVPV	EEDEDE	EEEDEEE	GGGGGGGA	PPPTPSPTP	GEEEEEEEEG
QTQ	QSSS	ELTGP	PTTPAP	ERERERE	GRSSSSRG	SPPPSPPPS	GRGNGNGRGN
AAP	TSKS	GGKEA	SGGGGG	PPSSPPS	QAQAQQAQ	EEEEDEEEE	GSSGSSGSSG
RER	DSES	PAAPP	ANPPNA	SSSPSSS	QRKTTKRQ	ERREKERRE	KLKKYYKKLK
EKE	EKAY	PAPTP	DGGDGG	PDPLPDP	SSGSSGSS	ESESESESE	PPPAPPAPPP
	NEAK	PPPQP	EPKPEE	PEPTPEP	TTTAATTT	KGKGKGKGK	PPPPPPSPPP
	PPVP	PTGGT	ESDDSE	RLEEELR	DDEEEEDD	PEPSPEPSP	PPPPSPPPPP
	QQPA	QAARE	KPKKPK	TTAPATT	DDSDDSDD	PPPSPSPPP	PPPSPPSPPP
	QTQA	QEEQE	SDEEDS	EEEAEEE	GGEGGEGG	QAAQAQAAQ	PPSSPPSSPP
	SFDD	SAGGG	SGTTGS	GGNSNGG	GGGAGAGG	SGSGSGSGS	PTTTTTTTTP
	SPTR	TTTPT	SRRRRS	GRGGGRG	GGTGGTGG	SPTPTPSPT	RRRRRRRRRR
	SQRS	AASSG	TSTTST	GSGSGSG	KKDKKDKK	CSSSSSSSC	SESESESESE
	TEPE	DPTTS	TTEETT	KKAPAKK	SCSSSSSS	DDDDGDDDD	SGGNGGNGGS
	EDSD	EDDED	DEEDEE	PEPAPEP	SDSSSSDS	EKEKEKEKE	AAAAPPAAAA
	EESE	EETSE	DEGGED	SGSGSGS	SRSRSRSR	KEKEKEKEK	AAAASAAAAK
	GQQA	GGGDG	DSSSSD	AAPAPAA	TTTPTTTT	PKPKPEPEP	ARAARAADAA
	KPKR	MEREM	EPEEPE	AKPQPKA	TTTTPTTT	PSPPPLPPP	ASSASSASSA
	NQNA	PAPAA	GGGGRG	APAYAPA	EDEEDEDE	PTPKPKPTP	DDDDDEDDDD
	PHPH	PAPEP	KKGGKK	DDDNDDD	EDEEEEDE	PTPSPTPSP	DDEDDDDEDD
	PSEP	PQRQP	REAAER	EDEGEDE	EEEAEEAE	RERERERER	GGNGGGGNGG
	PSRS	QRSRQ	SAASAA	GDDGDDG	EPEGDDDG	SAGAGAGAS	NNKKNNKKNN
	QQEQ	SATSS	AEEEEA	GGGQGGG	GDDDDDDG	SPSPSPSPS	PAPAPPAPAP
	RASQ	STTTT	DSDSDS	KEKEKEK	KEDKKDEK	ADADADADA	PPSPPPPSPS
	RPAR	DDEEE	EEKPEE	KGDPDGK	QPPQQPPQ	GSGSGSGSG	PTPPPTPSPT
	GKEE	KPEEP	EREERE	SSRRRSS	RGGGGGGR	PPSPPPSPP	QEEQEEQEEQ
	PPEE	PGPSP	QPPPPQ	SSSDSSS	RGRGRGRG	PTPPTPPTP	SAASAAASAA
	PTDD	SPSSS	RRQQRR	TTTATTT	SSPSSPSS	SDAKRKADS	SAASAASAAS

Table A21. Left components of characteristic inverse non-complementary repeats (material downloaded from NCBI) related to ordered regions

Table includes, for each repeat length, (at most) first 100 n-grams sorted according confidence, lift, and support, all in descending order.

