Response to corrections requested by Associate Editor

- Line 176: Changed "On the other hand" to "Conversely".
- Line 375: Changed "On the other hand" to "In contrast".
- No other changes have been made in this iteration.

Summary of Major Changes

- Fig. 1e: Added scatter plot of R^2 for $n_{\mathsf{H_2O}}$–Z_{C} fits vs R^2 for $n_{\mathsf{O_2}}$–Z_{C} fits for all possible combinations of basis species with three amino acids, $\mathsf{H_2O}$ and $\mathsf{O_2}$ to illustrate the criteria for choosing basis species.
- Removed the rQEC derivation (residual-corrected values of $n_{\mathsf{H_2O}}$); now values of $n_{\mathsf{H_2O}}$ are taken directly from the QEC basis species (glutamine, glutamic acid, cysteine, $\mathsf{H_2O}$, $\mathsf{O_2}$). This change affects the scale and appearance of the plots but does not alter the findings, except to point out that negative slopes on these plots are associated with the background correlation between $n_{\mathsf{H_2O}}$ and Z_{C} for amino acids.
- To visualize the background correlation between $n_{\mathsf{H_2O}}$ and Z_{C}, guidelines parallel to the fit for amino acids have been added to the plots in Figs. 3, 5, and 6.
- Added Figure 2 with schematic of Z_{C} and $n_{\mathsf{H_2O}}$ calculations.
- Redrew Fig. 7 to plot (a) time or (b) type of solute on horizontal axis.

Point-by-point Response to Anonymous Referee #1

Dick et al. have mined the biomolecular literature to show that the composition of proteins in microorganisms reflect the salinity of their environments. In particular, their results provide evidence that the stoichiometric hydration state of amino acids is lower in many saline settings than in freshwater environments. The authors use metagenomes, metatranscriptomes and proteomes of individual organisms resulting from environmental and laboratory studies. Their method of analysis includes a rather novel technique – they assess the difference in the stoichiometric hydration state (n_H2O) of theoretical formation reactions for the amino acids in different proteins (measured or inferred from metagenomes). These formation reactions are familiar to those who carry out geochemical modeling, though the choice of basis species is unusual. These formation reactions are familiar to those who carry out geochemical modeling, though the choice of basis species is unusual. H2O is used as a basis species in addition to O2 and three amino acids (glutamine, glutamic acid and cysteine).

The manuscript has been revised to show the reasons for this choice of basis species more clearly; in particular, Figure 1 now includes a plot comparing all possible choices of basis species that were considered within our constraints.

To help make sense of their results, the authors also compute and compare values of the oxidation state of carbon in amino acids/proteins as well as their hydropathicities and isoelectric points. Ultimately, the authors seek to show a quantitative relationship between the composition of organisms (their biomolecules) and their environments.

Thank you for the thorough review and constructive suggestions. We respond to each point below.

Because this work used techniques that are well known in one field (geochemical modeling) and applies them to another (biomolecular sequence analysis), it would be most helpful if the authors showed an example of the differing stoichiometric hydrations state of two proteins. Maybe this wouldn't work too well in a figure, but perhaps some combination of a table and schematic would go a long way towards explaining their methods.

Added Figure 2: Schematic of $n_{\mathsf{H_2O}}$ and Z_{C} calculations for one protein. The selected protein is chicken lysozyme (UniProt ID: LYSC_CHICK), which should be familiar to most protein chemists as it is historically one of the most extensively characterized proteins in the laboratory. The schematic represents the amino acid composition, chemical formula, and numerical results for this protein. It should be clear that the specific result depends on the amino acid composition of the protein, so we have included only one protein for clarity..

The title of Table 1 should spell out what rQEC is – especially since it is conceptually and acronymically very close to QEC.

The rQEC derivation was so named because it involved "residual-corrected" values of $n_{\mathsf{H_2O}}$ obtained from the QEC basis species (glutamine, glutamic acid, cysteine, $\mathsf{H_2O}$, $\mathsf{O_2}$). We have removed the rQEC derivation from the revised manuscript and instead just use the coefficients from the QEC basis species without modification (see below).

Some clarification is needed concerning the calculation of rQEC. In Table 1, the value of n_H2O for alanine is 0.369. The example for calculating n_H2O using the QEC formulation for alanine is 0.6. The correction noted in the caption for Fig. 1 to transform QEC to rQEC is 0.355. My calculator says that 0.6-0.355 = 0.245, not 0.369. Please explain.

The rQEC derivation was made in two steps: (1) computing the residuals of the linear fits between $n_{\mathsf{H_2O}}$ (from the QEC basis species) and Z_{C}; (2) subtracting a constant from the residuals. Step 1 can be thought of as a baseline or residual correction and Step 2 as a recentering operation. Therefore, the calculation for alanine is not 0.6 – 0.355, but rather [the residual between the fitted line and 0.6] – 0.355.

The criteria we consider in choosing the basis species are that (1) $n_{\mathsf{H_2O}}$ of amino acids should have very little correlation with Z_{C}, (2) $n_{\mathsf{O_2}}$ of amino acids should be strongly correlated with Z_{C}, and (3) the basis species should represent metabolites with high network connectivity.

The derivation of rQEC was meant to "fine-tune" the QEC basis species in order to satisfy criterion (1) above, but we realize in retrospect that this derivation is not theoretically justified, since rQEC loses the important quality that $n_{\mathsf{H_2O}}$ should directly quantify the stoichiometry of thermodynamic components (basis species) in overall chemical reactions.

We have added a new panel to Figure 1 that shows the R^2 values for $n_{\mathsf{H_2O}}$–Z_{C} and $n_{\mathsf{O_2}}$–Z_{C} fits for all possible combinations of three amino acids with $\mathsf{H_2O}$

and O_2. QEC is in the lower right corner of this plot and is nearly optimal. Although some other sets of basis species have even lower R^2 values for n_{H_2O}–Z_C fits, and slightly higher R^2 values for n_{O_2}–Z_C fits, they consist of amino acids (e.g. tryptophan and tyrosine) that are not central metabolites. On the other hand, glutamine and glutamic acid are more desirable because of their major roles in metabolism (criterion #3 above). Therefore, QEC appears to be the most reasonable choice of all the basis species we considered here.

We note, however, that QEC still carries a small negative correlation between n_{H_2O} and Z_C for amino acids. In the revised manuscript, we do not attempt to remove this background correlation, as was done previously with rQEC. Instead, we revised the description of Fig. 3 [emphasis indicates added text]:

The trends *of carbon oxidation state* described above are visible in the scatter plot in Fig. 3*, with an added dimension: stoichiometric hydration state. The guidelines in this plot are parallel to the n_{H_2O}–Z_C trend for amino acids (Fig. 2); their slope represents the background correlation between n_{H_2O} and Z_C that is inherent in the stoichiometric analysis. Sample data for Bison Pool and the submarine vents are distributed parallel to these guidelines. Therefore, the decrease of n_{H_2O} along these redox gradients can be attributed to the background correlation in the stoichiometric analysis, and the differences between samples within each dataset are specifically associated with changes in carbon oxidation state and not stoichiometric hydration state.* This is an expected outcome, as the redox gradients considered here do not have large changes in salinity. ...

Lines 195-196: The authors here refer to 8 amino acids by their three-letter abbreviations, but in Table 1 and in the naming of their basis species (QEC), they refer to amino acids by their one-letter abbreviations. Is there a particular reason for this difference?

The three-letter abbreviations seem more fitting for a sentence structure, but the one-letter abbreviations save space in the table and are more appropriate for forming acronyms. For consistency we have changed this sentence to use the one-letter abbreviations.

It seems like the text on lines 226-227 could be better represented by an equation. This would make it easier to look back on how the stoichiometric hydration state was calculated.

The equations for computing n_{H_2O} and Z_C from amino acid composition have been added here.

Section 3.5 needs more explanation. The title of this section suggests that it's about organisms containing the Nif gene, and the authors get around to talking about these organisms, but some explanation is needed about why this gene was used as a filter for which proteomes to select (data availability?). Also, start this section with 'what' and 'why', then tell us the 'how'. It starts with 'how,' making it hard to follow.

Added at the beginning of this paragraph: "[*what*] In a separate study, Poudel et al. (2018) used carbon oxidation state as a metric for comparing proteomes of organisms containing the nitrogenase gene (Nif). [*why*] The evolution of these organisms is associated with rising atmospheric oxygen through geolog-

ical history. In order to replicate their results, ..." [*how*: rest of the paragraph]

Section 3.6 The authors should state explicitly if they did or did not take into account how temperature effects values of the isoelectric point. The same goes for using GRAVY. Amino acid pKa's and the permittivity of water certainly change with temperature.

Added: "The pK values used for calculating pI (Bjellqvist et al., 1993, 1994) and transfer free energies used in the derivation of the GRAVY scale (Kyte and Doolittle, 1982) correspond to 25 °C and 1 bar and no attempt was made here to account for the temperature effects on these properties."

Section 3.7 Is the sum of the 100 subsamples equivalent to ∼50,000 amino acids for each sample? Then what is the typical subsample density?

No, each subsample (not the sum of them) has ca. 50,000 amino acids. Reworded this as: "The number of sequences included in each subsample was chosen to give a total length closest to 50,000 amino acids on average." Also added these lines: "The subsample density, or number of sequences included in each sample, depends on the average length of the metagenomic or metatranscriptomic sequences and is listed in Tables S1 and S2. This number ranges from 251 for the dataset with the highest mean protein fragment length (199.1; metagenome of hot-spring source of Bison Pool) to 1696 for the dataset with the lowest mean protein fragment length (29.5; metatranscriptome of site GS684 in the Baltic Sea)."

The beginning of Section 4.2, like in other parts of the manuscript, starts out with 'how', but should lead with what the section is all about. For instance, this paragraph should start by saying that the stoichiometric hydration state of proteins can be determined by more factors than just salinity. Instead, it starts with "Metagenomic and metatranscriptomic data for different filter size fractions are available for the Baltic Sea." This topic sentence does not reveal to the reader what this section is about and it fails to capture the point of the analyses described in the section.

Inserted a new "topic paragraph" for this section including the recommended topic sentence [emphasized text moved from Conclusion as also recommended]: "The stoichiometric hydration state of proteins can be influenced by factors other than just salinity. Previous authors have observed large differences between free-living and particle-associated microbial communities, which may be due in part to anoxic conditions arising from limited diffusion in particles (Simon et al., 2014). *As described below, we found a trend of relatively low* n_{H_2O} *in particles compared to free-living fractions in both the Baltic Sea and Amazon River. This effect is probably associated with phylogenetic differences among the size fractions, but reduced accessibility to bulk water may be a contributing factor. Further support for the possible influence of physical accessibility is reduced* n_{H_2O} *in the interior compared to upper layers of the Guerrero Negro microbial mat.*"

Line 291 notes the "0.1–0.8 mm size fraction," but what this means isn't explained until the next section. Either explain it where it first

appears or direct the reader to where it is explained. In general, the authors should be careful what they mean. When a filter fraction is noted, this could mean the DNA collected from the filtrate or that which doesn't pass through.

Added emphasized text: "*For the Baltic Sea metagenomes and metatranscriptomes,* the 0.1–0.8 μm and 0.8–3.0 μm size fractions *of particles that don't pass through the filter, which are used for subsequent DNA extraction and sequencing,* represent free living bacteria, while the 3.0–200 μm fraction contains particle-associated bacteria with average larger genome sizes and greater inferred metabolic and regulatory capacity (Dupont et al., 2014)."

Perhaps an explanation for why values of n_H2O in the Rodriguez-Brito et al., 2010 data set do not follow the expected trend is that fish nurseries are extremely nutrient rich and the associated microbial communities may not be responding as they would in a typical natural system that is less persistently copiotrophic.

Added: "Specifically, the microbial communities in the aquaculture ponds may not be responding as they would in a typical natural system that is less nutrient-rich."

Also added this text after the analysis of the differentially expressed proteins in laboratory experiments: "The large negative shift of $\Delta n_{\mathsf{H_2O}}$ associated with most organic solutes compared to NaCl lends support to the notion that high organic loading could contribute to the relatively low $n_{\mathsf{H_2O}}$ of protein sequences from metagenomes of freshwater aquaculture ponds (Fig. 6b)."

See also the related response to Referee #2; the suggestion was made that the lower $n_{\mathsf{H_2O}}$ could be associated with a greater abundance of heterotrophs (due to input of organic compounds), as noted previously in this paper for heterotroph-rich zones in other systems (Bison Pool, Guerrero Negro microbial mat).

Many of the sentence in the Section 5 (Conclusions) should be the first sentence of the sections whose results they summarize. This would make following the text in these sections more straightforward. Tell the reader the result, then explain the supporting evidence.

We have applied this recommendation by moving the summary about particle size to the beginning of the "Multifactorial hydration effects" section (see above) and the summary about laboratory experiments to the "Compositional analysis of differentially expressed proteins" section (see below). The remainder of the Conclusion has been revised to give a concise summary and synthesis.

Lines 371-372 – this lead sentence begins to summarize the paragraph, but then wanders away. It seems that the authors should simply note that in addition to spatial changes in salinity, there are temporal effects to changes that also merit study/consideration.

We have replaced the first two sentences of this paragraph with the topic sentence taken from the Conclusion: "While biomolecular data for environmental salinity gradients reflect phylogenetic differences and evolution, laboratory experiments provide information on the physiological effects of osmotic conditions on protein expression in particular organisms." Note that this lead paragraph

also alludes to temporal effects ("dynamic process"), but the section also includes data on different solutes and other experiments not specifically dealing with time-course changes, so the whole section is introduced with "physiological effects of osmotic conditions on protein expression in particular organisms".

Figure 1 – what is the difference between the blue-fuzz-halo and black rectangular/square shapes in panels e, f, h and i? I'm guessing that this is due to the large number of proteins in whole proteomes, but why the difference in symbols? Same question for Fig. 5.

According to the documentation for the "smoothScatter" function in R, the blue colors are a "smoothed color density representation of a scatterplot" and the black symbols are points in the low-density region, which can be used to identify outliers. These plots have been removed from Fig. 1 in the revision; likewise, the former Fig. 5 has been removed because it did not add much to the paper. (These scatter plots showed whole-proteome data for human and *E. coli*, which are not directly relevant to the environmental salinity gradients considered here.)

Figure 2. The caption says that the abbreviations and data sources for panel (a) are given in Fig 2. They are not.

Thanks for pointing this out; the abbreviations and data sources are now given here. In addition, an outline has been added to the point for proteomes from Nif-A organisms to indicate that they tend to occupy more oxidized environments compared to the other nitrogenase-bearing organisms (Poudel et al., 2018).

Panel (b) should be remade. The symbols differ in color, fill and direction, but the caption only notes what the directional difference means. Also, though I see that this plot is made at the same scale as panel (a), the result is a lot of white space and a bunch of cramped symbols connected by slightly different line styles. I've enlarged it on my external monitor and it's still hard to make sense of it.

Panel (b) has been made less crowded by splitting the data into two panels (surface samples: panel b; deeper samples: panel c) and the scale was adjusted to remove white space.

Figure 3. It would be helpful if there was something like "–> salinity" along the x-axis.

Added "→ higher salinity →" to the axis label.

Figure 4. Is the difference between the open and closed symbols in panels a, b, d and e that the open ones represent lower salinity samples and the closed ones higher salinity ones? If so, please state in the caption.

Yes, the open symbols represent river samples (lower salinity) and the closed ones represent plume samples (higher salinity). The words "river" and "plume" have been added to the legend to make this clear.

Figure 7. color coding time series data in panels c and e would be quite helpful. It should be noted somewhere in Table 2 that the ID and associated information are relevant to Figure 8.

The figure has been redrawn so that log(time, minutes) is now on the horizontal axis. This makes the multiple time series experiments easy to distinguish from each other. Color and symbol shape are used here to represent the proteomics experiments.

Table 2 and former Fig. 8 for halophiles have been removed. Now the data for protein expression in halophiles under hyperosmotic stress are highlighted in Fig. 7 (red triangles) and are referenced in Fig. S3.

The supplemental figures in S1 and S2 need captions.

Added captions:

Figure S1: Transcriptomics data for non-halophilic bacteria in hyperosmotic stress experiments. The plots show median differences of compositional metrics, GRAVY, and pI for proteins coded by the differentially expressed genes, [. . .]

Figure S2: Proteomics data for non-halophilic bacteria in hyperosmotic stress experiments. The plots show median differences of compositional metrics, GRAVY, and pI for the differentially expressed proteins, [. . .]

Figure S3: Proteomics data for halophilic archaea in osmotic stress experiments. For completeness, data for both hyperosmotic (circles) and hypoosmotic (squares) experiments, which are reported together in the proteomics studies, are shown here, but only hyperosmotic stress data are used in the manuscript. The plots show median differences of compositional metrics, GRAVY, and pI for the differentially expressed proteins, [. . .]

[. . . all captions . . .] i.e. median value for all up-regulated proteins minus median value for all down-regulated proteins in each dataset. Data sources, indicated by letters, are described in the following table and footnotes. Reference keys in the table, derived from the first letters of the authors' surnames and publication year, correspond to file names used for the datasets in the canprot package.

Other Changes

- Proteomes of Nif-bearing organisms are now made using RefSeq release 201 of July 2020, updated from release 95 of July 2019. The update decreases the number of matching organisms slightly (Nif-A: down 2 to 155; Nif-B: down 1 to 68), but does not noticeably alter the calculated Z_{C} and $n_{\mathsf{H_2O}}$ shown in Fig. 3.

- List specific proteins used for comparison of GRAVY and pI calculations with ProtParam (UniProt IDs: LYSC_CHICK, RNAS1_BOVIN, AMYA_PYRFU).

- Removed human and *E. coli* proteome plots (panels formerly in Fig. 2 and former Fig. 5).

- An additional bacterial proteomics dataset for hyperosmotic stress was included (Huang et al., 2018 referenced in Figure S2).

