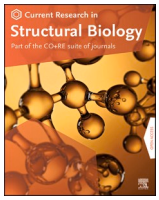
Current Research in Structural Biology 7 (2024) 100129

Contents lists available at ScienceDirect

Current Research in Structural Biology

journal homepage: www.journals.elsevier.com/current-research-in-structural-biology



Research Article

Living in trinity of extremes: Genomic and proteomic signatures of halophilic, thermophilic, and pH adaptation

Aidana Amangeldina [a,b], Zhen Wah Tan [a], Igor N. Berezovsky [a,b,*]

[a] Bioinformatics Institute (BII), Agency for Science, Technology and Research (A*STAR), 30 Biopolis Street, #07-01, Matrix, 138671, Singapore
[b] Department of Biological Sciences (DBS), National University of Singapore (NUS), 8 Medical Drive, 117579, Singapore

ARTICLE INFO

Handling editor: A Wlodawer

Keywords:
Thermophilic adaptation
Halophilic adaptation
PH adaptation
Genomes/proteomes
Protein physics
Molecular adaptation
Protein evolution
Archaea
Bacteria

ABSTRACT

Since nucleic acids and proteins of unicellular prokaryotes are directly exposed to extreme environmental conditions, it is possible to explore the genomic-proteomic compositional determinants of molecular mechanisms of adaptation developed by them in response to harsh environmental conditions. Using a wealth of currently available complete genomes/proteomes we were able to explore signatures of adaptation to three environmental factors, pH, salinity, and temperature, observing major trends in compositions of their nucleic acids and proteins. We derived predictors of thermostability, halophilic, and pH adaptations and complemented them by the principal components analysis. We observed a clear difference between thermophilic and salinity/pH adaptations, whereas latter invoke seemingly overlapping mechanisms. The genome-proteome compositional trade-off reveals an intricate balance between the work of base paring and base stacking in stabilization of coding DNA and r/tRNAs, and, at the same time, universal requirements for the stability and foldability of proteins regardless of the nucleotide biases. Nevertheless, we still found hidden fingerprints of ancient evolutionary connections between the nucleotide and amino acid compositions indicating their emergence, mutual evolution, and adjustment. The evolutionary perspective on the adaptation mechanisms is further studied here by means of the comparative analysis of genomic/proteomic traits of archaeal and bacterial species. The overall picture of genomic/proteomic signals of adaptation obtained here provides a foundation for future engineering and design of functional biomolecules resistant to harsh environments.

1. Introduction

The genomic-proteomic compositional determinants of molecular mechanisms of adaptation to extremes of pH, salinity, and temperature in unicellular organisms, archaea, and bacteria, are studied here. Cell viability is obviously dependent on proteins that perform various functions, whereas the passage of genetic information necessary for building the cell machinery also requires the integrity of the genetic material (DNA) and the transcriptional intermediate (RNA). Thus, DNA, RNA, and proteins are the most essential molecules in the cell, and their functioning and physical stability at any given environmental condition determine cell survival and reproduction. As directed by the central dogma of molecular biology, the integrity of the flow of genetic information cannot be guaranteed without stability of any element in the chain: from DNA to RNA to proteins. Some previous studies reported that nucleic acid bias shapes amino acid compositions and thus shapes the evolutionary landscape of proteins (Singer and Hickey, 2000; Bohlin et al., 2013; Tekaia and Yeramian, 2006) while others argue that there is more complex relationship (Goncearenco and Berezovsky, 2014), especially in relation to adaptation to extreme environments (Fukuchi et al., 2003; Goncearenco et al., 2014; Nakashima et al., 2003; Tekaia et al., 2002). Extreme environmental adaptation introduces certain trends for stability on both the nucleic acid (DNA/RNA) and protein levels, such that DNA with biases might encode further for different amino acids, and amino acid biases needed for protein stability may originate trends towards specific codons, affecting, thus, the nucleotide sequences (Goncearenco and Berezovsky, 2014). Therefore, a combination of the redundant genetic code and the availability of several types of amino acids with similar physicochemical properties leads to a complex trade-off working in the mutual environmental adaptation of DNA and proteins (Goncearenco and Berezovsky, 2014; Goncearenco et al., 2014).

* Corresponding author. Bioinformatics Institute (BII), Agency for Science, Technology and Research (A*STAR), 30 Biopolis Street, #07-01, Matrix, 138671, Singapore.
E-mail address: igorb@bii.a-star.edu.sg (I.N. Berezovsky).

https://doi.org/10.1016/j.crstbi.2024.100129
Received 29 November 2023; Received in revised form 16 January 2024; Accepted 16 January 2024
Available online 1 February 2024

Additionally, consideration of adaptation mechanisms should also include the evolutionary perspective (Goncearenco et al., 2014; Berezovsky and Shakhnovich, 2005; Tokuriki et al., 2009) that leaves its marks from the very Origin of Life, determining the genetic code emergence, codon chronology (Trifonov, 2000), and consensus temporal order of amino acids (Trifonov, 2000; Trifonov et al., 2001), The evolution is driven by the polymer nature of nucleic acids and proteins that establishes their basic units (Berezovsky et al., 1999, 2000a, 2017a; Koczyk and Berezovsky, 2008; Svedberg, 1929; Berezovsky, 2003; Berezovsky and Trifonov, 2001), and shapes them on different stages of evolution (Trifonov et al., 2001; Aziz and Caetano-Anolles, 2021), The emergence of functional diversity (Berezovsky et al., 2003, 2017a; Aziz et al., 2016; Tal et al., 2016; Zeldovich et al., 2006; Goncearenco and Berezovsky, 2015), and specific ways of its design (Berezovsky, 2019; Hocker, 2014; Yin et al., 2021) and regulation (Berezovsky et al., 2017b), including allosteric mechanisms (Tee et al., 2020, 2021, 2022; Guarnera and Berezovsky, 2019) are also results of the evolution.

Enzymes from extremophiles with enhanced balance between the stability, activity, and flexibility in order to function (Hou et al., 2023) at extreme pH, salinity, and temperature are instrumental in application in harsh industrial processes. The protein stability is a result of a mutual work of positive and negative components of design (Berezovsky et al., 2007), making as low as possible energy of the protein native state (positive design) and increasing energies of misfolded conformations (negative design). It results in a widening of the gap between the energies of the native state and non-native structures (Berezovsky et al., 2007). The case study of the work of negative and positive components of design in protein thermostability revealed so-called “from both ends of hydrophobicity scale” trend in the amino acid composition (Goncearenco et al., 2014; Berezovsky et al., 2007; Ma et al., 2010). Specifically, increasing usage of strong hydrophobic and charged residues at the expense of polar ones was observed upon increase of the organismal optimal growth temperature (OGT). This compositional bias was shown to contribute to the enthalpy of the protein, providing stronger van der Waals interactions (Berezovsky and Shakhnovich, 2005; Berezovsky, 2003; Berezovsky et al., 1997, 2000b) in the protein core and forming additional stabilizing ion pairs and hydrogen bonds on the protein surface – elements of the *positive design* (Berezovsky et al., 2007). At the same time, increase of the amount of positively charged amino acids contributes to the negative component of design by increasing repulsions between positive charges (Berezovsky et al., 2007) in misfolded conformations and in non-native protein-protein interfaces (Ma et al., 2010; Berezovsky, 2011). Obviously, the fraction of only one sign of charged residues can be used in the negative design to provide massive repulsion in non-native conformation. In protein thermostability of both monomeric proteins (Goncearenco et al., 2014; Berezovsky et al., 2007; Ma et al., 2010; Berezovsky, 2011) and protein-protein interfaces (Ma et al., 2010) the positive charges are apparently a key element of the negative design, increasing the energies of non-native conformations. As a result, the gap between the native state energy and those of misfolds is widening and thermodynamic stability of the protein is increasing (Goncearenco et al., 2014; Berezovsky et al., 2007; Ma et al., 2010; Berezovsky, 2011). Adaptation to extremes of temperature and to other harsh environments is provided by combined contributions of distinct stabilizing interactions (Berezovsky, 2003, 2011; Berezovsky et al., 1997, 1999, 2000b, 2005; Cambillau and Claverie, 2000; Folch et al., 2008; Jaenicke, 1999; Dyson et al., 2006; Pace et al., 2014a, 2014b; Pucci and Rooman, 2014; Pylaeva et al., 2018; Shakhnovich, 2006; Van Dijk et al., 2015; Kajander et al., 2000; Makhatadze et al., 2003; Bresler and Talmud, 1944a, 1944b), presence of which is reflected in a number of compositional and sequence/structure determinants (Goncearenco et al., 2014; Ma et al., 2010; Cambillau and Claverie, 2000; Shakhnovich, 2006; Van Dijk et al., 2015; Chakravarty and Varadarajan, 2000; Mamonova et al., 2013; Gromiha et al., 2013; Gromiha and Suresh, 2008). It was shown, for example, that specific interactions and characteristics of the structure may dominate in adaptation to certain extreme conditions, such as (hyper)thermophilic and psychrophilic adaptation (Pucci and Rooman, 2017), which require better packed and stabilized structures in the former and their less rigid, flexible homologs in the latter (Feller and Gerday, 2003; Goodchild et al., 2004).

A salty environment is characterized by the low water availability at high salt ionic concentrations, providing less contribution to protein folding by hydrophobic effect. It was shown that halophilic adaptation can happen through destabilization of unfolded state by cation exclusion in unfolded states, while electrostatic interactions between the cation and abundant acidic amino acid residues can contribute to the stability of the folded state (Ortega et al., 2015). Putting above results into the terminology of positive-negative design (Berezovsky et al., 2007) discussed above: depletion of positive charges leads to additional repulsion in misfolded conformations constituting the negative component of design, while contribution to the positive one is provided by electrostatic interactions between the environmental cations and negatively charged amino acids. The excess of acidic residues can also support the protein hydration, stabilizing the folded protein form via interactions with hydrated cations at high salt concentrations (Ortega et al., 2015; Deole et al., 2013; Ebrahimie et al., 2011). In the other study it was suggested that hydration of protein with various acidic residue content had the same hydration level, while the role of acidic residues could be in the prevention of protein aggregation (Daronkola and Verde, 2021), which might be another example of the work of negative charges in negative design. At the same time, however, in addition to documented role of acidic amino acid residues in high salt adaptation, it was proposed elsewhere that halophilic adaptation can be based on basic residues (Elevi Bardavid and Oren, 2012).

There is also a diversity of opinions on the determinants of pH adaptation, including opposite conclusions that increased (decreased) positively charged residues and decreased (increased) negatively charged residues are important for high/low pH adaptation (Daronkola and Verde, 2021). The specificity of protein adaptation to different pH environments for each class of proteins was also proposed (Dubnovitsky et al., 2005). Since pH adaptation is related to the change of the charge state of the amino acid, which might lead to disruption of the existing structure, ionizable residues and the ionization state of certain residues seem to play an important role in protein stability and function (Beliën et al., 2009; Xu et al., 2013). For example, replacing residues susceptible to charge-state change with non or less-susceptible ones was found to be a viable strategy for pH adaptation design (Suplatov et al., 2014). It was also shown that replacing a basic amino acid with an acidic one improved the protein stability and catalytic efficiency under an acidic environment (Yang et al., 2013). And vice versa, replacing acidic residues with less acidic or more basic (like glutamic acid with glutamine) led to a change of the pH optima (Suplatov et al., 2014; Fushinobu et al., 1998; Liu et al., 2009). Additional potential strategy for both pH adaptation and salinity adaptation is pKa modulation (Beliën et al., 2009; Xu et al., 2013; Gutteridge and Thornton, 2005; Francois et al., 2006; Andreeva and James, 1991).

The next level of complexity in the study of environmental adaptation is a situation when some extreme environments are coupled (Oren, 2002; Reed et al., 2013), in which certain compositional trends can be related to adaptation to several different extreme environments at the same time. For example, an increased ratio of acidic over basic amino acid residues has been linked not only to psychrophiles (Xia et al., 2018), but also to adaptation to high pH (alkalophilic proteins) (Mamo et al., 2009) and to high salt (halophilic proteins) (Daronkola and Verde, 2021). Another example is Asparagine (N) amino acid residue that is thermolabile and alkali susceptible (Gülich et al., 2002; Walden et al., 2004): increased fraction of charged amino acids and lower content of Asn residues is an adaptation trend common for both thermophilic and alkaliphilic protein stabilization (Manikandan et al., 2006). The elevated number of salt bridges is characteristic for thermophilic and salinity adaptations (Berezovsky et al., 2007; Ma et al., 2010; Dyson et al., 2006; Mamonova et al., 2013; Bandyopadhyay et al., 2007; Dym

et al., 1995; Nayek et al., 2014). So far, only limited number of studies account for poly-extremophilic adaptation, where adaptation to one factor might be affecting (enabling or disabling) the other type of adaptation (Alcaide et al., 2015; Popinako et al., 2017; Sriaporn et al., 2021). In particular, it was shown that adaptation to high salinity can be coupled with adaptation to a high pH (Manikandan et al., 2006). An increase in acidic residues and decrease in Lysine, which becomes unstable due to proton loss at high pH and has a too-long hydrophobic chain for the high salt environment, was observed for both alkaliphilic and halophilic proteins (Ortega et al., 2015; Popinako et al., 2017). Another drawback in previous studies of the polyextremophile adaptation is that they are typically performed on individual proteins, not allowing to infer generic trends reflecting mechanisms of adaptation characteristic for the whole proteomes.

We work here with large datasets of genomes/proteomes annotated with data on the optimal growth pH (OGP), salinity (OGS), and temperature (OGT) of corresponding organisms. It allows us to capture the most important signals of adaptation to each extreme environment, to follow trends reflecting the stability tuning of both nucleic acids and proteins, to delineate the causal relation between the compositional trends in amino and nucleic acids, and to find the relationship between adaptation mechanisms evolved in response to different environments. We also considered emergence and development of adaptation mechanisms from the evolutionary perspective, analysing the compositional trends characteristic for mechanisms of adaptation working in groups of archaeal and bacterial organisms.