Repeat length							
3	4	5	6	7	8	9	10
YLL	YLLY	VLLLV	NYVVYN	GLGAGLG	INLKKLNI	GLSVPVSLG	VVLALLALVV
FYF	CVVC	VVVVV	NLKKLN	EVIRIVE	GAGLLGAG	ELGNKNGLE	TGVTTTTVGT
YLY	YVVY	GVNVG	KLYYLK	ALAAALA	GKRHHRKG	VCVCVCVCV	LSILLLLISL
LWL	LCCL	ILSLI	VSVVSV	ASVDVSA	AGAGGAGA	GAAAGAAAG	AVGLLLLGVA
YFY	FYYF	IILII	ERYYRE	IERVREI	GAGAAGAG	GVGFGVGFG	CGFGCGFGCG
YVY	IINN	LGVGL	NKYYKN	GFGAGFG	KNLGGLNK	GLGAGLGAG	GCCGCCGCCG
IFI	YFFY	VLGLV	LLRRLL	NGDWDGN	YNLDDLNY	NELLSLLEN	IIIILLIIII
IIL	CIIC	ILFLI	GLAALG	RALDLAR	YQLLLLQY	IVKDRDKVI	LALLLLLAL
YIY	VPVL	LIKIL	VVVVVV	VAGSGAV	DVKTTKVD	AAWAAAWAA	STGGFFGGTS
FIF	MFFM	LIPIL	IILLII	LILKLIL	GGGLLGGL	CVCVCVCVC	VVVVVVVVVV
VWV	LLSL	LIDIL	IINNII	LLVRVLL	ILLIILLI	GVGVGVGVG	YNNYNNYNNY
FVF	CCCC	GLGAG	DELLED	VCVCVCV	LSELLESL	IGILLLIGI	AGAGAAGAGA
IYI	WAAW	IIDII	GTLLTG	VNRLRNV	SLFDDFLS	DGDLRLDGD	LLLLLLLLLL
CVC	LNIS	LLYLL	LLGGLL	AALALAA	AAYAAYAA	GYLSFSLYG	GAAGAAGAAG
LII	IMMI	IIFII	DVIIVD	AATQTAA	EILSSLIE	LLVLFLVLL	GAGAGGAGAG
ILI	CYYC	IIGII	FVLLVF	AVDEDVA	EVLEELVE	FIENFNEIF	
CLC	LVAL	VLTLV	IIKKII	HMSDSMH	SAFGGFAS	GSAFGFASG	
LLF	WEEW	TLQLT	IYKKYI	PIQVQIP	VNVGGVNV	HAILTLIAH	
III	AIGG	IINII	ALVVLA	AAGKGAA	IIIIIIII	LSDVGVDSL	
FLL	ALVA	DINID	ILLLLI	GGALAGG	NATAATAN	AAGLVLGAA	
LIL	LWWL	ILALI	ISFFSI	IKNKNKI	YDKAAKDY	ALDDADDLA	
GIV	SVRV	LIIIL	LVLLVL	AIEYEIA	GKCAACKG	ALNTFTNLA	
LFL	VAGL	LVIVL	TIDDIT	KATVTAK	AGATTAGA	EALLELLAE	
VYV	VCCV	NITIN	VLAALV	LDKIKDL	GGFGGFGG	EGSRIRSGE	
FLF	VGSV	IIAII	IKEEKI	LFSSSFL	GSLIILSG	FPKTVTKPF	
ICI	QWWQ	VVIVV	LFIIFL	RYLVLYR	ILILKLIL	GAGFGAGFG	
LCL	ICCI	LINIL	LFLLFL	WGCSCGW	KFGAAGFK	GVIPDPIVG	
VCV	AVGL	LVPVL	VLDDLV	CEVRVEC	LVKEEKVL	IEKFKFKEI	
LLV	FMMF	YPDPY	KKLLKK	ELLELLE	NNYNNYNN	LRLRLRLRL	
IIN	PLDI	VLFLV	LIIIIL	MMDYDMM	RIEGGEIR	LSNVGVNSL	
LYL	CHHC	FLLLF	LLVVLL	GVGFGVG	TGIAAIGT	LTASSSATL	
IIG	VVDG	IISII	GAGFGA	IVLLLVI	TIAIIAIT	NYNNYNNYN	
VVL	YTGL	IAFAI	IKNNKI	LKSASKL	WIEKKEIW	PALLNLLAP	
LIT	ALLL	LIAIL	LGAAGL	NATITAN	YNAIIANY	SIESASEIS	
ILL	ILGI	FLALF	NINNIN	VAGVGAV	AETTTTEA	STSTETSTS	
YAY	ASFV	VSFSV	RIGGIR	ADEIEDA	AIYKKYIA	TTAGTGATT	
VLL	TITI	LVTVL	VEDDEV	ELFNFLE	ALAGGALA	TVGSYSGVT	
YTY	TLTV	NVLVN	VLLLLV	IIIIIIII	ATAVVATA	YISISISIY	
YQY	TYLR	FILIF	VVAAVV	IVFTFVI	DDIDDIDD	AAGGIGGAA	
VFV	IGNG	ILVLI	AKNNKA	KLAVALK	FEKVVKEF	CGCCMCCGC	
VIV	LVVV	LYTYL	DIGGID	KTIDITK	GGSIISGG	CGFGCGFGC	
YRY	EPDY	ILTLI	ILGGLI	RAKLKAR	GILSSLIG	DSALHLASD	
YKY	FSGI	IRGRI	KAVVAK	YGGAGGY	IAAVVAAI	GAAGAGAAG	
VIG	IIKN	IVGVI	LLIILL	AAYEYAA	IYKNNKYI	GAGLGLGAG	
LVL	VDGT	LLCLL	YYGGYY	AFAGAFA	KIKIIKIK	GTSVWVSTG	
YGY	GCGL	VGYGV	GGIIGG	CVCVCVC	KNLTTLNK	GVGFGFGVG	
IAI	LALP	IVAVI	LAVVAL	GGYPYGG	LIIIIIIL	IIIIIIIII	
IGI	YRLF	LKKLN	LFFFFL	GKMLMKG	LLPLLPLL	IMKFLFKMI	
LLI	IILI	AIIIA	LGDDGL	GTGSGTG	LPGLLGPL	KNRNSNRNK	
IVV	IINK	IVDVI	LKVVKL	ILITILI	LSALLASL	LAALSLAAL	
IVI	LLLK	LFAFL	AVRRVA	LLLFLLL	NEYNNYEN	NIIKEKIIN	
FFF	GLLL	IVNVI	FFFFFF	LQFIFQL	NKRYYRKN	NNIINIINN	
FAF	GNIA	LFSFL	FILLIF	NIFEFIN	NNSVVSNN	SLWGNGWLS	
IDV	GVVS	LITIL	ITGGTI	QTEVETQ	RRWRRWRR	TIAGFGAIT	
VLV	KLNI	VIIIV	LLLGVA	YIESEIY	SQNIINQS	VQGLGLGQV	
LIN	LLLV	AIRIA	TLAALT	AGAFAGA	TEVKKVET	VTDTVTDTV	
LLL	LTGN	GFGVG	VALLAV	AGFGAGF	TLLTTLLT	AAAAHAAAA	
NII	AILS	LIYIL	VVGGVV	DGVEVGD	VVMVVMVV	AAARYRAAA	
IHI	AITG	LVYVL	DFIIFD	DVDYDVD	VVVVVVVV	AADAAAYAA	
YNY	INSN	LIFIL	DLTTLD	LQAAAQL	AALLLLAA	AAGLFLGAA	
AIV	LGLL	FLGLF	KIAAIK	VARHRAV	AATIITAA	AINLVLNIA	
IIK	LTTG	ITATI	KKSLIR	AAGNGAA	AIRAARIA	CGCCTCCGC	
ILT	NGTL	LFLFL	LFKKFL	AALYLAA	ALAAAALA	DELVLVLED	
IID	VLKT	LIGIL	NAIIAN	AAVVVAA	ELFGGFLE	DIDDIDDID	
FRF	GTVN	VIDIV	NIKKIN	ARVKVRA	ENYVVYNE	DLKGTGKLD	
IMI	LVPV	VISIV	NLEELN	AVAGAVA	FTFAAFTF	EVPFEFPVE	
ILA	SQTV	VCVCV	RLHHLR	DDGMGDD	GTLMMLTG	FGCGFGCGF	
VMV	VGVG	FTYTF	TKDDKT	GGLTLGG	IIAIIAII	FLSLCLSLF	
FTF	VLAL	GLILG	VGLGVG	GIGAGIG	IILLLLII	GAGFGFGAG	
INL	DIYS	IDLDI	VLGGLV	GITETIG	ILILLILI	GAGLGAGLG	
FPF	GTLN	IGVGI	AGFGAG	HFANAFH	ILLLLLLI	GFDLTLDFG	