- Removed table (former Table 2) and plots (former Fig. 8) for halophile protein expression datasets. The halophile proteomics data for hyperosmotic stress are now shown in Fig. 7, and Figure S3 has been added to give references for the data. Hypoosmotic stress experiments are no longer analyzed in the manuscript, but are included in Figure S3 for completeness.

- Added reference that urea permeates cells and is not hypertonic (Burg et al., 2007).

Point-by-point Response to Anonymous Referee #2

In general, I found the paper to be another interesting read from the primary author. However, as a microbiologist that is interested in understanding how energy availability and demand affect the distribution of microorganisms and their evolution, I would appreciate seeing a more robust effort to link the thermodynamics way of thinking (as presented here) to physiological process or mechanism that could then be used to gauge why such patterns may exist. More or less, I think this is a missed opportunity that, if executed effectively, could elevate the utility of this paper and this way of thinking. Thus, I strongly suggest the authors attempt to explain their observations at a level that makes sense to the more biologically oriented reader. As I was reading this, I could not help but think to myself how any one or several observations made sense from the level of phenotype and natural selection. The authors might consider asking themselves this same question and then speculating where possible to make this body of work a greater utility for the community.

Thank you for your detailed attention to the concepts and analysis in our paper and your suggestions for improving the work. We respond to the main critiques below:

1) a more robust effort to link the thermodynamics way of thinking (as presented here) to physiological process or mechanism

This is an ongoing challenge. An obstacle (which could also be seen as a "missed opportunity") is that the thermodynamic way of thinking deals with energetic differences between two states of a system; without further (i.e. extra-thermodynamic) constraints, it is not possible to explicitly deal with underlying mechanisms in a thermodynamic model. This paper does not attempt to build such a thermodynamic model, but uses thermodynamics as a guiding concept. A major application of thermodynamics in geochemistry is to describe and predict compositional changes in a system, e.g. the distribution of aqueous species and mineral phases with different chemical formulas. The aim of this paper is to develop a framework for describing compositional changes in geobiochemical systems, and one of the first challenges is to recognize that the most appropriate descriptive variables are probably different from inorganic geochemical systems. We present our conceptual arguments that oxidation and hydration state should be considered as primary variables, develop metrics that quantify them, and use the metagenomic data to explore how these metrics respond to environmental gradients of salinity and redox conditions. Clearly, this is far from the sophisticated applications of thermodynamics in geochemistry, but it serves as a step toward a broader appreciation that compositional changes are not random, but are aligned with environmental conditions. That should motivate the development of more rigorous thermodynamic models in future studies.

As a partial response to the request for a more mechanistic understanding, it can be noted that Fig. 7 has been redrawn to place time on the horizontal axis. With this change, it should be more apparent that the chemical composi-

tion of the differentially expressed proteins changes dynamically in laboratory experiments.

2) I strongly suggest the authors attempt to explain their observations at a level that makes sense to the more biologically oriented reader.

The paper uses some technical language from physical chemistry and thermodynamics by necessity, and these technical terms are defined when introduced. These concepts are used to quantitatively analyze metagenomic datasets that are chosen to represent well-known regional gradients. The analysis of laboratory data includes protein expression in response to salt and osmotic shock. Therefore, the core of the paper is concerned with biological phenomena in an environmental context. The mixing of biological data and physicochemical metrics is what makes this paper unique; removing the quantitative language would eliminate its main contribution.

We note that the entire section on "Conceptual background" was added in a previous revision (before submission to this journal) to make the paper more accessible to biologists. The paragraphs here deal with issues about intracellular conditions, amino acid composition, distinction with polymerization reactions, selection for structural stability of proteins, other variables like temperature and pH, and relation of the basis species to biosynthetic mechanisms. However, our intention is not to write a theoretical paper but rather to present a coherent set of data analyses to convince the reader that compositional differences of proteins have a basic significance in geobiochemical systems.

3) I could not help but think to myself how any one or several observations made sense from the level of phenotype and natural selection. The authors might consider asking themselves this same question and then speculating where possible to make this body of work a greater utility for the community.

We believe that the analysis of laboratory experiments of protein expression in salt and osmotic conditions does provide basic information about the effects of the environment on the observable characteristics of cells. Admittedly, this is only one aspect of the phenotype, and other types of experiments could be considered, like gene expression, metabolomes, and metabolic fluxes, but analysis of those types of data is out of the scope of this paper.

A relevant finding from a paper in preparation is that the stoichiometric hydration state of differentially expressed proteins is strongly decreased in 3D (tissue-like) compared to 2D (monolayer) culture conditions of eukaryotic cells (Dick, 2020). The lower $n_{\mathsf{H_2O}}$ in 3D culture has some similarity to the observation in this study that metagenome-inferred proteins in particles tend to have lower hydration state compared to free-living fractions. These responses could plausibly be associated with lower water accessibility in the interiors of particles in environmental samples and in spheroids in 3D cell culture.

Regarding the evolutionary implications, another paper is in early preparation that shows the hydration and oxidation state computed for whole proteomes of phylogenetic groups predicted from the RefSeq database. This tree-like view of the chemical composition no doubt would help solidify the relevance of the

physicochemical concepts used here to biological systems. That is being developed for a separate paper with its own set of data analysis of microbial community composition and it is too early to cite the results in this paper.

The following list of minor comments is meant to further improve this work: Line 1: For the average reader – what is the connection between thermodynamics and environmental variation. Lead in with this first.

Added text in Abstract: "Prediction of the direction of change of a system under specified environmental conditions is one reason for the widespread utility of thermodynamic models in geochemistry."

Line 8: Replace "behave" with something more valid. The metric does not correlate for XXX in hypersaline environments. . .

Changed "behave" to "respond".

Line 15: Communities do not adapt, populations of individuals do.

Changed "communities" to "populations".

Line 26: I would not call this complementary but rather an interrelated approach since selection (imposed as an argument in previous paragraph) can and should act on the energetic demand of protein synthesis.

Changed "complementary" to "interrelated".

Line 39-40: What about the authors own work on the communities inhabiting the out flow channel at Bison Pool, Yellowstone?

Added references and reworded the sentence for better context: "The oxidation state of proteins as well as lipids has been shown to be associated with oxidation-reduction (redox) gradients in a hot spring (Dick and Shock, 2011; Boyer et al., 2020), but so far energetic models have not been broadly adopted as a tool for relating metagenomic and geochemical data."

Line 44-45: While I don't disagree with this assumption, at least as a first order constraint, it would be useful to relate to the reader why this assumption is made. Perhaps to avoid this confusion, the authors move this statement to below where they describe and justify their approach.

This sentence has been moved down to the second point in the "Conceptual background" section, following the reference about missing hydrogen and oxidation state in stoichiometric models (Karl and Grabowski, 2017).

Line 58-62: This paragraph seems out of place. I suggest moving the discussion of what you did previously up in the introduction and add the last sentence of this paragraph to the end of the preceding paragraph.

The statement of previous work and what's new in this study has been moved up to the position of the former Lines 44-45 mentioned in the previous comment. The long-term research goal has been removed, because it doesn't seem to fit anywhere now.

Line 67: alternatives to what?

Each area of concern is summarized here as "X or Y", which seems consistent with the dictionary's definition of an alternative as "a choice between two

things". To avoid confusion, this has been reworded as "six areas of concern summarized as: 1) ... 2) ... ..."

Line 305-310: I don't understand the reasoning here? Why did eukaryotes start to become important in these systems? Are there actually eukaryotes in these systems? The authors have the data to evaluate this and should evaluate it to see if the logic makes sense.

This has been removed in the revision. The comparison of the average stoichiometric hydration state of human proteins with *E. coli* and the metagenomic data analyzed in this study provided preliminary support for the concept of a lower $n_{\mathrm{H_2O}}$ in eukaryotes, but a more targeted data analysis is needed to strengthen this claim. Also note that the human and *E. coli* proteomes have been removed in the revised description of the choice of basis species (Fig. 1).

Line 315: Why would heterotroph proteomes have a lower hydration state?

There might be something basically different about their metabolic pathways in terms of water requirements at the biochemical level. Apart from *E. coli*, there probably are not many existing metabolic models that could be used to test this speculation. Added sentence: "A better understanding of these trends would require more extensive phylogenetically resolved comparisons of the compositional differences as well as biochemical (or computational) analyses of water fluxes in metabolic pathways."

Line 315-317: is there an argument to be made about why a major evolutionary transition favors a shift from higher to lower dehydration state? i.e., is this an adaptive feature that allows the latter to compete with the former from an evolutionary perspective?

This is certainly a valid question, but we are unable to provide a convincing mechanistic reason for why lower hydration state might offer a selective advantage. Perhaps it should be considered not as adaptation but as physical constraint, similar in a way to Gould and Lewontin (1979)'s spandrels. Structures that are physically durable, such as macromolecular complexes in organelles or larger assemblages like tissues, might be those that are relatively dry. Physical dryness (i.e. lower water content) could be a selective force for lower stoichiometric hydration state of biomolecules, but the latter by itself may have no fitness advantage.

If lower $n_{\mathrm{H_2O}}$ turned out to characterize some evolutionary transitions, it would seem to be consistent with the postulate that "ontogeny recapitulates phylogeny" and the observation that progressive loss of water occurs in animal development through the stages of embryo, fetus, birth and growth (Moulton, 1923).

[These ideas are rather speculative, and don't specifically deal with the (non-eukaryotic) metagenomes that are analyzed here, so haven't been added to the text.]

Line 325: is it possible that diffusion limitation makes H2O less available to cells living nearer to a particle surface? Again, an explanation for what the observations might mean is warranted.

Particles likely provide opportunities for some amount of physical separation

from the bulk aqueous phase; it's harder to pin down the molecular mechanisms. Added: "Together with the lower n_{H_2O} for proteins inferred from metagenomes and metatranscriptomes in the larger size fractions from Baltic Sea samples, this could reflect a lower availability of H_2O to organisms living near the particle surface due to physical separation from the bulk aqueous phase and associated diffusion limitation or lower water activity (Wang et al., 2003)."

Line 350: proteins in metagenomes

Changed "plume metagenomes" to "proteins in plume metagenomes".

Line 360: Could this be due to aquaculture and introduction of more organic compounds/waste and its selection of heterotrophic taxa, that as stated earlier in the paper, tend to host proteomes with a lower hydration state

This seems very reasonable. Added: "The microbial communities in the aquaculture ponds may not be responding as they would in a typical natural system that is less nutrient-rich. As noted above for putative heterotroph-rich zones in other systems, the lower stoichiometric hydration state could be associated with the enrichment of heterotrophic taxa, in this case due to the addition of organic compounds to the aquaculture ponds."

See also the response to Referee #1 and the revised discussion of the analysis of differentially expressed proteins: "The negative shift of Δn_{H_2O} associated with most organic solutes compared to NaCl lends support to the notion that high organic loading could contribute to the relatively low n_{H_2O} of protein sequences from metagenomes of freshwater aquaculture systems."

Uncovering chemical signatures of salinity gradients through compositional analysis of protein sequences

Jeffrey M. Dick[1,2], Miao Yu[1], and Jingqiang Tan[1]

[1]Key Laboratory of Metallogenic Prediction of Nonferrous Metals and Geological Environment Monitoring, Ministry of Education, School of Geosciences and Info-Physics, Central South University, Changsha 410083, China
[2]State Key Laboratory of Organic Geochemistry, Guangzhou Institute of Geochemistry, Chinese Academy of Sciences, Guangzhou 510640, China

Correspondence: J. M. Dick (jeff@chnosz.net) or M. Yu (yumiao1987@pku.edu.cn)

Abstract. Prediction of the direction of change of a system under specified environmental conditions is one reason for the widespread utility of thermodynamic models in geochemistry. However, thermodynamic influences on the chemical compositions of proteins in nature have remained enigmatic despite much work that demonstrates the impact of environmental conditions on amino acid frequencies. Here, we present evidence that the dehydrating effect of salinity is detectable as chemical differences in protein sequences inferred from 1) metagenomes and metatranscriptomes in regional salinity gradients and 2) differential gene and protein expression in microbial cells under hyperosmotic stress. The stoichiometric hydration state (n_{H_2O}), derived from the number of water molecules in theoretical reactions to form proteins from a particular set of basis species (glutamine, glutamic acid, cysteine, O_2, H_2O), decreases along salinity gradients including the Baltic Sea and Amazon River and ocean plume and in particle-associated compared to free-living fractions. However, the proposed metric does not ~~behave~~respond as expected for hypersaline environments. Analysis of data compiled for hyperosmotic stress experiments under controlled laboratory conditions shows that differentially expressed proteins~~, as well as proteins coded by differentially expressed transcripts,~~ are on average shifted toward lower n_{H_2O}. Notably, the dehydration effect is stronger for most organic solutes compared to NaCl. This new method of compositional analysis can be used to identify possible thermodynamic effects in the distribution of proteins along chemical gradients at a range of scales from ~~biofilms~~microbial mats to oceans.

1 Introduction

How microbial ~~communities~~populations adapt to environmental gradients is a major challenge at the intersection of geochemistry, microbiology, and biochemistry. Patterns of amino acid usage in proteins are important indicators of microbial adaptation, and amino acid composition at the genome level is well known to depend on growth temperature (Zeldovich et al., 2007). Furthermore, measures of evolutionary distance and community composition based on protein sequences predicted from metagenomic sequencing are strongly associated with environmental temperature and pH (Alsop et al., 2014). It is widely acknowledged that the effect of amino acid substitutions on the structural stability of proteins is a major factor affecting amino acid usage in thermophiles (Sterner and Liebl, 2001; Zeldovich et al., 2007). Similarly, a large body of work has demonstrated amino acid signatures associated with proteins from halophilic organisms (Kunin et al., 2008; Paul et al., 2008; Oren, 2013;

Boyd et al., 2014). The most common interpretation of these trends is that particular amino acid substitutions are selected through evolution to increase the stability and solubility of the folded conformation and enhance other structural properties such as flexibility (Paul et al., 2008).

An ~~complementary~~interrelated approach to interpreting patterns of amino acid composition is based on the energetics of amino acid synthesis. Energetic costs in terms of ATP requirements have been used to model protein expression levels in bacterial and yeast cells (Akashi and Gojobori, 2002; Wagner, 2005). Although ATP demands depend on environmental conditions (Akashi and Gojobori, 2002), a limitation of ATP-based models is that they are derived for specific biosynthetic pathways, such as whether cells are grown in respiratory or fermentative (i.e. aerobic or anaerobic) conditions (Wagner, 2005). A different class of models, based on thermodynamic analysis of the overall Gibbs energy of reactions to synthesize metabolites from inorganic precursors, quantifies the energetics of the reactions in terms of temperature, pressure, and chemical activities of all the species in the reactions, including those that define pH and oxidation-reduction potential (Shock et al., 2010). Notably, the overall ~~energetics~~Gibbs energies for amino acid synthesis become more favorable, but to a different extent for each amino acid, between cold, oxidizing seawater and hot, reducing hydrothermal solution (Amend and Shock, 1998). A recent systems biology study demonstrates tradeoffs between Gibbs energy of alternative pathways for amino acid synthesis and cofactor use efficiency (which affects ATP costs) in the model organism *Escherichia coli* and suggests that pathway thermodynamics play a role in thermophilic adaptation (Du et al., 2018). ~~Nevertheless~~The oxidation state of proteins as well as lipids has been shown to be associated with oxidation-reduction (redox) gradients in a hot spring (Dick and Shock, 2011; Boyer et al., 2020), but so far energetic models have not ~~made much headway in~~been broadly adopted as a tool for relating metagenomic and geochemical data. This may be because few studies have asked whether specific changes in the chemical composition of biomolecules reflect specific environmental conditions.

To help close this gap, here we use compositional analysis of protein sequences to identify chemical signatures of two types of environmental conditions: redox and salinity gradients. ~~Because redox reactions are inherent in many aspects of metabolism, while hydration and dehydration reactions are essential for the synthesis of biomacromolecules (Braakman and Smith, 2013), our approach is shaped by the assumption that O_2 and H_2O are two primary components that link environmental conditions to the energetics of biomolecular synthesis.~~In a previous study (Dick et al., 2019), we compared one broad class of geochemical conditions (redox gradients) with one compositional metric for proteins (carbon oxidation state). Here, we expand the geobiochemical framework to two dimensions by considering another set of environments (salinity gradients) and another compositional metric (stoichiometric hydration state). Thermodynamic considerations predict that redox gradients supply a driving force for changes in the oxidation state of biomolecules (similar reasoning applies to the oxygen content of proteins; Acquisti et al., 2007), while salinity gradients, through the dehydrating potential associated with osmotic effects, exert a force that selectively alters the hydration state of biomolecules.

To test these predictions, we used two compositional metrics, the carbon oxidation state (Z_C) and stoichiometric hydration state (n_{H_2O}). Z_C is computed from the chemical formulas of organic molecules, and takes values between the extremes of -4 for CH_4 and +4 for CO_2, although the range for particular classes of biomolecules is much smaller (Amend et al., 2013). n_{H_2O} is derived from the number of water molecules in theoretical formation reactions of proteins from basis species (Dick,

2016, 2017). Through the compositional analysis of representative metagenomic and metatranscriptomic datasets, we show that Z_C and n_{H_2O} are most closely aligned with environmental redox and salinity gradients, respectively. These findings apply to freshwater and marine environments, but trends for hypersaline environments deviate from the thermodynamic predictions, most likely due to evolutionary optimizations of hydrophobicity and isoelectric point to stabilize the structures of proteins in halophilic organisms.