2. Materials and Methods

2.1. Environmental data

BacDive database was used as the primary source database to build the sets of organisms with available temperature, pH, or salinity and to obtain corresponding lists of BacDive IDs for each environmental factor. The original data consisted of 'general growth' and 'optimal growth' data, represented as 'numeric range' or 'numeric points'. Each environmental entry for one BacDive entry was accessed, and we have selected 'optimum' and 'growth' points. The entry to be used in the analysis for characterizing features of corresponding proteome was selected according to the following priority: the optimal growth (OG) number for each environmental factor is assigned as the average value of optimal growth interval. For those organisms that had entry points given as only one number we use this number as an optimal one (both minimum and maximum values are designated to this number). If only maximum or minimum optimum growth points was available (for example <105 °C), this available number is taken as an OG value. If optimal growth interval is not available, we use the average of the growth value interval, or the available general growth point. The salt concentration measurements in molarity or artificial seawater concentration were converted to NaCl salt percent concentration. Below are examples of the conversion of molarity and artificial seawater concentration into the NaCl percent concentration. First, given molarity = 0.14 M and molar weight of NaCl $MW_{NaCl} = 58.44\ \frac{g}{mol}$, one obtains $0.14\ \frac{Mol}{L} * 58.44\ \frac{g}{mol} = 8.18\ \frac{g}{L}$, or 0.82 % NaCl concentration. Second, given the reference protocol for preparation of the 35 % artificial seawater (AS (Kester et al., 1967),) that should contain 23.93 g/kg of NaCl (2.393 %, slight halophile), the AS = 100 % will show 6.8 % salt concentration corresponding to moderate halophile.

The organism annotation consists of an organism name, domain, oxygen tolerance, phylum, gram stain. Partially or fully unclassified species and eukaryotic organisms were discarded from the dataset. The species without domain or with several unconventional environmental factor annotations were corrected or cleared out from the dataset. The four datasets were obtained: full pH dataset (pH_set, Suppl. Table S1), full salinity dataset (S_set, Suppl. Table S2), full temperature data (T_set, Suppl. Table S3), and the dataset with all three environmental factor data available for each organism (Env_set, Suppl. Table S4). The classification of organisms according to their degree of adaptation to extreme environments is presented in Table 1. The statistics for all the above groups are provided in Table 2, and the information on over-representation of organisms on certain environmental conditions (e.g., OGTs 28, 30, 37 °C etc.) - in Suppl. Table S6.

2.2. Sequence data

The nucleotide coding sequences and protein sequences were downloaded from the NCBI database (Refseq (Tatusova et al., 2016) or GenBank (Clark et al., 2016)). tRNA and rRNA data has been extracted from GenBank. Non-codon-biased sequences were generated by assigning equal probability to each of the synonymous codons weighted by their encoding amino-acid weight in the proteome. Amino acid composition and dipeptide composition was calculated for each protein for the proteome of each organism.

Proteins were annotated as a membrane or nonmembrane proteins (Olivella et al., 2013), depending on the presence of 'membr' string in the protein name. Further, the missing membrane annotations were extracted using BLAST+, version 2.10 (Camacho et al., 2009). The BLAST search was built on a blastp-fast task with a parameter of e-value equal to 10^{-5} with minimum query coverage of 50 % per high-scoring segment pair. The proteins with a minimum 35 % identity with membrane proteins were annotated as membrane ones and excluded from further consideration because of the distinct amino acid composition between the membrane and nonmembrane proteins.

2.3. Proteomic predictors of adaptation to extremes of pH, salinity, and temperature

Amino acid composition frequencies were calculated for each proteome. The artificial proteomes were created via averaging compositions of proteomes of species having overrepresented temperature, pH, and salinity measurement points. We introduced here a "z score'' predictors of OGT, OGS, and OGP which take into account different variance of individual amino acids. We showed earlier that in different combinations of amino acids some of them can outweigh contribution of others to the adaptation signature, because of the larger variance of their frequencies (Ma et al., 2010). To account for different variances, the individual changes in amino acid frequencies and make the effects comparable we standardize the frequencies for each organism g: $z_{gi} = \frac{f_{gi} - <f_i>}{\sigma_{f_i}}$, $i = (1, 2, \ldots 20)$, where f_{gi} is the frequency of the amino aicd in the proteome, $< f_i >$ is average frequency of the amino acid in the dataset, and σ_{f_i} – the standard deviation of the amino acid frequency in the dataset. Suppl. Fig. S1 shows proteomic amino acid frequencies with

Table 1
Classification range criteria used for different environmental groups.

Classification Range	Classification group
Temperature	
(, 25] °C	psychrophiles
(25, 50) °C	mesophile
[50–80) °C	thermophile
[80) °C	hyperthermophile
Salinity	
(, 2] NaCl %	nonhalophile
(2, 5) NaCl %	slight halophile
[5, 20) NaCl %	moderate halophile
[20) NaCl %	extreme halophile
pH	
[0, 5.5] pH	acidophile
(5.5, 8) pH	neutrophile
[8, 14] pH	alkaliphile

Table 2
Summary of datasets for each environmental factor and for environmental set. Note: BacDive classifies 186 organisms from * as thermophilic, one organism from * as psychrophilic, two organisms from ** as hyperthermophilic.

Dataset type	*Counts*	*Bacteria*	*Archaea*
Full T data (T_set)	***9225***	***8884***	***341***
psychrophiles	648	645	3
mesophile	8052*	7830	222
thermophile	445**	400	45
hyperthermophile	80	9	71
Full S data (S_set)	***2568***	***2489***	***79***
nonhalophile	1239	1217	22
slight halophile	806	801	5
moderate halophile	496	469	27
extreme halophile	27	2	25
Full P data (P_set)	***2958***	***2869***	***89***
alkaliphile	497	488	9
neutrophile	2348	2274	74
acidophile	113	107	6
Env Dataset (Env_set)	**2236**	**2159**	**77**
psychrophiles	238	238	0
mesophile	1870	1807	63
thermophile	118	112	6
hyperthermophile	10	2	8
nonhalophile	1103	1082	21
slight halophile	699	694	5
moderate halophile	408	381	27
extreme halophile	26	2	24
neutrophile	1765	1700	65
alkaliphile	408	400	8
acidophile	63	59	4

their variances, natural (chart A) and Z-scored (B), obtained on the complete set of prokaryotic proteomes used in this work (Suppl. Table S5). Variance of standardized amino acid frequencies (Z-scored) reflects relative changes in composition and renders amino acids to be comparable to each other, eliminating the heteroscedasticity. To build every adaptation predictor, we calculated all meaningful combinations of amino acids (in total $2^{19} - 1$) and selected the one with the best correlation. The total fraction of amino acids comprising the combination for each organism composed the combination frequency value. Predictors are a linear regression model that estimated the degree of correlation between combination frequency and the corresponding environmental factor (OGP, OGS, and OGT).

2.4. Principal component analysis

The PCA analysis was performed using Python packages on the 'Env' dataset with multi-dimensional environmental annotation (pH, salinity, and temperature), and individually on each environmental factor and its respective dataset. Additionally, the predictors were generated separately for each domain. To analyse how temperature, pH and salinity factors work together, 'Env' set with annotation for all 3 environmental factors was analyzed for adaptation patterns within amino acid groups. Firstly, the PCA (using python package sklrearn (Bohlin et al., 2013)) have been performed to compare adaptation patterns expressed through groups of amino acid residues depending on the physical-chemical characteristics: negatively charged 'DE', positively charged 'KR', polar 'QNST', small, weakly hydrophobic residues 'AG', aromatic residues 'FWYH', strong hydrophobic residues 'LVIMPC'. Further, all 400 dipeptides counts were calculated for each proteome. The PCA with all 3 environmental variables has been further performed for normalized homopeptide and heteropeptide frequencies, and for all dipeptides together. Dipeptide frequencies were calculated as in (Pe'er et al., 2004). The frequencies of six groups of amino acids were also analyzed in relation to the GC content and the R/Y ratio.

2.5. Proteomic amino acid depth

The amino acid depth is a parameter that reflects proper compactness and ratio between the hydrophobic core and hydrophilic surface in the native protein globule (Chakravarty and Varadarajan, 1999; Pintar et al., 2003). It is possible to calculate a proteomic average of depth, as it can be deduced purely from the amino acid compositions. We used proteome-average amino acid depth for the whole-proteome characterization of proteins in the individual organism that reflects molecular mechanisms of protein adaptation in corresponding organism. The proteomic depths values were plotted against the GC content.

2.6. Genomic signals of adaptation to pH, salinity, and temperature

To explore biases in nucleotide sequences, we compared the natural (nat) sequences (genomes with coding sequences obtained from NCBI) and non-codon-bias (ncb) sequences (synonymous codons assigned with equal probability). The correlation of general (non-position-dependent) composition elements with environmental factors have been calculated. T-test (one-tailed) has been performed for a percentage of each nucleotide against 25 % (average value if all 4 nucleotides had equal weight in a genome) to observe abundant amino acids. For each statistical test, the significant p-values ($^{+}$<0.001, ** <0.01, *<0.05) have been shown. Next, the purine and pyrimidine ratio has been calculated and compared with the general G + C % composition. The nucleotide frequencies at each position in the codons were calculated for both nat and ncb sequences. The correlation between nucleotide position in codon and environmental factor has been calculated for DNA (nat and ncb). The general A, T, G, C composition have been calculated for coding-/non-coding DNA, rRNA, and tRNA. The dinucleotide counts were also calculated and correlations of their normalized values with environmental factors were analyzed for coding-/non-coding DNA and r-/tRNAs. generated. All purine and pyrimidine combinations have been counted and plotted against GC composition.

3. Results

We explore here molecular mechanisms of adaptation emerging in bacterial genomes and proteomes in response to harsh environments, namely extremes of the pH, salinity, and temperature existing in nature. Our goal is to find major trends in amino acid and nucleotide compositions working in adaptation mechanisms of corresponding biomolecules. In some cases, prokaryotes considered here thrive under combination of several extreme environments, which is further complicated by the possible evolutionary distinction of these organism –archaea or bacteria. As a result, one may observe a pretty complex interplay between mechanisms of adaptation and corresponding to them non-trivial combinations of compositional signatures. Additionally, it is always a challenging task to understand a causality of these compositional characteristics and their signatures (Goncearenco and Berezovsky, 2014; Goncearenco et al., 2014). Indeed, the biases in amino acid compositions can be governed by the adaptation on nucleotide level and vice versa, disguising the real reasons for observed effects. Therefore, we implement here a "divide and conquer" strategy, attempting first to determine trends that chiefly reflect adaptation to certain extreme environments, then trying to understand how our observations reflect mechanisms competing and complementing each other in adaption to combinations of extreme environments, as well as distinct evolutionary history of archaea and bacteria (Goncearenco and Berezovsky, 2014; Goncearenco et al., 2014).

3.1. Proteomic signals of adaptation to extreme environments

3.1.1. Proteomic predictors of thermo-/pH-/salinity

We started from calculating the Z-scored predictors of adaptation to three extreme environmental conditions, temperature, pH, and salinity.

The predictors are Z-scored fractions of the sets of amino acids yielding the highest R (correlation coefficient) with the corresponding environmental characteristic – Optimal Growth Temperature (OGT, from now on T), Optimal pH (OGP), and Optimal Salinity (OGS). We prepared and annotated sets of organisms with available data for a certain environmental factor: Temperature dataset containing organisms with known OGT – 9225 organisms; salinity and pH datasets with 2568 and 2958 organisms, respectively. We also used a dataset of 2236 organisms, for which all three optimal environmental factors are known (See also Materials and Methods, Tables 3 and 4, and Suppl. Table S4). Fig. 1 contains pairs of correlations for the most optimal predictors obtained on the sets of individual environmental factors (T, pH, and S, left column) and on the Environmental set (Env_set) of organisms with all three known environmental characteristics (right column). The signal of temperature adaptation in form of predictors EIKPRVY and EIKPRVYL (Fig. 1A) was found consistently for T_set and Env_set datasets with strong correlation coefficients equal to 0.85 (p-value<0.15e-299) and 0.81 (p-value<1.7e-159), respectively. The difference from the original IVYWREL signature of thermal adaptation is explained by the usage of Z-scored predictors (Goncearenco et al., 2014; Ma et al., 2010), which more correctly account for contribution of amino acid residues to the predictor, considering their proteomic variances (Ma et al., 2010). Another strong correlation was revealed in salinity adaptation for predictor combination DEGHN (R-value = 0.79) for both S_set and Env_set (Fig. 1C, p-values: <1.2e-160 and < 7.3e-139 for S_set and Env_set, respectively). The weakest signal was found in case of pH adaptation with predictors (DEFGIPQ, R-value = 0.44, p-value<1.3e-28) and (DEFGILMNPRT, R = 0.44, p-value<7.8e-20) obtained on P_set and Env_set (Fig. 1B), respectively. The DEFGIP combination is common for both P_set and Env_set.

Considering specifics of amino acid trends, charts A (Fig. 1) show that predictor of thermophilic adaptation contains: (i) strong hydrophobes – isoleucine/proline/valine (IPV) and ILPV (with additional leucine) revealed by the T and Env sets of proteomes, respectively; (ii) charged residues of both signs – negatively charged glutamic acid and positively charged lysine (K) and arginine (R); and (iii) polar tyrosine (Y). Noteworthy, in addition to earlier described importance of Y for protein-protein interactions (Goncearenco et al., 2014; Berezovsky et al., 2007; Ma et al., 2010; Berezovsky, 2011), its bulky side-chain can also contribute to van der Waals interactions – the major component of the protein enthalpy (Berezovsky et al., 2000b). Predictors of adaptation to pH and salinity show distinct trends, including groups of negatively charged residues, aspartic (D) and glutamic (E) acids. The importance of these residues for adaptation to corresponding factors is further corroborated by their presence in predictors obtained for all groups of organisms: specific factors set (pH_set and S_set), Env_set, as well as for groups of archaea and bacteria considered in the above sets (Suppl. Fig. S2). Additionally, strong hydrophobes are present in predictors of pH with isoleucine (I) and leucine (L) observed in predictors for all considered groups of adaptation factors and archaea/bacteria evolutionary groups of organisms (Fig. 1B and Suppl. Fig. S1B). Glycine (G) and polar residues (glutamine (Q) and asparagine (N)) are observed in pH (Fig. 1B and Suppl. Fig. S2B) and S (Fig. 1C) predictors, respectively, with exception for archaea/bacteria groups (Suppl. Fig. S2C).