TLI	IIDK	INYNI	AIFFIA	INLKKLN	ILVVVVLI	GGSAGASGG	
FNF	ISSV	LTMTL	GKIIKG	LTTITTL	INNIINNI	HYHYHYHYH	
IDI	ITIT	ISISI	GVAAVG	SIYRYIS	ITRFFRTI	IDEIEIEDI	
LVV	LLGL	IVKVI	ILIILI	TGVHVGT	IWNNNNWI	IGAATAAGI	
LNI	NNID	NIVIN	ISIISI	VGAKAGV	LAALLAAL	IILIIILII	
LFG	NNIL	IVLVI	IVLLVI	VLLSLLV	LDRYYRDL	ISDVYVDSI	
CGC	VGLP	VIGIV	KDVVDK	VTEEETV	LKQYYQKL	IYIYIYIYI	
IIT	YFGN	VIVIV	LKLLKL	AAAWAAA	LLRLLRLL	KEVFEFVEK	
INI	FAVG	VVNVV	NVFFVN	AGGWGGA	LNTLLTNL	KINNYNNIK	
VVV	IDLV	DLALD	NYPPYN	ANFTFNA	LRRRRRRL	LAVGAGVAL	
ITI	IIGN	FLSLF	VIVVIV	AVFGIVA	LTNIINTL	LCPCLCPCL	
VIA	LINN	GVGFG	AIGGIA	AVQYQVA	NLEIIELN	LFDEMEDFL	
GIL	LLLD	HGFGH	AWAAWA	DGVLVGD	NQLLLLQN	LKTKNKTKL	
INN	NGTT	IFEFI	DNVVND	DVSGSVD	NVNAANVN	LLLLTLLLL	
IVA	NINI	IGIGI	ETVVTE	ELLPLLE	QGELLEGQ	LLPLLLPLL	
ILN	NNIK	ILQLI	EVNNVE	GADVDAG	SDAIIADS	MLLLSLLLM	
YDY	YTTT	ILYLI	IISSII	GNFAFNG	SLVGGVLS	NDLMSMLDN	
IVN	ALVL	ISYSI	LDTTDL	IAGGGAI	TEAIIAET	NLKKLKKLN	
LVT	AVVL	IVIVI	LGAGAG	LIIIIIL	TGVAAVGT	PCLCPCLCP	
YSY	GLVA	IVVVI	LLTTLL	MFISIFM	TGYTTYGT	RLCCFCCLR	
IVG	LGQI	LLWLL	LTLLTL	PFVNVFP	TLSAASLT	SIDDEDDIS	
VTI	LLKG	VGNGV	LVGGVL	RGIEIGR	TVSGGSVT	SLALMLALS	
ITV	LPGK	VILIV	NCNNCN	RGSFSGR	VGLAALGV	SRLYRYLRS	
YEY	LVLA	VLKLV	NLTTLN	RTGVGTR	VLLLLLLV	VDDVVVDDV	
FKF	TDLY	YAVAY	RNHHNR	RVACAVR	VNTAATNV	VIVIVIVIV	
LML	TGDG	FGAGF	RVPVTE	TELFLET	VQLEELQV	VLPLCLPLV	
TIV	VGVV	KYLYK	VLSSLV	TTGHGTT	YARLLRAY	VVVVVVVVV	
NIL	VSGT	LFIFL	VTNNTV	VVVAVVV	YGAVVAGY	YAYDADYAY	
ILG	WALN	VLYLV	VVSSVV	YLLNLLY	YYNYYNYY	YRPDADPRY	

