Editor's Comments

I suggest not to consider the experimental data (to test the outcome of the oxidative stress predictions on 3D cultures by mass spectrometry) requested from one of the reviewers since they are not pertinent.

Comments of Reviewer #1

Basic reporting

The abstract is concise and descriptive. The document presents several typos (see below). I suggest having a native English speaker revising the grammar and the sentence flow throughout the manuscript.

Examples:
Line 4: myriad (of) molecular-level

- No change made here. Many Q&A websites state that myriad can be used as an adjective or a noun.

Lines 34, 41: incorrect use of the word 'for'

- Line 34: Changed "for" to "of" ("ranges of other physical-chemical variables").
- Line 41: No change here. I think "proteomic datasets for laboratory hypoxia" is the correct wording. Compare with e.g. "The dataset for each donor ..." used in one of the cited references (p. 5 of Riis et al., 2016).

Line 46: 'the' is missing – 'the complexity of (the) underlying…'

- Inserted "the" ("the underlying regulatory mechanisms").

Line 56: 'is inherent to the' and not 'is inherent in the'

- Changed as suggested.

Line 123: Consider change 'must take account of the gain' to 'must consider the gain' …

- Changed to "account for" ("a calculation ... must account for the gain").
- Fixed other typos:
  - Line 69: changed "In this view" to "In this situation".
  - Line 110: changed "affecting decomposition" to "affecting the decomposition".
  - Line 141: changed "are divided" to "can be divided".
  - Line 353: chnaged "diagrams have been made" to "diagrams were made".
  - Line 375: changed "Tables 3" to "Table 3".
  - Line 380: changed "induces the up-expression" to "induces the expression".

– Line 456: fixed repeated "to".

The author stated that there are few studies available addressing the thermodynamic potential related to proteome changes, though none is cited. Available literature should be discussed and cited properly.

- I have trouble finding any paper to cite here other than a self-citation, which would appear gratuitous. Changed "few studies have investigated" to "little attention has been given to".

Tables 1-4 are well structured and their captions are adequate. Still, the meaning of n1 and n2 in the table legend is lacking.

- A sentence has been added to the caption of Table 1 to define these variables.

A good explanation for the thermodynamic concepts of average state of carbon and of water demand per residue is provided. Still, I suggest to introduce the commonly used term of 'hydration shell' to discern 'protein hydration' from 'hydration state'.

- Modified a sentence here: "These dynamically interacting molecules form a hydration shell ..."

Equation R1 only accounts for Cys, Glu and Gln from the 20 amino acids (aa) that compose proteins. While this is a mathematical model, a brief comment is expected regarding the choice of these aa or the exclusion of the remaining 17 aa.

- Added: "A basis consists of a minimum number of species whose compositions can be linearly combined to represent the composition of any protein. The primary amino acid sequence of any protein encodes for molecules with five elements (C, H, N, O, S), so five basis species are needed."

- Concerning the specific choice of these amino acids, see the additional text added in response to Reviewer #3.

The mathematical model was applied to understand the thermodynamic basis of the proteomic changes in several pathological and experimental conditions such as colorectal cancer, pancreatic cancer, hypoxia and hyperosmotic stress. Even though these are provided as working examples, some references are needed. For instance, line 157 (colorectal cancer) and line 222 (hypoxia) require referencing.

- Line 157 (colorectal cancer): Added references to Jimenez el al., 2010 and Wiśniewski et al., 2015.

- Line 222 (hypoxia): Added references to Datta et al., 2010, Li et al., 2012, and Fuhrmann et al., 2013.

Lines 168-169: Are you sure that n1 and n2 are always the number of down- and up-regulated proteins in cancer versus normal tissue? For instance, sets b, c, and p from Table 1 compare carcinoma samples with adenoma samples. Can adenoma be regarded as a normal tissue? Shouldn't it be classified as a benign tumor? Please revise and rephrase it accordingly.

- Added this sentence where the Table is introduced in the text: "For datasets comparing different stages of cancer progression, groups n_1 and n_2 correspond to the down- and up-expressed proteins in the more advanced stage (e.g. carcinoma) compared to the less advanced stage (e.g. adenoma)."