Overall, T-adaptation's trend is distinct from those of pH and salinity, corroborating earlier conclusion on the "from both ends of hydrophobicity scale" strategy of thermostabilization facilitated by the increase of strong hydrophobes and charged residues both necessary for stabilization of the hydrophobic core and hydrophilic surface (Berezovsky et al., 2007). Adaptation to changing amount of water ions (pH) and salt is apparently regulated by the variations in fractions of charged and polar residues: they are increased upon shift of pH from basic to acidic and increase of salinity, respectively. Specifically, negatively charged aspartic (D) and glutamic (E) acids work in adaptation to both environments, and histidine (H) - in salt adaptation. Additionally, there is some stabilization of the protein core by hydrophobic residues (IL) in case of pH increase. Growing fraction of glycine (G) points to its potential role in optimization of the globular structure packing and flexibility. For example, in pH adaptation it may facilitate packing of the globule provided by increase amounts of strongly hydrophobic residues. In halophilic adaptation, while the fractions of hydrophobic residues are not changed, G may contribute to flexibility of the globule making it easier for protein's charged/polar residues to interact with an environment. Looking for potential differences in mechanisms of adaptation determined by the specifics in the evolutionary history of Archaea and Bacteria, we built corresponding predictors for groups of organisms belonging to each of these branches of the Tree of Life. All predictors are practically like those obtained on T, pH, S, and Env sets. Among few interesting distinct details are: (i) observation that predictors are stronger for archaeal species in case of T and pH adaptation (EKRILVPY with R = 0.91/p-value<1.15e-65 and DEGILQ with R = 0.63/p-value = 0.0003, respectively); (ii) the same and simplified predictor of pH, DEGILQ, for both kingdoms; (iii) only negative charges (ED) and H in both predictors of salinity.

3.1.2. Principal component analysis reveals specific contributions of different groups of amino acid in adaptation to different extreme environments

Predictors of adaptation to extreme temperature, pH, and salinity described above provide a clear hint of the potential types of interactions that determine adaptation to corresponding extreme environments. Obtained predictors show that in some cases the same amino acids are apparently working in predictors to several environments. Therefore, we decided to perform Principal Component Analysis (PCA), anticipating to see specific trends archetypal to one or another type of adaptation in addition to common biases present in all of them. Additionally, compositional biases and adaptation trends related to phylogenetic differences and evolutionary history of organisms, specifically, differences between characteristics for Archaea and Bacteria could be also observed. To perform PCA analysis, which would allow us to compare trends in different types of adaptation (temperature, salinity, and pH) and evolutionary history (Archaea versus Bacteria), one should operate with a generic set of characteristics, groups of amino acids, then to relate PCA-based observations to those from the analysis of predictors. Therefore, we first perform eleven-dimension/feature PCA, considering three environmental factors (T, pH, and S), GC content, purine/pyrimidine (R/Y) ratio, as well as groups of negatively (DE) and positively (KR) charged, polar (QNST), small/weakly polar (AG), aromatic (FWYH), and hydrophobic (LVIMPC) residues as features determining dimensionality. The placement of amino acids in corresponding groups, including alanine and glycine in a separate group of small/weakly polar and histidine in the group of aromatic amino acids, is decided on the basis of our earlier studies of the molecular adaptation mechanisms (Goncearenco and Berezovsky, 2014; Berezovsky et al., 2007; Ma et al., 2010; Berezovsky, 2011; Zeldovich et al., 2007). These studies revealed specific characteristics and contributions of amino acids to distinct stabilizing interactions and mechanisms, determining and justifying their respective grouping. The eleven-feature analysis was performed on the environmental dataset (Env_set, 2336 organisms) with available data on all environmental factors – OGT/OGS/OGP. Next we turned to the nine-feature analysis, which was executed on corresponding temperature (T_set, 9225 organisms), salinity (S_set, 2568), and pH (pH_set, 2958) groups organisms with data available for corresponding environmental factor. Fig. 2 presents results of the PCA obtained for the Env_set of proteomes and shown for pairs of PC s 1–2, 2–3, and 3–4, respectively. Four charts in each PC consider archaea/bacteria groups (top chart), psychro-/meso-/thermo-/hyperthermophilic, species in thermal adaptation, non-/slight/moderate/extreme halophilic organisms in halophilic adaptation, and acido-/neutron-/alkaliphiles groups in adaptation to pH. Consideration of corresponding pairs of principal components in the eleven-factor analysis of environmental set (PCs 1–2, 2–3, and 3–4, Fig. 2) reveals adaptation to salinity manifested

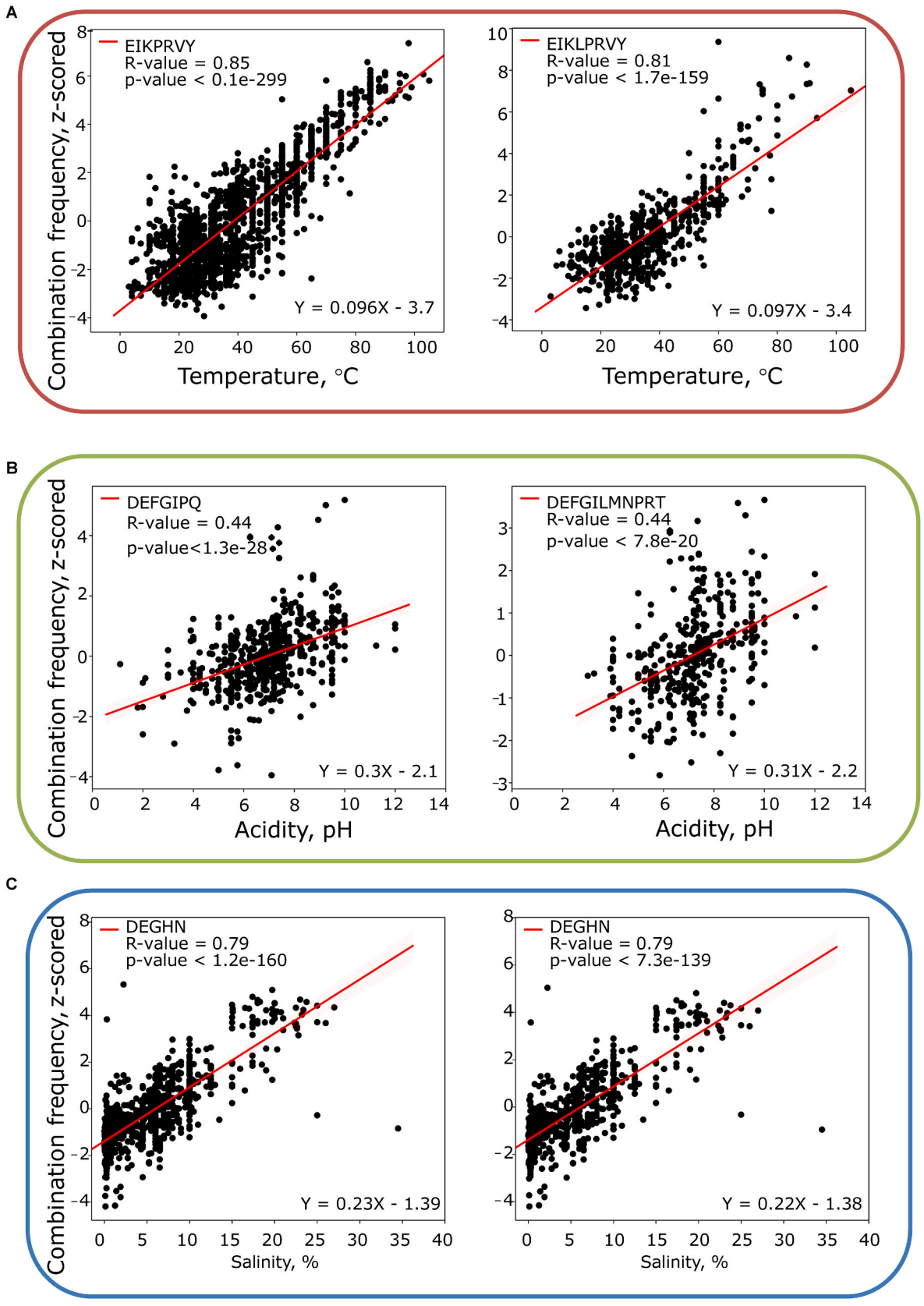


Fig. 1. Predictors for environmental factor adaptation for datasets of organisms with annotated environmental data. The datasets include both archaeal and bacterial proteomes.

(A) Temperature adaptation predictors for T_set (9225 proteomes, left chart); **(B)** pH adaptation predictors for P_set (2958, left chart); **(C)** Salinity adaptation predictors for S_set (2568, left chart). The right chart present analysis of the Env_set (2336 organisms) with all three factors, OGP, OGS, and OGT, annotated for each proteome.

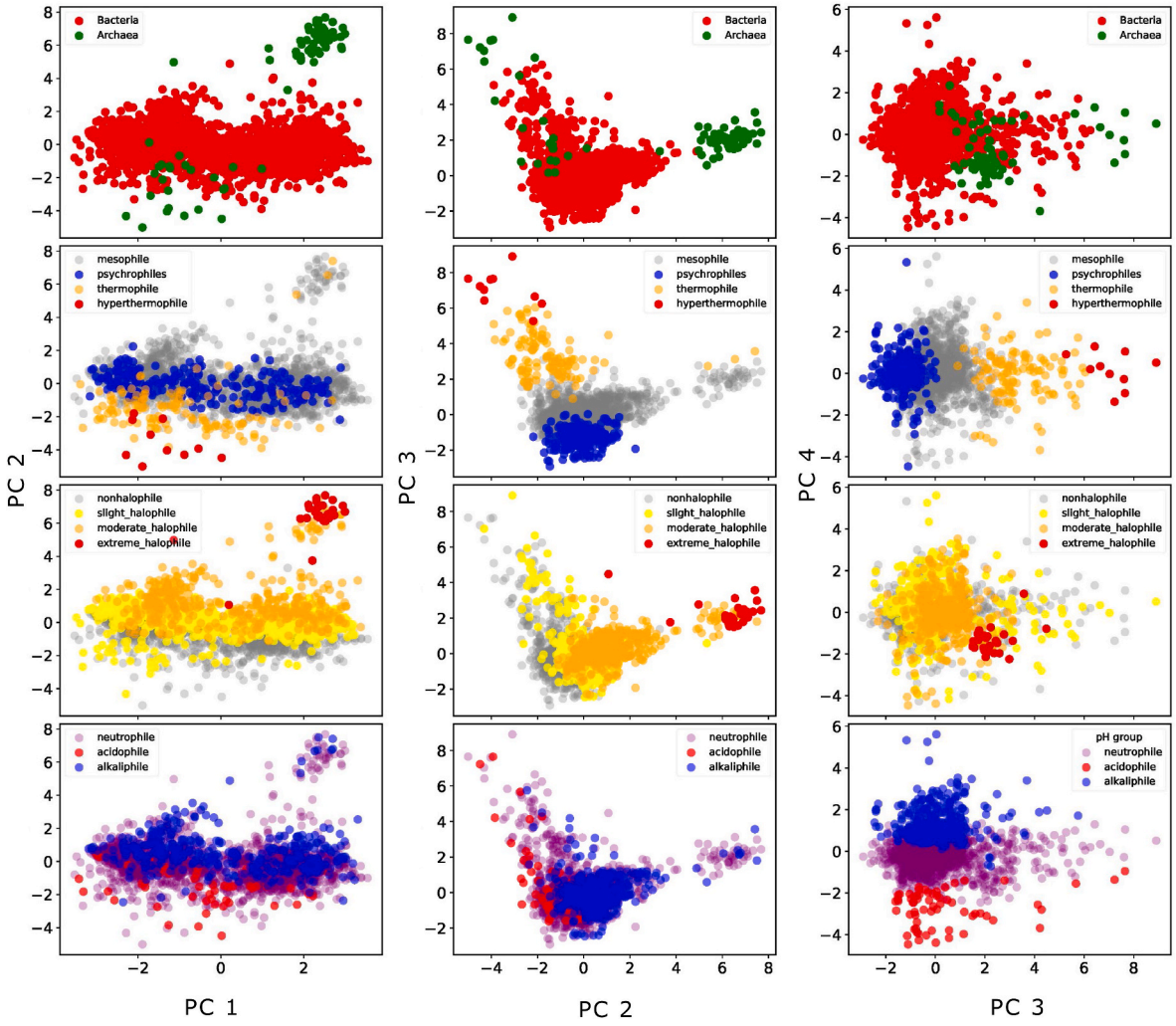


Fig. 2. Detection of adaptation patterns for each environmental factor with PCA analysis with 11 features. First row: archaea (green) and bacteria (red). Second-fourth rows: temperature, salinity, and pH, respectively. First-third columns: PC1-PC2; PC2-PC3; PC3-PC4.

in separation into groups of non/slight/moderate/extreme halophiles along the second principal component (second charts from the bottom in the left and central columns, Fig. 2). The adaptation to different salinity is also indicated in separation along the second principal component in nine-factor analysis performed on S and pH sets of organisms (second charts from the bottom in the central and right columns, Suppl. Fig. S3) and along the third principal component in the analysis of the T_set of organisms (Suppl. Fig. S3). The third principal component in the eleven-component analysis of Env_set shows temperature adaptation, yielding groups of meso-/thermo-/hyperthemophilic organisms (second charts from the top in the central and right columns, Fig. 2). Corresponding groups of organisms are also obtained along the second PC in the nine-component analysis of T (second chart from the top, left column, Suppl. Fig. S3) and along the third PC in S and pH adapted species (second chart from the top, central and right columns, Suppl. Fig. S3). These results show that adaptation to temperature is detectable in all groups of organisms with fully characterized adaptation to one of the environmental conditions – temperature (T), salinity (S), and pH. Finally, the fourth PC on the eleven-component PCA performed on the Env set shows grouping of organisms according to the pH of their living environments (bottom chart in the right column, Fig. 2), which is also indicated by the PC3 in the nine-component analysis of the pH adaptation (bottom chart in the right column, Suppl. Fig. S3).