Table A22. Left components of characteristic inverse non-complementary repeats (material downloaded from NCBI) related to borderline regions

Table includes, for each repeat length, (at most) first 100 n-grams sorted according confidence, lift, and support, all in descending order.

Repeat length						
4	5	6	7	8	9	10
PVRV	GMWMG	RMSSMR	TPDFDPT	SPIGGIPS	LTPPTPPTL	RLRGLLGRLR
GDIA	MKWKM	QKIIKQ	TSAVAST	DNLEELND	AGDKIKDGA	DEVVEEVVED
AENR	MAWAM	GPWWPG	NAWGWAN	FSLEELSF	KSFKEKFSK	PVVPVVPVVP
MWWM	QMAMQ	GDKKDG	RVAQAVR	IELPPLEI	VKTSPSTKV	
RDES	MYYYM	MLEELM	DIAEAID	SAGGGGAS	AAAAEAAAA	
	ALAGS	MVTTVM	EITETIE	HFHHHHFH	AAAAGAAAA	
	HEFEH	RGEEGR	KFLDLFK	IEDEEDEI	LKLSDSLKL	
	MFQFM	EFFFFE	PRVFVRP	TDDRRDDT	LQMKLKMQL	
	MYMYM	ERGGRE	KLTITLK	AKLTTLKA	MMKMTMKMM	
	PFWFP	ETIITE	REEFEER	DKGAAGKD	PAAAAAAAP	
	QDFDQ	PRDDRP	RFQVQFR	EFKKKKFE	RTQVKVQTR	
	QVFVQ	SIQQIS	RSQPQSR	NKEFFEKN	VAEAEAEAV	
	SNLKK	AAAAPA	ADLMLDA	SLVEEVLS	VLEELEELV	
	WDWDW	ADMMDA	AEREREA	YEVRRVEY	VPVVPVVPV	
	WEEEW	ARTTRA	EANHNAE	SSAAAASS	YADAMADAY	
	YMMMY	DHDDHD	LSGSGSL	KELEELEK	DIDIDIDID	
	EYFYE	GNEENG	AEATAEA	AAASSAAA	RRRRWRRRR	
	ELRAL	ITEETI	AVARAVA	MKMEEMKM	ADAADAADA	
	ECECE	NHDDHN	EEIWIEE	AALAALAA		
	QRFRQ	RIEEIR	ELIAILE	AIKLLKIA		
	EYHYE	RRIIRR	GAMLMAG	LDLDDLDL		
	DWNWD	RDIIDR	GKIDIKG	LTQGGQTL		
	MNYNM	VVPPVV	HEEREEH			
	MVTVM	KKYYKK	INGNGNI			
	RYDYR	LNKKNL	LGNFNGL			

	QGDGQ	LNRRNL	LKDGDKL			
	YRSRY	MQMMQM	MKKRKKM			
	ASMSA	NVSSVN	PATPTAP			
	LHHHL	RRFFRR	QVSGSVQ			
	EFMFE	VKSSKV	RLKCKLR			
	HDADH	EVIIVE	SPFFFPS			
	CVFVC	ASEESA	SVRARVS			
	DGCGD	AAAAGA	TKVPVKT			
	HTDTH	DTDDTD	TSNNNST			
	IRWRI	MLLLLM	TSSGSST			
	PWSWP	PYPPYP	ADIVIDA			
	KDTDK	QVSSVQ	AEKMKEA			
	QSASQ	REPPER	AMQEQMA			
	YKLKY	SDFFDS	ARAVARA			
	KSWSK	TKSSKT	AVGTGVA			
	EIWIE	VEPPEV	DEIDIED			
	MVGVM	VINNIV	DIQNQID			
	QHYHQ	AHAAHA	DLPKPLD			
	KYSYK	EEYYEE	ELTKTLE			
	RPGPR	LDVVDL	ESIIISE			
	HPHPH		EYDDDYE			
	NRTRN		GAPIPAG			
	PDIDP		GEEDEEG			
			GHRNRHG			
			GPIRIPG			
			GQAVAQG			
			GSGKGSG			
			GTNPNTG			
			HANKNAH			
			IIEGEII			
			IKNINKI			
			KALDLAK			
			KELILEK			
			KFKLKFK			
			KIPQPIK			
			KNNHNNK			
			KVIYIVK			
			KVTCTVK			
			KYTETYK			
			LERAREL			
			LQQTQQL			
			LRQAQRL			
			MITLTIM			
			MKMTMKM			
			MTIVITM			
			PKKLKKP			
			PPVPVPP			
			QATSTAQ			
			QKYFYKQ			
			RADEDAR			
			REGRGER			
			RNRKRNR			
			SAFVFAS			
			SLEGELS			
			STNKNTS			
			SVSRSVS			
			SYKAKYS			
			TASSSAT			
			TAVAVAT			
			TIADAIT			
			TKVLVKT			
			TVEKEVT			
			VDDLDDV			
			VDVTVDV			
			VEEQEEV			
			VETKTEV			
			VPSKSPV			
			VSGDGSV			
			VTAGATV			
			EKVKVKE			
			LDKAKDL			
			EVTITVE			
			LREQERL			
			GGASAGG			
			TSTATST			