~~In a previous study (Dick et al., 2019), we compared one broad class of geochemical conditions (redox gradients) with one compositional metric for proteins (carbon oxidation state). Here, we expand the geobiochemical framework to two dimensions by considering another set of environments (salinity gradients) and another compositional metric (stoichiometric hydration state). A long-term research goal is to extend this framework to as many dimensions as there are thermodynamic components plus temperature and pressure.~~

2 Conceptual background

In this study we use compositional analysis to uncover environmental imprints in protein sequences. Analysis of compositional data is used by geochemists to study processes such as water-rock interaction and ore deposition, and is often one of the first steps in constructing thermodynamic models, but its application to living systems is relatively uncommon. Therefore, it is important to describe the conceptual basis for our methods. To do this, we identified six areas of concern ~~posed~~summarized as ~~alternatives~~: 1) intracellular or environmental conditions, 2) amino acids or atoms, 3) condensation or theoretical formation reactions, 4) chemical composition or conformational stability, 5) oxidation and hydration state or temperature and pH, and 6) mathematical or biosynthetic models.

A first concern is that intracellular conditions are maintained within physiological ranges, so the influence of external conditions on the composition of microbial biomolecules may be limited. However, cell membranes are permeable to uncharged species such as hydrogen (Slonczewski et al., 2009), supporting the argument that the oxidation state of the cytoplasm, and therefore the energetics of metabolic reactions, are influenced by the external environment (Poudel et al., 2018; Canovas and Shock, 2020). Likewise, oxygen diffuses rapidly through lipid membranes, depending on their composition and structure, and rates of diffusion increase with temperature (Möller et al., 2016). Cell membranes are also permeable to water (Record et al., 1998). For *E. coli*, which grows most rapidly at about 0.3 OsM (osmolarity), increasing the extracellular osmotic strength from 0.1 to 1.0 OsM ~~[~~(approximately the osmotic concentration of seawater; BioNumbers BNID 100802 (Milo et al., 2010)~~]~~) reduces the amount of free cytoplasmic water by more than half (Record et al., 1998). Halophiles, which thrive at even higher salinities, accumulate inorganic salts or organic solutes to maintain osmotic balance with the environment (Garner and Burg, 1994; Oren, 2013). The result is that, with few exceptions, intracellular conditions must be isosmotic with the environment, or somewhat higher to maintain turgor pressure (Gunde-Cimerman et al., 2018). Water activity is lower in more concentrated solutions, and intracellular water activity estimated from freezing point and cell composition data closely follows that of the growth medium, but is often offset to lower values (Chirife et al., 1981), perhaps due to macromolecular crowding effects (Gar-

ner and Burg, 1994). ~~In other words~~To summarize, high osmotic strength causes a decrease in hydration potential, measured as water activity, both outside and inside cells.

This brief review suggests that oxidation and hydration potentials in cell interiors, at least under experimental conditions, are influenced by, but not equal to, environmental conditions. Ideally, we would like to compare the compositions of biomolecules to conditions actually measured inside cells or in the immediate surroundings of cells, but these measurements are generally not available for microbial communities in their natural environments, so we make comparisons with large-scale geochemical gradients, except for different layers of the Guerrero Negro microbial mat, where metagenomic and chemical data are available on the scale of millimeters.

Second, previous authors have emphasized the importance of changes in elemental stoichiometry – that is, atomic composition – and not only amino acid composition in the molecular evolution of proteins (Baudouin-Cornu et al., 2001). Although stoichiometric predictions are amenable to experimental tests, such as the long-term evolution of ~~*Escherichia*~~*E. coli* in the laboratory (Turner et al., 2017), the omission of a major bioelement, hydrogen, and the oxidation state of organic matter from most stoichiometric models (Karl and Grabowski, 2017) means that there are also significant opportunities for theory development. Because redox reactions are inherent in many aspects of metabolism, while hydration and dehydration reactions are essential for the synthesis of biomacromolecules (Braakman and Smith, 2013), our approach is shaped by the assumption that O_2 and H_2O are two primary components that link environmental conditions to the energetics of biomolecular synthesis.

The third point follows from the previous one. The polymerization of amino acids is a condensation reaction that releases one H_2O per bond formed, independent of the particular amino acids that are involved. By contrast, our analysis depends crucially on the concept of a "formation reaction", which in the thermodynamic literature represents the composition of a chemical species, either in terms of elements (Warn and Peters, 1996), or in terms of other species (May and Rowland, 2018). When these other species are restricted in number to the minimum needed to represent the composition of all possible species in the system, they constitute a set of "basis species", which can be thought of as the building blocks of the system, similar to the concept of thermodynamic components (Anderson, 2005). Therefore, a formation reaction from basis species is a mass-balanced, but non-unique, stoichiometric representation of the chemical composition of the protein. This type of reaction in general does not correspond to amino acid biosynthesis or polymerization, so to avoid confusion, we refer to these formation reactions as "theoretical formation reactions"; the number of water molecules in the theoretical formation reactions, normalized by the protein length, is the "stoichiometric hydration state".

From a mechanistic standpoint, an analysis using any set of basis species is inadequate, since the number of basis species (five, corresponding to the elements C, H, N, O, and S) is smaller than the number of biochemical precursors and inorganic species that are actually involved in amino acid synthesis (Du et al., 2018). The use of O_2, H_2O, and other basis species to represent the composition of proteins reflects the hypothesis that they are conjugate to thermodynamically meaningful descriptive variables (specifically, chemical potentials) even if they are not directly involved in the biosynthetic mechanisms for amino acids. The projection of amino acid composition (20-D) into the compositional space represented by basis species (5-D) is a type of dimensionality reduction, but the variables are chosen based on a physicochemical hypothesis, unlike principal components analysis (PCA) or other unsupervised methods, where the projection is determined by the data.

A fourth concern is that this analysis is based on the hypothesis that thermodynamic forces affect the chemical compositions of proteins over evolutionary time, which is different from the more common hypothesis of optimization of structural stability. Thermodynamic models define the "cost" of a protein as a function of not only amino acid composition but also environmental conditions. Conceptually, this follows from Le Chatelier's principle, in that increasing the chemical activity of a reactant (on the left-hand side of a reaction) drives the reaction toward the products.~~, or~~Stated in more general terms, ~~that~~ the overall Gibbs energy of a reaction depends on the activities of species in the reaction (Shock et al., 2010; Amend and LaRowe, 2019). Consider two proteins with different amino acid compositions, and therefore also different chemical compositions and theoretical formation reactions, which should be normalized by the number of residues in order to compare proteins of different length. The formation of the protein with more water as a reactant is theoretically favored by increasing the water activity, whereas the formation of the protein with more oxygen as a reactant is favored by increasing the oxygen activity. The water and oxygen activity are thermodynamic measures of hydration and oxidation potential and can be converted to other scales, such as oxidation-reduction potential (ORP).

This reasoning provides the theoretical justification for using chemical composition as an indicator of molecular adaptation to specific environmental conditions, but does not replace interpretations based on structural considerations. Halophilic organisms exhibit well-documented patterns of amino acid usage, including lower hydrophobicity and higher abundance of acidic residues, that impart greater stability, solubility, and flexibility of proteins (Paul et al., 2008). These adaptations are reflected in lower values of the GRAVY hydrophobicity scale (Paul et al., 2008; Boyd et al., 2014) and/or isoelectric point of proteins (pI) (Oren, 2013). In Sect. 4.3 and 4.4, we compare the compositional metrics with GRAVY and pI for the same datasets.

Fifth, temperature, pH, and other environmental parameters besides redox and salinity might influence the oxidation and hydration state of proteins. For instance, the redox gradients in hydrothermal systems are also temperature gradients, due to the mixing of seawater and hydrothermal fluid, and we have not attempted to disentangle the effects of temperature and redox conditions. However, our previous analysis of other redox gradients, including stratified hypersaline lakes, indicates that carbon oxidation state of biomolecules can vary even in systems where temperature changes are much smaller (Dick et al., 2019). It is an axiomatic statement that changes in oxidation state can be associated with one thermodynamic component of a system; our objective in the present study is to explore the differences between this and one other component, represented by hydration state. Future work should also account for the effects of pH and temperature, which is possible using thermodynamic models for proteins (Dick and Shock, 2011).

Finally, it should be noted that the basis species used in the stoichiometric analysis are chosen primarily for mathematical convenience, not because of evolutionary or biosynthetic requirements. ~~The basis species we use for deriving the stoichiometric hydration state of proteins are cysteine, glutamine, glutamic acid, O_2, and H_2O (designated "QEC"). The primary reason for choosing these~~The main criterion we consider for the choice of basis species is to reduce the covariation between the metrics for oxidation and hydration state~~; that covariation is~~, which arises as a mathematical consequence of projecting the atomic formulas of proteins into a particular compositional space, and may not reflect meaningful differences of chemical composition. ~~There is nothing implied by the choice of basis species about evolutionary or biosynthetic mechanisms, and any set of basis species is thermodynamically valid, as long as they are the minimum number needed to represent the chemical composition of all the~~

~~species in the system (Anderson, 2005). However, it is most convenient to select basis species that correspond to the controlling variables of the system. The QEC basis species has a biological rationale since glutamine and glutamic acid are often identified as highly abundant metabolites and have been characterized as "nodal point" metabolites (Walsh et al., 2018). Other~~Additional considerations are described in Sect. 3.2.

3 Methods

3.1 Carbon oxidation state

The most common metric used in geochemistry for the oxidation state of organic molecules is the average oxidation state of carbon (Z_C), which also goes by other names such as nominal oxidation state of carbon (NOSC) (LaRowe and Van Cappellen, 2011). This quantity measures the average degree of oxidation of carbon atoms in organic molecules. For a protein for which the primary sequence has the chemical formula $C_cH_hN_nO_oS_s$, the value of Z_C can be calculated from (Dick and Shock, 2011; Dick, 2014)

$$Z_C = \frac{-h + 3n + 2o + 2s}{c} \tag{1}$$

The derivation of Eq. (1) is based on the relative electronegativities of the elements, expressed as oxidation numbers (e.g. Kauffman, 1986; Minkiewicz et al., 2018). When bonded to carbon, H is assigned an oxidation number of +1, and N, O, and S have oxidation numbers of -3, -2, and -2. Eq. (1) gives the remaining charge that must be present on each C atom, on average, to satisfy overall neutrality. Because of the relatively simple structures of amino acids and the primary structure of proteins, in which N, O, and S are bonded to only H and C, it is possible to calculate the average oxidation state of carbon using Eq. (1). However, this equation is not necessarily valid for other classes of organic molecules or some types of post-translational modifications of proteins, including the formation of disulfide bonds. An important relation ~~given by~~ inherent in Eq. (1) is the redox neutrality of hydration and dehydration reactions; any pair of hypothetical (or real) proteins whose formulas differ only by some amount of H_2O have ~~identical~~ equal carbon oxidation states.

3.2 Choice of basis species: theoretical considerations

A major premise of this study is that oxidation state and hydration state are two primary variables in geobiochemical systems. Accordingly, when choosing the basis species that can be combined to make the proteins, O_2 and H_2O are the only fixed requirements. This leaves three basis species that when combined with each other and with O_2 and H_2O must be able to give any possible formula written as $C_cH_hN_nO_oS_s$. ~~Note again~~ We reiterate that this analysis refers to the chemical formulas of polypeptide sequences, that is, the primary structure of proteins, not post-translational modifications or H_2O molecules in the hydration shell of folded proteins.

Eq. (1) is derived from electronegativity relations and therefore allows the calculation of the carbon oxidation state from a given chemical formula, independent of any chemical reactions. In contrast, there is no way to count the number of H_2O molecules in a chemical formula; H_2O appears only in chemical reactions. But it is important to note that any particular reaction

that involves only H_2O is redox-neutral. Conversely, the coefficient of O_2 in redox reactions is closely related to the number of electrons transferred. Let us consider the 20 protein-forming amino acids as a baseline for compositional analysis; the numbers of H_2O and O_2 in the formation reactions of the amino acids from a particular set of basis species are denoted by n_{H_2O} and n_{O_2}. The choice of basis species in our study is guided by the dual objectives that 1) n_{H_2O} of amino acids should have very little correlation with Z_C and 2) n_{O_2} of amino acids should be strongly correlated with Z_C. It should be emphasized that these are not criteria for "correctness", since basis species, like thermodynamic components, only have to be the minimum number needed to represent the chemical composition of all the species that can be formed from them (Anderson, 2005). Instead, basis species selected using these conditions yield a convenient mathematical projection of elemental composition; that is, nearly horizontal or vertical trends on n_{H_2O}–Z_C scatterplots for proteins from environmental gradients specifically reflect changes in oxidation state or hydration state, respectively. ~~Extrapolation of this principle to the general case gives the criterion that a metric for hydration state should be disconnected from redox effects. In other words, when applied to a population of target molecules, such as all the proteins in a genome, the correlation between metrics for oxidation state and hydration state should be minimized.~~

~~Accordingly, we aim to find a projection of the elemental composition of primary protein sequences that clearly separates Z_C and the stoichiometric number of H_2O. There are no thermodynamic restrictions on the choice of basis species, but~~ An additional consideration is that a biologically meaningful set of basis species is likely to comprise metabolites that have high network connectivity, that is, are involved in reactions with many other metabolites. Reactions involving glutamine and glutamic acid, or its ionized form, glutamate, are major steps of nitrogen metabolism (Morowitz, 1999; DeBerardinis and Cheng, 2010), and these amino acids have been characterized as "nodal point" metabolites (Walsh et al., 2018). Either methionine or cysteine would provide the sulfur required for the system, but cysteine is relevant as a constituent of the glutathione molecule, which has important roles in cellular redox chemistry (Walsh et al., 2018). These considerations support the proposal of the amino acids glutamine, glutamic acid, and cysteine (collectively abbreviated QEC) together with O_2 and H_2O as a biologically relevant set of basis species for describing the chemical compositions of proteins (Dick, 2016). These three amino acids are among the top eight amino acids ranked by number of reactions in a metabolic model for ~~*Escherichia*~~*E. coli* (Feist et al., 2007) (~~Glu~~E: 52, ~~Ser~~S: 25, ~~Asp~~D: 23, ~~Gln~~Q: 18, ~~Ala~~A: 15, ~~Gly~~G: 15, ~~Met~~M: 15, ~~Cys~~C: 13).

3.3 ~~Derivation of stoichiometric hydration state~~Choice of basis species: stoichiometric analysis

Here we compute the stoichiometric hydration state by analyzing the compositions of the 20 proteinogenic amino acids in detail. ~~Using the~~We start with a "default" set of basis species chosen for their common occurrence in overall catabolic reactions (Amend and LaRowe, 2019): CO_2, NH_3, H_2S, H_2O, and O_2. Using these basis species (designated CHNOS), the theoretical formation reaction of alanine ($C_3H_7NO_2$) is

$$3CO_2 + 2H_2O + NH_3 \rightarrow C_3H_7NO_2 + 3O_2 \quad \text{(R1)}$$

and the oxygen and water content of the amino acid (i.e, n_{O_2} = -3 and n_{H_2O} = 2) are the opposite of the coefficients on O_2 and H_2O in the reaction. ~~Similar~~Analogous reactions for the other amino acids were used to make Fig. 1a–b. Using glutamine

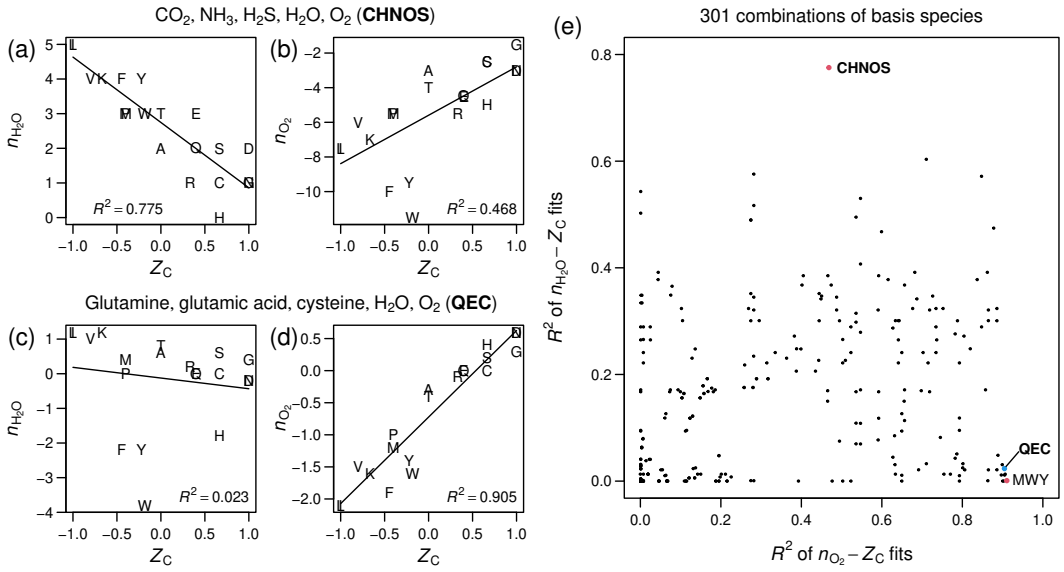

Figure 1. Stoichiometric ~~values~~numbers of H_2O and O_2 for theoretical formation reactions of amino acids computed with different sets of basis species, plotted against carbon oxidation state (Z_C), which is computed from the elemental formula and does not depend on the choice of basis species. Linear regressions and R^2 values were calculated using the `lm` function in R (R Core Team, 2020). **(a–b)** CO_2, NH_3, H_2S, H_2O, O_2 (CHNOS). **(c–d)** Glutamine, glutamic acid, cysteine, H_2O, O_2 (QEC). **(e)** Scatterplot of R^2 values for n_{H_2O}–Z_C fits against R^2 values for n_{O_2}–Z_C fits for all combinations of basis species consisting of H_2O, O_2 and three amino acids (including the points labeled QEC and MWY (methionine, tryptophan, tyrosine)), or CO_2, NH_3, H_2S, H_2O, and O_2 (CHNOS).~~(CHNOS and QEC) and derivation of the residual correction (rQEC). **(a–b)** Number of H_2O and O_2 in the theoretical formation reactions of amino acids from CO_2–NH_3–H_2S–H_2O–O_2 (CHNOS) are plotted against carbon oxidation state (Z_C), which is also computed from the chemical formula but does not depend on the choice of basis species. Linear models and R^2 values were calculated using the `lm` function in R (R Core Team, 2020). **(c–d)** Changing the basis species to glutamine–glutamic acid–cysteine–H_2O–O_2 (QEC) strengthens the association between Z_C and n_{O_2} and decreases that between Z_C and n_{H_2O}. However, there is still a noticeable negative correlation between Z_C and n_{H_2O}, which is also visible in scatterplots of all proteins in **(e)** *H. sapiens* and **(f)** *E. coli* K12 [UniProt reference proteomes UP000005640 and UP000000625 (The UniProt Consortium, 2019)]. **(g)** Residuals from the linear model in **(d)** minus a constant of 0.355 yield values for the stoichiometric hydration state (rQEC) of amino acids. **(h–i)** Stoichiometric hydration states of proteins calculated with the rQEC values. The constant was defined so that the mean n_{H_2O} for human proteins equals zero.~~

Table 1. Values of stoichiometric hydration state (n_{H_2O}) of amino acids ~~residues~~ calculated with the ~~rQEC~~QEC ~~derivation~~basis species (glutamine, glutamic acid, cysteine, H_2O, O_2) and average oxidation state of carbon (Z_C) and number of carbon atoms (n_C). Standard one-letter abbreviations for the amino acids (AA) are used.