Turning to the specific trends in amino acid compositions characteristic for adaptation to extreme environments, the Suppl. Fig. S4 presents values of PC coefficients reflecting contributions of eleven factors to separation along corresponding PCs. PC1 shows a clear separation between the groups of small/weakly polar (AG), positively (KR) and negatively (ED) charged, polar (QNST), and aromatic (FWYH) amino acids, indirectly corroborating their grouping. Noteworthy, in addition to physical-chemical characteristic of amino acids contributing to the above grouping, there is an important evolutionary twist originating from the very emergence and temporal order of codons and amino acids (Trifonov, 2000). The specifics of the evolutionary temporal order (Trifonov, 2000; Trifonov et al., 2001) place alanine and glycine in the same group, showing them important determinants of the first stage of protein evolution (Trifonov et al., 2001). The PC2 in Fig. S4 reflects adaptation to salinity (PC Coefficient, $PCC_S = 0.58$, for salinity feature), showing contribution of negatively charge residues ($PCC_{DE} = 0.57$) and opposite effect of hydrophobes ($PCC_{LVIMPC} = -0.47$). This trend is corroborated by the analysis of the S_set (Suppl. Fig. S5B), showing following PCC for salinity, negatively charged, and hydrophobic residues, respectively: 0.62, 0.6, and −0.45. Temperature adaptation characterized by PC3 in the analysis of Env_set (Suppl. Fig. S4, $PCC_T = 0.71$) is apparently supported by the hydrophobes ($PCC_{LVIMPC} = 0.35$), charged amino acids ($PCC_{DE} = 0.34$ and $PCC_{KR} = 0.26$) and by the selection against polar once ($PCC_{QNST} = -0.33$). This picture is confirmed, but slightly modified by the data obtained on the T_set: $PCC_{LVIMPC} = 0.49$, $PCC_{KR} = 0.31$, and $PCC_{QNST} = -0.42$, pointing to the role of only positively charged residues in thermostability in agreement with

predictors obtained here (Fig. 1 and Suppl. Fig. S2) and in our earlier observations (Goncearenco et al., 2014; Berezovsky et al., 2007; Ma et al., 2010). Finally, extremes of pH reflected in PC4 ($PCC_{pH} = 0.93$) are associated with increase of hydrophobes ($PCC_{LVIMPC} = 0.33$) obtained in the analysis of Env_set. This observation is further diversified according to PC2 ($PCC_{pH} = 0.45$) and PC3 ($PCC_{pH} = 0.45$) in the analysis of pH set (Suppl. Fig. S5C). The PC2 shows contribution of only negative charges ($PCC_{DE} = 0.6$), while selection against hydrophobes ($PCC_{LVIMPC} = -0.58$). The PC3, at the same time, points to the mutual contribution of hydrophobes ($PCC_{LVIMPC} = 0.55$) and charges of both signs ($PCC_{DE} = 0.39$ and $PCC_{KR} = 0.3$). Noteworthy, the separation in two groups, acidophiles and alkaliphiles, takes place in the group of Bacteria depicted in the bottom chart/right columns: (i) in Fig. 2 along the PC4 (Env_set); (ii) in Suppl. Fig. S3 along the PC3 (pH set). Thus, the difference obtained for PC2 and PC3 in the analysis of pH_set may hint to distinct ways of adaptation emerged in acidophiles and alkaliphiles. The former is apparently relying only on negative charges ($PCC_{DE} = 0.6$ and $PCC_{LVIMPC} = -0.58$), while the latter - on the contribution from both hydrophobes and charges in the latter ($PCC_{LVIMPC} = 0.55$, $PCC_{DE} = 0.39$ and $PCC_{KR} = 0.3$; Suppl. Fig. S5C). Archaea also shows splitting into two groups adapted to different extremes of temperature and salinity: it is easy to see how two separated green groups in the first row in Fig. 2 and Suppl. Fig. S3 correspond to group of hyperthermophiles (red dots, second raw from the top) and extreme halophiles (red dots, third raw from the top). Interestingly, in this case two groups of Archaea show distinction in terms of adaptation to different environmental factors, temperature, and salinity. In case of Bacteria considered above, we found two groups adapted to extreme of the same factor pH: acidophiles and alkaliphiles (bottom right charts in Fig. 2 and Suppl. Fig. S3).

We also performed PCA on dipeptides of analyzed proteomes, using the same eleven-and nine-feature approach and applying it to only homo-/heteropepetides and for combination thereof, using the Env_set and T_set as in input data. Fig. 3 shows PC1-2 projection of the nine-feature PCA obtained on a T_set for homo-/heteropeptides (left/right charts, respectively) with archaeal and bacterial organisms shown by green and red dots, respectively. We start here from the T_set: (i) while it contains only OGT data for every organism and only partial data for OGS and OGP, it allows to have a reliable separation in the largest group with annotated environmental traits, and this way (ii) it provides a robust layout for mapping groups of organisms adapted to other environmental extremes, unravelling hidden intricate relationships between them. For example, like PCA on amino acid compositions (Fig. 2 and Suppl. Fig. S3), we observed overlapping patterns of archaea and bacteria, but the former provide even more detailed picture of separation according to adaptation to distinct extreme environments. Specifically, grouping of archaeal proteomes on the PCA1-2 for homopeptides (Fig. 3, left chart) reveals location of organisms with adaptation to environmental salinity (Group 1 includes 24 and 20 moderate and extreme halophiles, respectively). This analysis also shows groups of mesophiles (43 organisms in Group 2), as well as location of species with adaptation to hight temperature (49 hyperthermophiles and 11 thermophiles in Group 3; see also Suppl. Table S7). The analysis of heteropeptides (Fig. 3, right chart) points to a set of species adapted to increased salinity (Group 4: 24 and 20 moderate and extreme halophiles) and to high environmental temperature (Group 5: 42 and 38 hyperthermophiles and thermophiles; see also Suppl. Table S7). Fig. 4 presents more details of this analysis, showing data obtained separately for homo-/heteropeptides and their combinations and marked according to each individual environmental factor. Obviously, overall pictures obtained for all (chart A) and heteropeptides (chart C) are very similar, as the latter dominates the set of dipeptides. More importantly, consideration of individual factors – T, S, and pH in second, third, and fourth columns respectively – clearly shows the relationship between the organisms adapted to environmental extremes and those thriving in corresponding milder or even neutral environments. The eleven-feature PCA on the Env_set confirms above grouping (Suppl. Fig. S6), providing a chance to analyse combined adaptation to more than one extreme environment. Though Env_set is much smaller and the number of "multi-extremophiles" is rather limited, it is still possible to use them for exploring combined mechanisms of adaptation in the future work. Noteworthy, dipeptide analysis does not discriminate groups between organisms living under different pH, while PCA on groups of amino acids showed a clear separation between acidophiles and alkaliphiles (Fig. 2 and Suppl. Fig. S3).

3.2. Linking proteomic and genomic characteristics of adaptation

3.2.1. The base paring/stacking balance determined by nucleotide/dinucleotide compositions

The study of molecular mechanisms of adaptation in archaea and bacteria should always include consideration of the relationship between two major types of biomolecules, nucleic acids, and proteins (Goncearenco and Berezovsky, 2014; Goncearenco et al., 2014). Based on previously found trade-off in the genome-proteome compositions and using a complete collection of 9306 organisms (Suppl. Table S5) considered in the work, we illustrate the concerted work of mechanisms of adaptation in nucleic acids and proteins using the GC content as a reference characteristic. Fig. 5 (top chart) shows a connection between the GC content and the purine/pyrimidine ratio, in which decrease of

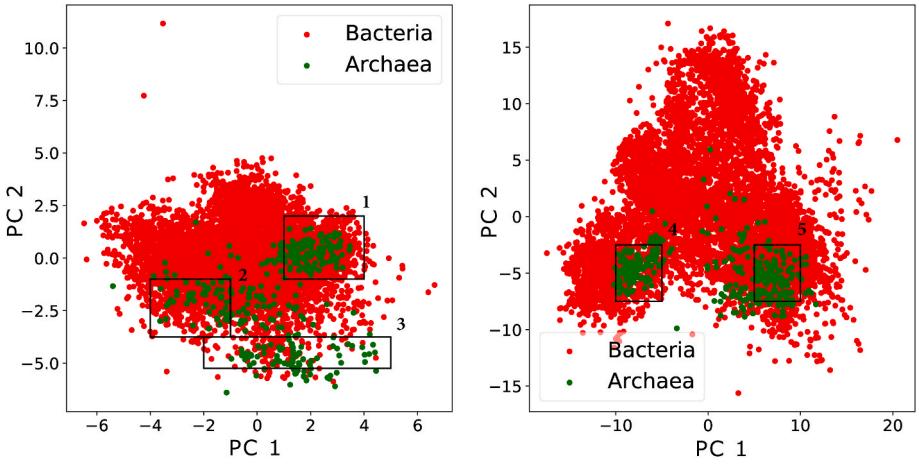


Fig. 3. Distinct groups of Archaea revealed by PCA dipeptides. Left chart – homopeptides; right – heteropeptides. T_set was used in calculations. Groups 1, 2, and 3 are separated along PC2, indicating the presence of halophiles, non-extremophiles, and thermophiles, respectively. Groups 4 and 5 along PC1 show halophiles and thermophiles, respectively. Detailed group information can be found in Supplementary Table S7.

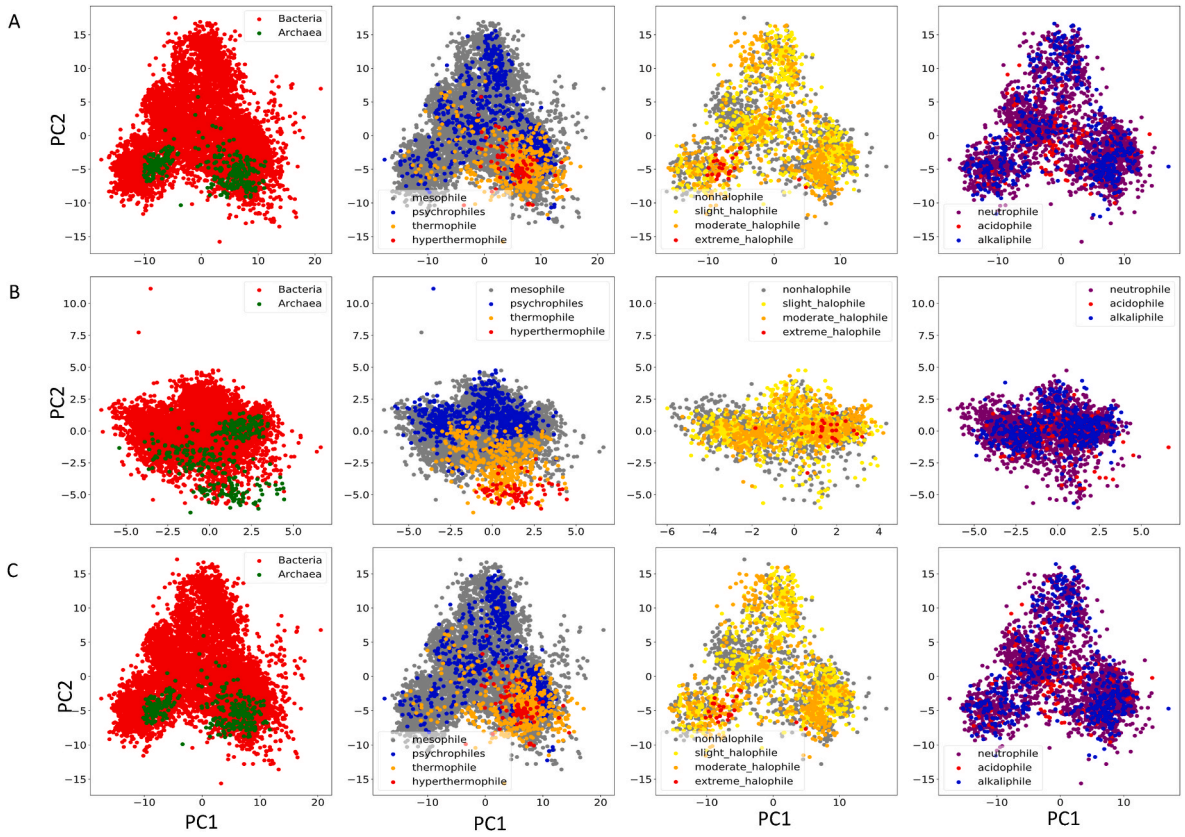


Fig. 4. PCA analysis of dipeptides (T_set). (A) all dipeptides; **(B)** homopeptides; **(C)** heteropeptides. First column: domain of Life – archaea and bacteria, second - temperature, third – salinity, and fourth - pH.

the genomic GC content is accompanied by the increase of the purine (A + G) load in the sense strand of the DNA. The purine-purine base stacking, thus, can be a very important, if not a dominating factor of DNA stability in genomes with low GC content, while the base pairing is apparently the major contributor to the DNA stability throughout most of the GC range. This observation agrees with our earlier conclusions made on much smaller dataset of 1364 organisms (Goncearenco and Berezovsky, 2014). We also used statistics of dipeptides to further explore the balance between stacking and paring in coding DNA, rRNA, and tRNA, depending on their nucleotide compositions. It appears that stacking can contribute to stability of both strands in double-stranded coding DNA, as the dsDNA is enriched with purine-purine and pyrimidine-pyrimidine dinucleotides in the sense and anti-sense strands at low GC (Suppl. Fig. S7A). In one-stranded rRNA that rarely makes stems, the pyrimidine-pyrimidine dinucleotides do not show specific trend, leaving it to only purine-purine ones to work for its stability (Suppl. Fig. S7B). The tRNA also shows a specificity in work of dinucleotides: purine-purine ones contribute to stability of the chain at low GC, but pyrimidine-pyrimidine apparently prevent potential overstabilization of double-stranded stems formed in its structures (Suppl. Fig. S7C).