Experimental design

The goal is well presented and the paper addresses a relevant, yet obscure, question that may have impact in the (bio)medical/health sciences.

- Thank you.

Line92: Name the other sources.

- Changed "other sources" to "the NCBI website (for one study reporting GI numbers; see Table 4)".

I am not sure about the inclusion of samples of chronic pancreatitis, autoimmune pancreatitis and diabetes mellitus (set d, l and q from Table 2) in the analysis due to the different pathological backgrounds. At least, they should be regarded as potential outliers and not just set s and, thus, they should be discussed accordingly in the R&D section.

- Many previous studies consider pancreatic cancer and pancreatitis together as part of a spectrum of related disease. The titles of some of the papers cited here show that the authors treat them jointly:
  - A proteomic comparison of formalin-fixed paraffin-embedded pancreatic tissue from autoimmune pancreatitis, chronic pancreatitis, and pancreatic cancer (Pan et al., 2013)
  - Proteomic analysis of chronic pancreatitis and pancreatic adenocarcinoma (Crnogorac-Jurcevic et al., 2005)
- This statement from Pan et al., 2011 is also relevant: "It has previously been shown that patients with chronic pancreatitis have an increased risk of developing pancreatic cancer [4–7]; thus it is not surprising that many molecular features presented in pancreatic cancer are also presented in chronic pancreatitis [4–9]."
- Therefore, rather than necessarily being outliers, the pancreatitis datasets are relevant to the process of cancer progression. Of the 23 datasets considered here in the context of pancreatic cancer (Table 2), one compares chronic pancreatitis to normal tissue, and two from a single study compare pancreatic cancer to autoimmune or chronic pancreatitis. These datasets are easily identified in Table 2. The first of these datasets is colored red in Figure 1B to highlight the similarities with a dataset for low-grade cancer; both of these show changes in chemical composition that are different from most of the datasets for pancreatic cancer, as was discussed in Lines 213–218. Not including the pancreatitis datasets would eliminate the opportunity to look at the data and make this observation.

Also, regarding 'hypoxia and 3D culture', I do not feel that cells cultured as spheroids or other 3D models exactly replicate hypoxia models per se. While oxygen tension can be small in the core of the organoid/spheroid, there are other confounders that can explain changes in ZC and water demand per residue, such as nutrient deprival (and thus oxidation of unusual substrates), different extracellular architectures and even light penetration (which may affect redox reactions).

- No claim has been made in this paper that the spheroid/3D cultures "replicate hypoxia". Note that the spheroid datasets are distinguished from the hypoxia datasets in Table 3 ("SPH") and in Figure 2 (red points). In the text, these datasets are introduced as "a related situation" with the following features:
  - Interior regions of 3D cultures are often diffusion-limited (McMahon et al., 2012).
  - There are some overlaps, but also many differences, between gene expression in hypoxia and 3D cultures (DelNero et al., 2015).
- I welcome the Reviewer's comments on the effects of different processes in the 3D cultures. To clarify the purpose for the inclusion of these datasets, the following sentences have been added: "These studies emphasize that growth in 3D culture is associated with heterogeneous oxygen concentrations and have found an interdependence between the effects of hypoxia and 3D growth on gene expression. The proteomic changes likely reflect not only oxygen limitation but also other processes connected with 3D growth (e.g. nutrient deprivation, extracellular architecture, and even light penetration). Although the comparisons made here do not address these individual factors, they do provide information on whether hypoxia and 3D culture lead to similar changes in the overall chemical composition of proteomes."
- To make it more clear that both types of datasets (hypoxia and 3D culture) are included, but separately labeled, the title of Table 3 has been changed to read "Selected proteomic datasets for hypoxia and reoxygenation experiments or growth in 3D culture".

Validity of the findings

My concern is with data heterogeneity. For instance, for pancreatic cancer, inclusion of diabetes mellitus, chronic pancreatitis and autoimmune pancreatitis may have biased diagrams from Figure 3. Additionally, inclusion of 3D cell culture models to mimic hypoxia is somehow controversial. Thus, I suggest to exclude these data from the thermodynamic analysis or discuss it in Supplementary data.

- I argued above that "looking at the data" is needed to pose questions about chemical composition (Figures 1–2) that are relevant to previous studies that themselves have compared the biology of pancreatic cancer with pancreatitis and of hypoxia with 3D culture. I understand the Reviewer's concerns with including these data in the thermodynamic analysis (Figure 3). However, no changes have been made, as justified by the following considerations:
  - Owing to the criteria for inclusion in the thermodynamic analysis (based on signs and sizes of compositional changes), neither the datasets for chronic pancreatitis nor for pancreatic cancer with diabetes mellitus make it into the analysis shown in Figure 3B.
  - Contributing to Figure 3E are two datasets for comparisons of pancreatic cancer with autoimmune pancreatitis or with chronic pancreatits rather than with normal tissue.
    - To clarify the differences between these datasets, the phrase "in cancer compared to non-cancerous (normal or pancreatitis) tissue" has been added to the compositional description of these datasets in the Results.

    - A trial calculation of Figure 3E was made excluding the two datasets for comparison to pancreatitis; the positions of the resulting equipotential lines differ somewhat from that shown in the manuscript, in a direction that looks more like the CRC diagram. These differences are not large enough to affect the discussion or conclusions.
  - The number of datasets that meet the inclusion criteria are 4 for hypoxia and 4 for 3D growth (these datasets are identified on page 3 of Figure S3). No large differences between hypoxia and 3D growth can be identified in the individual potential diagrams in Figure S3, so Figure 3C is justifiably representative of the features of these datasets as a whole.
  - As a further recognition that datasets for both hypoxia and 3D culture are both present and are not identical conditions, the subfigure captions in Figure 2A and 3C have been changed from "hypoxia" to "hypoxia or 3D culture".

Comments for the author

The author provides an interesting topic on thermodynamic variables driving the proteomic changes in a variety of pathological and experimental conditions. If the concerns exposed above are duly addressed or justified I support the publication.

Comments of Reviewer #2

Basic reporting

no comment

Experimental design

The paper is substantially a good mathematical work based on deposited biological data, but in my opinion there is not a perfect match with the aim and scope of the journal that does not publish Mathematical Sciences paper.

- I am aware of some previous papers in PeerJ that present mathematical models and computational analysis of existing datasets:
  - Microbial metabolism: optimal control of uptake versus synthesis (doi:10.7717/peerj.267)
  - Amino-acid site variability among natural and designed proteins (doi:10.7717/peerj.211)
  - Abundance-weighted phylogenetic diversity measures distinguish microbial community states and are robust to sampling depth (doi:10.7717/peerj.157)
  - PhyloSift: phylogenetic analysis of genomes and metagenomes (doi:10.7717/peerj.243)
  - System wide analyses have underestimated protein abundances and the importance of transcription in mammals (doi:10.7717/peerj.270)

Validity of the findings

The reported results may represent a new point of view of cellular and tissues behaviour dynamics, but they are based only on literature datasets without any direct examination of the formulated hypothesis, with a consequential speculation bias.

- Yes, much of this study is based on exploratory data analysis. The generation of new hypotheses through exploration of data is one type of scientific endeavor.

Comments for the Author

I suggest the authors to improve the paper with some experimental data performed by themselves. For example they could test the outcome of the oxidative stress predictions on 3D cultures by mass spectrometry.

- Thank you for this suggestion. I do look forward to working with experimentalists in the future to test and refine the predictions made here.

Response to Reviewer #2

Comments of Reviewer #3

Basic reporting

The paper is clearly written. I commend the author for sharing the raw data in an R package.

- Thank you.

1. I recommend to shorten the abstract.

- The length of the submitted abstract is 304 words; PeerJ places the limit at approx. 500 words. Reviewer #1 found the abstract to be "concise and descriptive". The length of the abstract is justified by presenting a context for compositional and thermodynamic analysis (which is not a standard approach) and to ease the reader into terminology that may be unfamiliar.

- One sentence in the abstract was simplified ("Diagrams summarizing the relative potential ...").

2. Considering 'self-contained' policy of the journal, I think it may not be appropriate to mention preliminary analysis of breast cancer, for which no data is shown, even if it is in the discussion section.

- This statement has been removed.

Figures are appropriately labeled.

3. Perhaps Figure 4 could be moved to supplementary information.

- Figure 4 is essential to the main text because it cites the references for the numerical values used in the estimation of the change in lipid-to-protein ratio given certain assumptions.

- The code snippet shown in this figure outlines the actual sequence of calculations, with relevant comments and results shown on the right. I was motivated to present stepwise calculations in this way after seeing the book by Milo and Phillips, 2015 (Cell Biology by the Numbers). Although Figure 4 is not as pretty as the blackboard-style mathematical equations in their book, it shows the same type of information.

- I think it is important to consider this type of "back of the envelope" calculation (as mentioned in the figure title) as a way to gauge the interdependence of compositional changes in the cell.

4. Figure 3 has to be placed after equation 3.

- The placement of Figure 3 in the submitted manuscript was higher than Equation 3, and on the page before the first reference to the figure. The figure has been moved back in the revised manuscript; the actual placement of the figure in the final version will depend on the journal's own layout.

5. I also think subfigures in Figure 3 could be arranged in a regular grid.

- Done.

Experimental design

1. Is total protein charge Z missing from the equation 1? It is part of the equation in Dick, 2014.

- Good point. This equation has been changed to be consistent with that in Dick, 2014. However, the note made here still applies ("ionization by gain or loss of protons alters charge and the number of H equally, so has no effect on the value of Z_C"). I've added the additional note that "therefore, the calculation of Z_C is made here for proteins in their completely non-ionized forms".

2. It may be mentioned in Dick, 2016 but I think it is necessary to explain why QEC amino acids are sufficient for assessing Zc of the protein.

- I wish to remind the reviewer that Z_C is calculated from the chemical formula and does not depend on the choice of basis species. However, the relationship between Z_C and the variables inherent in the basis species is an important issue that was not fully explained in the submitted manuscript, and may not be apparent in the description given by Dick, 2016. I have added a more detailed explanation, based on a comparison of scatterplots of $\overline{n}_{H_2O}$ vs Z_C and $\overline{n}_{O_2}$ vs Z_C for human proteins using either the QEC basis or a basis composed of inorganic species (more common in geochemical models). These plots have been included as Figure S1, and a paragraph has been added to describe this comparison. The essence is that QEC is a convenient choice because it gives $\overline{n}_{O_2}$ that is highly correlated with Z_C (so is a useful indicator of oxidation state) while showing $\overline{n}_{H_2O}$ that has very little correlation with Z_C (and therefore represents a different compositional dimension).

3. Is there any known functional bias (GO enrichment) in QEC+ biased proteins?

- Whether there is a general relationship between function and chemical composition of proteins is an interesting question. Many of the cited papers do describe functional biases for differentially expressed proteins, and the finding here that cancer often exhibits a higher Z_C (or $\overline{n}_{O_2}$ in the QEC basis) for up-expressed compared to down-expressed proteins does suggest the possibility of a functional association. However, it seems of limited value to point to individual instances now, and an in-depth investigation of the possible relationships between function and chemical composition is outside the scope of the current study.

4. A Figure connecting Figure 1 and Figure 2 and drawing parallels between cancer cells and hypoxic/hyperosmotic cells could be helpful.

- Thank you for this suggestion. I think the content and design choices for such a conceptual figure are very important. I would not feel comfortable presenting my own choices after a single iteration; a satisfying result would take many more iterations, along with feedback from others.

5. Although it is mentioned in the caption, it is not that clear how to interpret the sign of mean differences, if it is up or down regulated proteins that have increase of Zc and NH2O.

- The correct answer to your question appears to be: neither! An increase of Z_C and $\overline{n}_{H_2O}$ can not be attributed to either the up- or down-regulated proteins alone; it is the difference between them.

- I have been unable to find a definitive statement about the proper English usage, but note some comments found online:

  - "Word expression: The difference between x and 7; Algebraic expression: x -7" http://www.leeward.hawaii.edu/files/mathlab/handouts/translating_words_into_algebra.pdf
  - ""the difference between x and y" is x-y." https://www.algebra.com/algebra/homework/expressions/expressions.faq.question.1055573.html
  - "D is the difference between each pair of X and Y scores (i.e. X - Y)" (Statistics in Plain English, Third Edition, p. 100 [Google Books])

- If the above is accepted, "the difference between <up> and <down>" implies the mathematical expression <up> - <down>, so that a positive sign means <up> is greater than <down>.

- Accordingly, I have revised the figure captions and other sentences in the text so they now read e.g. "differences between the mean for up-expressed and the mean for down-expressed proteins ...". Also, the words "for the means of up- minus down-expressed groups" have been inserted after the first mention of ΔZ_C and $\Delta\overline{n}_{H_2O}$ in the text.

6. It is not explained in the text why oxygen fugacity is added to equation 2.

- I would not say that oxygen fugacity is "added to" equation 2. That equation is a specific statement of the general expression for chemical affinity, aka negative Gibbs energy of reaction ($-\Delta G = A = 2.303RT\log(K/Q)$). Here, K is the equilibrium constant and Q is the activity quotient of the reaction. All of the species in Reaction R1 have activity or fugacity terms in Q. Because O_2 is one of the basis species, it appears in Reaction R1, and must therefore appear in Equation 2.

- Based on the above, one answer to your question is that the fugacity of O_2 appears in Equation 2 because O_2 is included among the basis species. As described previously, O_2 is the basis species that is representative of oxidation state.

- If the gist of your question is why *fugacity* rather than activity of O_2 is used, it is that gaseous rather than dissolved O_2 is the reference physical state most often used in previous thermodynamic models (mostly by geochemists). If a_{O_2} were used instead, its values would be offset from f_{O_2} according to the solubility of oxygen but otherwise the two models would be thermodynamically equivalent.

- The immediately preceding explanation has been inserted after Equation 2.

Validity of the findings

1. My major concern is disregarding the effect of metabolome changes on Zc and N_H2O since metabolic changes is one of the hallmarks of cancer.

- Please view this study as a starting point. The description of the chemical composition of the proteins has brought in new concepts (Z_C and $\overline{n}_{H_2O}$) to provide a fresh perspective on datasets from many previously published proteomic studies. This characterization of proteomic data is a large task that yields a new set of quantitative findings, but also opens the door to new problems.

- I do not mean to disregard other aspects of cells. The Discussion presents a possible scenario for coupling of oxidation of the proteome to increased synthesis of lipids. The possible effects of metabolic coupling between stromal and epithelial cells are also mentioned (Line 466; discussed in more detail in Dick, 2016). I'm sure there will be more better models for the interaction between chemical changes of the proteome and those other metabolites as well as with the microenvironment. But now there are very few biologists who are even concerned with the chemical changes that do occur in proteomes! We still need to take that first step.

2. I think there is a discussion point missing on whether the changes in protein concentrations are the result of physical constraints on protein synthesis/degradation or a manifestation of cellular response in terms of gene expression levels.

- I've added this statement to the Discussion: "The analysis done here is primarily concerned with top-down causal factors (physical constraints on protein synthesis and degradation), but does not preclude bottom-up factors (e.g. regulation of gene expression)."

- Both types of causes coexist. E.g., "higher level conditions influence what happens at the lower levels, even if the lower levels do the work. ... Note that the claim is not that the environment is the only relevant factor; rather that it is one of the causally effective factors. There will always be multiple causal factors, some bottom-up and some top-down; the final result comes from the confluence of these effects." (Ellis, 2012. Top-down causation and emergence: some comments on mechanisms. *Interface Focus* (2012) 2, 126–140. doi:10.1098/rsfs.2011.0062)

- Regarding the distinction between physical constraints and gene expression, a reasonable answer to the Reviewer's question is: "both". More biological evidence, and a review of the literature, would be needed to defend this position; such an extended discussion would be out of place here.

3. Speculative statements in conclusions should be moved to Discussion section.

- Changed as requested.

Comments for the Author

Other changes

- The **canprot** package has been updated to version 0.0.5. Compared to the submitted version, this involved a minor reorganization of code, without any changes in data or results.