AA	n_{H_2O}	Z_C	n_C	AA	n_{H_2O}	Z_C	n_C
A	0.6	0	3	M	0.4	-2/5	5
C	0.0	2/3	3	N	-0.2	1	4
D	-0.2	1	4	P	0.0	-2/5	5
E	0.0	2/5	5	Q	0.0	2/5	5
F	-2.2	-4/9	9	R	0.2	1/3	6
G	0.4	1	2	S	0.6	2/3	3
H	-1.8	2/3	6	T	0.8	0	4
I	1.2	-1	6	V	1.0	-4/5	5
K	1.2	-2/3	6	W	-3.8	-2/11	11
L	1.2	-1	6	Y	-2.2	-2/9	9

($C_5H_{10}N_2O_3$), glutamic acid ($C_5H_9NO_4$), cysteine ($C_3H_7NO_2S$), H_2O, and O_2 (the QEC basis species), the theoretical formation reaction of alanine is

$$0.4C_5H_{10}N_2O_3 + 0.2C_5H_9NO_4 + 0.6H_2O \rightarrow C_3H_7NO_2 + 0.3O_2 \quad \text{(R2)}$$

showing that the oxygen and water content are n_{O_2} = -0.3 and n_{H_2O} = 0.6. Calculations for all the amino acids using the QEC basis were used to make Fig. 1c–~~f~~d.

As measured by R^2 in linear regressions, the CHNOS basis yields a strong negative correlation between Z_C and n_{H_2O} for the amino acids (Fig. 1a), but a relatively weak correlation between Z_C and n_{O_2} (Fig. 1b). The QEC basis provides a ~~much~~ stronger association between Z_C and n_{O_2} and ~~greatly~~ reduces the correlation between Z_C and n_{H_2O} (Fig. 1c–d). However, there is still a small negative correlation for amino acids (Fig. 1~~d~~c). A plot with the R^2 values for all possible combinations of H_2O, O_2, and 3 amino acids indicates that QEC has relatively low R^2 of n_{H_2O}–Z_C and high R^2 of n_{O_2}–Z_C (Fig. 1e). Therefore, it is a suitable candidate to meet the objectives described above. Although another combination of amino acids – methionine, tryptophan, and tyrosine (MWY) – has even lower R^2 for the n_{H_2O}–Z_C fit (Fig. 1e), tryptophan and tyrosine are not highly connected metabolites and therefore are less preferable as basis species.

~~which is also visible in whole-proteome data for humans and *E. coli* (Fig. 1e–f). We calculated residual-corrected values of n_{H_2O} by taking the residuals of a linear model for amino acids (Fig. 1d), then subtracting a constant, defined such that the mean n_{H_2O} for all human proteins equals zero. This derivation, which we refer to as "rQEC", gives the residual-corrected stoichiometric hydration state for each amino acid, which is plotted in Fig. 1g and listed in Table 1. Even with the residual correction for amino acids, there remain slightly positive and negative correlations for human and *E. coli* proteins (Fig. 1h–i).~~

~~As noted above, the mean n_{H_2O} for human proteins was defined to be zero; the mean for proteins in *E. coli* is somewhat greater, at 0.014.~~

By strengthening the association between Z_C and n_{O_2}, which ~~can both be interpreted as~~ represent alternative metrics for oxidation state, and reducing the correlation between Z_C and n_{H_2O}, the QEC basis species provides a more convenient projection of ~~chemical~~elemental composition than a "default" choice of inorganic species, such as CO_2, NH_3, H_2S, H_2O, and O_2, which commonly appear in overall catabolic reactions (Amend and LaRowe, 2019). The selection of basis species is an evolving method, and further analysis with other metabolites may lead to a more convenient set of basis species to project the elemental composition of proteins into chemical variables.~~Furthermore, the residual correction allows the identification of horizontal or vertical trends on n_{H_2O}–Z_C scatterplots to be associated with changes in only oxidation state or hydration state, respectively.~~

3.4 Compositional metrics for proteins and metagenomes

For a given protein, the stoichiometric hydration state was calculated ~~by taking the sum of (number of each amino acid multiplied by the respective value of n_{H_2O} in Table 1), then dividing the result by the number of amino acids.~~from

$$n_{H_2O} = \frac{\sum n_i \left(n_{H_2O,i} - 1\right)}{\sum n_i} + 1 \tag{2}$$

where n_i is the frequency of the ith amino acid (i = 1 to 20) in the protein and $n_{H_2O,i}$ is the stoichiometric hydration state of that amino acid (Table 1). The "-1" in the numerator accounts for the loss of H_2O in the polymerization of amino acids, and the "+1" after the fraction accounts for the N-terminal H and C-terminal OH of the polypeptide.

~~The average oxidation state of carbon was also calculated from the values for the amino acids [see Table 1 of Dick and Shock (2011)].~~

Unlike n_{H_2O}, ~~averages for~~ Z_C for proteins must be weighted by the number of carbon atoms in each amino acid, i.e.

$$Z_C = \frac{\sum n_i n_{C,i} Z_{C,i}}{\sum n_i n_{C,i}} \tag{3}$$

where $n_{C,i}$ and $Z_{C,i}$ are the number of carbon atoms and carbon oxidation state of the ith amino acid (see Table 1). For example, Z_C of the dipeptide Ala-Gly can be calculated as (3 × 0 + 2 × 1) / (3 + 2), where 3 and 2 are the numbers of carbon atoms and 0 and 1 are the Z_C of Ala and Gly, respectively. The result, 0.4, can be checked by applying Eq. 1 to the chemical formula of alanylglycine ($C_5H_{10}N_2O_3$). The methods for calculating n_{H_2O} and Z_C from elemental composition and amino acid composition are shown schematically in Fig. 2.

3.5 Amino acid composition of proteomes of Nif-bearing organisms

In a separate study, Poudel et al. (2018) used carbon oxidation state as a metric for comparing proteomes of organisms containing the nitrogenase gene (Nif). The evolution of these organisms is associated with rising atmospheric oxygen through geological history. In order to approximately replicate their results, amino acid compositions of all proteins for each bacterial, archaeal, and viral taxon in the NCBI Reference Sequence (RefSeq) database (O'Leary et al., 2016) were compiled from RefSeq release ~~95~~201 (July ~~2019~~2020). Scripts to do this, and the resulting data file of amino acid compositions of ~~36,425~~42,787

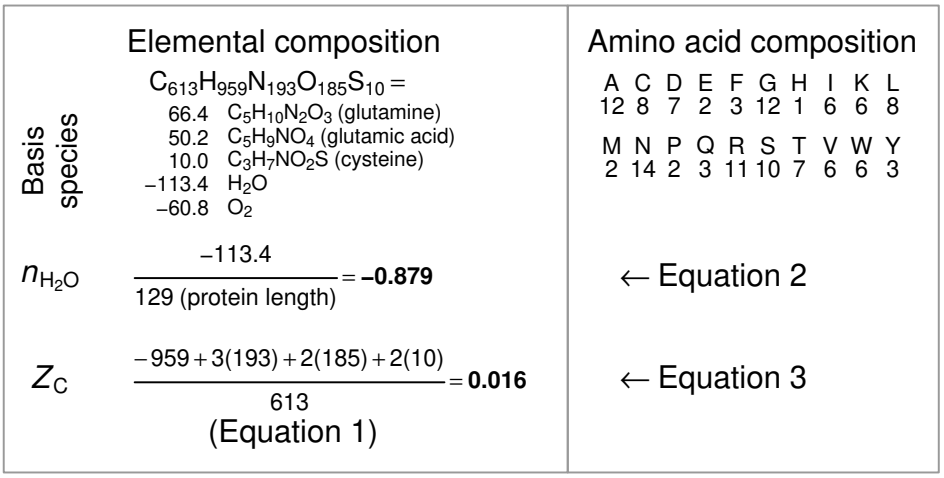

Figure 2. Schematic of calculations of n_{H_2O} and Z_C for a single protein. The selected protein is chicken egg white lysozyme (UniProt ID: LYSC_CHICK), which is historically an extensively characterized protein in the laboratory. The protein sequence was used to tabulate the amino acid composition (right column), which in turn was used to generate the elemental composition (left column). The coefficients on the basis species are determined from the elemental composition by mass-balance constraints. Dividing the number of H_2O in the basis species by the protein length gives the stoichiometric hydration state (n_{H_2O}). Independent of the basis species, the elemental composition yields the average oxidation state of carbon (Z_C) according to Eq. (1). To reduce computing steps, in this study the amino acid compositions of proteins (obtained e.g. from metagenomic sequences) were used to calculate n_{H_2O} and Z_C with Eqs. (2) and (3) and the values for amino acids in Table 1.

taxa, are available in the JMDplots R package (see *Code and data availability*). Names of organisms containing different nitrogenase (Nif) homologs were extracted from Supplemental Table 1A of Poudel et al. (2018). These names were matched to the closest organism name in RefSeq. Duplicated species (represented by different strains) were removed, as were matching organisms with fewer than 1000 RefSeq protein sequences. As a result, the numbers of organisms included in the present calculations (Nif-A: ~~157~~155, Nif-B: ~~69~~68, Nif-C: 14, Nif-D: 7) are less than those identified in Poudel et al. (2018). Note that values of Z_C calculated here (Fig. 3a) are lower than those shown in Fig. 5 of Poudel et al. (2018). This difference is associated with the weighting by carbon number (described above), which was not performed by Poudel et al. (2018).

3.6 GRAVY and pI

The grand average of hydropathicity (GRAVY) was calculated using published hydropathy values for amino acids (Kyte and Doolittle, 1982). The isoelectric point (pI) was calculated using published p*K* values for terminal groups (Bjellqvist et al., 1993) and sidechains (Bjellqvist et al., 1994); however, the calculation does not implement position-specific adjustments (Bjellqvist et al., 1994). The p*K* values used for calculating pI (Bjellqvist et al., 1993, 1994) and transfer free energies used in the derivation of the GRAVY scale (Kyte and Doolittle, 1982) correspond to 25 °C and 1 bar and no attempt was made here to account for the temperature effects on these properties. The charge for each ionizable group was precalculated from pH 0 to 14 at intervals of

0.01, and the isoelectric point was computed as the pH where the sum of charges of all groups in the protein is closest to zero. These calculations were implemented as new functions in the canprot R package (Dick, 2017) (see *Code and data availability*). Comparisons for selected proteins (UniProt IDs: LYSC_CHICK, RNAS1_BOVIN, AMYA_PYRFU) show that the calculated values of GRAVY and pI are equal to those obtained with the ProtParam tool (Gasteiger et al., 2005).

3.7 Prediction of protein sequences

Protein sequences were predicted from metagenomic reads using a previously described workflow (Dick et al., 2019). Briefly, reads were trimmed, filtered, and dereplicated using scripts adapted from the MG-RAST pipeline (Keegan et al., 2016). For metatranscriptomic datasets, ribosomal RNA sequences were removed using SortMeRNA (Kopylova et al., 2012). Protein-coding sequences were identified using FragGeneScan (Rho et al., 2010), and the amino acid sequences of the predicted proteins were used in further calculations. For large datasets, only a portion of the available reads was processed (at least 500,000 reads; see Supplementary Tables S1 and S2). This reduces the computational requirements without noticeably affecting the calculated average compositions (Dick et al., 2019).

Means and standard deviations of Z_C, n_{H_2O}, GRAVY, and pI were calculated for 100 random subsamples of protein sequences from each metagenomic or metatranscriptomic dataset. The numbers of sequences included in theeach subsamples werewas chosen to give a total length closest to 50,000 amino acids on average. The subsample density, or number of sequences included in each sample, depends on the average length of the metagenomic or metatranscriptomic sequences and is listed in Tables S1 and S2. This number ranges from 251 for the dataset with the highest mean protein fragment length (199.1; metagenome of hot-spring source of Bison Pool) to 1696 for the dataset with the lowest mean protein fragment length (29.5; metatranscriptome of site GS684 in the Baltic Sea).

4 Results and discussion

4.1 Comparison of redox and salinity gradients

To search for the hypothesized dehydration signal in metagenomic data, we began with redox gradients as a negative control. Submarine hydrothermal vents are zones of complex interactions between reduced endmember fluids and relatively oxidized seawater (Reeves et al., 2014; Ooka et al., 2019). Terrestrial hydrothermal systems, such as the hot springs in Yellowstone National Park, USA, provide a source of reduced fluids that are oxidized by degassing and mixing with air and surface groundwater as well as biological activity including sulfide oxidation (Lindsay et al., 2018). Redox gradients can also develop over smaller length scales. The surface of the Guerrero Negro microbial mat (Baja California Sur, Mexico) is exposed to ca. 1 m deep hypersaline, oxygenated water (approximately 200 μM O_2), but in the mat, oxygen rises during the daytime and is depleted within a few millimeters, giving way to anoxic, then sulfidic conditions (Ley et al., 2006).

Using metagenomic data for these redox gradients (Kunin et al., 2008; Havig et al., 2011; Swingley et al., 2012; Reveillaud et al., 2016; Fortunato et al., 2018), Dick et al. (2019) showed that the carbon oxidation states of DNA, messenger RNA, and

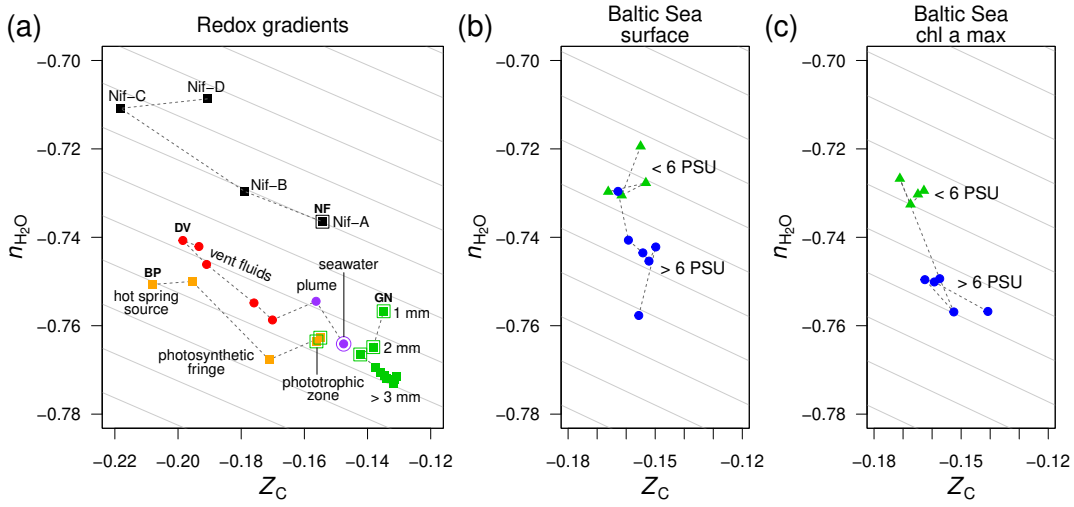

Figure 3. Compositional analysis of proteins in redox gradients and the Baltic Sea salinity gradient. **(a)** Redox gradients. Abbreviations and data sources: ~~are given in Fig. 3~~BP (Bison Pool hot spring; Havig et al., 2011; Swingley et al., 2012), DV (diffuse submarine vents; Reveillaud et al., 2016; Fortunato et al., 2018), GN (Guerrero Negro microbial mat; Kunin et al., 2008), NF (nitrogenase-bearing organisms; Poudel et al., 2018). The NF data are based on reference proteomes (see Methods); all others are for protein sequences predicted from metagenomic data. Outlined symbols indicate samples ~~in~~from relatively oxidizing conditions. **(b)** Surface and **(c)** deeper samples (chl a max: chlorophyll a maximum, 9–30 m deep) from the Baltic Sea transect. Metagenomes as described in Dupont et al. (2014) were downloaded from iMicrobe (Youens-Clark et al., 2019); the plots show data for the 0.1–0.8 μm size fraction ~~are plotted here. Upward- and downward-pointing symbols, connected by dashed and dotted lines, represent surface and deeper samples, respectively,~~collected from stations along the transect at low salinity (< 6 PSU) and high salinity (> 6 PSU). Background guidelines have slopes equal to that of the n_{H_2O}–Z_C linear regression for amino acids in Fig. 1c.

proteins increase down the outflow channel of Bison Pool and between fluids from diffuse hydrothermal vents and relatively oxidizing seawater. ~~Notably~~Moreover, intact polar lipids extracted from the microbial communities of Bison Pool and other alkaline hot springs also exhibit downstream increases in carbon oxidation state (Boyer et al., 2020), ~~confirming~~revealing that ~~similar~~parallel compositional trends characterize ~~multiple classes~~all major types of biomacromolecules in these hot springs. The Z_C of proteins increases more subtly toward the surface in the upper few millimeters of the Guerrero Negro microbial mat; it also increases at greater depths, perhaps due to heterotrophic degradation and/or horizontal gene transfer (Dick et al., 2019). Furthermore, an evolutionary trajectory associated with the occurrence of different homologs of nitrogenase (Nif) in anaerobic and aerobic organisms is characterized by increasing Z_C of the proteomes of these organisms (Poudel et al., 2018).

The trends of carbon oxidation state described above are visible in the ~~n_{H_2O}–Z_C~~ scatter plot in Fig. 3a, with an added dimension: stoichiometric hydration state. The guidelines in this plot are parallel to the n_{H_2O}–Z_C trend for amino acids (Fig. 1c); their slope represents the background correlation between n_{H_2O} and Z_C that is associated with the choice of basis species. Sample data for Bison Pool and the submarine vents are distributed parallel to these guidelines. Therefore, the decrease of n_{H_2O}

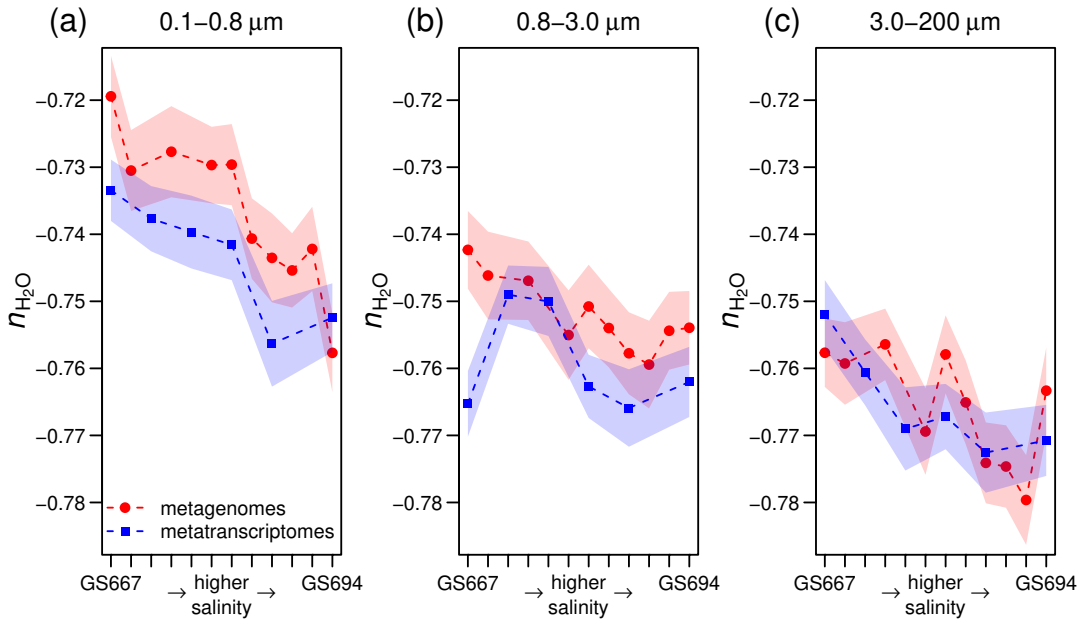

Figure 4. Stoichiometric hydration state of proteins in metagenomes (Dupont et al., 2014) and metatranscriptomes (Asplund-Samuelsson et al., 2016) of surface water samples in the Baltic Sea with increasing particle size: **(a)** 0.1–0.8 µm, **(b)** 0.8–3.0 µm, **(c)** 3.0–200 µm. From left to right, the samples on the ~~x~~ horizontal axis (some IDs omitted for clarity) are arranged from freshwater to marine conditions in the Sorcerer II Global Ocean Sampling Expedition (Dupont et al., 2014); all sample IDs are GS667, GS665, GS669, GS673, GS675, GS659, GS679, GS681, GS683, GS685, GS687, GS694. Width of shading represents ±1 standard deviation in subsampled sequences (see Methods).

along these redox gradients can be attributed to the background correlation in the stoichiometric analysis, and the differences between samples within each dataset are specifically associated with changes in carbon oxidation state and not stoichiometric hydration state.~~with the exception of Guerrero Negro, these datasets exhibit larger changes in carbon oxidation state than stoichiometric hydration state.~~ This is an expected outcome, as the redox gradients considered here do not have large changes in salinity. In particular, concentrations of Cl^-, a conservative ion, increase by less than 10% (6.1 to 6.6 mM) in the outflow of Bison Pool due to evaporation (Swingley et al., 2012). The diffuse vents considered here have concentrations of Cl^- between 515 and 624 mM, not greatly different from bottom seawater at 545 mM ~~[~~(Dataset S1 of Reeves et al. (2014)~~]~~).

As a well-known example of a regional salinity gradient, the Baltic Sea exhibits a freshwater to marine transition over 1800 km, but dissolved oxygen at the surface is at or near saturation with air (Dupont et al., 2014), so this transect does not represent a redox gradient. For protein sequences derived from metagenomes in the 0.1–0.8 µm size fraction, there are large changes in stoichiometric hydration state along the Baltic Sea transect, but relatively small differences in the carbon oxidation state (Fig. 3b). This pattern holds for samples from both the surface and chlorophyll a maximum (9–30 m deep; Fig. 3c).

4.2 Multifactorial hydration effects

~~Metagenomic and metatranscriptomic data for different filter size fractions are available for the Baltic Sea.~~ The stoichiometric hydration state of proteins can be influenced by factors other than just salinity. Previous authors have observed large differences in microbial community composition between free-living and particle-associated fractions, which may be due in part to anoxic conditions arising from limited diffusion in particles (Simon et al., 2014). As described below, we found a trend of relatively low n_{H_2O} in particles compared to free-living fractions in both the Baltic Sea and Amazon River. This effect is probably associated with phylogenetic differences among the size fractions, but reduced accessibility to bulk water may be a contributing factor. Further support for the possible influence of physical accessibility is the reduced n_{H_2O} in the interior compared to upper layers of the Guerrero Negro microbial mat.

For the Baltic Sea metagenomes and metatranscriptomes, the 0.1–0.8 μm and 0.8–3.0 μm size fractions of particles that don't pass through the filter, which are used for subsequent DNA extraction and sequencing, represent free living bacteria, while the 3.0–200 μm fraction contains particle-associated bacteria with average larger genome sizes and greater inferred metabolic and regulatory capacity (Dupont et al., 2014). ~~Figure~~Fig. 4a–c shows that proteins inferred from metagenomes for larger particles have lower n_{H_2O} than those for the smallest size fraction. The Guerrero Negro microbial mat offers another opportunity to compare exposed and interior environments. Unlike Z_C, which reaches a minimum a few millimeters into the mat, n_{H_2O} decreases throughout the mat, but the changes are most pronounced in the upper few millimeters (Fig. 3a).

One hypothesis that could explain these findings is that the interiors of particles and the mat are sequestered to some extent from the surrounding aqueous environment. If limited accessibility to the aqueous phase were manifested as lower water activity, ~~[~~perhaps due to surface effects associated with geological nanomaterials (Wang et al., 2003) and/or higher concentrations of solutes~~]~~, it would provide a thermodynamic drive that favors lower n_{H_2O} of proteins. However, it should be noted that particles are also suitable habitats for multicellular and eukaryotic populations (Simon et al., 2014). Therefore, the trends in stoichiometric hydration state may require an explanation in terms of both physical and phylogenetic differences, which should be explored in future studies. ~~A lower average n_{H_2O} in one eukaryotic organism, humans, is apparent in comparison to *E. coli* (Sect. 3.3) and in the positive values of n_{H_2O} for most of the metagenomic and metatranscriptomic datasets considered here (see Figs. 3–5) (recall that the mean for human proteins was defined to be zero). These preliminary observations suggest that the evolution of multicellularity may be accompanied by an overall decrease in stoichiometric hydration state.~~

An~~other~~ important evolutionary transition is the emergence of heterotrophic metabolism, which is a later innovation than autotrophic core metabolism (Morowitz, 1999; Braakman and Smith, 2013). It is notable that the deeper layers of the Guerrero Negro mat show greater evidence for heterotrophic metabolism (Kunin et al., 2008); likewise, heterotrophs in the "photosynthetic fringe" in Bison Pool may outcompete the autotrophs that dominate at higher and lower temperatures (Swingley et al., 2012). These putative heterotroph-rich zones show locally lower values of n_{H_2O} (Fig. 3a). If decreasing stoichiometric hydration state is a common theme across ~~these major~~some evolutionary transitions, then the relatively high n_{H_2O} in the proteomes of organisms carrying the ancestral nitrogenase Nif-D (Fig. 3a) is not unexpected. A better understanding of these trends would require more extensive phylogenetically resolved comparisons of the compositional differences as well as quantitative analyses of water fluxes in different metabolic pathways.

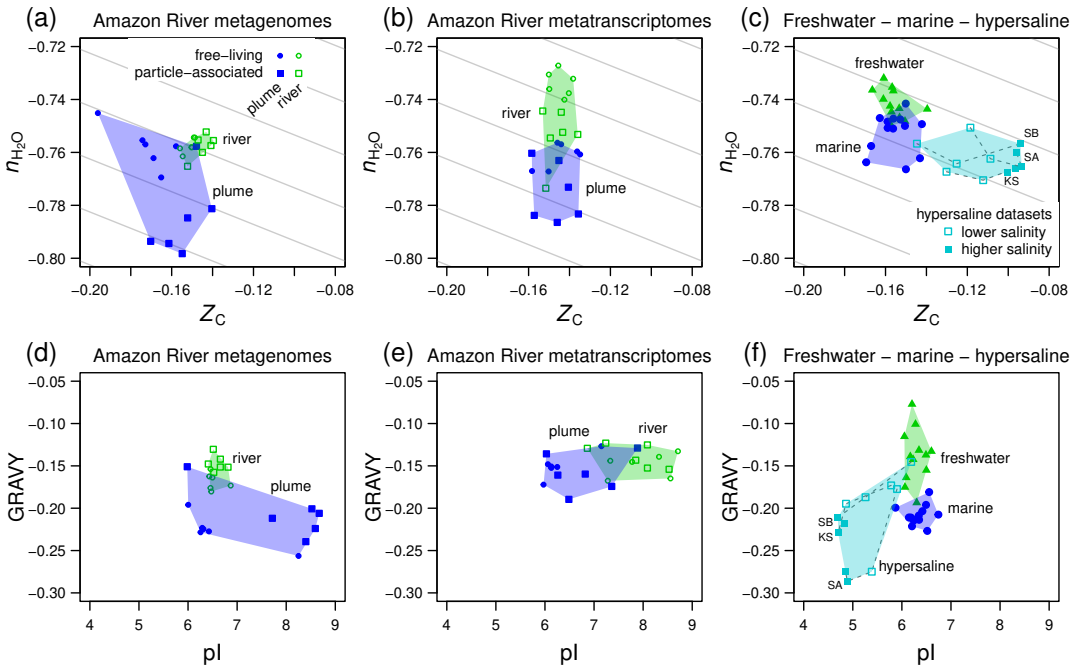

Figure 5. Compositional analysis and hydropathicity and isoelectric point calculations for proteins from the Amazon River and plume and other metagenomes. Samples representing freshwater, marine, and hypersaline environments are indicated by the colored convex hulls. **(a)** Metagenomic and **(b)** metatranscriptomic data for particle-associated and free-living fractions from the lower Amazon River (Satinsky et al., 2015) and plume in the Atlantic Ocean (Satinsky et al., 2014). **(c)** Freshwater (lakes in Sweden and USA) and marine metagenomes considered in a previous comparative study (Eiler et al., 2014) and metagenomes from hypersaline environments including Kulunda Steppe soda lakes in Siberia, Russia (Vavourakis et al., 2016) (KS), Santa Pola salterns in Spain (Ghai et al., 2011; Fernandez et al., 2013) (SA), and salterns in the South Bay of San Francisco, CA, USA (Kimbrel et al., 2018) (SB). Plots **(d-f)** show values of average hydropathicity (GRAVY) and isoelectric point (pI) of proteins for the same datasets. Background guidelines have slopes equal to that of the n_{H_2O}–Z_C linear regression for amino acids in Fig. 1c.

4.3 Compositional trends in rivers, lakes, and hypersaline environments

The Amazon river and ocean plume provide another example of a freshwater to marine transition, with salinities that range from below the scale of practical salinity units (PSU) in the river to 23–36 PSU in the plume (Satinsky et al., 2014, 2015). We used published metagenomic and metatranscriptomic data for filtered samples classified as free-living (0.2 to 2.0 μm) and particle-associated (2.0 to 156 μm) (Satinsky et al., 2014, 2015). River samples form a tight cluster on a plot of stoichiometric hydration state against carbon oxidation state of proteins, and the ~~free-living size fraction of~~ plume samples ~~is~~are scattered over lower Z_C ~~whereas the particle-associated fraction shows very~~ and low values of n_{H_2O}, particularly for the particle-associated fraction (Fig. 5a). For metatranscriptomes, there is a noticeable decrease of n_{H_2O} from the river to the ocean plume but little difference in carbon oxidation state (Fig. 5b), and the particle-associated samples again exhibit a generally lower n_{H_2O} than

the free-living samples. Together with the lower n_{H_2O} for proteins inferred from metagenomes and metatranscriptomes in the larger size fractions from Baltic Sea samples, this could reflect a lower availability of H_2O to organisms living near the particle surface due to physical separation from the bulk aqueous phase and associated diffusion limitation or lower water activity (Wang et al., 2003).

~~To continue the investigation,~~ We also considered data used in a previous comparative study and data for hypersaline environments including evaporation ponds (salterns) and lakes in desert areas. Eiler et al. (2014) characterized microbial communities using metagenomic data for various freshwater samples (lakes in the USA and Sweden) and marine locations. For hypersaline settings, we used metagenomic data from the Santa Pola salterns in Spain (Ghai et al., 2011; Fernandez et al., 2013), natural soda lakes of the Kulunda Steppe in Serbia (Vavourakis et al., 2016), and South Bay salterns in California, USA (Kimbrel et al., 2018). The compositional analysis reveals a relatively low n_{H_2O} of proteins inferred from the marine metagenomes compared to freshwater samples in the Eiler et al. dataset (Fig. 5c). Surprisingly, hypersaline metagenomes have ranges of n_{H_2O} of proteins that are similar to marine environments, but considerably higher Z_C (Fig. 5c). To interpret these results, we considered other factors that are known to influence the amino acid compositions of proteins in halophiles.

"Salt-in" halophilic organisms have proteins with relatively low isoelectric point that remain soluble at high salt concentrations (Ghai et al., 2011). ~~Notably,~~ It should be noted that proteins with a lower pI also tend to have relatively high Z_C due to higher abundances of aspartic acid and glutamic acid, which are relatively oxidized (see Amend and Shock, 1998, Dick, 2014, and Fig. 1). Consequently, the lower pI characteristic of "salt-in" organisms is also associated with an increase of carbon oxidation state. Because of the large pI differences (Fig. 5f), the increase of Z_C in hypersaline environments can not be interpreted as an indicator of an environmental redox gradient. Some halophilic organisms are also ~~noted~~ known to have proteins that are less hydrophobic, with lower values of GRAVY (Paul et al., 2008; Boyd et al., 2014). Because hydrophobic amino acids have relatively low values of Z_C (Dick, 2014), a negative correlation between GRAVY and Z_C is also expected.

~~for proteins are negatively correlated., as shown in Fig. 6a for all proteins in the *E. coli* genome. On the other hand, there is very little correlation in these proteins between GRAVY and n_{H_2O} (Fig. 6b). A small correlation between pI and Z_C is also apparent in the *E. coli* genome, in contrast to no correlation with n_{H_2O} (Fig. 6c–d). Therefore, it seems likely that selection for hydrophobicity or isoelectric point are not largely responsible for trends of n_{H_2O} in environmental samples.~~

Consistent with these well-known features of halophilic adaptation, marine metagenomes exhibit lower hydrophobicity than most of the freshwater samples, and hypersaline metagenomes are shifted to both lower GRAVY and pI (Fig. 5f). However, there are irregular trends in the Amazon River data. Compared to the river, the proteins in plume metagenomes exhibit lower GRAVY and either higher or lower pI (Fig. 5d). Similarly, other authors have reported that although lower pI is a signature of many hypersaline environments, it does not clearly distinguish marine from lower-salinity environments (Rhodes et al., 2010). ~~On the other hand~~In contrast, the plume metatranscriptomes do show decreased pI but no major difference in GRAVY compared to river samples (Fig. 5e).