3.2.2. *Persistence of the protein foldability and stability*

Stability of proteins (Shakhnovich, 2006) requires adherence to the optimal ratio between the interior and exterior of the protein globule (Bresler and Talmud, 1944a, 1944b). We resorted here to composition-based characteristic that describes this ratio, amino acid depth (Chakravarty and Varadarajan, 1999; Pintar et al., 2003): a distance between the protein's atom and the nearest bulky water molecules surrounding the protein. Since depth reflects a proper compactness and ratio between the hydrophobic core and hydrophilic surface in the native protein globule, we used the genome-averaged amino acid depths as a compositional criterion of the proteome-wide protein foldability and stability. It appeared that values of the averaged proteomic depth are confined within a narrow interval from 0.96 to 1.02 for all 9306 proteomes (Fig. 5, bottom chart), and persistence of the depth value is a characteristic feature of both archaeal and bacterial proteomes (Suppl. Fig. S8; right column, bottom, and middle charts). Noteworthy, there as apparently a separate group of archaea with high (from 60 to 70 %) GC content, showing notably lower average proteomic depth value. If presence of this group is not a trivial result of relatively small number of archaeal species (it should be further explored), the latter might be indicative of stronger packed globules typical for proteomes of ancient hyperthermophilic archaea following the structure-based packing-driven strategy of thermostabilization proposed in our earlier work (Berezovsky and Shakhnovich, 2005). The current support for this conclusion is provided by observation of two subgroups of archaea adapted to high temperature and salinity observed in PCA analysis (Figs. 2–4 and Suppl. Figs. S3 and S6). This hypothesis also agrees with our earlier observation on the specifics of genome-proteome trade-off that results in mutual pressure between nucleotide and amino acid compositions in some cases of their extremes (Goncearenco and Berezovsky, 2014). All the above motivated us to consider further details of nucleotide compositions, their potential biases determined by the adaptation to the "trinity of extremes", and the causality in their emergence and mutual evolution with amino acid compositions in response to environmental pressure.

3.3. *Genomic signals of adaptation to extreme environments*

3.3.1. *Nucleotide composition biases and adaptation to thermo-/pH-/salinity extremes*

Considering genomic nucleotide compositions from the perspective of adaptation to different extreme environments, one can see

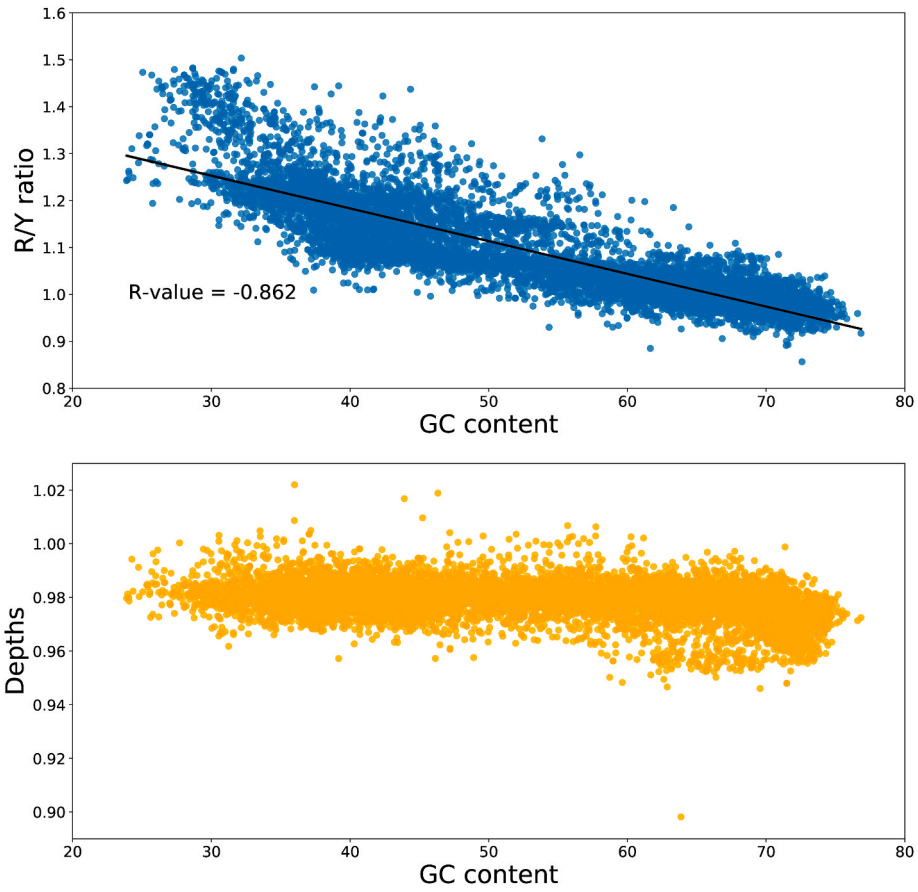


Fig. 5. The relationship between the whole genome-proteome characteristics of nucleotide and amino acid compositions. (A) GC content and purine/pyrimidine (R/Y) ratio; **(B)** GC content and depth. p-value < e−300 (in both cases).

distinctions manifested in GC content and purine load (A + G), pointing to differences in mechanisms of adaptation. Table 3 shows that while the purine load (A + G) in coding sequences is correlated with OGT and anti-correlated with OGP and OGS, the GC content apparently can work in adaptation to salinity and pH, but temperature. Suppl. Table S8 provides further details of the contribution of individual nucleotides, showing that A and T are anticorrelated with OGP and OGT, while G and C do correlate. The correlation is stronger in case of halophilic adaptation. Interestingly, there is a specific picture of contributions from individual nucleotides in non-coding DNA sequences: (i) T and G contents are correlated with OGP and OGS, but A and C – anti-correlated, the effects are stronger in salinity case; (ii) T/C are anticorrelated/correlated with OGT. In both rRNA and tRNA, there is a clear preference for G and C upon increase of OGT, while A and U are selected out. Thus, there is an apparent directionality in the role for A and G nucleotides: (i) the former works for thermophilic adaptation, but selected out in case of pH and salinity adaptation of double stranded DNA (stronger in coding one) and in thermophilic adaptation of rRNA and tRNA; it also weakly contributes to adaptation to pH and salinity, mostly in case of tRNA; (ii) the latter is contribution to halophilic and pH adaptation of coding/noncoding double-stranded DNA and thermophilic adaptation of rRNA/tRNA (Suppl. Table 8). It should be noted that all the biases described above were obtained for archaeal proteomes, while bacterial ones do not show any significant trends (Suppl. Table 9).

3.3.2. *Codon-position dependent nucleotide composition biases and adaptation to thermo-/pH-/salinity*

Further details on the relation between the nucleotide and amino

Table 3
Correlation between Nucleic Acid Stability Parameters and Environmental Factors in Archaea.
Correlations between nucleic acid stability parameters and environmental factors in Archaea and Bacteria are shown. It specifically explores the interplay between the stabilization achieved by the purine-to-pyrimidine ratio (R/Y ratio) and the stabilization achieved by the G + C composition (G + C %) in DNA and RNA. Additionally, it investigates the correlation of these stability parameters with environmental factors such as temperature, pH, and salinity.

	Natural DNA sequence				Non-codon-biased DNA sequence			
		R OGT	R OGP	R OGS		R OGT	R OGP	R OGS
GC %	53.09	−0.26^{+}	**0.42**$^{+}$	**0.66**$^{+}$	50.88^{+}	0.04	**0.29****	**0.53**$^{+}$
R/Y	1.17^{+}	**0.58**$^{+}$	*−0.42*$^{+}$	*−0.59*$^{+}$	1.2^{+}	0.09	0.25*	0.11
	rRNA				tRNA			
		R OGT	R OGP	R OGS		R OGT	R OGP	R OGS
GC %	**57.98**	**0.84**$^{+}$	−0.15	−0.09	**63.26**	**0.91**$^{+}$	−0.3*	−0.29*
R/Y	1.08^{+}	0.29^{+}	0.04*	0.06*	1.15^{+}	0.32^{+}	0.1^{+}	0.17^{+}

Note: Significance of correlation given as + *p-value* < 0.001, ** *p-value* < 0.01, * *p-value* ≤ 0.05, with no asterisk *p-value* > 0.05.

acid trends characteristic for environmental extremes are encrypted in nucleotide biases of individual codon positions in coding DNA sequences. Like in the case of nucleotide compositions, the most pronounced biases (presented in Table 4) were observed only for archaeal proteomes. We use here correlation coefficients between the nucleotide frequencies on a codon position (natural, N_{nat}, and the ratio of natural over non-codon biased frequencies, N_{nat}/N_{ncb}) and the corresponding environmental factor as indicators of the importance of nucleotide for tuning of nucleotide composition and local patterns. A correlation of the non-codon biased frequencies (N_{ncb}) on a position with the corresponding environmental factor serves as an indicator of the potential contribution of nucleotide on this position to the tuning of amino acid composition.

The first codon position is characterized by the preference for adenine important for adjustment of nucleotide composition (Table 4) and competition between preference for cytosine in nucleotide composition ($R_{C\_nat} = 0.42$, p-value<0.001), but selection against G in the first position of codons defining the amino acid ($R_{G\_ncb} = -0.37$, p-value<0.001), respectively. Signals of halophilic and pH adaptation are very similar on the first codon position: there is an indicator of importance of G selection against C for tuning of the amino acid composition, while C contributes to tuning of the nucleotide composition (Table 4). In both adaptations there is a selection against A complemented by selection against T in case of pH adaptation in relation to nucleotide composition. Also, both halophilic and pH adaptations show anti-correlation of A with both environmental factors with $R_{A\_ncb} = -0.48$ (p-value<0.01) and $R_{A\_ncb} = -0.47$ (p-value<0.05), respectively. The second codon position shows a demand for guanine and thymine for nucleotide composition, while selection against cytosine in thermophilic adaptation. At the same time C may contribute to adjustment of amino acid composition required for thermostability, while T is selected out (Table 4). Trends on the second codon position obtained for halophilic and pH adaptations are very similar with only additional selection against G for nucleotide compositions in case of halophilic adaptation. Otherwise, there is a clear demand for C necessary for nucleotide composition, but, at the same time, selection against it in adjustment of amino acid frequencies. The T shows reverse picture: importance for tuning of the amino acid composition, while bias against it in the nucleotide composition. The third position is least demanding, showing different trends in all adaptations: (i) preference for A and G and, respectively, selection against C and T for amino acid composition in thermophilic adaption; (ii) preference for C for nucleotide composition in pH adaptation; (iii) selection against A in changing amino acid composition in halophilic adaptation (Table 4). As we already mentioned above, codon-position dependent nucleotide signals associated with adaptation of bacterial genomes/proteomes are only detectable for the case of thermophilic adaptation, where they show like archaea, but much weaker signals. Specifically, first position reveals very weak preference for A necessary for nucleotide composition tuning ($R_{A\_nat/ncb} = 0.21$). Some increase of T ($R_{T\_nat/ncb} = 0.37$ and $R_{T\_nat} = 0.36$) paired with selection against C ($R_{C\_nat/ncb} = -0.3$ contributing to changes of nucleotide composition was detected on second codon position. At the same time demands on the tuning of amino acid composition are opposite: weak preference for C ($R_{C\_ncb} = 0.36$) and selection against T ($R_{T\_ncb} = -0.31$). Third position shows a strong correlation with A ($R_{A\_ncb} = 0.71$) and weak with G ($R_{G\_ncb} = 0.22$) along with selection against T ($R_{T\_ncb} = -0.49$) and C ($R_{C\_ncb} = -0.49$) reflecting potential adjustment of the amino acid composition. In all correlations above obtained for thermal adaptation in Bacteria, the p-value<0.001. Overall, generalized triplet in archaea determining the amino acid (protein adaptation) with a codon-position signature of thermophilic adaptation looks like $[C]_1\,[C]_2\,[A, G]_3$, while for both halophilic and pH adaptations - $[G]_1\,[T]_2\,[x]_3$, where x designates no preference for any nucleotide. The generalized triplets for adaptation of nucleic acids are: $[G]_1\,[T]_2\,[C, T]_3$ for thermophilic adaptation and $[C]_1\,[C]_2\,[x]_3$ and $[C]_1\,[C]_2\,[C]_3$ for adaptation to salinity and pH, respectively. Corresponding generalized triplet for thermophilic adaptation in bacteria read $[x]_1\,[C]_2\,[A, G]_3$ and $[A]_1\,[T]_2\,[x]_3$ for amino acid and nucleotide compositions, respectively.

Table 4
Position-dependent correlation of nucleotides in DNA of each domain with temperature, pH, and salinity.

Temperature					
Codon Position 1	*R-value*	Codon Position 2	*R-value*	Codon Position 3	*R-value*
A_nat/ncb	0.776^{+}	C_ncb	0.674^{+}	A_ncb	0.797^{+}
		T_ nat	0.674^{+}	G_ncb	0.521^{+}
		G nat/ncb	0.596^{+}	C_ncb	-0.71^{+}
		T nat/ncb	0.59^{+}	T_ncb	-0.71^{+}
		T_ncb	-0.54^{+}		
		C_ nat	-0.6^{+}		
		C nat/ncb	-0.62^{+}		
pH					
Codon Position 1	*R-value*	Codon Position 2	*R-value*	Codon Position 3	*R-value*
G_ nat	0.578^{+}	C_nat/ncb	0.53^{+}	C_ nat	0.529^{+}
G_ncb	0.578^{+}	T_ncb	0.499**	C_nat/ncb	0.519**
C_nat/ncb	0.548^{+}	C_ nat	0.492**		
C_ nat	0.496**	T_nat/ncb	-0.54^{+}		
T_nat/ncb	−0.51**	C_ncb	-0.57^{+}		
A_nat/ncb	−0.52**	T_ nat	-0.57^{+}		
A_ nat	-0.54^{+}				
T_ nat	-0.58^{+}				
C_ncb	-0.67^{+}				
Salinity					
Codon Position 1	*R-value*	Codon Position 2	*R-value*	Codon Position 3	*R-value*
G_ nat	0.595**	C_nat/ncb	0.676^{+}	A_ncb	−0.53**
G_ncb	0.595**	C_ nat	0.622**		
C_nat/ncb	0.522*	T_ncb	0.577**		
A_ nat	−0.57**	G_nat/ncb	−0.54**		
C_ncb	-0.66^{+}	T_nat/ncb	−0.62**		
A_nat/ncb	-0.67^{+}	T_ nat	-0.71^{+}		
		C_ncb	-0.71^{+}		

Note: Significance of correlation given as $+$ *p-value* < 0.001, ** *p-value* < 0.01, * *p-value* ≤ 0.05, with no asterisk *p-value* > 0.05.
Correlation values lower than 0.49 were removed.