Table A23. Left components of characteristic inverse non-complementary repeats (material downloaded from DisProt) related to disordered regions

Table includes, for each repeat length, (at most) first 100 n-grams sorted according confidence, lift, and support, all in descending order.

Repeat length							
3	4	5	6	7	8	9	10
YTP	EQQE	PSYSP	SPSYSP	PAPAPAP	PQQPQQPF	PAPAPAPAP	APAPAPAPAP
PSY	DSDS	EKSEV	VPKKPV	PQQPQQP	QQPFPQQP	GGGGQGGGG	PQQPQQPFPQ
EQQ	NDDK	KKPVP	EEEEKE	QPFPQQP	QPQQPFPQ	PFPQQPQQP	QPFPQQPQQP
YSP	VPVP	PKKPV	EVQQVE	QPQQPFP	KPKAAKPK	QQPQQPFPQ	DDDDDDDDDD
QQE	PPPG	PVPKK	GVVVVG	GGGWGGG	AGAAAAGA	DEDEDEDED	GGGGGGGGGG
FSF	QQPF	QPQQP	QPQQPF	KHKDKHK	AAAAAAAA	QPQLPFPQQ	
GWG	QQVE	EVEVE	PPPPPG	SPSYSPS	DDDDDDDD	TPTPTPTPT	
EQK	GPPP	QQPFP	APAAPA	DEDEDED	EEEEEEEE	APAPAPAPA	
KLK	PEVP	PEEEE	EDEEDE	EAEAEAE		DDDDDDDDD	
ADA	GFSF	VPKKP	LVEEEE	GGAPAGG		SDSDSDSDS	
EEP	KKEP	PQQPQ	EDDDDE	GLFDFLG			
DSD	KPVP	QPFPQ	EEYEEY	ITSNSTI			
PVP	GHHG	EEPEE	EPKKPV	KKAPAKK			
SPS	EVQQ	PFPQQ	GLGGLG	KKPVPVP			
DDK	PQQS	KAPPA	KAEEAK	LPTGTPL			
DKD	FPQQ	PPPPG	KGKKGK	PKPEPKP			
VPV	FSFS	PVKVP	KSLLSK	RDRDRDR			
SEV	PFPQ	KGKGK	NNNNNN	SDSHSDS			
VES	QPFP	KPPPP	PFPQQP	TPTPTPT			
QSQ	QQPQ	VAVAV	DDDDDD	SDSDSDS			
APA	VKKP	EEGEE	AAAAAA	DDDDDDD			
GTG	AAAE	EEKKP	RRRRRR	PPPPPPP			
GYG	EKKP	EKGKE	PPPPPP	AAAAAAA			
KEK	EVEE	GHHGP	AAPPAA	APAPAPA			
NNN	PKKV	KAPPP	EEKKEE	DEEDEED			
QQP	EAPP	KKPEV	TAAAAT	EPEPEPE			
PKK	EPVP	KKPPP	EEEEEE	PKKAKKP			
PPP	VQQE	LVEEE	SSSSSS	EEEEEEE			
DED	FEEE	PVAKK					
QEL	KKAV	PVPVP					
PEE	PKVP	QQGYS					
	TNTG	RRQRR					
	YTPS	VEAPP					
	EDKD	AAEAA					
	EEKV	AQAQA					
	EGAA	DKHKD					
	FSFG	EDWDE					
	GGQG	EEWEE					
	KKIV	GGQGG					
	KNDK	KKAKK					
	PVKP	KKVKK					
	QGYS	PEEPV					
	SAEK	PEVPP					
	SEEN	PRPRP					
	TDDT	PVPEE					
	TFSF	QNNNQ					
	YSQQ	RRNRR					
	AAES	RRSRR					
	APPV	SFGSG					
	DKDE	VEPPP					
	EDED	VPPPK					
	EPPP	VPVPK					
	GCCG	VVEEK					
	KKPK	AAAPA					
	PKKE	AAKAA					
	PPKE	AAPAE					
	PPPE	AGAGA					
	QGGG	AGPGA					
	QNNQ	AVPVP					
	SAAP	DDSDD					
	SIIS	EDKDE					
	VEAE	EEEPP					
	VIKK	EEPVP					
	VPKE	EEYEE					
	AAEA	EGKGE					
	AQTT	EKKPV					
	DSKE	EPEEV					
	DTTD	EPNPE					
	DYEE	EPQPE					
	EAAP	EQKQE					
	EAVV	ESESE					

	EEAG	GGTGG					
	EHHE	GPPPP					
	EREE	GQQSQ					
	GGMG	GYGYG					
	KDSA	KAKKP					
	KDVE	KKAPP					
	KGKG	KKDKK					
	KPAA	KPPPA					
	KSFG	NASAN					
	NEEN	NDDKK					
	NNNQ	NNNNN					
	NPPN	NYQQY					
	NRTP	PGAGP					
	PAAA	PPPPK					
	PPAE	PVVVP					
	PTFS	QAQAQ					
	PVEL	QPLPQ					
	PVPT	QQPYP					
	QNNN	RDRDR					
	QSQQ	RRKRR					
	QSYG	SDEDS					
	RKEE	SDVDS					
	RRSR	SEVSK					
	SDED	SPAPS					
	SDKD	SQQPF					
	SEED	VEPEV					
	SKSD	VPAPA					
	SNQG	VPEEP					
	SNSN	VPPPV					

Table A24. Left components of characteristic inverse non-complementary repeats (material downloaded from DisProt) related to ordered regions

Table includes, for each repeat length, (at most) first 100 n-grams sorted according confidence, lift, and support, all in descending order.

Repeat length							
3	4	5	6	7	8	9	10
DSG	PPGP	GPPGP	PGPPGP	GPPGPPG	PGPPGPPG	SSSSSSSSS	SDSDSDSDSD
NEA	PGPP	PGPPG	PTPPTP	DSDSDSD	KKKIIKKK	DSDSDSDSD	SSSSSSSSSS
NVA	GAPG	GGRGG	GRGRGR	GPPGPAG	AEVEAAKK	PGPPGPPGP	EEEEEEEEEE
AEL	EKQK	GAPGP	KKSAAE	AEATAEA	GPPGPPGP	GGGGSGGGG	QQQQQQQQQQ
VRF	KQKE	PGPAG	ASKKAA	AAKKSAA	HHHHHHHH	QQQQQQQQQ	
VAT	GTPP	PGAPG	EATAEA	AASKKAA	SDSDSDSD	EEEEEEEEE	
FRV	EKRE	ASKKA	LLLLLL	GGGGSGG	SSSSSSSS	SSSSSSSSS	
TLK	ERKE	AGAPG	PPDIPD	GPSSSPG	GGGGGGGG	DSDSDSDSD	
VAS	AAKK	KKAAE	PPGPPS	GQPGPAG	PGPPGPPG	PGPPGPPGP	
TVR	PPSF	EAAKK	SPPGPP	GVPFPVG	KKKIIKKK	GGGGSGGGG	
AAN	VPGP	GQPGP	VEAAKK	PEPSPEP	AEVEAAKK	QQQQQQQQQ	
AVN	ENEA	VEKRE	AGPPGA	QQQAQQQ	GPPGPPGP	EEEEEEEEE	
LEI	GSPG	AANVA	EAASKK	QVEGEVQ	HHHHHHHH		
TTK	PGPV	ADAVK	GGRRGG	SLSSLSS	SDSDSDSD		
GDY	GRGG	AKKSA	KKVVKK	VVASAVV	SSSSSSSS		
KTT	PTGP	ASSSA	PGPAGA	GGGSGGG	GGGGGGGG		
ATN	ARVR	AYRYA	SLSSLS	GGGGGGG			
KNV	KLTV	GPAGA	AAKKSA	GGRGRGG			
ITA	KAAE	LDADL	AGAPGP	PEPEPEP			
TAV	KEVI	AGPPG	APPGPP	HHHHHHH			
NKA	KGSD	GSPGP	EEELKL	QQQQQQQ			
SVR	GGRG	KVADA	FLAALF				
YKG	NVAS	RGPPG	GAGGAG				
LTA	VTLT	LLALL	GPPGPV				
IKS	YDGG	SSISQ	GRGGRG				
IGE	ASKK	AGKPG	ISTTSI				
YGI	KKKV	ATSTA	KKNNKK				
TLV	PKKK	ELEKQ	LFEEFL				
PRG	EDSG	GPPGA	LVGGVL				
SDA	ESEA	IENEA	LVVVVL				
RSV	KDGK	IRSGG	PGGPGG				

IGY	KVTG	PPDPP	PGPAGP				
NVK	PGTP	QKLES	PPGPPK				
VLT	KKKP	RSRSR	RREERR				
AET	LTVK	TIKAG	SDSDSD				
RVS	NVAT	VAKAV	QQQQQQ				
GKY	SPPG	VPGPV	GGGGGG				
TAK	ASVR	WTHTW	KKKKKK				
TYT	AVNK	AAVAA	AGAAGA				
EGI	GPPD	EELKL	HHHHHH				
PGT	KNVA	ELPEG					
PKD	LTVT	EPPGP					
TDL	PGEP	GGFGG					
DRK	TVVA	GPAGP					
QLE	AVKK	SITIS					
VSN	DSGK	SPPSP					
LEC	GPRG	ALLLA					
AVI	PGKP	EKKEV					
IKI	PKDL	ERREV					
WTV	ELLK	GKPGD					
ATI	GPPS	GPEGA					
TFE	GQPG	KAGAK					
LLT	KKSK	KEREK					
TYK	LEKQ	KEYEK					
GDD	NTAT	LAQAL					
IVS	NYIV	LEAEI					
REI	TVKL	PGSPG					
GPV	ATVT	PKKKV					
KTW	GEPG	SASSS					
NTA	GPSG	SATAS					
GTY	GSDS	SPPGP					
TKL	GSKI	TVKVT					
SDP	KTTD	AASKK					
IEG	SKTV	AESEA					
NSV	VKVT	AGYGA					
VTF	AESE	AKAKA					
KAT	GKPG	ATNTA					
DIT	GPQG	AVVVA					
VTW	GRPG	CPEPC					
FEV	KVAD	EELEE					
LPG	LSVE	EEPKE					
TDV	PEGV	EPKEE					
IAT	SDSG	GGIGG					
TGP	SSSA	GPAGE					
VNE	TLTV	GPPGR					
GLI	VATV	GPRGP					
RVT	VRFQ	GSKII					
TKN	DGKK	ININI					
IVA	KLEA	KEKQK					
TAR	QKEL	KKKPE					
VIN	SNEA	KLQLK					
VKF	TEKS	KVKTL					
LRS	TVTS	LADAL					
SID	VERR	LEPPE					
SKQ	ETLG	LKTKL					
VDL	FRNE	LSDSL					
DPK	GPKG	LSQSL					
KGT	GTPG	LTKVK					
GEY	ILEK	PLDLP					
YGV	KIIK	PQQQP					
IRS	LTIK	QPPPQ					
NVP	LVVL	SEAES					
RVN	PPDP	SHSHS					
ADS	SSTS	SISIS					
LRI	STAT	SSFSS					
SAD	TEEV	TKLKT					
AIN	VGAA	TVQSG					
ANV	VKVL	VEFEV					
IPD	VPPS	VKKKP					
IPG	VTLK	VPGPP					

Table A25. Order levels and lengths of homorepeats found in association rules

Order level	Amino acid	Homorepeat length	Rule lift	Rule confidence
DD	A	6	0.863	53.18
		7	0.863	55.39
		8	0.703	59.04
		9	0.863	70.73
		10	0.877	80.00
	D	3	3.063	68.65
		4	2.574	85.52
		5	2.157	93.82
		6	1.568	96.65
		7	1.559	100.00
		8	1.191	100.00
		9	1.220	100.00
		10	1.096	100.00
	E	3	3.303	74.02
		4	2.758	91.63
		5	2.268	98.64
		6	1.619	99.79
		7	1.559	100.00
		8	1.191	100.00
		9	1.220	100.00
		10	1.096	100.00
	G	3	2.558	57.33
		4	2.493	82.84
		5	2.299	100.00
		6	1.623	100.00
		7	1.559	100.00
		8	1.191	100.00
		9	1.220	100.00
		10	1.096	100.00
	H	4	3.010	100.00
		5	2.299	100.00
		6	1.623	100.00
		7	1.559	100.00
		8	1.191	100.00
		9	1.220	100.00
		10	1.096	100.00
	K	3	2.587	57.98
		4	2.368	78.68
		5	2.299	100.00
		6	1.623	100.00
		7	1.559	100.00
		8	1.191	100.00
		9	1.220	100.00
	N	4	1.901	63.17
		5	2.299	100.00
		6	1.623	100.00
		7	1.559	100.00
		8	1.191	100.00
		9	1.220	100.00
		10	1.096	100.00
	P	3	4.313	96.65
		4	2.993	99.45
		5	2.299	100.00
		6	1.623	100.00
		7	1.559	100.00
		8	1.191	100.00
		9	1.220	100.00
		10	1.096	100.00
	Q	3	4.026	90.23
		4	2.919	97.00
		5	2.276	98.97
		6	1.619	99.74
		7	1.559	100.00
		8	1.191	100.00
		9	1.220	100.00
		10	1.096	100.00
	R	3	2.886	64.68
		4	2.208	73.37
		5	1.863	81.02
		6	1.298	80.00
		7	1.357	87.03
		8	1.128	94.73
		9	1.220	100.00

		10	1.096	100.00
	S	3	3.353	75.14
		4	2.830	94.04
		5	2.264	98.48
		6	1.620	99.82
		7	1.558	99.90
		8	1.191	100.00
		9	1.220	100.00
		10	1.096	100.00
	T	4	2.183	72.52
		5	2.141	93.13
		6	1.623	100.00
		7	1.559	100.00
		8	1.191	100.00
		9	1.220	100.00
		10	1.096	100.00
OO	C	4	1.953	100.00
		5	2.024	100.00
		6	3.028	100.00
		7	3.209	100.00
	F	3	1.615	96.00
		4	1.889	96.70
		5	1.984	98.03
		6	3.028	100.00
		7	3.209	100.00
	I	3	1.644	97.70
		4	1.930	98.82
		5	1.968	97.22
		6	2.954	97.56
		7	3.209	100.00
		8	7.755	100.00
		9	6.629	100.00
	L	3	1.609	95.61
		4	1.901	97.31
		5	1.981	97.86
		6	2.987	98.64
		7	3.109	96.87
		8	7.238	93.33
		9	5.800	87.50
		10	11.998	87.50
	M	5	2.024	100.00
		6	3.028	100.00
	V	3	1.577	93.72
		4	1.843	94.33
		5	2.024	100.00
		6	3.028	100.00
		7	3.209	100.00
		8	7.755	100.00
		9	6.629	100.00
		10	13.712	100.00
	W	4	1.953	100.00
		5	2.024	100.00
		6	3.028	100.00

Biography

Samira Almokhtar Alshafah was born on 29th December 1978 in Zawia, Libya. She finished primary school in Almajed School (Harsha-Libya) 1993, high school in Almajed School, (Harsha-Libya) 1996, and Bachelor studies in computer science at Zawia University, Faculty of Engineering, Department of Electronic Engineering (Zawia-Libya) 2001. She was on master studies in Computer science at Libya Academy of Graduate Studies (Tripoli) from 2004 to 2007. and graduate at 2007. Samira enrolled PhD studies in 2010 at Faulty of Mathematics, University of Belgrade, Serbia, as a candidate from Zawia University for PhD studies.

She worked as a teacher in computer science in high school (Harsha- Libya) from 2002 to 2004, and as a lecturer at the Faculty of Engineering, Department of Electronic Engineering (Zawia-Libya) from 2004. Member of the examination committee in the Faculty of Engineering (Zawia-Libya) became 2004 and 2005, and a faculty member (professor) at Faculty of Engineering, Department of Electronic Engineering (Zawia-Libya) after graduate of the Master studies 2007. Samira also taught courses in Java language at the Institute of Higher education of Computer Technologies (Enjela-Libya) 2007-2008. She was the supervisor of the project Graduated bachelor degree at the Higher Institute of Computer Technologies (Enjela-Libya) 2009.

Her research interest was in developing DC motor drive using computer, Handwritten Arabic Characters Recognition, and after enrolled in PhD studies bioinformatics and data mining.

Samira Published two research papers in International Journals and two papers in conferences.

She is married and has three children.

Изјава о ауторству

Име и презиме аутора Самира Алсхафах

Број индекса 2042/2009

Изјављујем

да је докторска дисертација под насловом

Истраживање података на протеинским нискама: н-грамска анализа уређених и неуређених региона протеина

- резултат сопственог истраживачког рада;
- да дисертација у целини ни у деловима није била предложена за стицање друге дипломе према студијским програмима других високошколских установа;
- да су резултати коректно наведени и
- да нисам кршио/ла ауторска права и користио/ла интелектуалну својину других лица.

Потпис аутора

У Београду, 15.05.2018.

_____

Изјава о истоветности штампане и електронске верзије докторског рада

Име и презиме аутора Самира Алсхафах

Број индекса 2042/2009

Студијски програм Информатика

Наслов рада Истраживање података на протеинским нискама: н-грамска анализа уређених и неуређених региона протеина

Ментор Ненад Митић

Изјављујем да је штампана верзија мог докторског рада истоветна електронској верзији коју сам предао/ла ради похрањена у **Дигиталном репозиторијуму Универзитета у Београду.**

Дозвољавам да се објаве моји лични подаци везани за добијање академског назива доктора наука, као што су име и презиме, година и место рођења и датум одбране рада.

Ови лични подаци могу се објавити на мрежним страницама дигиталне библиотеке, у електронском каталогу и у публикацијама Универзитета у Београду.

Потпис аутора

У Београду, 15.05.2018.

_____

Изјава о коришћењу

Овлашћујем Универзитетску библиотеку „Светозар Марковић“ да у Дигитални репозиторијум Универзитета у Београду унесе моју докторску дисертацију под насловом:

Истраживање података на протеинским нискама: н-грамска анализа уређених и неуређених региона протеина

која је моје ауторско дело.

Дисертацију са свим прилозима предао/ла сам у електронском формату погодном за трајно архивирање.

Моју докторску дисертацију похрањену у Дигиталном репозиторијуму Универзитета у Београду и доступну у отвореном приступу могу да користе сви који поштују одредбе садржане у одабраном типу лиценце Креативне заједнице (Creative Commons) за коју сам се одлучио/ла.

1. Ауторство (CC BY)

2. Ауторство – некомерцијално (CC BY-NC)

3. Ауторство – некомерцијално – без прерада (CC BY-NC-ND)

4. Ауторство – некомерцијално – делити под истим условима (CC BY-NC-SA)

5. Ауторство – без прерада (CC BY-ND)

6. Ауторство – делити под истим условима (CC BY-SA)

(Молимо да заокружите само једну од шест понуђених лиценци.
Кратак опис лиценци је саставни део ове изјаве).

Потпис аутора

У Београду, 15.05.2018.

_____

1. **Ауторство**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце, чак и у комерцијалне сврхе. Ово је најслободнија од свих лиценци.

2. **Ауторство – некомерцијално**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела.

3. **Ауторство – некомерцијално – без прерада**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела. У односу на све остале лиценце, овом лиценцом се ограничава највећи обим права коришћења дела.

4. **Ауторство – некомерцијално – делити под истим условима**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца не дозвољава комерцијалну употребу дела и прерада.

5. **Ауторство – без прерада**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца дозвољава комерцијалну употребу дела.

6. **Ауторство – делити под истим условима**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца дозвољава комерцијалну употребу дела и прерада. Слична је софтверским лиценцама, односно лиценцама отвореног кода.