

**Combing transcriptomes for secrets of deep-sea survival:
Environmental diversity reveals patterns of protein
evolution**

Journal:	<i>Integrative and Comparative Biology</i>
Manuscript ID	Draft
Manuscript Type:	Symposium article
Date Submitted by the Author:	n/a
Complete List of Authors:	Winnikoff, Jacob; Monterey Bay Aquarium Research Institute; University of California Santa Cruz, Ecology and Evolutionary Biology Francis, Warren; University of Southern Denmark Thuesen, Erik; Evergreen State College Haddock, Steven; Monterey Bay Aquarium Research Institute; University of California Santa Cruz, Ecology and Evolutionary Biology
Keywords:	convergence, biochemical adaptation, deep sea, ctenophore, transcriptome

SCHOLARONE™
Manuscripts

**Combing transcriptomes for secrets of deep-sea survival:
Environmental diversity reveals patterns of protein evolution**

Running title: Convergent protein adaptation in ctenophores

J.R. Winnikoff^{1,2}, W.R. Francis³, E.V. Thuesen⁴, and S.H.D. Haddock^{*1,2}

¹Monterey Bay Aquarium Research Institute, 7700 Sandholdt Rd., Moss Landing, CA 95039, USA

²University of California at Santa Cruz, 1156 High St., Santa Cruz, CA 95064, USA

³University of Southern Denmark, Campusvej 55, 5230 Odense, Denmark

⁴The Evergreen State College, 2700 Evergreen Pkwy. NW, Olympia, WA 98505, USA

* Corresponding author:

Steven H. D. Haddock

(831) 775-1793

haddock@mbari.org

ABSTRACT

Ctenophores, also known as comb jellies, live across extremely broad ranges of temperature and hydrostatic pressure in the ocean. Because various ctenophore lineages adapted independently to similar environmental conditions, Phylum Ctenophora is an ideal system for the study of protein adaptation to extreme environments in a comparative framework. We present such a study here, using a phylogenetically informed method to compare sequences of four essential metabolic enzymes across gradients of habitat depth and temperature. This method predicts convergent adaptation to these environmental parameters at the amino acid level, providing a novel view of protein adaptation to extreme environments and demonstrating the power and relevance of phylogenetic comparison applied to multi-species transcriptomic datasets from early-diverging metazoa.

INTRODUCTION

Organisms like ctenophores, cnidarians, and sponges are notable for having colonized a vast diversity of habitats. Ctenophores, for instance, can be found from the equator to polar seas and from sea level to over 7000 m deep. Since early-diverging metazoa tend to lack physical protection for their cells, much of their prodigious environmental adaptation is biochemical. As such, it implicates changes in certain vital proteins, such as metabolic “housekeeping” enzymes, whose performance is sensitive to the environment. These changes should in turn be traceable to organisms’ transcriptomes. Despite the growing abundance of transcriptomes available for early-diverging taxa from a wide variety of environments, techniques for correlating protein-coding sites to phenotypes in multi-species datasets remain underdeveloped.

Here we demonstrate the power of such an analysis by applying an evolutionary model-based tool to metabolic enzyme alignments from a set of 34 different ctenophore transcriptomes. Ctenophores, having evolved repeatedly to colonize shallow, deep, warm and cold habitats, present an attractive opportunity to explore patterns of protein adaptation to hydrostatic pressure and to temperature. In particular, repeated evolution of environmental tolerance phenotypes allows us to dissect the convergent nature of these adaptations and assess the size of the adaptive solution space within which biochemical evolution works.

Pressure and temperature: fundamental parameters of chemistry and biology alike

Each organism is a complex system of chemical reactions, with enzyme catalysts (Dong and Somero 2009) and substrates (Hochachka 1985) adaptively tuned to the environment that the organism inhabits. Pressure and temperature influence the rate and equilibrium of all chemical reactions, which is the principle underlying differences in environmental tolerance across species and motivating the study of pressure and temperature as evolutionary selectors. Effects of temperature on ectotherm metabolism are relatively well understood: enzymatic reaction rates and therefore respiration increase exponentially with environmental temperature until proteins begin to unfold, at which point respiration drops precipitously

and the organism typically dies (Pörtner 2001). This relationship between metabolic flux, protein folding, and temperature subjects enzymes to an evolutionary trade-off resulting in “marginal stability”, a state in which an enzyme is rigid enough not to denature, but flexible enough to achieve catalysis at its native temperature (Somero 2003). Cold-adapted enzymes tend to be highly efficient (high reaction velocity at low temperature) but unstable; vice versa for their warm-adapted orthologs (Fields et al. 2006; Lockwood and Somero 2012; Dong et al. 2018).

Metabolic effects of increased pressure have been generalized as mirroring those of low temperature (Menzies et al. 1974), because in some cases elevated temperature has been reported to rescue organisms from hyperbaric metabolic depression (Brown and Thatje 2011). In reality, organismal responses to pressure are highly taxon-specific. Metabolic effects of both pressure *and* the temperature-pressure interaction vary in direction. At unnatural pressures, respiration may decrease (Bailey et al. 1994) or increase (Morita 1957; Bartlett 2002) and in some cases, behavior appears to be unaffected (Cottin et al. 2012). When organismal metabolism is separated into its constituent chemical reactions, pressure effects appear more consistent and volume change of the system becomes the explanatory parameter (Somero 1990). Intuitively, reactions that increase in volume are inhibited by high pressure, while contractile reactions are enhanced. For a given enzymatic reaction, e.g. pyruvate to lactate, the net volume change cannot be altered, but the transient volume changes associated with substrate binding and conversion (i.e. “binding volume” $\Delta V_{E+S \rightarrow E \cdot S}$ and “transition volume” $\Delta V^\ddagger$) can be and often are. These values tend to be minimized in deep-adapted enzyme orthologs (Low and Somero 1976). Strong metabolic responses to pressure at the organismal and single-reaction levels (Teal and Carey 1967), along with adaptive structural modifications in a variety of enzymes and cytoskeletal proteins (Siebenaller and Somero 1978; Somero and Siebenaller 1979; Swezey and Somero 1985; Morita 2008), demonstrate that hydrostatic pressure exerts considerable selective pressure. It is at least as pervasive a selector as temperature, since >98% of Earth’s habitable volume lies between, 200 and 11000 m ocean depth (Dawson 2012) and therefore under 20 to 1100 bar of pressure.

Function to structure: sparsely charted waters

Since the unifying functional concepts of marginal stability and volume change were implicated in temperature- and pressure-adaptation of proteins, the open question has been how these characteristics map to protein-coding loci in the genome. Most progress has been made in the realm of thermal stability, where ubiquitous structural themes have been identified. These include substitution of uncharged amino acid residues with charged ones, which form hydrogen bonds to hold domains or subunits together (Dong and Somero 2009; Lockwood and Somero 2012) and substitution with bulky hydrophobic amino acids to facilitate stabilizing van der Waals interactions in the protein core (Haney et al. 1999). Substitution with proline can also introduce turns at critical sites in the protein backbone, limiting its flexibility (Bogin et al. 1998).

There are two leading structural hypotheses for pressure-adaptation of proteins. The first implicates hydrophilicity, positing that since water potential is higher under pressure, hydrophilic configurations are favored and pressure-tolerant proteins either (a) evolve to maximize surface hydration or (b) in the case of enzymes, evolve to minimize *change* in surface hydration between the free and substrate-bound conformations (Somero 1990). To a first approximation, this implies that high-pressure-adapted proteins should have more hydrophilic residues in contact with solvent than their low-pressure orthologs. Indirect support for this hypothesis comes from work with the small molecule trimethylamine *N*-oxide (TMAO), which protects proteins against pressure perturbation in deep-sea fishes (Yancey et al. 2004). TMAO alters the structure of water (Sharp et al. 2001) so that it more thoroughly hydrates macromolecules (Schroer et al. 2016). If this effect can stabilize proteins against pressure, it stands to reason that intrinsically hydrophilic surfaces might as well.

The second structural hypothesis concerns buried amino acid residues rather than solvent-exposed ones. It proposes simply that bulkier, less compressible amino acid side chains can fill “voids” in a protein’s interior, making it resistant to structural perturbation by pressure (Chalikian and Macgregor 2009). An analysis of globular protein compressibilities (Gromiha and Ponnuswamy 1993) supported this

hypothesis by finding the strongest determinant of protein compressibility to be compressibilities of its constituent amino acids. The predictive model that arose from that study is an attractive step toward describing the structural basis of pressure tolerance, but requires improvement in two major areas. First, being composition-driven, the model cannot account for the sequence context of amino acids nor effects of secondary structure, which are known to be significant (Gross and Jaenicke 1994). Second, its relevance to enzymes remains unclear because a link has yet to be established between adiabatic compressibility and biologically relevant properties like substrate binding affinity and maximal reaction velocity.

Thanks to next-gen sequencing, we are able to use a comparative transcriptomic approach to map organismal environmental tolerance phenotypes to protein-coding loci. As a comparative method, this requires sampling organisms across the broadest possible ranges of pressure and temperature. It is equally imperative that our comparative analysis take phylogeny into account and derive its power from multiple independent (i.e. convergent) adaptive events (Dunn et al. 2018). Phylogeny-agnostic methods are prone to “phylogenetic pseudoreplication” (Garland 2001), where statistical power is derived from multiple species even though they share homologs of a single adaptive innovation.

Enter the ctenophore

Ctenophores, or “comb jellies” are a clade of marine animals fulfilling both of the above requirements. They can be found living at -2 to 30°C, from the surface to over 7000 m depth (#CTENODIVERSITYA; pressure increases 1 bar for every 10 m sea water) and their phylogeny reveals repeated colonization of deep and shallow waters. In addition, as an early-diverging lineage with hundreds of millions of years of distinct evolutionary history (Dunn et al. 2015), they present an attractive opportunity to probe convergent molecular evolution on a grand scale by way of eventual comparison to Cnidaria. Having assembled transcriptomes for 34 ctenophore species, we have sufficient data both to infer a working ctenophore phylogeny (#CTENODIVERSITYB) and to analyze the amino acid sequences of pressure- and temperature-sensitive proteins.

Dissecting the layers of convergent adaptation

It is apparent in #CTENODIVERSITY that several distantly related ctenophores are convergently adapted to live at similar depths. Such homoplasy at the organismal level may be the most common meaning of “convergent adaptation,” but homoplasy likely extends further into the pathways and genes associated with environmental tolerance. Detecting convergence at the amino acid site scale allows us to resolve the level at which convergent adaptation is taking place. #CONVERGELEVELS illustrates three different levels of convergence that could explain the same organism-level phenotypic pattern. Consider a hypothetical environmental tolerance phenotype and a metabolic pathway featuring three enzymes. (1) Convergence can exist at the **pathway level** when a derived phenotype is produced by adaptations in the same pathway, but not necessarily in the same gene. For example, in species B, the first enzyme compensates for decreased binding affinity in the second enzyme by feeding it more substrate, whereas this is unnecessary in species A because the second enzyme has evolved a higher binding affinity. Total flux through the pathway is the same in both species. (2) Convergence at the **gene level** implies that the same gene is responsible for the derived phenotype in both species, but the adapted sites within this gene and its enzyme product are not necessarily the same in species A and B. (3) Convergence at the **site level** requires the same sites in the same gene to account for the derived phenotype. Extensive site-level convergence implies a small evolutionary solution space for the adaptive challenge in question. Note that other authors (Stern 2013; Lee and Coop 2017) have referred to (1) above as *convergent adaptation* and (3) as *parallel adaptation*. Viewing convergence as a matter of degree rather than kind (Arendt and Reznick 2007) motivates the terms used here. All three of the illustrated patterns contribute to other kinds of environmental adaptation (Riehle et al. 2001) and so may all play a role in the cases of pressure and temperature as well.

Structure-functional hypotheses: needles in a transcriptomic haystack

Using a phylogenetically informed comparative method to resolve convergent adaptation at the amino acid site scale is conceptually straightforward, but impractical without post-genomic methods. Traditional comparative biochemistry would prescribe choosing pairs of orthologous proteins from differently-adapted congeners, then reciprocally mutating every residue that differs between the orthologs and assaying the mutants under relevant environmental conditions (Morita 2008). With 60-100 sites differing for such a pair of typical 500-residue globular proteins, this would amount to a heroic mutagenesis effort (Kazlauskas and Bornscheuer 2009). Expanding the scope of convergence to consider multiple interacting proteins becomes intractable. Transcriptome analysis alleviates this problem by identifying a manageable number of high-likelihood hypotheses that can be tested by site-directed mutagenesis. Once a convergence-detection algorithm has been experimentally validated at the site scale, it can be applied to complete transcriptomes to identify additional genes associated with a phenotype and to assess the multi-gene architecture of convergent phenotypes as discussed above.

METHODS

Specimen and data collection

Ctenophores were collected using blue water SCUBA techniques (0-20 m depth) and ROVs *Ventana* and *Doc Ricketts* (200-4000 m depth). Median depth and temperature of occurrence was determined for each species using the MBARI VARS database (Schlining and Stout 2006). For species not collected on MBARI expeditions, environmental parameters were estimated using CTD data from the collection locale. $\log_{10}(\text{depth})$ was used in our analyses, consistent with similar previous investigations (Porter et al. 2016) and with the adiabatic gas law.

Primary metabolism was chosen as the biological system on which to test our hypotheses. Metabolic pathways like glycolysis and the Krebs cycle are essential to survival and involve a set of known interactions between well-characterized enzymes. Many of these enzymes exhibit pressure effects in other marine taxa (Siebenaller and Somero 1978; Somero and Siebenaller 1979; Dahlhoff and Somero

1991; Gerringer et al. 2017). Amino acid sequences for the enzymes D-lactate dehydrogenase (LDH, EC 1.1.1.28), cytosolic and mitochondrial malate dehydrogenase (cMDH and mMDH, EC 1.1.1.37) and pyruvate kinase (PK, EC 2.7.1.40) aggregated using OrthoFinder (Emms and Kelly 2015) from 34 transcriptomes assembled using Trinity (Grabherr et al. 2013) and translated with TransDecoder (Haas et al. 2013). Transcriptomes were derived from comb row tissue where possible, since this tissue exhibits high activity for all the above enzymes (Thuesen, unpublished). The guide phylogeny for comparative analysis was inferred using RaXML on an alignment of 248 single-copy orthogroups.

Site-level convergence detection

Convergently adapted sites were detected using the PCOC (Profile Change-One Change) algorithm (Rey et al. 2018). In brief, this method works by applying a 60-profile CAT model of protein evolution in a maximum likelihood framework (Quang et al. 2008) along a species tree and determining, on a site-wise basis, whether branches designated as having a “convergent” phenotype conform to a different amino acid profile than branches with the “ancestral” phenotype. Also incorporated in PCOC is the frequency of amino acid changes on transition branches between the ancestral and convergent phenotypes. To accommodate our analysis of continuous environmental tolerance phenotypes, several extensions were made to the existing PCOC workflow and implemented as a Python wrapper script available here: github.com/octopode/continuous-converge.

Convergent scenario formulation

The PCOC method as published requires the investigator to reconstruct the trait of interest along the species tree using binary states. This is straightforward in the case of a naturally dichotomous character like C4 metabolism in plants (Rey et al. 2018), but non-trivial when dealing with continuous characters like temperature and depth of occurrence. We used the following routine to propose convergent scenarios consistent with trait values and phylogeny:

First, a simple ancestral trait reconstruction was performed along the species tree. Trait values (temperature or $\log_{10}(\text{depth})$) were recorded for each internal and terminal node, then binned using a set width of 1.75°C or $0.2 \log_{10}(\text{m})$. Finally, trait cutoffs were placed between each bin and each cutoff was used to place a set of transition events on the tree (#PCOCOCONTINUOUSA), comprising a “convergent scenario” in which the continuous trait of interest was rendered binary. Because PCOC only registers transitions in one direction, this binary trait was inverted for all scenarios in which the root reconstructed as “convergent”. In biological terms, this amounted to calling the “shallow” or “cold” state (low trait values) convergent and the “deep” or “warm” state (high trait values) ancestral. We entertained both possibilities because reconstructing the actual habitat of the ctenophore LCMA is outside the scope of this analysis. The resultant set of scenarios was filtered to ensure that each is unique and that a minimum of 4 independent transition events are present. The latter mitigated false positives due to phylogenetic pseudoreplication. The PCOC algorithm was run on each enzyme alignment under 18 depth convergence scenarios and 6 temperature convergence scenarios (#PCOCOCONTINUOUSB). The maximum PCOC posterior probability (PP) across all scenarios was recorded for each site, provided it fell above a threshold of 0.5 (#PCOCOCONTINUOUSC). This rule was used to reduce computational load in the following step.

Evolutionary simulation-based bootstrapping

The posterior probability value returned by PCOC gives a likelihood-based relative measure of whether an amino acid site evolved convergently in association with a phenotype. It does not provide a controlled level of statistical confidence. As this extra measure of confidence is desirable for designing mutagenesis experiments, we used the simulation feature included in the PCOC toolkit to implement an automatic bootstrap test (#PCOCOCONTINUOUSD). For each site, the maximum-likelihood CAT amino acid profile was recovered for both the ancestral and convergent branches. The CAT model was then used to repeatedly simulate evolution along the species tree. To obtain false positive rate α (1 - specificity), non-convergent evolution was simulated using only the ancestral profile with noise added. To obtain false

negative rate β (1 - sensitivity), convergent evolution was simulated using both profiles, with transitions placed according to the maximum-PP convergent scenario. 1000 of each non-convergent and convergent simulations were run per site.

This bootstrap test improves upon existing PCOC simulation facilities by providing positive and negative calling confidence thresholds based on a more site-specific model. Where the previous tool allowed simulation of non-convergent and convergent evolution along the actual species tree using random pairs of amino acid profiles, our extension specifies the profiles of best fit to the alignment column in question. If more conservative bootstrapping is desired, it still permits incorporation of branch-length uncertainty and other sources of noise into the model.

Protein structural analysis

Though PCOC has yet to be empirically validated with respect to temperature- and pressure-adaptation, we performed some simple hypothesis tests downstream of convergent site prediction to illustrate the types of structure-functional relationships comparative transcriptomic analysis can be used to explore. B-factor profile, a measure of predicted local conformational mobility, was transformed using Ordered Quantile Normalization and modeled as a function of predicted depth- and temperature-adaptiveness using 2-way ANOVA (Type III sum of squares). To ascertain biases in secondary structure composition and local environment of predicted adaptive sites, Fisher's Exact test was used. Unresolved sites were omitted from the analysis because they are subject to sampling bias: inconclusivity is an artifact caused by insufficient taxon sampling. All statistical tests were performed in R. To model 3D structures from which B-factor profile, secondary structure, and local environment data could be extracted, the I-TASSER suite (Roy et al. 2010) was applied to enzyme sequences from *Bolinopsis infundibulum*. The crystal structure used as a threading template was specified manually for LDH to ensure that its EC number agreed with that of the query protein.

RESULTS

Frequency of adaptive sites, overlap of pressure- and temperature-adaptation

Convergent adaptation associated with habitat depth was predicted at 2.0-13.5% of sites in each enzyme. Convergent adaptation associated with temperature was predicted at 1.8-11.5% of sites and 0.9-7.8% of sites were associated with both environmental parameters. For brevity, sites predicted by PCOC to be adaptive to depth or temperature are referred to here as adaptive sites. #VENN illustrates the overlap of depth- and temperature-adaptive sites across all four enzymes.

Local environment of adaptive sites

#SITESTATS details the distribution of predicted adaptive sites within predicted enzyme structures. Adaptive sites were overrepresented in solvent-exposed regions, but no bias was found in their distribution among helix, sheet and coil secondary structures except in LDH. In LDH, depth-adaptive sites occurred in significantly more flexible parts of the protein; temperature did not correlate significantly to B-factor profile in any of the four enzymes. Residue counts and more detailed statistics can be found in Table S1. In LDH and cMDH, several adaptive sites were predicted in the NADH binding pocket, a Rossman fold motif at the N-terminus of the dehydrogenases. These and other adaptive sites are indicated in #MANHATTAN. Similar figures for all four of the enzymes and both of the environmental parameters analyzed can be found in Supplementary Materials.

DISCUSSION

Consistency with historical high-pressure enzymology results

The four enzymes analyzed here were chosen because they are essential to primary metabolism and because their functional responses to pressure have been characterized in other taxa. PK exhibits decreased maximal velocity under pressure, but the mechanism for this is obscure beyond that the enzyme

has positive activation volume, undergoing a transient expansion to convert its substrate. Around 1000 bar, subunits of rabbit PK are known to dissociate (de Felice et al. 1999), but absence of depth-adaptive sites in the subunit interface suggests that this may not occur in ctenophores under biologically relevant pressures. LDH and MDH are both NAD⁺-dependent dehydrogenases. Members of this enzyme family are highly pressure-sensitive in shallow-living fishes (Somero and Siebenaller 1979), crustaceans, molluscs and annelids, due in part to pressure-inhibited binding of their cofactor NADH (Somero and Siebenaller 1979; Dahlhoff and Somero 1991). The Rossmann fold domain responsible for binding this cofactor constitutes roughly 30 amino acids at the N-terminus of each enzyme. This is the only region where depth-adaptive sites were predicted in all three of the Rossmann-fold dehydrogenases analyzed. None of the depth-adaptive sites were in direct contact with NADH; contact amino acids are highly conserved, but the depth-adaptive sites could affect compressibility of the Rossmann fold and thus its ability to bind NADH under pressure.

Potential implications for environmental adaptation of proteins

There are two possible explanations for the large overlap of sites predicted to be adaptive to pressure and temperature. First and most conservative is that depth and temperature of encounter covary too strongly in our dataset. In this case, sampling greater temperature diversity among shallow taxa will better resolve sites adapted to each environmental parameter and may decrease their overlap. Second, it is possible that these sites are actually adapted to both pressure and temperature, or adapted to one and exapted to the other. If expanded taxon sampling fails to support the first explanation, then the second must be tested experimentally (see further work).

Three trends emerged from analysis of adaptive sites' local environment in the protein. First, adaptive sites tended to occur in parts of the protein with above-average conformational flexibility, i.e. B-factor profile. Second, the secondary structure composition of adaptive sites was not significantly different from that of non-adaptive sites, and where this signal was strongest it biased towards alpha-

helices. Third, adaptive sites were more likely than non-adaptive sites to be in contact with solvent and in LDH, *all* subunit interface sites were predicted to be adaptive.

The first pattern is consistent with the marginal stability concept: If it is functionally critical that an enzyme’s mobile regions remain mobile, then it follows that these regions would adapt to counteract immobilizing effects of elevated pressure and reduced temperature. The second finding fails to support a relevant observation by Gross and Jaenicke (1994) that formation of the alpha-helix is favored by pressure over that of the beta-sheet. If this difference in compressibility is evolutionarily relevant, one might expect to find disproportionately many depth-adaptive sites in vulnerable beta-sheet regions. We did not observe a strong signal in the opposite direction, and it remains possible that pressure adaptation has a beta-sheet bias in larger protein sets. The third finding that adaptive sites are disproportionately common on the exterior of protein subunits was unsurprising, since buried residues tend to be more conserved. Still, it underscores that hydration and ligand interactions may be some of the more pressure-sensitive aspects of enzymes and raises the possibility that in these proteins, compression of internal voids may play a secondary role. All of the patterns identified here can be more rigorously evaluated and generalized by applying the same analysis to a much larger orthoset.

Applications: from evolutionary biology to protein engineering

Environmental tolerances of proteins and pathways are of obvious interest to biologists seeking to understand organisms’ distributions in space and time. Abiotic factors like temperature and pressure often dictate a species’ fundamental niche. Knowledge of the number and type of mutations required to access a particular niche could lead to a better understanding of evolutionary lineages’ ecological history. It could also help to model how quickly various lineages might be able to adapt with global change.

Beyond basic science, temperature and pressure tolerance of proteins is a focus of biotechnology. Surveying the thermal diversity of life has yielded such popular biotech tools as thermostable *Taq* polymerase and heat-labile Shrimp Alkaline Phosphatase. Many chemical synthesis processes are conducted under pressure (Kunugi 1992), so exploration of bathyal diversity could yield analogous

advances. At present, however, bioengineers are largely limited to what proteins natural and artificial selection provide. They can optimize these proteins for various applications using rational design, but extensive structural knowledge is needed to make mutations that maintain function, let alone adjust it toward a specification. Molecular convergence-detection methods can narrow the vast parameter space associated with site-directed mutagenesis. With just an alignment and phylogeny from an appropriate set of taxa, a protein engineer might generate a tractable number of candidate mutations to achieve a desired behavior at particular temperature, pressure, pH, ROS level, etc.

Further work

It is essential that comparative transcriptomic techniques continue to be developed, tested, and refined. While the method reported here expands upon an innovative and promising technique introduced just last year, it can already be improved in several ways: First, convergent scenario formulation should be refined using a more sophisticated trait evolution model than Brownian motion. While it provides a useful approximation of ancestral traits, numerous problems with this model have been identified (Pennell et al. 2015). As implemented, it does not account for uncertainty in trait values and may facilitate placement of spurious convergent trait transitions. Though PCOC was found to be fairly robust to such errors (Rey et al. 2018), they increase the risk of false positives in principle. Second, when working with a transcriptome-scale dataset, it is appropriate to train the evolutionary model being used (in this case, the CAT model) on the dataset itself, rather than some external large alignment (Quang et al. 2008). Third, it would be useful to apply an evolutionary model with deletion and insertion parameters. PCOC uses tree pruning to handle gapped sites, but this strategy is not able to predict convergently adaptive insertions or deletions, features that are biologically interesting and useful to protein engineers.

At this point in time, comparative transcriptomic analyses are best used as powerful hypothesis generators. Algorithms like PCOC can identify patterns that have a very low likelihood of arising by chance, yet there are myriad covarying selective factors that could lure the overzealous investigator into a correlation-causation pitfall. Only controlled experiments can mitigate this risk and ultimately validate

any predictive method. To this end, we will mutate predicted adaptive sites in the enzymes analyzed here and assay the mutant gene products across gradients of pressure and temperature. If these mutations cause the expected changes in pressure and temperature tolerance, then groundwork will be laid to apply comparative transcriptomics to countless other dimensions of biochemical diversity.

FUNDING

This work was supported by the National Science Foundation [DEB-1542679]; National Institutes of Health [ROI-GMO87198]; and the David and Lucile Packard Foundation.

ACKNOWLEDGEMENTS

The authors would like to thank Prof. Joseph Ryan and Dr. Melissa DeBiasse for performing OrthoFinder analysis. Lynne Christianson and Shannon Johnson provided indispensable support in collecting and cataloging the animals. Highly skilled ship and ROV crews aboard *R/Vs Rachel Carson* and *Western Flyer* were likewise essential to the collection effort. The presenting author is grateful to Darrin T. Schultz for his ready supply of programming and editorial advice.

REFERENCES

Arendt J, Reznick D. 2007. Convergence and parallelism reconsidered : what have we learned about the genetics of adaptation? TREE 23:26–32.

Bailey TG, Torres JJ, Youngbluth MJ, Owen GP. 1994. Effect of decompression on mesopelagic gelatinous zooplankton: a comparison of in situ and shipboard measurements of metabolism. Mar Ecol Prog Ser 113:13–27.

Bartlett DH. 2002. Pressure effects on in vivo microbial processes. Biochim Biophys Acta 1595:367–381.

- Bogin O, Peretz M, Hacham Y, Korkhin Y, Frolow F, Gilboa AJK, Burstein Y. 1998. Enhanced thermal stability of *Clostridium beijerinckii* alcohol dehydrogenase after strategic substitution of amino acid residues with prolines from the homologous thermophilic *Thermoanaerobacter brockii* alcohol dehydrogenase. *Prot Sci* 7:1156–1163.
- Brown A, Thatje S. 2011. Respiratory Response of the Deep-Sea Amphipod *Stephonyx biscayensis* Indicates Bathymetric Range Limitation by Temperature and Hydrostatic Pressure. *PLoS ONE* 6:e28562.
- Chalikian TV, Macgregor RBJ. 2009. Origins of Pressure-Induced Protein Transitions. *J Mol Biol* 394:834–842.
- Cottin D, Brown A, Oliphant A, Mestre NC, Ravaux J, Shillito B, Thatje S. 2012. Sustained hydrostatic pressure tolerance of the shallow water shrimp *Palaemonetes varians* at different temperatures: Insights into the colonisation of the deep sea. *Comp Biochem Physiol Mol Integr Physiol* 162:357–363.
- Dahlhoff EA, Somero GN. 1991. Pressure and temperature adaptation of cytosolic malate dehydrogenases of shallow and deep-living marine invertebrates: evidence for high body temperatures in hydrothermal vent animals. *J Exp Biol* 159:473–487.
- Dawson MN. 2012. Species richness, habitable volume and species densities in freshwater, the sea, and on land. *Front Biogeog* 4:105–116.
- de Felice FG, Soares VC, Ferreira ST. 1999. Subunit dissociation and inactivation of pyruvate kinase by hydrostatic pressure Oxidation of sulfhydryl groups and ligand effects on enzyme stability. *Eur J Biochem* 266:163–169.
- Dong Y, Liao M, Meng X, Somero GN. 2018. Structural flexibility and protein adaptation to temperature: Molecular dynamics analysis of malate dehydrogenases of marine molluscs. *Proc Nat Acad Sci USA* 115:1274–1279.
- Dong Y, Somero GN. 2009. Temperature adaptation of cytosolic malate dehydrogenases of limpets (genus *Lottia*): differences in stability and function due to minor changes in sequence correlate

- with biogeographic and vertical distributions. *J Exp Biol* 212:169–177.
- Dunn CW, Leys SP, Haddock SHD. 2015. The hidden biology of sponges and ctenophores. *Trends Ecol Evol* 30:282–291.
- Dunn CW, Zapata F, Munro C, Siebert S, Hejnal A. 2018. Pairwise comparisons across species are problematic when analyzing functional genomic data. *Proc Nat Acad Sci USA* 115:409–417.
- Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol* 16.
- Fields PA, Rudomin EL, Somero GN. 2006. Temperature sensitivities of cytosolic malate dehydrogenases from native and invasive species of marine mussels (genus *Mytilus*): sequence-function linkages and correlations with biogeographic distribution. *J Exp Biol* 209:656–667.
- Garland T. 2001. Phylogenetic comparison and artificial selection: Two approaches in evolutionary physiology. In: *Hypoxia: From Genes to the Bedside* p. 107–132.
- Gerringer ME, Drazen JC, Yancey PH. 2017. Metabolic enzyme activities of abyssal and hadal fishes: pressure effects and a re-evaluation of depth-related changes. *Deep-Sea Res* 125:135–146.
- Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, Palma F, Birren BW, Nusbaum C, Lindblad-toh K, Friedman N, Regev A. 2013. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nat Biotechnol* 29:644–652.
- Gromiha MM, Ponnuswamy PK. 1993. Relationship between amino acid properties and protein compressibility.
- Gross M, Jaenicke R. 1994. Proteins under pressure. The influence of high hydrostatic pressure on structure, function and assembly of proteins and protein complexes. *Eur J Biochem* 221:617–630.
- Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, MacManes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, LeDuc RD, Friedman N, Regev A. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis.

- Nat Protoc 8:1494–1512.
- Haney PJ, Badger JH, Buldak GL, Reich CI, Woese CR, Olsen GJ. 1999. Thermal adaptation analyzed by comparison of protein sequences from mesophilic and extremely thermophilic *Methanococcus* species. *Proc Nat Acad Sci USA* 96:3578–3583.
- Hochachka PW. 1985. Fuels and pathways as designed systems for support of muscle work. *J Exp Biol* 115:149–164.
- Kazlauskas RJ, Bornscheuer UT. 2009. Finding better protein engineering strategies. *Nat Chem Biol* 5:526–529.
- Kunugi S. 1992. Enzyme Reactions under High Pressure and Their Applications. *Ann NY Acad Sci* 672:293–304.
- Lee KM, Coop G. 2017. Distinguishing among modes of convergent adaptation using population genomic data. *Genetics* 207:1591–1619.
- Lockwood BL, Somero GN. 2012. Functional determinants of temperature adaptation in enzymes of cold- versus warm-adapted mussels (Genus *Mytilus*). *Mol Biol Evol* 29:3061–3070.
- Low PS, Somero GN. 1976. Adaptation of Muscle Pyruvate Kinases to Environmental Temperatures and Pressures. *J Exp Zool* 198:1–12.
- Menzies RJ, George RY, Paul AZ. 1974. The Effects of Hydrostatic Pressure on Living Aquatic Organisms. *Int Rev Ges Hydrobiol* 59:287–197.
- Morita RY. 1957. Effect of hydrostatic pressure on succinic, formic, and malic dehydrogenases in *Escherichia coli*. *J Bacteriol* 74:251–255.
- Morita T. 2008. Comparative sequence analysis of myosin heavy chain proteins from congeneric shallow- and deep-living rattail fish (genus *Coryphaenoides*). *J Exp Biol* 211:1362–1367.
- Pennell MW, Fitzjohn RG, Cornwell WK, Harmon LJ. 2015. Model Adequacy and the Macroevolution of Angiosperm Functional Traits. *Am Nat* 186.
- Porter ML, Roberts NW, Partridge JC. 2016. Molecular Phylogenetics and Evolution Evolution under pressure and the adaptation of visual pigment compressibility in deep-sea environments. *Mol*

- Phylogenet Evol 105:160–165.
- Pörtner HO. 2001. Climate change and temperature-dependent biogeography: oxygen limitation of thermal tolerance in animals. *Naturwissenschaften* 88:137–146.
- Quang LS, Gascuel O, Lartillot N. 2008. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics* 24:2317–2323.
- Rey C, Gu L, Marie S, de Lyon ENS, Umr C, Inserm U, Biologie LD, Lyon D. 2018. Accurate Detection of Convergent Amino-Acid Evolution with PCOC. *Mol Biol Evol* 35:2296–2306.
- Riehle MM, Bennett AF, Long AD. 2001. Genetic architecture of thermal adaptation in *Escherichia coli*. *Proc Nat Acad Sci USA* 98:525–530.
- Roy A, Kucukural A, Zhang Y. 2010. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc* 5:725–738.
- Schlining BM, Stout NJ. 2006. MBARI's Video Annotation and Reference System. In: *IEEE OCEANS* 2006.
- Schroer MA, Michalowsky J, Fischer B, Gru G. 2016. Stabilizing effect of TMAO on globular PNIPAM states: preferential attraction induces preferential. *Phys Chem Chem Phys* 18:31459–31470.
- Sharp KA, Madan B, Manas E, Vanderkooi JM. 2001. Water structure changes induced by hydrophobic and polar solutes revealed by simulations and infrared spectroscopy. *J Chem Phys* 114:1791–1796.
- Siebenaller JF, Somero GN. 1978. Pressure-Adaptive Differences in Lactate Dehydrogenases of Congeneric Fishes Living at Different Depths. *Science* 201:255–257.
- Somero GN. 1990. Life at low volume change: Hydrostatic pressure as a selective factor in the aquatic environment. *Am Zool* 30:123–135.
- Somero GN. 2003. Protein adaptations to temperature and pressure: complementary roles of adaptive changes in amino acid sequence and internal milieu. *Comp Biochem Physiol B Biochem Mol Bio* 136:577–591.
- Somero GN, Siebenaller JF. 1979. Inefficient lactate dehydrogenases of deep-sea fishes. *Nature* 282:100–

1
2
3 102.
4

5 Stern DL. 2013. The genetic causes of convergent evolution. *Nat Rev Genet* 14:751–764.
6

7 Swezey RR, Somero GN. 1985. Pressure Effects on Actin Self-Assembly: Interspecific Differences in the
8
9 Equilibrium and Kinetics of the G to F Transformation *. *Biochemistry* 24:852–860.
10

11 Teal JM, Carey FG. 1967. Effects of pressure and temperature on the respiration of euphausiids. *Deep-*
12
13 *Sea Res* 14:725–733.
14

15 Yancey PH, Rhea MD, Kemp KM, Bailey DM. 2004. Trimethylamine oxide, betaine and other osmolytes
16
17 in deep-sea animals: depth trends and effects on enzymes under hydrostatic pressure. *Cell Molec*
18
19 *Biol* 50:371–376.
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

FIGURE CAPTIONS

#CTENODIVERSITY

Ctenophore habitat diversity:

- A. Depth vs. temperature for collected species, overlaid with typical depth-temperature profiles for collection locales. Range of the Ctenophora encompasses three main types of habitat: shallow-warm, shallow-cold and deep-cold. No species have yet been reported from deep hydrothermal vents.
- B. Brownian motion ancestral reconstruction of habitat depth across sequenced ctenophore species. All three orders of ctenophore (Cydippida, green; Lobata, violet; and Beroida, orange) are well-represented in each of the three habitat types. Cydippida is polyphyletic (Podar et al., 2001).

#CONVERGELEVELS

Convergent adaptation observed at the organism level can be caused by convergence at various sub-organismal scales. Three possible explanations for organism-level convergence are illustrated here using a hypothetical metabolic pathway.

#PCOCONTINUOUS

A schematic of PCOC implementation for a continuous trait, beginning with convergent scenario formulation, proceeding to identify the highest-posterior probability (PP) trait cutoff at each site and finally running a bootstrap test on a high-PP site (site 3 shown) to yield PP confidence thresholds as reported in a column of the Manhattan plot. PP (represented by the black diamond) within the lower/green area indicates a convergently adapted site to a confidence of 95%, a PP within the upper/red bar indicates a non-convergently-adapted site to a confidence of 95% and a PP in the white area or in the overlap of red and green indicates unresolved site status.

#VENN

Visual representation of the number of amino acid sites predicted adaptive to habitat depth, temperature or both across all four enzymes analyzed.

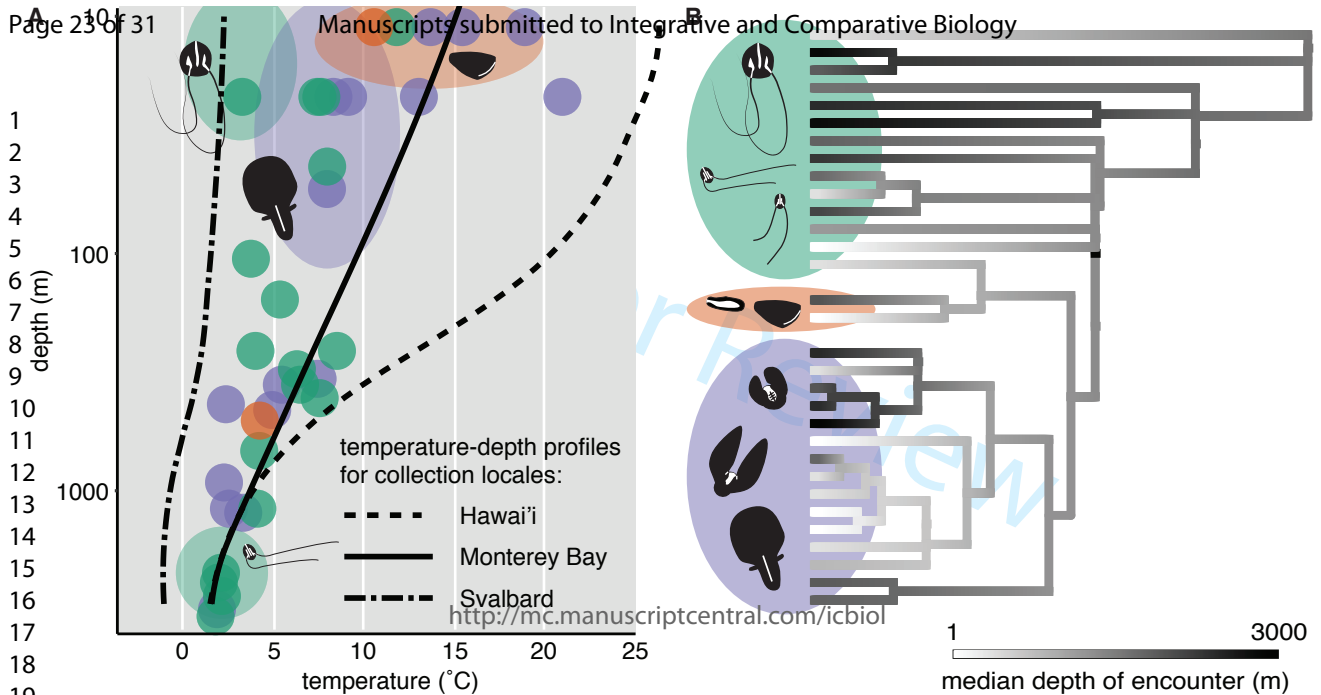
#SITESTATS

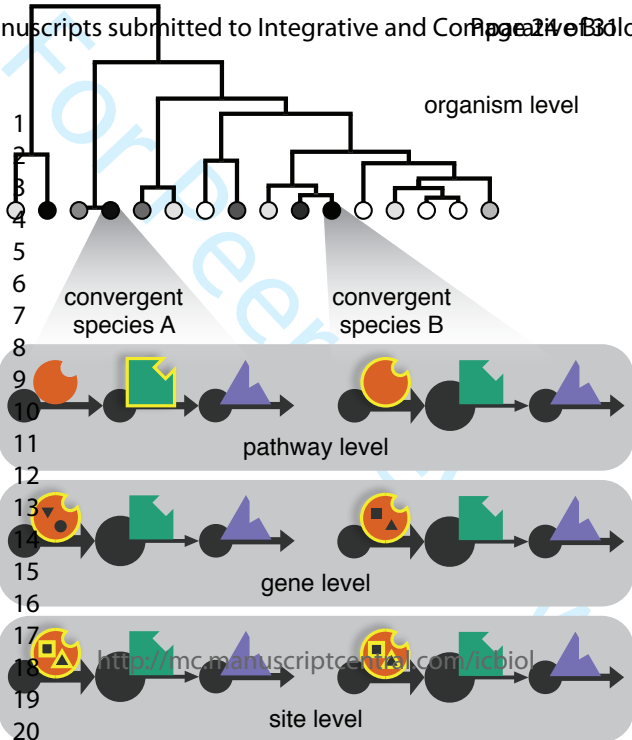
A comparison of the local environments of adaptive and non-adaptive sites within the enzymes pyruvate kinase (PK), cytosolic malate dehydrogenase (cMDH) and lactate dehydrogenase (LDH). Sites predicted to be adaptive to depth, temperature and both parameters are compared to those predicted to be adaptive to neither parameter. Comparisons found to be significant to a level of $\alpha < 0.05$ are labeled in red text. Distribution of sites are considered (A) across a range of backbone mobility (B-factor profile values) as well as among (B) Helix, Sheet and Coil (H, S, C) secondary structures (SS) and (C) solvent-Exposed, subunit Interface and Buried (E, I, B) locations (loc.).

#MANHATTAN

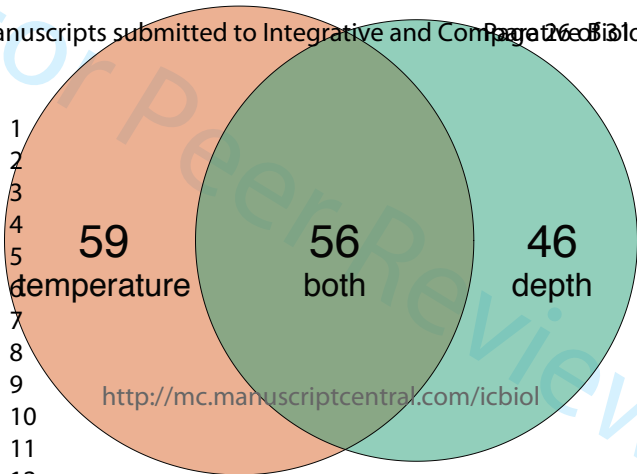
Alignments of cMDH (top) and LDH (bottom) enzymes: predicted depth-adaptive sites are highlighted and indicated with a red pointer. The black partition with red dot in the alignment column indicates the trait discretization cutoff used to propose the convergent evolutionary scenario. PCOC posterior probability and bootstrap confidence thresholds are plotted above the alignment as in

#PCOCONTINUOUS.

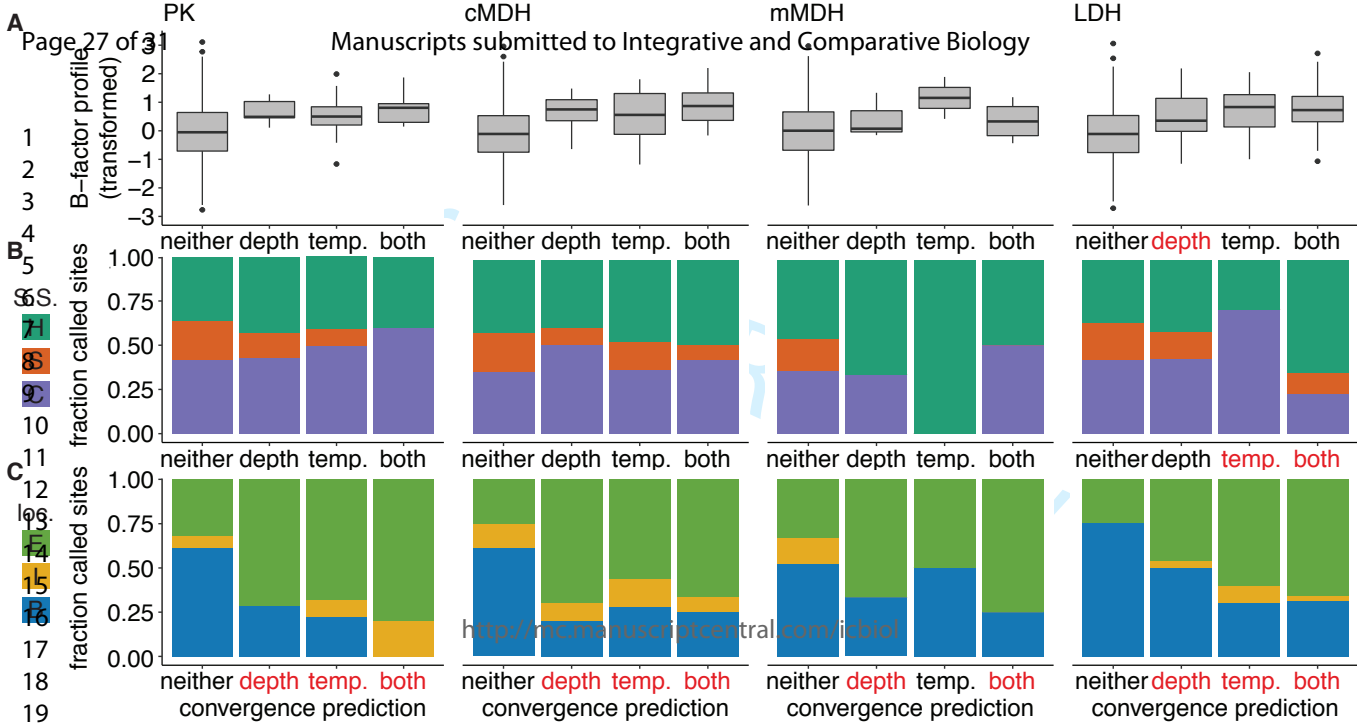


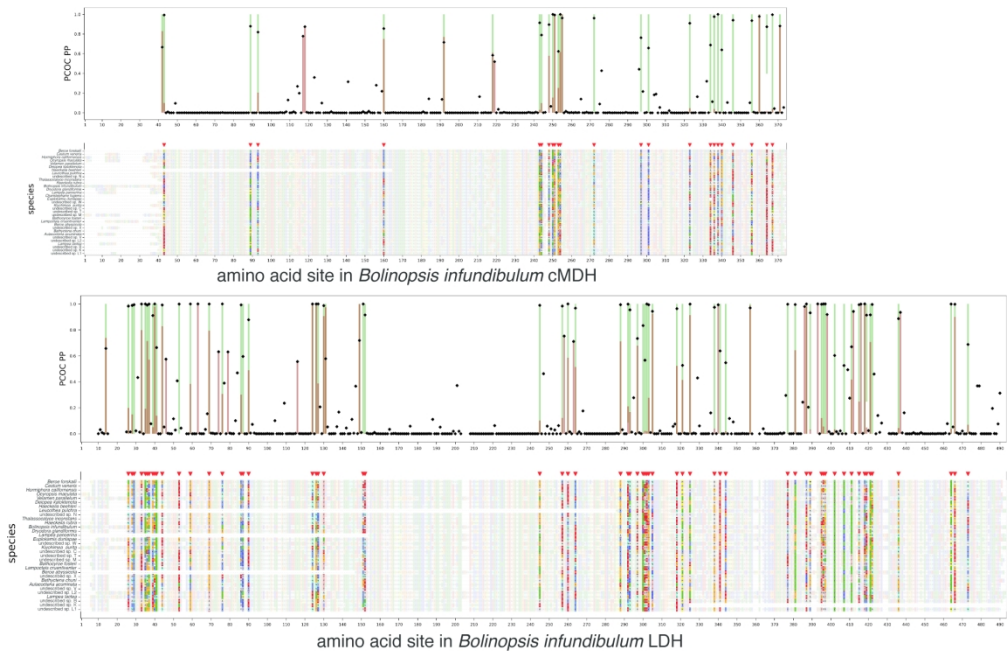






<http://mc.manuscriptcentral.com/icbiol>





Alignments of cMDH (top) and LDH (bottom) enzymes: predicted depth-adaptive sites are highlighted and indicated with a red pointer. The black partition with red dot in the alignment column indicates the trait discretization cutoff used to propose the convergent evolutionary scenario. PCOC posterior probability and bootstrap confidence thresholds are plotted above the alignment as in #PCOCONTINUOUS.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

residue location econdary structur B-factor profile

PK					cMl	
	SS	df	F	p	SS	df
Depth	1.46	1	1.49	0.22	0.07	1
Temp.	0.08	1	0.08	0.78	2.26	1
Depth x Temp.	0.66	1	0.67	0.41	0.03	1
Residuals	521.73	532.00			307.12	320.00
	neither	depth	temp.	both	neither	depth
Coil	210	3	11	3	97	5
Helix	180	3	9	2	119	4
Sheet	111	1	2	0	60	1
Total	501	7	22	5	276	10
	Fisher's $p = 0.75$				Fisher's $p = 0.89$	
Buried	308	2	5	0	170	2
Exposed	161	5	15	4	69	7
Interface	32	0	2	1	37	1
Total	501	7	22	5	276	10
	Fisher's $p = 7.6E-5^*$				Fisher's $p = 6.1E-5^*$	

DH			mMDH				LD	
F	p		SS	df	F	p	SS	df
0.07	0.79		0.25	1	0.26	0.61	7.22	1
2.36	0.13		0.14	1	0.14	0.70	0.04	1
0.03	0.87		0.05	1	0.05	0.82	3.22	1
			328.43	335.00			423.74	447.00
temp.	both		neither	depth	temp.	both	neither	depth
9	5		117	1	0	2	159	11
12	6		153	2	2	2	142	11
4	1		59	0	0	0	78	4
25	12		329	3	2	4	379	26
			Fisher's $p = 0.89$			Fisher's $p = 0.02^*$		
7	3		173	1	1	1	286	13
14	8		109	2	1	3	93	12
4	1		47	0	0	0	0	1
25	12		329	3	2	4	379	26
*			Fisher's $p = 0.59$			Fisher's $p = 5.2E-10$		

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

DH

F	p
7.61	0.01 *
0.05	0.83
3.40	0.07 .

temp.	both
7	8
3	23
0	4
10	35
3	11
6	23
1	1
10	35

0*

For Peer Review