4. Discussion

We explore here molecular mechanisms of adaptation emerging in prokaryotic genomes and proteomes in response to harsh environments, aiming to establish major nucleotide and amino acid compositional determinants of adaptation to extremes of the pH, salinity, and temperature. To this end, we have performed a series of computational experiments from deriving predictors of protein adaptation to optimal growth pH (OGP), salinity (OGS), and temperature (OGT), to the PCA analysis of the contribution of key groups of amino acids and dipeptides in adaptation to distinct extreme environments. It was followed by the analysis of corresponding genomic sequences aiming to determine trends in the major molecules storing the genetic information and facilitating its translation to protein sequences/structures (coding DNA, tRNA, and rRNA), and to establish a causality between related biases in the nucleotide and amino acid compositions. We also analyzed the effect of the evolutionary history on mechanisms of adaptation, exploring its fingerprints in compositional trends and sequence/structure determinants in genomes/proteomes of organisms belonging to two major branches of the Tree of Life – Archaea and Bacteria. We used here a dataset of genomes/proteomes (total 9306 organisms), for which we were able to annotate environmental conditions they are thriving in. Three sets of organisms contain exhaustively annotated data on optimal growth pH (pH_set, 2958 organisms), salinity (S_set, 2568), and

temperature (T_set, 9225), as well as so-called *environmental set* (Env_set, 2336) possesses an annotation of all three optimal growth conditions for each organism.

The difference between predictors of optimal temperature, pH, and salinity adaptation trends is obvious (Fig. 1), suggesting existence of molecular mechanisms specific to each environment that act in the framework of the positive and negative design components providing thermodynamics stability of the native protein structure (Berezovsky et al., 2007). Remarkably, while predictors of adaptation to pH, salinity, and temperature and PCA show that adaptation mechanisms strictly follow the positive-negative design paradigm, the mechanisms themselves yield interesting specific distinctions. For example, in agreement with our earlier works, we observed the work of "from both ends of hydrophobicity scale" strategy of thermostability provided by the IVY-PREK signature consisting of two major groups of amino acids - strong hydrophobes (IVP) and charges (EKR). The fractions of these amino acids are increasing upon the OGT increase at the expense of polar residues. The positively charged amino acids, lysine, and arginine, play a distinctive role in negative design increasing energies of misfolded conformations because of repulsion between their positive charges. The pH and salinity predictors and PCA data (see also below for details) reveal the key role of negatively charged residues, glutamic and aspartic acids, in negative design working for adaptation to high pH and salinity. The repulsion between negative charges apparently works for increasing the energy of misfolded conformation and prevent destructive role of hydroxy ions upon pH increase. This conclusion is indirectly supported by the observation on cation exclusion in case of halophilic adaptation, which facilitates repulsion between negative charges in misfolded conformations (Ortega et al., 2015). In case of pH adaptation, negative design is supported by the selection against histidine in addition to lysine/arginine, which should exclude any stabilizing interactions between opposite charges in non-native conformations. At the same time, there is some increase of His in predictors of halophilic adaptation, pointing to its role in positive design, the same role that Arg and Lys play in positive design working in thermal stabilization (Goncearenco et al., 2014; Berezovsky et al., 2007; Ma et al., 2010). Additionally, despite generally weak signal observed for pH-trends there is an overall indication of the potential for better compactization of more flexible globule (increase of G fraction) upon increase of pH/salinity, which is further facilitated by more massive interactions with environments provided by the charged (DE for pH and complemented by H in halophilic adaptation) and polar (Q inpH and N in halophilic adaptation) amino acids. Presence of strong hydrophobes (IL) in the pH predictor hints that flexibility originated by glycine residues can lead to more efficient packing in pH adaptation, while structures can become more flexible and prone to solvation in S adaptation. Overall, the major distinctions observed here in relation to positive-negative design paradigm is the key role of positively charged amino acids in negative design working in thermostability, while negatively charged amino acids work in negative design in pH and halophilic adaptations. Remarkably, a certain similarity in the nature of pH and salinity extreme environments, the key role of ions and charges, leads to a similar solution for negative design via using negatively charged residues. It distinguishes pH and halophilic adaptations form the thermal adjustment, in which positively charged residues, Arg and Lys, are the key players in the negative design. The distinction between the thermal and halophilic adaptations is further complemented by usage of His in the latter for positive design, contrary to Arg and Lys in the former (Goncearenco et al., 2014; Berezovsky et al., 2007; Ma et al., 2010). The major role of polar amino acids in pH and halophilic adaptation is in the interactions with environment critical for these adaptations, while in the thermal adaptation these interactions are not beneficial for stability, hence reduced (Goncearenco et al., 2014; Berezovsky et al., 2007; Ma et al., 2010).

Analysing predictors of environmental factors, we observe major trends in amino acid compositions characteristic for adaptation to corresponding environment. Then, using Principal Component Analysis to go into details of these trends, delineating major contributors to certain environments and potential synergism in their work in case of combined extreme conditions. Contributions to PC components in eleven-feature PCA (groups of amino acids as below, three environmental factors, GC content, and R/Y ratio) discriminate between groups of hydrophobic (LVIMPC), small/weakly polar (AG), polar (QNST), aromatic (FHWY), and positively (KR) and negatively (DE) charged residues. We found that first principal component (PC1) selects out the AG group, showing that while it contributes negatively, similar to GC content (trivial result of the GC-rich codons). The opposite, positive contribution, is provided by several major groups, such as (KR), (QNST), and (FWYH). This observation apparently points to the relevance of AG group to the origin of Life manifested in two alphabets Alanine and Glycine, which determined origin of the genetic code and encoded amino acids (Trifonov, 2000; Trifonov et al., 2001). Second-to-fourth PCs reflect an adaptation and corresponding compositional trends to salinity, temperature, and pH. While these results mostly agree with those of corresponding predictors, more detailed PCA should be performed on the groups of annotated individual factors. Comparison of the eleven-feature/component PCA analysis of Env_set with nine-feature ones for T, S, and pH sets, indeed, showed that analysis for latter provides more specific details. For example, nine-component PCA on the T set (Suppl. Fig. S5A) shows the role of hydrophobes and positive charges, while negative trend for polars in thermal adaptation in agreement with predictors of thermostability obtained in this work (Fig. 1 and Suppl. Fig. S2) and earlier (Goncearenco et al., 2014; Berezovsky et al., 2007; Ma et al., 2010; Zeldovich et al., 2007). It also shows absence of the contribution form negative charges ($PCC_{DE} = 0.14$, Suppl. Fig. S5A) obtained, at the same time, in the eleven-component PCA ($PCC_{DE} = 0.34$, Suppl. Fig. S4). The nine-feature analysis for pH and salinity (Suppl. Figs. S5B and C) reveals, in agreement with conclusions on predictors, similarity between major contributors, negatively charged residues (DE), in both adaptations. It is also observed that there is a selection against strong hydrophobes (LVIMPC), allowing more flexible globule with exposed parts involved in interactions with solvent and counter-ions. Thus, the PCA data corroborates the picture obtained in predictors of adaptation, providing further details, and hinting on potential solution for adaptation for combined environmental extremes (Goncearenco and Berezovsky, 2014; Ma et al., 2010).

The PCA analysis of dipeptides was performed on the largest T_set (9225 organisms) and Env_set of organisms (2336). The nine-feature PCA performed on the T_set (Figs. 3 and 4 and Suppl. Table S3) revealed a wide and high-density layout formed by organisms with fully annotated OGT data. It allowed to see grouping of archaeal organisms with different degree of thermal adaptation in relation to grouping of the same organisms according to their adaptation to extreme pH and salinity (Groups 1–5 in Fig. 3). Remarkably, we saw similar separation of archaeal proteomes into three groups in the nine-component PCA analysis of amino acid groups (Suppl. Fig. S3): we detected group 1 (PC2>4) with domination of thermophilic proteomes – 28 thermophilic and 71 hyperthermophilic; group 2 (0<PC2>2; PC3<2) – with 79 mesophilic proteomes; group 3 (PC3>4) – with 27 halophiles and 24 extreme halophiles. These groups should be further studied by considering individual environmental factors, revealing corresponding grouping in the analysis of homo-/heteropepetides (Fig. 4B and C) and their combinations (Fig. 4A). Despite smaller number of organisms (2236) in the Env_set, the eleven-feature PCA confirms above grouping, since all organisms in this set are characterised by complete data on three environmental factors, pH, S, and T, is available (Suppl. Fig. S4). It allows even deeper study of the relationship between adaptation to combination of extreme environments. Noticeably, adaptation to T and S is clearly detected for groups of respective organisms, while pH environment is not reflected in the data. It agrees with rather weaker signals obtained for predictors of pH adaptation, while PCA performed on groups of amino acids was sensitive enough to detect separate groups of acido- and alkaliphiles (Fig. 2 and Suppl. Fig. S3). Further analysis of

the contributions of individual homo- and heteropeptides and relationship between them is also of interest, but it should be a topic of a separate study. While, in general, adaptation to extreme salinity and pH are apparently driven by the use of charged residues (mostly negatively charged residues D and E), there is an indication of the potential role of hydrophobes in alkaliphilic adaptation: it is reflected in the $PCC_{LVIMPC} = 0.33$ of the PC4 in Suppl. Fig. S4 and $PCC_{LVIMPC} = 0.55$ of the PC3 in Suppl. Fig. S5 (see also bottom chart in the right columns of Fig. 2 and Suppl. Fig. S2 for illustration). It is interesting to further explore, therefore, whether difference between PC3 and PC4 in in the analysis of pH set reflects two ways of adaptation: only using negative charges for acidophiles and both hydrophobes and negative/positive charges for alkaliphiles.

Next, we investigate an interplay between the biases and trends in nucleic acid and protein compositions, observing the trade-off between them that reflects different adaptation molecular mechanisms acting on corresponding biomolecules. The question was how stability of nucleic acids is being adapted to environmental pressure and how this adaptation takes place in relation to adjustment of amino acid compositions working in protein adaptation. We started from the whole-genome analysis of two basic characteristics of nucleotide composition, GC content and purine/pyrimidine (R/Y) ratio, which are used as compositional determinants of nucleic acids' stability (Goncearenco and Berezovsky, 2014; Goncearenco et al., 2014). Specifically, the GC pairing provides three hydrogen bonds interactions, which are stronger than two hydrogen bonds in the AT pairing (Marmur and Doty, 1962; Saenger, 1984). At the same time, the purine-purine (RpR) stacking (for all possible dinucleotide combinations of A and G) has lower energy than stacking of other dinucleotides (Saenger, 1984; Friedman and Honig, 1995). We revealed that in genomes with low GC content the R/Y ratio is increased, and there is an excess of purine-purine (RpR) dinucleotides in both strands of the double-stranded DNA (Suppl. Fig. S7). This dinucleotide bias is directly related to the contribution of purine-purine stacking to stability, pointing to a potential switch from the base paring to base stacking as the dominant mechanism of DNA stability in genomes with low GC content. The above relationship between the GC content and purine load (accounted for in form of R/Y ratio here) shows an intricate balance between the work of two mechanisms that secure stability of the double-stranded DNA (Goncearenco and Berezovsky, 2014; Zeldovich et al., 2007; Friedman and Honig, 1995; Yakovchuk et al., 2006): base stacking (Friedman and Honig, 1995; Yakovchuk et al., 2006) provided by the purine load and base paring (Marmur and Doty, 1962; Yakovchuk et al., 2006) determined by the GC content.

To decipher the genome-proteome connection between trends in the nucleotide and amino acid compositions we considered a relationship between the GC content and the amino acid depth. The former is a genomic characteristic and the latter – a composition-based characteristic of the protein foldability and stability, which can be used as a whole-proteome average parameter. It appeared that depth as a characteristic of protein stability does not depend on the nucleotide composition (Fig. 5, bottom chart), being strongly governed by the requirement on the optimal ratio between the interior and exterior of the protein globule (Bresler and Talmud, 1944a, 1944b). The conserved values of depth are also obtained for both archaeal and bacterial proteins, though there is some 30 % of halophilic organisms in the small group of archaeal proteomes, which are bottom-right outliers in the Depth-GC content dependence in Supp. Fig. S8C (right chart). All together above picture raised a question how the codon usage is working in keeping the compromise between the genomic and proteomic compositions, which we answered considering the nucleotide compositions and their codon position-dependent components in relation to amino acid frequencies that they may encode.