There is not enough space here to comprehensively examine all the available metagenomic data for environmental salinity gradients. However, we have identified one dataset that gives a contradictory result, and therefore offers more perspective on the compositional relationships of proteins coded by metagenomes in salinity gradients. This dataset was generated in a time-

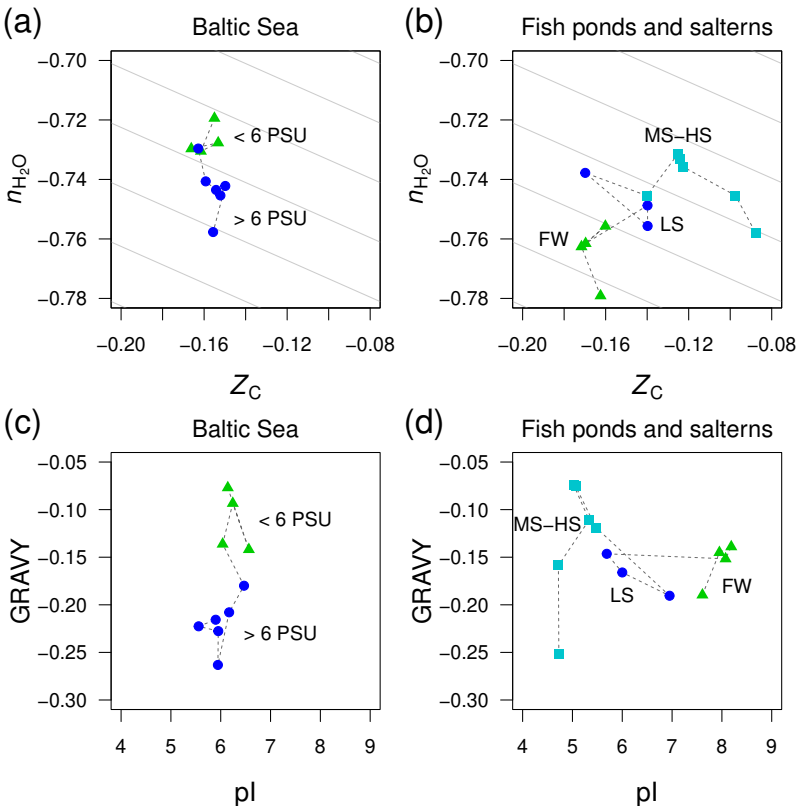

Figure 6. Divergent trends of n_{H_2O} and Z_C of proteins from metagenomes for **(a)** the Baltic Sea and **(b)** freshwater and higher-salinity samples from southern California (Rodriguez-Brito et al., 2010). The datasets from Rodriguez-Brito et al. (2010) are classified according to salinity: freshwater (FW; 3 samples at different times from the "tilapia channel" and 1 sample from the "prebead pond"), low salinity (LS; 3 samples at different times from the low salinity saltern), and hypersaline (MS–HS; 4 samples from a medium salinity and 2 from a high salinity saltern). Plots **(c)** and **(d)** show GRAVY and pI computed for the same datasets. Background guidelines have slopes equal to that of the n_{H_2O}–Z_C linear regression for amino acids in Fig. 1c.

series study of microbial and viral community dynamics in a freshwater aquaculture facility ("tilapia channel" and "prebead bond") and low-, medium-, and high-salinity salterns in southern California (Rodriguez-Brito et al., 2010). Here, we have used only the reported microbial sequences (not the viral dataset) and considered all time points together. Contrary to our starting hypothesis, the stoichiometric hydration state of proteins is lowest in the freshwater samples, which is the reverse of the trend from the Baltic Sea (Fig. 6a–b). A side-by-side comparison of the Baltic Sea and Rodriguez-Brito et al. datasets shows large changes of GRAVY in the former, but pI in the latter (Fig. 6c–d), which is another indication that these variables ~~respond as expected~~ are responsive only in certain ranges of salinity.

This counterexample demonstrates that the sign of differences of n_{H_2O} is not predictable in all environments; however, the large negative offset in the freshwater samples may be a signal of some other influence, perhaps related to the human control

of these ponds, which are used as fish nurseries. Specifically, the microbial communities in the aquaculture ponds may not be responding as they would in a typical natural system that is less nutrient-rich. As noted above for putative heterotroph-rich zones in other systems, the lower stoichiometric hydration state could be associated with the enrichment of heterotrophic taxa, in this case due to the addition of organic compounds to the aquaculture ponds.

Considering all the datasets shown in Figs. 5 and 6, there appears to be no globally consistent metric for environmental salinity gradients that can be derived from amino acid composition. If we exclude the Rodriguez-Brito et al. (2010) dataset, then n_{H_2O} exhibits a consistent decreasing trend in marine compared to freshwater samples. However, this trend does not continue into hypersaline environments.

4.4 Compositional analysis of differentially expressed proteins

~~Coming away from a picture of salinity gradients as only spatial phenomena, there is much interest in the impact of changing salinities on microbial organisms. To cite one example relevant to environmental studies, cyanobacteria respond to salt shock through stages including cell shrinkage, influx of external salts, synthesis of compatible solutes, changes in gene and protein expression, and acclimation (Qiao et al., 2013).~~ While biomolecular data for environmental salinity gradients reflect both ecological and evolutionary differences, laboratory experiments provide information on the physiological effects of osmotic conditions on protein expression in particular organisms. It is also important to recognize that osmotic stress can be imposed by solutes other than NaCl; the effects of organic solutes differ in relation to their ability to permeate or depolarize cell membranes and to be sensed by cellular osmoregulatory systems (Kanesaki et al., 2002; Shabala et al., 2009; Withman et al., 2013). ~~It is clear that~~Because microbial adaptation to changes in osmotic conditions is a dynamic process, ~~so~~ it is helpful to look at gene and protein expression data for a range of times and conditions that can be controlled in the lab.

We ~~performed multiple literature searches~~ searched the literature to compile data for differential gene and protein expression in non-halophilic bacteria in NaCl or other osmotic stress conditions. As a general rule, we only included ~~only~~ datasets with a minimum of 20 down-regulated and 20 up-regulated genes or proteins; however, smaller datasets were included if they are part of a study with larger datasets. This compilation consists of 49 transcriptomics and ~~29~~30 proteomics datasets from ~~35~~36 studies (note that different time points and treatments are considered as separate datasets); descriptions and references for all datasets are given in Figures S1 and S2. In addition, four datasets for differential expression of proteins in halophilic archaea in hyperosmotic stress were located (Leuko et al., 2009; Zhang et al., 2016; Lin et al., 2017; Jevtić et al., 2019) (see Figure S3). ~~We assembled the lists of up- and down-regulated proteins in each dataset or, for gene expression studies, the proteins corresponding to the up- and down-regulated genes, and converted gene names or accession numbers to UniProt accessions using the UniProt mapping tool (Huang et al., 2011). The compiled data are available as CSV files in R packages (see *Code and data availability*).~~ This is a major update to an earlier compilation of data for hyperosmotic stress experiments (Dick, 2017), but we have limited the present compilation to data for bacteria or archaea; data for osmotic stress induced by NaCl or glucose in eukaryotic cells are considered in a separate paper (Dick, 2020a).

We assembled the lists of up- and down-regulated proteins in each dataset or, for gene expression studies, the proteins corresponding to the up- and down-regulated genes, and converted gene names or accession numbers to UniProt accessions

Table 2. ~~Halophilic organisms, growth conditions, number of differentially expressed proteins, and sources of data for hypoosmotic and hyperosmotic stress experiments. Units for NaCl concentrations are taken from the references; approximate conversions between molarity and weight percent are 1 M NaCl ≈ 6%, 2.5 M NaCl ≈ 13%, 4 M NaCl ≈ 20%.~~

~~Data sources: (a, b) Tables 1 and 2 of Leuko et al. (2009). (c, d) Table S-1 of Zhang et al. (2016). Values of reporter intensities at each condition (6%, 10%, and 17.5% NaCl) were quantile normalized and used to compute intensity ratios (6% / 10% NaCl and 17.5% / 10% NaCl). Only proteins with expression ratios > 1.3 in either direction (Zhang et al., 2016), p-values < 0.05, and at least 2 peptides were included. (e, g) Tables S2 and S3 of Lin et al. (2017). (g, h) Supporting Table 1C of Jevtić et al. (2019). Only proteins with at least 2-fold expression difference and marked as significant were included.~~

using the UniProt mapping tool (Huang et al., 2011). The compiled data are available as CSV files in R packages (see *Code and data availability*). After removing genes or proteins with unavailable or duplicated UniProt IDs and those with ambiguous differences (appearing in both the down- and up-regulated groups), the amino acid compositions computed for protein sequences downloaded from UniProt (The UniProt Consortium, 2019) were used for the compositional analysis of carbon oxidation state and stoichiometric hydration state. Median differences (i.e. Δn_{H_2O} and ΔZ_C) were calculated as the median value for all up-regulated proteins minus the median value for all down-regulated proteins in each dataset. ~~In Fig. 7, the values of ΔZ_C and Δn_{H_2O} represented by empty and lettered symbols refer to median differences in individual datasets; that is, the median value for all up-regulated proteins minus the median value for all down-regulated proteins. Although there is obvious scatter in values, the Δn_{H_2O} for proteins in transcriptomic and proteomic experiments is negative on average (Fig. 7a–b), but the differences are non-significant to marginally significant [p = 0.215 and 0.052, respectively; all p-values were calculated for paired two-sided Student's t-tests using R (R Core Team, 2020)]. The compilations of gene and protein expression data also show small average ΔZ_C, with p = 0.088 and 0.666, respectively.~~

Figure 7~~e~~a shows results for ~~selected~~ time-course experiments for hyperosmotic stress. Note that all values are differences calculated relative to the same control (~~starting condition~~initial time point) in a given study. In transcriptomic experiments for a commensal species (*Enterococcus faecalis*), a soil bacterium (*Methylocystis* sp. strain SC2), and two pathogens (*E. coli* O157:H7 and *Salmonella enterica* serovar Typhimurium) (Solheim et al., 2014; Han et al., 2017; Kocharunchitt et al., 2014; Finn et al., 2015), there is a marked progression toward lower Δn_{H_2O} of the associated proteins with time. In a transcriptomic experiment for salt stress in *Synechocystis* sp. PCC 6803 (Qiao et al., 2013), Δn_{H_2O} is shifted negatively between 24 and 48 h, but rises to a ~~less negative~~slightly positive value at 72 h. Proteomic data are available from two of these studies, indicating that the differentially expressed proteins in *E. coli* (Kocharunchitt et al., 2014) also show decreasing Δn_{H_2O} with time ~~(Fig. 7d)~~, but in the proteomic experiment for *Synechocystis* sp. PCC 6803 (Qiao et al., 2013), Δn_{H_2O} changes sign from negative to positive between 24 and 48 h (Fig. 7a).

Perhaps the most striking result to emerge from this analysis is the strong dehydrating signal associated with osmotic stress imposed by organic solutes. We compared pairs of datasets from the same study for NaCl and another solute at concentrations that give similar total osmolalities. Transcriptomic data for sorbitol (Kanesaki et al., 2002; Han et al., 2005), sucrose (Kohler et al., 2015), and glycerol (Finn et al., 2015) compared to controls all show a lower Δn_{H_2O} of the associated proteins than for

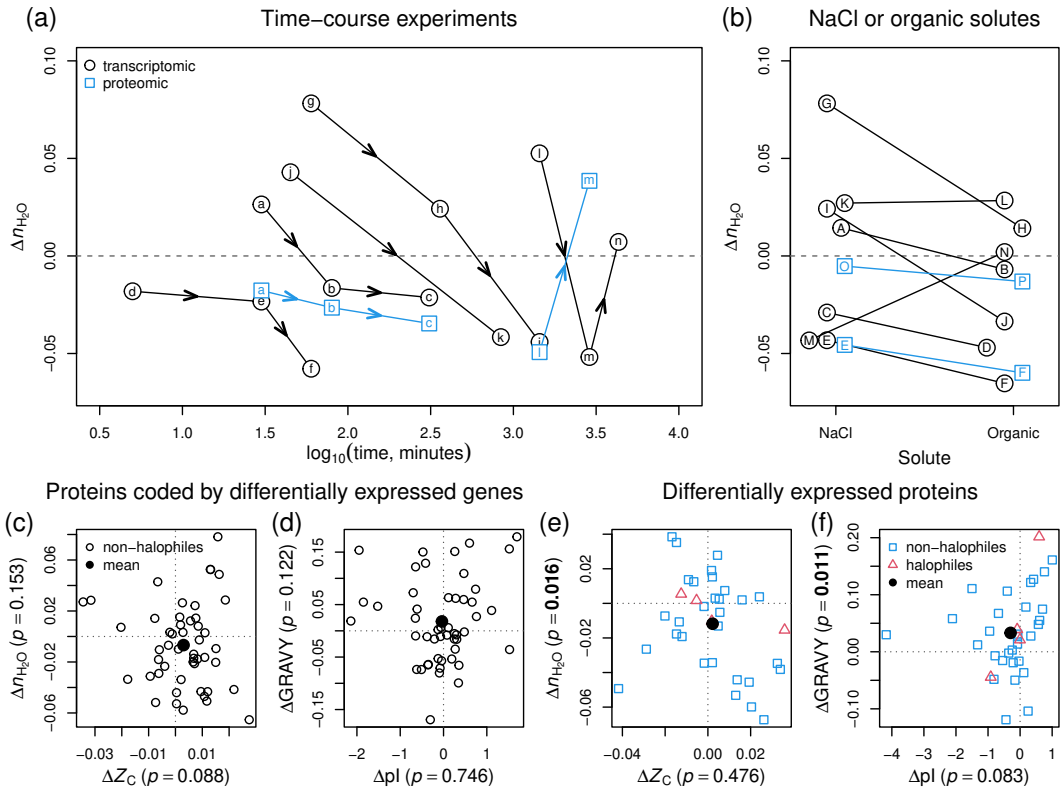

Figure 7. Compositional analysis of proteins in hyperosmotic stress experiments for non-halophil~~ices~~ bacteria and halophilic archaea. ~~All datasets and mean value for all datasets in each compilation are shown for **(a)** proteins coded by differentially expressed genes and **(b)** differentially expressed proteins. See Figures S1 and S2 for references for all datasets. Selected time course experiments are highlighted in **(c)** and **(d)**.~~ **(a)** Time-course experiments for bacteria; black circles represent datasets for proteins coded by differentially expressed genes (transcriptomics experiments) and blue squares represent datasets for differentially expressed proteins (proteomics experiments). ~~Points connected by lines show~~Lettered symbols represent the progression in each experiment: a–c (30, 80, 310 min; Kocharunchitt et al., 2014) (transcriptomes and proteomes), d–f (5, 30, 60 min; Solheim et al., 2014), g–i (1, 6, 24 h; Finn et al., 2015), j–k (45 min, 14 h; Han et al., 2017), l–n (24, 48, 72 h; Qiao et al., 2013) (transcriptomes and proteomes; no proteomic data available at 72 h). **(~~c–f~~b)** Pairs of experiments for bacteria under hyperosmotic stress imposed by NaCl or organic solutes. The sources of data are: A–B (sorbitol; Kanesaki et al., 2002), C–D (sorbitol; Han et al., 2005), E–F (sucrose; Kohler et al., 2015) (transcriptomes and proteomes), G–H (glycerol at 1 h; Finn et al., 2015), I–J (glycerol at 6 h; Finn et al., 2015), K–L (sucrose; Shabala et al., 2009), M–N (urea; Withman et al., 2013), O–P (glucose; Schmidt et al., 2016) (only proteomes). **(c–f)** Plots of median differences of n_{H_2O} and Z_C or GRAVY and pI for all compiled transcriptomic and proteomic data for hyperosmotic stress, including datasets shown in **(a)** and **(b)** together with data for other experiments. In each panel, open symbols represent individual datasets and filled symbols represent the mean for all datasets. The axis labels include the *p*-values for the mean difference for all datasets in each plot; *p*-values less than 0.05 are shown in bold. References for all datasets are in Figures S1 (transcriptomics for non-halophilic bacteria), S2 (proteomics for non-halophilic bacteria), and S3 (proteomics for halophilic archaea).

Figure 8. ~~Compositional analysis of differentially expressed proteins in halophiles under hypoosmotic and hyperosmotic stress. **(a)** Median differences of n_{H_2O} and Z_C between up- and down-regulated proteins in hypoosmotic compared to optimal growth conditions and hyperosmotic compared to optimal growth conditions. See Table 2 for experimental conditions and references. **(b)** Median differences of GRAVY and pI for the same datasets. **(c)** Median differences of GRAVY and pI for all compiled proteomics data for hyperosmotic stress in halophiles and non-halophiles.~~

NaCl compared to controls (Fig. 7~~e~~b). Data from the study of Finn et al. (2015) are plotted at 1 and 6 h in the experiment, indicating a time-dependent decrease of Δn_{H_2O} under both NaCl and glycerol treatment as well as more negative values for glycerol than NaCl. Experiments with different strains of *E. coli* show a ~~smaller negative difference between NaCl and sucrose~~slightly more positive value for sucrose than NaCl (Shabala et al., 2009) and ~~the only positive~~a much larger positive difference for ~~an organic solute (~~urea~~)~~ compared to NaCl (Withman et al., 2013). The available proteomic data also show lower n_{H_2O} for sucrose (Kohler et al., 2015) and glucose (Schmidt et al., 2016) compared to NaCl (Fig. 7~~f~~b). Note that the latter dataset is actually a comparison between growth on glucose and glucose with NaCl; growth on glucose alone produces a lower Δn_{H_2O} of the differentially expressed proteins.

The marked decrease of Δn_{H_2O} induced by solutes such as sorbitol, which does not permeate the plasma membrane, could ~~follow~~result from a higher effective osmotic pressure compared to NaCl (Kanesaki et al., 2002). Because it permeates cells, solutions of urea are not considered hypertonic (Burg et al., 2007), which may be one reason for the higher Δn_{H_2O} for urea compared to NaCl. ~~However, s~~Sucrose, which permeates but unlike NaCl does not depolarize the plasma membrane (Shabala et al., 2009), produces a slightly higher Δn_{H_2O} than NaCl in one transcriptomics dataset for *E. coli* (Shabala et al., 2009), ~~also exhibits a strong~~ but has a more marked dehydrating effect in both transcriptomics and proteomics datasets for *Caulobacter crescentus* (Kohler et al., 2015). The negative shift of Δn_{H_2O} associated with most organic solutes compared to NaCl lends support to the notion that high organic loading could contribute to the relatively low n_{H_2O} of protein sequences from metagenomes of freshwater aquaculture systems (Fig. 6b).

~~We also considered the changes in protein expression when halophilic organisms are exposed to hyperosmotic conditions in the laboratory. Proteomic data were found for four halophilic species of bacteria and archaea for hypo- and hyperosmotic stress under changing NaCl concentrations (Leuko et al., 2009; Zhang et al., 2016; Lin et al., 2017; Jevtić et al., 2019) (Table 2). The combined data are plotted in Fig. 7a. A negative Δn_{H_2O} of the differentially expressed proteins characterizes most of the hyperosmotic stress experiments; only *Tetragenococcus halophilus* shows a small positive value. Unexpectedly, growth at NaCl concentrations below the optimal concentrations (i.e. hypoosmotic stress) in three of these organisms – the archaeon *Halobacterium salinarium* and bacteria *Nocardiopsis xinjiangensis* and *Tetragenococcus halophilus* – induces an even larger loss of n_{H_2O} in the differentially expressed proteins (points labeled a, c, and e in Fig. 7a).~~

~~The median difference of GRAVY increases for differentially expressed proteins in three of the four halophilic organisms under hyperosmotic stress (Fig. 7b).~~ Considering all transcriptomic datasets together (see Figure S1 for references), the proteins coded by differentially expressed genes in non-halophilic bacteria under hyperosmotic stress do not show significant differences in Z_C, n_{H_2O}, pI, or GRAVY (Fig. 7c–d). However, the average difference of n_{H_2O} would become more negative if

the early time points in individual time-course experiments were excluded from the average (see Fig. 7a). Unlike the results for transcriptomes, ~~the data for hyperosmotic stress in both halophiles and non-halophiles,~~ the average value of GRAVY for all proteomics datasets (see Figures S2 and S3 for references) increases significantly (Fig. 7~~e~~f; p = ~~0.010~~0.011). The proteomic data also exhibit a small decrease of pI (p = ~~0.100~~0.083), which is expected for halophiles, but the increase of GRAVY – that is, higher hydrophobicity – is the opposite of the evolutionary trend for proteomes of halophilic organisms (Paul et al., 2008) and the metagenomic comparisons described above. Overall, the proteomic experiments record a significant decrease of n_{H_2O} in hyperosmotic stress (Fig. 7e; p = 0.016). We therefore ~~propose~~conclude that n_{H_2O} is a ~~more consistent~~ metric with consistent behavior for field and laboratory datasets, since it records decreasing hydration state of proteins with increasing salinity in the Baltic Sea and Amazon River and plume, and ~~in~~ of differentially expressed proteins ~~of both halophiles and non-halophiles~~in microbial cells grown under hyperosmotic stress.

5 Conclusions

This study was focused on describing the chemical compositions of proteins in a geochemical context. The theoretical novelty of this study is the derivation of a compositional metric for stoichiometric hydration state (n_{H_2O}) that is largely decoupled from changes in oxidation state (Z_C) of proteins. Therefore, based on mass-action effects in thermodynamics, ~~we predicted that the stoichiometric hydration state of proteins (n_{H_2O}) should~~ is predicted to decrease toward higher salinity but be mostly insensitive to redox gradients. We found that protein sequences inferred from metagenomes in regional salinity gradients, including the Baltic Sea freshwater-marine transect and Amazon River and plume, are characterized by changes of n_{H_2O} in the predicted direction. ~~However, the~~Although this trend does not continue into hypersaline environments, ~~and there are conflicting results derived from metagenomic data used in previous comparative studies: n_{H_2O} decreases between freshwater lakes and marine samples (Eiler et al., 2014) but increases between freshwater aquaculture ponds and salterns (Rodriguez-Brito et al., 2010). While biomolecular data for environmental salinity gradients reflect phylogenetic differences and evolution, laboratory experiments provide information on the physiological effects of osmotic conditions on protein expression in single organisms.~~ the applicability of the compositional analysis to microbial cells is supported by compilations of transcriptomic and proteomic data, ~~for non-halophilic organisms~~ which indicate ~~a small decrease of~~ decreasing n_{H_2O} on average for the differentially expressed proteins in hyperosmotic stress experiments. The dehydration signal becomes larger during many time-course experiments and is stronger for most organic solutes ~~(except urea)~~ than for NaCl. ~~Differentially expressed proteins in halophiles show a more complex response: for three of four organisms with available data, Δn_{H_2O} is much lower in hypoosmotic compared to hyperosmotic conditions, which is an unexpected finding.~~

~~We were also surprised to find a pattern of relatively low n_{H_2O} in the interior compared to upper layers of the Guerrero Negro microbial mat and in particles compared to free-living fractions in both the Baltic Sea and Amazon River. This effect is probably associated with phylogenetic differences among the size fractions, but reduced accessibility to bulk water may be a contributing factor. The latter possibility can be further investigated through compositional analysis of differentially expressed proteins between single-species biofilms and planktonic growth in the laboratory.~~

The central message of this study is that geochemical and laboratory conditions can influence, but naturally do not completely determine, the chemical compositions of proteins. ~~The compositional analysis establishes the feasibility and the limits of using thermodynamic models to predict the biomolecular makeup of organisms in new environments. The usefulness of~~ As a step toward constructing multidimensional chemical-thermodynamic models ~~is also apparent~~of microbial communities, ~~since~~ the present results provide evidence that different compositional metrics, representing the oxidation state and hydration state of molecules, can ~~in some cases~~ be associated specifically with redox and salinity gradients, respectively. The findings of this study underscore an opportunity for the integration of hydration state into evolutionary models that already consider changes in oxidation state or oxygen content of proteins (Acquisti et al., 2007; Poudel et al., 2018).

Code and data availability.

All metagenomic and metatranscriptomic data analyzed here were obtained from public databases using the accession numbers listed in Supplementary Table S1 for salinity gradients and Table S2 for redox gradients. The amino acid compositions of subsampled sequences from the metagenomic and metatranscriptomic data are available in the JMDplots R package, version ~~1.2.2~~1.2.4 (https://github.com/jedick/JMDplots), which is archived on Zenodo (Dick, 2020b). Specifically, the data are contained in the file `inst/extdata/gradH2O/MGP.rds`, which can be read using the R function readRDS (minimum R version: 2.3.0).

The compilation of differential gene expression data is available in the JMDplots package as xz-compressed CSV files in the directory `inst/extdata/expression/osmotic/`. The compilation of differential protein expression data is in the corresponding directory of the canprot R package, version ~~1.0.0~~1.1.0 (https://cran.r-project.org/package=canprot), which is also archived on Zenodo (Dick, 2020c). The results of the compositional analysis of differential expression data, which are used for Fig~~s~~. 7, are in the `inst/vignettes/` directories of the JMDplots and canprot packages.

The code used to make all of the figures and perform statistical testing is in the JMDplots package. The `gradH2O.Rmd` vignette in the package ~~contains~~demonstrates the function~~s calls~~ used ~~for~~to make the figures.

Author contributions. JMD designed and carried out the analysis. JMD, MY and JT interpreted the results. JMD wrote the manuscript with editing input from MY and JT.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. We are grateful to Saroj Poudel for commenting on an earlier version of the manuscript. This work was supported by funding from the State Key Laboratory of Organic Geochemistry (Grant No. SKLOG-201928 to JD).

References

Acquisti, C., Kleffe, J., and Collins, S.: Oxygen content of transmembrane proteins over macroevolutionary time scales, Nature, 445, 47–52, https://doi.org/10.1038/nature05450, 2007.

Akashi, H. and Gojobori, T.: Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*, Proceedings of the National Academy of Sciences, 99, 3695–3700, https://doi.org/10.1073/pnas.062526999, 2002.

Alsop, E. B., Boyd, E. S., and Raymond, J.: Merging metagenomics and geochemistry reveals environmental controls on biological diversity and evolution, BMC Ecology, 14, 16, https://doi.org/10.1186/1472-6785-14-16, 2014.

Amend, J. P. and LaRowe, D. E.: Mini-review: Demystifying microbial reaction energetics, Environmental Microbiology, 21, 3539–3547, https://doi.org/10.1111/1462-2920.14778, 2019.

Amend, J. P. and Shock, E. L.: Energetics of amino acid synthesis in hydrothermal ecosystems, Science, 281, 1659–1662, https://doi.org/10.1126/science.281.5383.1659, 1998.

Amend, J. P., LaRowe, D. E., McCollom, T. M., and Shock, E. L.: The energetics of organic synthesis inside and outside the cell, Philosophical Transactions of the Royal Society, B: Biological Sciences, 368, 20120 255, https://doi.org/10.1098/rstb.2012.0255, 2013.

Anderson, G. M.: Thermodynamics of Natural Systems, Cambridge University Press, Cambridge, 2nd edn., http://www.worldcat.org/oclc/474880901, 2005.

Asplund-Samuelsson, J., Sundh, J., Dupont, C. L., Allen, A. E., McCrow, J. P., Celepli, N. A., Bergman, B., Ininbergs, K., and Ekman, M.: Diversity and expression of bacterial metacaspases in an aquatic ecosystem, Frontiers in Microbiology, 7, 1043, https://doi.org/10.3389/fmicb.2016.01043, 2016.

Baudouin-Cornu, P., Surdin-Kerjan, Y., Marlière, P., and Thomas, D.: Molecular evolution of protein atomic composition, Science, 293, 297–300, https://doi.org/10.1126/science.1061052, 2001.

Bjellqvist, B., Hughes, G. J., Pasquali, C., Paquet, N., Ravier, F., Sanchez, J.-C., Frutiger, S., and Hochstrasser, D.: The focusing positions of polypeptides in immobilized pH gradients can be predicted from their amino acid sequences, Electrophoresis, 14, 1023–1031, https://doi.org/10.1002/elps.11501401163, 1993.

Bjellqvist, B., Basse, B., Olsen, E., and Celis, J. E.: Reference points for comparisons of two-dimensional maps of proteins from different human cell types defined in a pH scale where isoelectric points correlate with polypeptide compositions, Electrophoresis, 15, 529–539, https://doi.org/10.1002/elps.1150150171, 1994.

Boyd, E. S., Hamilton, T. L., Swanson, K. D., Howells, A. E., Baxter, B. K., Meuser, J. E., Posewitz, M. C., and Peters, J. W.: [FeFe]-hydrogenase abundance and diversity along a vertical redox gradient in Great Salt Lake, USA, International Journal of Molecular Sciences, 15, 21 947–21 966, https://doi.org/10.3390/ijms151221947, 2014.

Boyer, G. M., Schubotz, F., Summons, R. E., Woods, J., and Shock, E. L.: Carbon oxidation state in microbial polar lipids suggests adaptation to hot spring temperature and redox gradients, Frontiers in Microbiology, 11, 229, https://doi.org/10.3389/fmicb.2020.00229, 2020.

Braakman, R. and Smith, E.: The compositional and evolutionary logic of metabolism, Physical Biology, 10, 011 001, https://doi.org/10.1088/1478-3975/10/1/011001, 2013.

Burg, M. B., Ferraris, J. D., and Dmitrieva, N. I.: Cellular response to hyperosmotic stresses, Physiological Reviews, 87, 1441–1474, https://doi.org/10.1152/physrev.00056.2006, 2007.

Canovas, Peter A., I. and Shock, E. L.: Energetics of the citric acid cycle in the deep biosphere, in: Carbon in Earth's Interior, edited by Manning, C. E., Lin, J.-F., and Mao, W. L., chap. 25, pp. 303–327, American Geophysical Union, https://doi.org/10.1002/9781119508229.ch25, 2020.

Chirife, J., Fontan, C. F., and Scorza, O. C.: The intracellular water activity of bacteria in relation to the water activity of the growth medium, Journal of Applied Bacteriology, 50, 475–479, https://doi.org/10.1111/j.1365-2672.1981.tb04250.x, 1981.

DeBerardinis, R. J. and Cheng, T.: Q's next: The diverse functions of glutamine in metabolism, cell biology and cancer, Oncogene, 29, 313–324, https://doi.org/10.1038/onc.2009.358, 2010.

Dick, J. M.: Average oxidation state of carbon in proteins, Journal of the Royal Society Interface, 11, 20131 095, https://doi.org/10.1098/rsif.2013.1095, 2014.

Dick, J. M.: Proteomic indicators of oxidation and hydration state in colorectal cancer, PeerJ, 4, e2238, https://doi.org/10.7717/peerj.2238, 2016.

Dick, J. M.: Chemical composition and the potential for proteomic transformation in cancer, hypoxia, and hyperosmotic stress, PeerJ, 5, e3421, https://doi.org/10.7717/peerj.3421, 2017.

Dick, J. M.: Water as a reactant in the differential expression of proteins in cancer, bioRxiv, https://doi.org/10.1101/2020.04.09.035022, 2020a.

Dick, J. M.: JMDplots 1.2.4, Zenodo, https://doi.org/10.5281/zenodo.4111016, 2020b.

Dick, J. M.: canprot 1.1.0, Zenodo, https://doi.org/10.5281/zenodo.4105653, 2020c.

Dick, J. M. and Shock, E. L.: Calculation of the relative chemical stabilities of proteins as a function of temperature and redox chemistry in a hot spring, PLOS One, 6, e22 782, https://doi.org/10.1371/journal.pone.0022782, 2011.

Dick, J. M., Yu, M., Tan, J., and Lu, A.: Changes in carbon oxidation state of metagenomes along geochemical redox gradients, Frontiers in Microbiology, 10, 120, https://doi.org/10.3389/fmicb.2019.00120, 2019.

Du, B., Zielinski, D. C., Monk, J. M., and Palsson, B. O.: Thermodynamic favorability and pathway yield as evolutionary tradeoffs in biosynthetic pathway choice, Proceedings of the National Academy of Sciences, 115, 11 339–11 344, https://doi.org/10.1073/pnas.1805367115, 2018.

Dupont, C. L., Larsson, J., Yooseph, S., Ininbergs, K., Goll, J., Asplund-Samuelsson, J., McCrow, J. P., Celepli, N., Allen, L. Z., Ekman, M., Lucas, A. J., Hagström, Å., Thiagarajan, M., Brindefalk, B., Richter, A. R., Andersson, A. F., Tenney, A., Lundin, D., Tovchigrechko, A., Nylander, J. A. A., Brami, D., Badger, J. H., Allen, A. E., Rusch, D. B., Hoffman, J., Norrby, E., Friedman, R., Pinhassi, J., Venter, J. C., and Bergman, B.: Functional tradeoffs underpin salinity-driven divergence in microbial community composition, PLOS One, 9, 1–9, https://doi.org/10.1371/journal.pone.0089549, 2014.

Eiler, A., Zaremba-Niedzwiedzka, K., Martínez-García, M., McMahon, K. D., Stepanauskas, R., Andersson, S. G. E., and Bertilsson, S.: Productivity and salinity structuring of the microplankton revealed by comparative freshwater metagenomics, Environmental Microbiology, 16, 2682–2698, https://doi.org/10.1111/1462-2920.12301, 2014.

Feist, A. M., Henry, C. S., Reed, J. L., Krummenacker, M., Joyce, A. R., Karp, P. D., Broadbelt, L. J., Hatzimanikatis, V., and Palsson, B. Ø.: A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information, Molecular Systems Biology, 3, 121, https://doi.org/10.1038/msb4100155, 2007.

Fernandez, A. B., Ghai, R., Martin-Cuadrado, A. B., Sanchez-Porro, C., Rodriguez-Valera, F., and Ventosa, A.: Metagenome sequencing of prokaryotic microbiota from two hypersaline ponds of a marine saltern in Santa Pola, Spain, Genome Announcements, 1, 6, https://doi.org/10.1128/genomea.00933-13, 2013.

Finn, S., Rogers, L., Händler, K., McClure, P., Amézquita, A., Hinton, J. C. D., and Fanning, S.: Exposure of *Salmonella enterica* serovar Typhimurium to three humectants used in the food industry induces different osmoadaptation systems, Applied and Environmental Microbiology, 81, 6800–6811, https://doi.org/10.1128/AEM.01379-15, 2015.

Fortunato, C. S., Larson, B., Butterfield, D. A., and Huber, J. A.: Spatially distinct, temporally stable microbial populations mediate biogeochemical cycling at and below the seafloor in hydrothermal vent fluids, Environmental Microbiology, 20, 769–784, https://doi.org/10.1111/1462-2920.14011, 2018.

Garner, M. M. and Burg, M. B.: Macromolecular crowding and confinement in cells exposed to hypertonicity, American Journal of Physiology, 266, C877–C892, https://doi.org/10.1152/ajpcell.1994.266.4.C877, 1994.

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R. D., and Bairoch, A.: Protein identification and analysis tools on the ExPASy server, in: The Proteomics Protocols Handbook, edited by Walker, J. M., pp. 571–607, Humana Press, Totowa, NJ, https://doi.org/10.1385/1-59259-890-0:571, 2005.

Ghai, R., Pašić, L., Fernández, A. B., Martin-Cuadrado, A.-B., Mizuno, C. M., McMahon, K. D., Papke, R. T., Stepanauskas, R., Rodriguez-Brito, B., Rohwer, F., Sánchez-Porro, C., Ventosa, A., and Rodríguez-Valera, F.: New abundant microbial groups in aquatic hypersaline environments, Scientific Reports, 1, 135, https://doi.org/10.1038/srep00135, 2011.

Gunde-Cimerman, N., Plemenitaš, A., and Oren, A.: Strategies of adaptation of microorganisms of the three domains of life to high salt concentrations, FEMS Microbiology Reviews, 42, 353–375, https://doi.org/10.1093/femsre/fuy009, 2018.

Han, D., Link, H., and Liesack, W.: Response of *Methylocystis* sp. strain SC2 to salt stress: Physiology, global transcriptome, and amino acid profiles, Applied and Environmental Microbiology, 83, e00 866–17, https://doi.org/10.1128/AEM.00866-17, 2017.

Han, Y., Zhou, D., Pang, X., Zhang, L., Song, Y., Tong, Z., Bao, J., Dai, E., Wang, J., Guo, Z., Zhai, J., Du, Z., Wang, X., Wang, J., Huang, P., and Yang, R.: Comparative transcriptome analysis of *Yersinia pestis* in response to hyperosmotic and high-salinity stress, Research in Microbiology, 156, 403–415, https://doi.org/10.1016/j.resmic.2004.10.004, 2005.

Havig, J. R., Raymond, J., Meyer-Dombard, D. R., Zolotova, N., and Shock, E. L.: Merging isotopes and community genomics in a siliceous sinter-depositing hot spring, Journal of Geophysical Research, 116, G01 005, https://doi.org/10.1029/2010JG001415, 2011.

Huang, H., McGarvey, P. B., Suzek, B. E., Mazumder, R., Zhang, J., Chen, Y., and Wu, C. H.: A comprehensive protein-centric ID mapping service for molecular data integration, Bioinformatics, 27, 1190–1191, https://doi.org/10.1093/bioinformatics/btr101, 2011.

Jevtić, v., Stoll, B., Pfeiffer, F., Sharma, K., Urlaub, H., Marchfelder, A., and Lenz, C.: The response of *Haloferax volcanii* to salt and temperature stress: A proteome study by label-free mass spectrometry, Proteomics, 19, 1800 491, https://doi.org/10.1002/pmic.201800491, 2019.

Kanesaki, Y., Suzuki, I., Allakhverdiev, S. I., Mikami, K., and Murata, N.: Salt stress and hyperosmotic stress regulate the expression of different sets of genes in *Synechocystis* sp. PCC 6803, Biochemical and Biophysical Research Communications, 290, 339–348, https://doi.org/10.1006/bbrc.2001.6201, 2002.

Karl, D. M. and Grabowski, E.: The importance of H in particulate organic matter stoichiometry, export and energy flow, Frontiers in Microbiology, 8, 826, https://doi.org/10.3389/fmicb.2017.00826, 2017.

Kauffman, J. M.: Simple method for determination of oxidation numbers of atoms in compounds, Journal of Chemical Education, 63, 474–475, https://doi.org/10.1021/ed063p474, 1986.

Keegan, K. P., Glass, E. M., and Meyer, F.: MG-RAST, a metagenomics service for analysis of microbial community structure and function, in: Microbial Environmental Genomics (MEG), edited by Martin, F. and Uroz, S., pp. 207–233, Springer, New York, https://doi.org/10.1007/978-1-4939-3369-3_13, 2016.

Kimbrel, J. A., Ballor, N., Wu, Y.-W., David, M. M., Hazen, T. C., Simmons, B. A., Singer, S. W., and Jansson, J. K.: Microbial community structure and functional potential along a hypersaline gradient, Frontiers in Microbiology, 9, 1492, https://doi.org/10.3389/fmicb.2018.01492, 2018.

Kocharunchitt, C., King, T., Gobius, K., Bowman, J. P., and Ross, T.: Global genome response of *Escherichia coli* O157:H7 Sakai during dynamic changes in growth kinetics induced by an abrupt downshift in water activity, PLOS One, 9, e90 422, https://doi.org/10.1371/journal.pone.0090422, 2014.

Kohler, C., Lourenço, R. F., Bernhardt, J., Albrecht, D., Schüler, J., Hecker, M., and Gomes, S. L.: A comprehensive genomic, transcriptomic and proteomic analysis of a hyperosmotic stress sensitive α-proteobacterium, BMC Microbiology, 15, 1–15, https://doi.org/10.1186/s12866-015-0404-x, 2015.

Kopylova, E., Noé, L., and Touzet, H.: SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data, Bioinformatics, 28, 3211–3217, https://doi.org/10.1093/bioinformatics/bts611, 2012.

Kunin, V., Raes, J., Harris, J. K., Spear, J. R., Walker, J. J., Ivanova, N., von Mering, C., Bebout, B. M., Pace, N. R., Bork, P., and Hugenholtz, P.: Millimeter-scale genetic gradients and community-level molecular convergence in a hypersaline microbial mat, Molecular Systems Biology, 4, 198, https://doi.org/10.1038/msb.2008.35, 2008.

Kyte, J. and Doolittle, R. F.: A simple method for displaying the hydropathic character of a protein, Journal of Molecular Biology, 157, 105–132, https://doi.org/10.1016/0022-2836(82)90515-0, 1982.

LaRowe, D. E. and Van Cappellen, P.: Degradation of natural organic matter: A thermodynamic analysis, Geochimica et Cosmochimica Acta, 75, 2030–2042, https://doi.org/10.1016/j.gca.2011.01.020, 2011.

Leuko, S., Raftery, M. J., Burns, B. P., Walter, M. R., and Neilan, B. A.: Global protein-level responses of *Halobacterium salinarum* NRC-1 to prolonged changes in external sodium chloride concentrations, Journal of Proteome Research, 8, 2218–2225, https://doi.org/10.1021/pr800663c, 2009.

Ley, R. E., Harris, J. K., Wilcox, J., Spear, J. R., Miller, S. R., Bebout, B. M., Maresca, J. A., Bryant, D. A., Sogin, M. L., and Pace, N. R.: Unexpected diversity and complexity of the Guerrero Negro hypersaline microbial mat, Applied and Environmental Microbiology, 72, 3685–3695, https://doi.org/10.1128/AEM.72.5.3685-3695.2006, 2006.

Lin, J., Liang, H., Yan, J., and Luo, L.: The molecular mechanism and post-transcriptional regulation characteristic of *Tetragenococcus halophilus* acclimation to osmotic stress revealed by quantitative proteomics, Journal of Proteomics, 168, 1–14, https://doi.org/10.1016/j.jprot.2017.08.014, 2017.

Lindsay, M. R., Amenabar, M. J., Fecteau, K. M., Debes II, R. V., Fernandes Martins, M. C., Fristad, K. E., Xu, H., Hoehler, T. M., Shock, E. L., and Boyd, E. S.: Subsurface processes influence oxidant availability and chemoautotrophic hydrogen metabolism in Yellowstone hot springs, Geobiology, 16, 674–692, https://doi.org/10.1111/gbi.12308, 2018.

May, P. M. and Rowland, D.: JESS, a Joint Expert Speciation System – VI: thermodynamically-consistent standard Gibbs energies of reaction for aqueous solutions, New Journal of Chemistry, 42, 7617–7629, https://doi.org/10.1039/C7NJ03597G, 2018.

Milo, R., Jorgensen, P., Moran, U., Weber, G., and Springer, M.: BioNumbers—the database of key numbers in molecular and cell biology, Nucleic Acids Research, 38, D750–D753, https://doi.org/10.1093/nar/gkp889, 2010.

Minkiewicz, P., Darewicz, M., and Iwaniak, A.: Introducing a simple equation to express oxidation states as an alternative to using rules associated with words alone, Journal of Chemical Education, 95, 340–342, https://doi.org/10.1021/acs.jchemed.7b00322, 2018.

Morowitz, H. J.: A theory of biochemical organization, metabolic pathways, and evolution, Complexity, 4, 39–53, https://doi.org/10.1002/(SICI)1099-0526(199907/08)4:6<39::AID-CPLX8>3.0.CO;2-2, 1999.

Möller, M. N., Li, Q., Chinnaraj, M., Cheung, H. C., Lancaster, J. R., and Denicola, A.: Solubility and diffusion of oxygen in phospholipid membranes, Biochimica et Biophysica Acta (BBA) - Biomembranes, 1858, 2923–2930, https://doi.org/10.1016/j.bbamem.2016.09.003, 2016.

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C. M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V. S., Kodali, V. K., Li, W., Maglott, D., Masterson, P., McGarvey, K. M., Murphy, M. R., O'Neill, K., Pujar, S., Rangwala, S. H., Rausch, D., Riddick, L. D., Schoch, C., Shkeda, A., Storz, S. S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R. E., Vatsan, A. R., Wallin, C., Webb, D., Wu, W., Landrum, M. J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T. D., and Pruitt, K. D.: Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation, Nucleic Acids Research, 44, D733–D745, https://doi.org/10.1093/nar/gkv1189, 2016.

Ooka, H., McGlynn, S. E., and Nakamura, R.: Electrochemistry at deep-sea hydrothermal vents: Utilization of the thermodynamic driving force towards the autotrophic origin of life, ChemElectroChem, 6, 1316–1323, https://doi.org/10.1002/celc.201801432, 2019.

Oren, A.: Life at high salt concentrations, intracellular KCl concentrations, and acidic proteomes, Frontiers in Microbiology, 4, 315, https://doi.org/10.3389/fmicb.2013.00315, 2013.

Paul, S., Bag, S. K., Das, S., Harvill, E. T., and Dutta, C.: Molecular signature of hypersaline adaptation: insights from genome and proteome composition of halophilic prokaryotes, Genome Biology, 9, R70, https://doi.org/10.1186/gb-2008-9-4-r70, 2008.

Poudel, S., Colman, D. R., Fixen, K. R., Ledbetter, R. N., Zheng, Y., Pence, N., Seefeldt, L. C., Peters, J. W., Harwood, C. S., and Boyd, E. S.: Electron transfer to nitrogenase in different genomic and metabolic backgrounds, Journal of Bacteriology, 200, e00 757–17, https://doi.org/10.1128/JB.00757-17, 2018.

Qiao, J., Huang, S., Te, R., Wang, J., Chen, L., and Zhang, W.: Integrated proteomic and transcriptomic analysis reveals novel genes and regulatory mechanisms involved in salt stress responses in *Synechocystis* sp. PCC 6803, Applied Microbiology and Biotechnology, 97, 8253–8264, https://doi.org/10.1007/s00253-013-5139-8, 2013.

R Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org, https://www.R-project.org, 2020.

Record, Jr., M. T., Courtenay, E. S., Cayley, D. S., and Guttman, H. J.: Responses of *E. coli* to osmotic stress: Large changes in amounts of cytoplasmic solutes and water, Trends in Biochemical Sciences, 23, 143–148, https://doi.org/10.1016/S0968-0004(98)01196-7, 1998.

Reeves, E. P., McDermott, J. M., and Seewald, J. S.: The origin of methanethiol in midocean ridge hydrothermal fluids, Proceedings of the National Academy of Sciences, 111, 5474–5479, https://doi.org/10.1073/pnas.1400643111, 2014.

Reveillaud, J., Reddington, E., McDermott, J., Algar, C., Meyer, J. L., Sylva, S., Seewald, J., German, C. R., and Huber, J. A.: Subseafloor microbial communities in hydrogen-rich vent fluids from hydrothermal systems along the Mid-Cayman Rise, Environmental Microbiology, 18, 1970–1987, https://doi.org/10.1111/1462-2920.13173, 2016.

Rho, M., Tang, H., and Ye, Y.: FragGeneScan: Predicting genes in short and error-prone reads, Nucleic Acids Research, 38, e191, https://doi.org/10.1093/nar/gkq747, 2010.

Rhodes, M. E., Fitz-Gibbon, S. T., Oren, A., and House, C. H.: Amino acid signatures of salinity on an environmental scale with a focus on the Dead Sea, Environmental Microbiology, 12, 2613–2623, https://doi.org/10.1111/j.1462-2920.2010.02232.x, 2010.

Rodriguez-Brito, B., Li, L., Wegley, L., Furlan, M., Angly, F., Breitbart, M., Buchanan, J., Desnues, C., Dinsdale, E., Edwards, R., Felts, B., Haynes, M., Liu, H., Lipson, D., Mahaffy, J., Martin-Cuadrado, A. B., Mira, A., Nulton, J., Pašić, L., Rayhawk, S., Rodriguez-Mueller, J.,

Rodriguez-Valera, F., Salamon, P., Srinagesh, S., Thingstad, T. F., Tran, T., Thurber, R. V., Willner, D., Youle, M., and Rohwer, F.: Viral and microbial community dynamics in four aquatic environments, ISME Journal, 4, 739–751, https://doi.org/10.1038/ismej.2010.1, 2010.

Satinsky, B. M., Zielinski, B. L., Doherty, M., Smith, C. B., Sharma, S., Paul, J. H., Crump, B. C., and Moran, M. A.: The Amazon continuum dataset: quantitative metagenomic and metatranscriptomic inventories of the Amazon River plume, June 2010, Microbiome, 2, 17, https://doi.org/10.1186/2049-2618-2-17, 2014.

Satinsky, B. M., Fortunato, C. S., Doherty, M., Smith, C. B., Sharma, S., Ward, N. D., Krusche, A. V., Yager, P. L., Richey, J. E., Moran, M. A., and Crump, B. C.: Metagenomic and metatranscriptomic inventories of the lower Amazon River, May 2011, Microbiome, 3, 39, https://doi.org/10.1186/s40168-015-0099-0, 2015.

Schmidt, A., Kochanowski, K., Vedelaar, S., Ahrné, E., Volkmer, B., Callipo, L., Knoops, K., Bauer, M., Aebersold, R., and Heinemann, M.: The quantitative and condition-dependent *Escherichia coli* proteome, Nature Biotechnology, 34, 104–110, https://doi.org/10.1038/nbt.3418, 2016.

Shabala, L., Bowman, J., Brown, J., Ross, T., McMeekin, T., and Shabala, S.: Ion transport and osmotic adjustment in *Escherichia coli* in response to ionic and non-ionic osmotica, Environmental Microbiology, 11, 137–148, https://doi.org/10.1111/j.1462-2920.2008.01748.x, 2009.

Shock, E. L., Holland, M., Meyer-Dombard, D. R., Amend, J. P., Osburn, G. R., and Fischer, T. P.: Quantifying inorganic sources of geochemical energy in hydrothermal ecosystems, Yellowstone National Park, USA, Geochimica et Cosmochimica Acta, 74, 4005–4043, https://doi.org/10.1016/j.gca.2009.08.036, 2010.

Simon, H. M., Smith, M. W., and Herfort, L.: Metagenomic insights into particles and their associated microbiota in a coastal margin ecosystem, Frontiers in Microbiology, 5, 466, https://doi.org/10.3389/fmicb.2014.00466, 2014.

Slonczewski, J. L., Fujisawa, M., Dopson, M., and Krulwich, T. A.: Cytoplasmic pH measurement and homeostasis in bacteria and archaea, in: Advances in Microbial Physiology, edited by Poole, R. K., vol. 55, pp. 1–79, Academic Press, New York, https://doi.org/10.1016/S0065-2911(09)05501-5, 2009.

Solheim, M., La Rosa, S. L., Mathisen, T., Snipen, L. G., Nes, I. F., and Brede, D. A.: Transcriptomic and functional analysis of NaCl-induced stress in *Enterococcus faecalis*, PLOS One, 9, 1–13, https://doi.org/10.1371/journal.pone.0094571, 2014.

Sterner, R. and Liebl, W.: Thermophilic adaptation of proteins, Critical Reviews in Biochemistry and Molecular Biology, 36, 39–106, https://doi.org/10.1080/20014091074174, 2001.

Swingley, W. D., Meyer-Dombard, D. R., Shock, E. L., Alsop, E. B., Falenski, H. D., Havig, J. R., and Raymond, J.: Coordinating environmental genomics and geochemistry reveals metabolic transitions in a hot spring ecosystem, PLOS One, 7, e38 108, https://doi.org/10.1371/journal.pone.0038108, 2012.

The UniProt Consortium: UniProt: A worldwide hub of protein knowledge, Nucleic Acids Research, 47, D506–D515, https://doi.org/10.1093/nar/gky1049, 2019.

Turner, C. B., Wade, B. D., Meyer, J. R., Sommerfeld, B. A., and Lenski, R. E.: Evolution of organismal stoichiometry in a long-term experiment with *Escherichia coli*, Royal Society Open Science, 4, 170 497, https://doi.org/10.1098/rsos.170497, 2017.

Vavourakis, C. D., Ghai, R., Rodriguez-Valera, F., Sorokin, D. Y., Tringe, S. G., Hugenholtz, P., and Muyzer, G.: Metagenomic insights into the uncultured diversity and physiology of microbes in four hypersaline soda lake brines, Frontiers in Microbiology, 7, 211, https://doi.org/10.3389/fmicb.2016.00211, 2016.

Wagner, A.: Energy constraints on the evolution of gene expression, Molecular Biology and Evolution, 22, 1365–1374, https://doi.org/10.1093/molbev/msi126, 2005.

Walsh, C. T., Tu, B. P., and Tang, Y.: Eight kinetically stable but thermodynamically activated molecules that power cell metabolism, Chemical Reviews, 118, 1460–1494, https://doi.org/10.1021/acs.chemrev.7b00510, 2018.

Wang, Y., Bryan, C., Xu, H., and Gao, H.: Nanogeochemistry: Geochemical reactions and mass transfers in nanopores, Geology, 31, 387–390, https://doi.org/10.1130/0091-7613(2003)031<0387:NGRAMT>2.0.CO;2, 2003.

Warn, J. R. W. and Peters, A. P. H.: Concise Chemical Thermodynamics, CRC Press, 2nd edn., http://www.worldcat.org/oclc/36624543, 1996.

Withman, B., Gunasekera, T. S., Beesetty, P., Agans, R., and Paliy, O.: Transcriptional responses of uropathogenic *Escherichia coli* to increased environmental osmolality caused by salt or urea, Infection and Immunity, 81, 80–89, https://doi.org/10.1128/IAI.01049-12, 2013.

Youens-Clark, K., Bomhoff, M., Ponsero, A. J., Wood-Charlson, E. M., Lynch, J., Choi, I., Hartman, J. H., and Hurwitz, B. L.: iMicrobe: Tools and data-driven discovery platform for the microbiome sciences, GigaScience, 8, giz083, https://doi.org/10.1093/gigascience/giz083, 2019.

Zeldovich, K. B., Berezovsky, I. N., and Shakhnovich, E. I.: Protein and DNA sequence determinants of thermophilic adaptation, PLOS Computational Biology, 3, 62–72, https://doi.org/10.1371/journal.pcbi.0030005, 2007.

Zhang, Y., Li, Y., Zhang, Y., Wang, Z., Zhao, M., Su, N., Zhang, T., Chen, L., Wei, W., Luo, J., Zhou, Y., Xu, Y., Xu, P., Li, W., and Tao, Y.: Quantitative proteomics reveals membrane protein-mediated hypersaline sensitivity and adaptation in halophilic *Nocardiopsis xinjiangensis*, Journal of Proteome Research, 15, 68–85, https://doi.org/10.1021/acs.jproteome.5b00526, 2016.