Genome sequences hot and cold: a database of organisms with defined optimal growth temperatures

Authors

Karla Helena-Bueno1, Charlotte R. