Biogeosciences Discuss.,
https://doi.org/10.5194/bg-2020-146-RC1, 2020

Biogeosciences
Discussions
Open Access

Interactive comment on "Uncovering chemical signatures of salinity gradients through compositional analysis of protein sequences" *by* Jeffrey M. Dick et al.

Anonymous Referee #1

Received and published: 16 June 2020

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. H2O is used as a basis species in addition to O2 and three amino acids (glutamine, glutamic acid and cysteine). 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.

I support publication of this work after some clarifying text is added in the areas noted 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.

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

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.

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?

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.

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.

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.

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?

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.

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.

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.

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.

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.

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.

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

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

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.

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 supplemental figures in S1 and S2 need captions.

Interactive comment on Biogeosciences Discuss., https://doi.org/10.5194/bg-2020-146, 2020.