Nucleotide compositions clearly show the work of the GC content in halophilic and pH adaptation with stronger signal in the former in both coding and non-coding DNA. The purine load contributes to thermophilic adaptation in coding DNA (Table 3). At the same time GC content plays a key role in thermal stabilization of both rRNA and tRNA, apparently pointing to the importance of pairing in small stems and other, perhaps transient, double-stranded sections temporary formed in these molecules. Remarkably, A and G dominantly contribute in adaptation to different environments, targeting two types of molecules DNA and RNA: A helps to adapt to temperature in coding DNA, and weakly to adjust the RNA stability to extremes of pH and salinity (mostly tRNA); the latter works in halophilic and pH adaptation in DNA, but in thermophilic adaptation in RNA. The quantity of G is changing the strongest, reflecting that it chiefly determines changes of both GC content and purine load, and it is further facilitated by the decreased amount of A in rRNA and tRNA (Table 3 and Suppl. Table 8). Given our understanding of the role of GC content and purine load as determinants of two mechanisms of DNA/RNA stability, base paring and base stacking, the compositional changes correlated with distinct extreme environments reveal a fine balance in work of these two mechanisms in stabilization of nucleic acids. Indeed, the role of GC content is very clear in thermal stabilization of rRNA/tRNA (Marmur and Doty, 1962), pointing to the importance of base pairing in formation of short stems and double-stranded segments in these relatively small molecules. At the same time, the base stacking provided by the purine load is apparently present in thermostabilization of double-stranded DNA, while playing a relatively small role in halophilic and pH adaptation of rRNA/tRNA (Supp. Table S8).

Further, the codon-position dependent nucleotide biases shed a light on potential connections between the nucleotide and amino acid compositions and their mutual adjustment in response to environmental extremes. The differences between nucleotide biases in the first and second codon position between thermophilic adaptation and the pair halophilic-pH adaptations points to a clear discrimination between their adaptation mechanisms that should be considered in future engineering efforts. The third position is not surprisingly least demanding in relation to the tuning of amino acid composition, while it can be important for adjusting the nucleotide composition to deal with other environmental challenges. For example, the importance of GC content in preventing damaging effects of oxidation in aerobicity was shown elsewhere (Goncearenco et al., 2014). In general, codon-dependent trends, expressed via generalized triplets archetypal for adaptation of proteins and nucleic acids, confirmed the similarity of adaptation mechanisms in cases of pH and salinity. The generalized triplets show, however, distinct ways of stabilization in case of thermophilic adaptation. The trends and their compositional determinants are in a good agreement with those observed in predictors and PCA analysis for corresponding extreme environments. Notably, all the adjustments in nucleotide compositions were observed for archaeal proteomes, while bacterial show no or very weak signals with the only exception for thermophilic adaptation where we were able to obtain a signature of preferred codon content (Suppl. Table S9).

To conclude, a wealth of genomic/proteomic data available nowadays, allows us to consider more and more aspects of molecular adaptation to different extreme environments. We showed that considering several environmental factors at the same time reveal additional details and modified or even new/alternative mechanisms of adaptation typical for specific extreme conditions. In this work, for example, all the data on both nucleotide and amino acid level clearly pointed to a distinction between adaptation to temperature and to pH/salinity. The latter shows a certain level of similarity between their compositional trends, suggesting some similarity and connections between the corresponding mechanisms of adaptation. We also found that structure-based strategy of protein stabilization may work not only in thermophilic (Berezovsky and Shakhnovich, 2005), but also in adaptation to high salinity (Suppl. Fig. S8) – both apparently relics of the ancient nature of archaeal proteins (Goncearenco et al., 2014; Berezovsky and Shakhnovich, 2005). There are still many outstanding questions, which can be illustrated by Fig. 6 where 3D representations of PCA analysis show intriguing

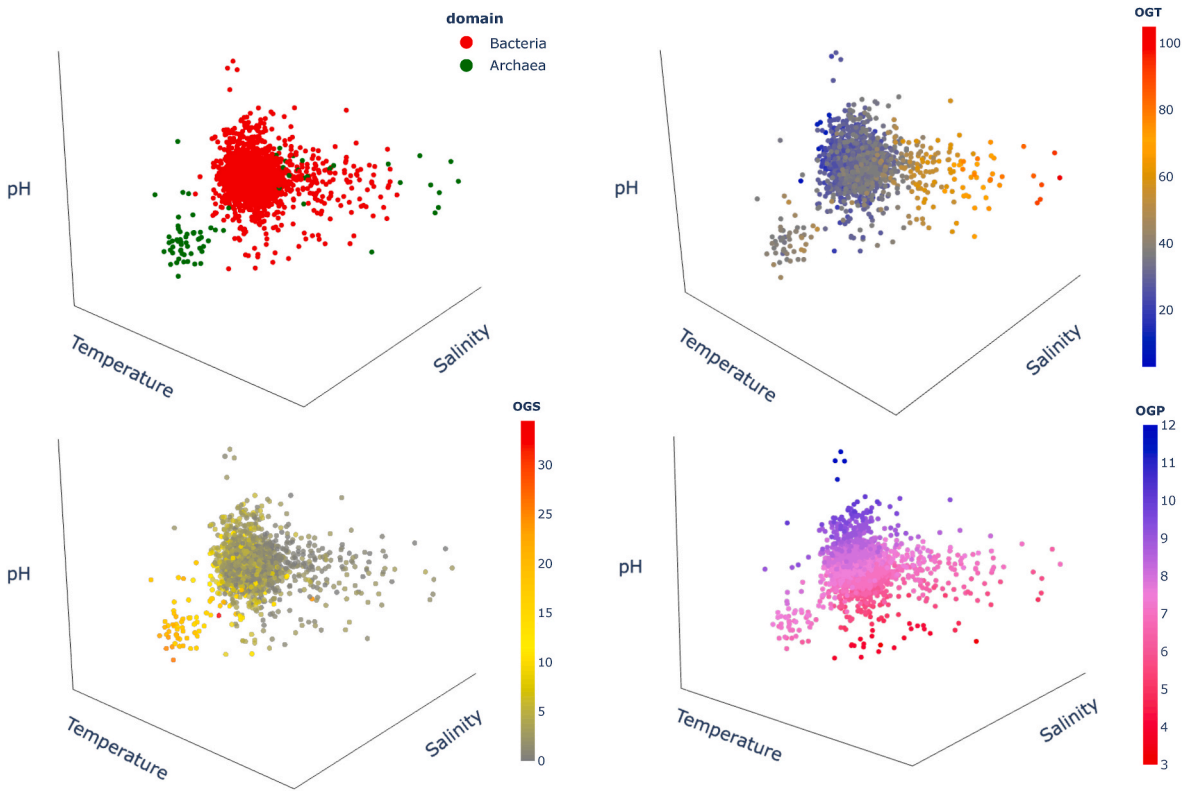


Fig. 6. The 3D PCA graph illustrating the complexity of adaptation to extreme environments and their combinations. The evolutionary perspective is reflected in the top left chart, showing archaeal (green) and bacterial (red) proteomes.

overlaps between environmental factors affecting archaeal (green, top left) and bacterial (red, top left) genomes/proteomes and working simultaneously on some of them. It will be important, therefore, to move from the studies of adaptation to individual extreme environments to their combinations. It may also require to consider in details proteins from different species moving from the analysis of whole-proteome trends to consideration of structural homologs from different organisms. The synergetic approach based on complementing the organismal trends as a generic foundation for the strategy of adaptation with more detailed consideration of specific targets will eventually allow us to engineer and design desirable functions and mechanisms of their regulation in individual or combined extreme environments.

CRediT authorship contribution statement

Aidana Amangeldina: performed the work, analyzed data, wrote paper. **Zhen Wah Tan:** performed the work, analyzed data. **Igor N. Berezovsky:** supervised the work, analyzed data, wrote and edited paper.

Declaration of competing interest

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

Data availability

Data will be made available on request.

Acknowledgements

This work was supported by the core funding provided by the Biomedical Research Council (BMRC) of the Agency for Science Technology, and Research (A*STAR), Singapore. INB was also partially supported by the NMRC MOH-001402-00 grant.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.crstbi.2024.100129.

References

Alcaide, M., Stogios, P.J., Lafraya, Á., Tchigvintsev, A., Flick, R., Bargiela, R., Chernikova, T.N., Reva, O.N., Hai, T., Leggewie, C.C., 2015. Pressure adaptation is linked to thermal adaptation in salt-saturated marine habitats. Environ. Microbiol. 17, 332–345.

Andreeva, N.S., James, M.N., 1991. Why does pepsin have a negative charge at very low pH? An analysis of conserved charged residues in aspartic proteinases. In: Structure and Function of the Aspartic Proteinases. Springer, pp. 39–45.

Aziz, M.F., Caetano-Anolles, G., 2021. Evolution of networks of protein domain organization. Sci. Rep. 11, 12075.

Aziz, M.F., Caetano-Anolles, K., Caetano-Anolles, G., 2016. The early history and emergence of molecular functions and modular scale-free network behavior. Sci. Rep. 6, 25058.

Bandyopadhyay, A.K., Krishnamoorthy, G., Padhy, L.C., Sonawat, H.M., 2007. Kinetics of salt-dependent unfolding of [2Fe–2S] ferredoxin of Halobacterium salinarum. Extremophiles 11, 615–625.

Beliën, T., Joye, I.J., Delcour, J.A., Courtin, C.M., 2009. Computational design-based molecular engineering of the glycosyl hydrolase family 11 B. subtilis XynA endoxylanase improves its acid stability. Protein Eng. Des. Sel. 22, 587–596.

Berezovsky, I.N., 2003. Discrete structure of van der Waals domains in globular proteins. Protein Eng. 16, 161–167.

Berezovsky, I.N., 2011. The diversity of physical forces and mechanisms in intermolecular interactions. Phys. Biol. 8, 035002.

Berezovsky, I.N., 2019. Towards descriptor of elementary functions for protein design. Curr. Opin. Struct. Biol. 58, 159–165.

Berezovsky, I.N., Shakhnovich, E.I., 2005. Physics and evolution of thermophilic adaptation. Proc. Natl. Acad. Sci. USA 102, 12742–12747.

Berezovsky, I.N., Trifonov, E.N., 2001. Van der Waals locks: loop-n-lock structure of globular proteins. J. Mol. Biol. 307, 1419–1426.

Berezovsky, I.N., Tumanyan, V.G., Esipova, N.G., 1997. Representation of amino acid sequences in terms of interaction energy in protein globules. FEBS Lett. 418, 43–46.

Berezovsky, I.N., Namiot, V.A., Tumanyan, V.G., Esipova, N.G., 1999. Hierarchy of the interaction energy distribution in the spatial structure of globular proteins and the problem of domain definition. J. Biomol. Struct. Dyn. 17, 133–155.
Berezovsky, I.N., Grosberg, A.Y., Trifonov, E.N., 2000a. Closed loops of nearly standard size: common basic element of protein structure. FEBS Lett. 466, 283–286.
Berezovsky, I.N., Esipova, N.G., Tumanyan, V.G., Namiot, V.A., 2000b. A new approach for the calculation of the energy of van der Waals interactions in macromolecules of globular proteins. J. Biomol. Struct. Dyn. 17, 799–809.
Berezovsky, I.N., Kirzhner, A., Kirzhner, V.M., Rosenfeld, V.R., Trifonov, E.N., 2003. Protein sequences yield a proteomic code. J. Biomol. Struct. Dyn. 21, 317–325.
Berezovsky, I.N., Chen, W.W., Choi, P.J., Shakhnovich, E.I., 2005. Entropic stabilization of proteins and its proteomic consequences. PLoS Comput. Biol. 1, e47.
Berezovsky, I.N., Zeldovich, K.B., Shakhnovich, E.I., 2007. Positive and negative design in stability and thermal adaptation of natural proteins. PLoS Comput. Biol. 3, e52.
Berezovsky, I.N., Guarnera, E., Zheng, Z., 2017a. Basic units of protein structure, folding, and function. Prog. Biophys. Mol. Biol. 128, 85–99.
Berezovsky, I.N., Guarnera, E., Zheng, Z., Eisenhaber, B., Eisenhaber, F., 2017b. Protein function machinery: from basic structural units to modulation of activity. Curr. Opin. Struct. Biol. 42, 67–74.
Bohlin, J., Brynildsrud, O., Vesth, T., Skjerve, E., Ussery, D.W., 2013. Amino acid usage is asymmetrically biased in AT-and GC-rich microbial genomes. PLoS One 8, e69878.
Bresler, S.E., Talmud, D.L., 1944a. On the nature of globular proteins. Compt Rend Acad Sci URSS 43, 310–314.
Bresler, S.E., Talmud, D.L., 1944b. On the nature of globular proteins. II A few consequences of the new hypothesis. Compt Rend Acad Sci URSS 43, 349–350.
Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L., 2009. BLAST+: architecture and applications. BMC Bioinf. 10, 421.
Cambillau, C., Claverie, J.M., 2000. Structural and genomic correlates of hyperthermostability. J. Biol. Chem. 275, 32383–32386.
Chakravarty, S., Varadarajan, R., 1999. Residue depth: a novel parameter for the analysis of protein structure and stability. Structure 7, 723–732.
Chakravarty, S., Varadarajan, R., 2000. Elucidation of determinants of protein stability through genome sequence analysis. FEBS Lett. 470, 65–69.
Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Sayers, E.W., 2016. GenBank. Nucleic Acids Res. 44, D67–D72.
Daronkola, H.G., Verde, A.V., 2021. Proteins maintain hydration at high [KCl] concentration regardless of content in acidic amino acids. Biophys. J. 120, 2746–2762.
Deole, R., Challacombe, J., Raiford, D.W., Hoff, W.D., 2013. An extremely halophilic proteobacterium combines a highly acidic proteome with a low cytoplasmic potassium content. J. Biol. Chem. 288, 581–588.
Dubnovitsky, A.P., Kapetaniou, E.G., Papageorgiou, A.C., 2005. Enzyme adaptation to alkaline pH: atomic resolution (1.08 Å) structure of phosphoserine aminotransferase from Bacillus alcalophilus. Protein Sci. 14, 97–110.
Dym, O., Mevarech, M., Sussman, J.L., 1995. Structural features that stabilize halophilic malate dehydrogenase from an archaebacterium. Science 267, 1344–1346.
Dyson, H.J., Wright, P.E., Scheraga, H.A., 2006. The role of hydrophobic interactions in initiation and propagation of protein folding. Proc. Natl. Acad. Sci. USA 103, 13057–13061.
Ebrahimie, E., Ebrahimi, M., Sarvestani, N.R., Ebrahimi, M., 2011. Protein attributes contribute to halo-stability, bioinformatics approach. Saline Syst. 7, 1–14.
Elevi Bardavid, R., Oren, A., 2012. Acid-shifted isoelectric point profiles of the proteins in a hypersaline microbial mat: an adaptation to life at high salt concentrations? Extremophiles 16, 787–792.
Feller, G., Gerday, C., 2003. Psychrophilic enzymes: hot topics in cold adaptation. Nat. Rev. Microbiol. 1, 200–208.
Folch, B., Rooman, M., Dehouck, Y., 2008. Thermostability of salt bridges versus hydrophobic interactions in proteins probed by statistical potentials. J. Chem. Inf. Model. 48, 119–127.
Francois, J.A., Starks, C.M., Sivanuntakorn, S., Jiang, H., Ransome, A.E., Nam, J.-W., Constantine, C.Z., Kappock, T.J., 2006. Structure of a NADH-insensitive hexameric citrate synthase that resists acid inactivation. Biochemistry 45, 13487–13499.
Friedman, R.A., Honig, B., 1995. A free energy analysis of nucleic acid base stacking in aqueous solution. Biophys. J. 69, 1528–1535.
Fukuchi, S., Yoshimune, K., Wakayama, M., Moriguchi, M., Nishikawa, K., 2003. Unique amino acid composition of proteins in halophilic bacteria. Journal of molecular biology 327, 347–357.
Fushinobu, S., Ito, K., Konno, M., Wakagi, T., Matsuzawa, H., 1998. Crystallographic and mutational analyses of an extremely acidophilic and acid-stable xylanase: biased distribution of acidic residues and importance of Asp37 for catalysis at low pH. Protein Eng. 11, 1121–1128.
Goncearenco, A., Berezovsky, I.N., 2014. The fundamental tradeoff in genomes and proteomes of prokaryotes established by the genetic code, codon entropy, and physics of nucleic acids and proteins. Biol. Direct 9, 29.
Goncearenco, A., Berezovsky, I.N., 2015. Protein function from its emergence to diversity in contemporary proteins. Phys. Biol. 12, 045002.
Goncearenco, A., Ma, B.G., Berezovsky, I.N., 2014. Molecular mechanisms of adaptation emerging from the physics and evolution of nucleic acids and proteins. Nucleic Acids Res. 42, 2879–2892.
Goodchild, A., Saunders, N.F., Ertan, H., Raftery, M., Guilhaus, M., Curmi, P.M., Cavicchioli, R., 2004. A proteomic determination of cold adaptation in the Antarctic archaeon, Methanococcoides burtonii. Mol. Microbiol. 53, 309–321.
Gromiha, M.M., Suresh, M.X., 2008. Discrimination of mesophilic and thermophilic proteins using machine learning algorithms. Proteins 70, 1274–1279.
Gromiha, M.M., Pathak, M.C., Saraboji, K., Ortlund, E.A., Gaucher, E.A., 2013. Hydrophobic environment is a key factor for the stability of thermophilic proteins. Proteins 81, 715–721.
Guarnera, E., Berezovsky, I.N., 2019. Toward comprehensive allosteric control over protein activity. Structure 27, 866–878 e861.
Gülich, S., Linhult, M., Ståhl, S., Hober, S., 2002. Engineering streptococcal protein G for increased alkaline stability. Protein Eng. 15, 835–842.
Gutteridge, A., Thornton, J.M., 2005. Understanding nature's catalytic toolkit. Trends Biochem. Sci. 30, 622–629.
Hocker, B., 2014. Design of proteins from smaller fragments-learning from evolution. Curr. Opin. Struct. Biol. 27, 56–62.
Hou, Q., Rooman, M., Pucci, F., 2023. Enzyme stability-activity trade-off: new insights from protein stability weaknesses and evolutionary conservation. J. Chem. Theor. Comput. 19, 3664–3671.
Jaenicke, R., 1999. Stability and folding of domain proteins. Prog. Biophys. Mol. Biol. 71, 155–241.
Kajander, T., Kahn, P.C., Passila, S.H., Cohen, D.C., Lehtiö, L., Adolfsen, W., Warwicker, J., Schell, U., Goldman, A., 2000. Buried charged surface in proteins. Structure 8, 1203–1214.
Kester, D.R., Duedall, I.W., Connors, D.N., Pytkowicz, R.M., 1967. Preparation of artificial seawater 1. Limnol. Oceanogr. 12, 176–179.
Koczyk, G., Berezovsky, I.N., 2008. Domain Hierarchy and closed Loops (DHcL): a server for exploring hierarchy of protein domain structure. Nucleic Acids Res. 36, W239–W245.
Liu, L., Wang, B., Chen, H., Wang, S., Wang, M., Zhang, S., Song, A., Shen, J., Wu, K., Jia, X., 2009. Rational pH-engineering of the thermostable xylanase based on computational model. Process Biochemistry 44, 912–915.
Ma, B.G., Goncearenco, A., Berezovsky, I.N., 2010. Thermophilic adaptation of protein complexes inferred from proteomic homology modeling. Structure 18, 819–828.
Makhatadze, G.I., Loladze, V.V., Ermolenko, D.N., Chen, X., Thomas, S.T., 2003. Contribution of surface salt bridges to protein stability: guidelines for protein engineering. Journal of molecular biology 327, 1135–1148.
Mamo, G., Thunnissen, M., Hatti-Kaul, R., Mattiasson, B., 2009. An alkaline active xylanase: insights into mechanisms of high pH catalytic adaptation. Biochimie 91, 1187–1196.
Mamonova, T.B., Glyakina, A.V., Galzitskaya, O.V., Kurnikova, M.G., 2013. Stability and rigidity/flexibility—two sides of the same coin? Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics 1834, 854–866.
Manikandan, K., Bhardwaj, A., Gupta, N., Lokanath, N.K., Ghosh, A., Reddy, V.S., Ramakumar, S., 2006. Crystal structures of native and xylosaccharide-bound alkali thermostable xylanase from an alkalophilic Bacillus sp. NG-27: structural insights into alkalophilicity and implications for adaptation to polyextreme conditions. Protein Sci. 15, 1951–1960.
Marmur, J., Doty, P., 1962. Determination of the base composition of deoxyribonucleic acid from its thermal denaturation temperature. J. Mol. Biol. 5, 109–118.
Nakashima, H., Fukuchi, S., Nishikawa, K., 2003. Compositional changes in RNA, DNA and proteins for bacterial adaptation to higher and lower temperatures. J. Biochem. 133, 507–513.
Nayek, A., Sen Gupta, P.S., Banerjee, S., Mondal, B., Bandyopadhyay, A.K., 2014. Salt-bridge energetics in halophilic proteins. PLoS One 9, e93862.
Olivella, M., Gonzalez, A., Pardo, L., Deupi, X., 2013. Relation between sequence and structure in membrane proteins. Bioinformatics 29, 1589–1592.
Oren, A., 2002. Diversity of halophilic microorganisms: environments, phylogeny, physiology, and applications. J. Ind. Microbiol. Biotechnol. 28, 56–63.
Ortega, G., Diercks, T., Millet, O., 2015. Halophilic protein adaptation results from synergistic residue-ion interactions in the folded and unfolded states. Chemistry & biology 22, 1597–1607.
Pace, C.N., Fu, H., Lee Fryar, K., Landua, J., Trevino, S.R., Schell, D., Thurlkill, R.L., Imura, S., Scholtz, J.M., Gajiwala, K., 2014a. Contribution of hydrogen bonds to protein stability. Protein Sci. 23, 652–661.
Pace, C.N., Scholtz, J.M., Grimsley, G.R., 2014b. Forces stabilizing proteins. FEBS Lett. 588, 2177–2184.
Pe'er, I., Felder, C.E., Man, O., Silman, I., Sussman, J.L., Beckmann, J.S., 2004. Proteomic signatures: amino acid and oligopeptide compositions differentiate among phyla. Proteins 54, 20–40.
Pintar, A., Carugo, O., Pongor, S., 2003. Atom depth in protein structure and function. Trends Biochem. Sci. 28, 593–597.
Popinako, A., Antonov, M., Tikhonov, A., Tikhonova, T., Popov, V., 2017. Structural adaptations of octaheme nitrite reductases from haloalkaliphilic Thioalkalivibrio bacteria to alkaline pH and high salinity. PLoS One 12, e0177392.
Pucci, F., Rooman, M., 2014. Stability curve prediction of homologous proteins using temperature-dependent statistical potentials. PLoS Comput. Biol. 10, e1003689.
Pucci, F., Rooman, M., 2017. Physical and molecular bases of protein thermal stability and cold adaptation. Curr. Opin. Struct. Biol. 42, 117–128.
Pylaeva, S., Brehm, M., Sebastiani, D., 2018. Salt bridge in aqueous solution: strong structural motifs but weak enthalpic effect. Sci. Rep. 8, 1–7.
Reed, C.J., Lewis, H., Trejo, E., Winston, V., 2013. Evilia C: protein adaptations in archaeal extremophiles. Archaea 2013.
Saenger, W., 1984. Principles of Nucleic Acid Structure. Springer-Verlag, New York.
Shakhnovich, E., 2006. Protein folding thermodynamics and dynamics: where physics, chemistry, and biology meet. Chem Rev 106, 1559–1588.
Singer, G.A., Hickey, D.A., 2000. Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. Mol. Biol. Evol. 17, 1581–1588.
Sriaporn, C., Campbell, K.A., Van Kranendonk, M.J., Handley, K.M., 2021. Genomic adaptations enabling Acidithiobacillus distribution across wide-ranging hot spring temperatures and pHs. Microbiome 9, 1–17.

Suplatov, D., Panin, N., Kirilin, E., Shcherbakova, T., Kudryavtsev, P., Švedas, V., 2014. Computational design of a pH stable enzyme: understanding molecular mechanism of penicillin acylase's adaptation to alkaline conditions. PLoS One 9, e100643.
Svedberg, T., 1929. Mass and Size of protein molecules. Nature 123, 871.
Tal, G., Boca, S.M., Mittenthal, J., Caetano-Anolles, G., 2016. A dynamic model for the evolution of protein structure. J. Mol. Evol. 82, 230–243.
Tatusova, T., DiCuccio, M., Badretdin, A., Chetvernin, V., Nawrocki, E.P., Zaslavsky, L., Lomsadze, A., Pruitt, K.D., Borodovsky, M., Ostell, J., 2016. NCBI prokaryotic genome annotation pipeline. Nucleic Acids Res. 44, 6614–6624.
Tee, W.V., Guarnera, E., Berezovsky, I.N., 2020. Disorder driven allosteric control of protein activity. Curr Res Struct Biol 2, 191–203.
Tee, W.V., Tan, Z.W., Lee, K., Guarnera, E., Berezovsky, I.N., 2021. Exploring the allosteric territory of protein function. J. Phys. Chem. B 125, 3763–3780.
Tee, W.V., Tan, Z.W., Guarnera, E., Berezovsky, I.N., 2022. Conservation and diversity in allosteric fingerprints of proteins for evolutionary-inspired engineering and design. J. Mol. Biol. 434, 167577.
Tekaia, F., Yeramian, E., 2006. Evolution of proteomes: fundamental signatures and global trends in amino acid compositions. BMC Genom. 7, 307.
Tekaia, F., Yeramian, E., Dujon, B., 2002. Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: a global picture with correspondence analysis. Gene 297, 51–60.
Tokuriki, N., Oldfield, C.J., Uversky, V.N., Berezovsky, I.N., Tawfik, D.S., 2009. Do viral proteins possess unique biophysical features? Trends Biochem. Sci. 34, 53–59.
Trifonov, E.N., 2000. Consensus temporal order of amino acids and evolution of the triplet code. Gene 261, 139–151.
Trifonov, E.N., Kirzhner, A., Kirzhner, V.M., Berezovsky, I.N., 2001. Distinct stages of protein evolution as suggested by protein sequence analysis. J. Mol. Evol. 53, 394–401.
Van Dijk, E., Hoogeveen, A., Abeln, S., 2015. The hydrophobic temperature dependence of amino acids directly calculated from protein structures. PLoS Comput. Biol. 11, e1004277.
Walden, H., Taylor, G.L., Lorentzen, E., Pohl, E., Lilie, H., Schramm, A., Knura, T., Stubbe, K., Tjaden, B., Hensel, R., 2004. Structure and function of a regulated archaeal triosephosphate isomerase adapted to high temperature. Journal of molecular biology 342, 861–875.
Xia, Y.-L., Sun, J.-H., Ai, S.-M., Li, Y., Du, X., Sang, P., Yang, L.-Q., Fu, Y.-X., Liu, S.-Q., 2018. Insights into the role of electrostatics in temperature adaptation: a comparative study of psychrophilic, mesophilic, and thermophilic subtilisin-like serine proteases. RSC Adv. 8, 29698–29713.
Xu, H., Zhang, F., Shang, H., Li, X., Wang, J., Qiao, D., Cao, Y., 2013. Alkalophilic adaptation of XynB endoxylanase from Aspergillus Niger via rational design of pKa of catalytic residues. J. Biosci. Bioeng. 115, 618–622.
Yakovchuk, P., Protozanova, E., Frank-Kamenetskii, M.D., 2006. Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. Nucleic Acids Res. 34, 564–574.
Yang, H., Liu, L., Shin, H-d, Chen, R.R., Li, J., Du, G., Chen, J., 2013. Structure-based engineering of histidine residues in the catalytic domain of α-amylase from Bacillus subtilis for improved protein stability and catalytic efficiency under acidic conditions. J. Biotechnol. 164, 59–66.
Yin, M., Goncearenco, A., Berezovsky, I.N., 2021. Deriving and using descriptors of elementary functions in rational protein design. Front Bioinform 1, 657529.
Zeldovich, K.B., Berezovsky, I.N., Shakhnovich, E.I., 2006. Physical origins of protein superfamilies. J. Mol. Biol. 357, 1335–1343.
Zeldovich, K.B., Berezovsky, I.N., Shakhnovich, E.I., 2007. Protein and DNA sequence determinants of thermophilic adaptation. PLoS Comput. Biol. 3, e5.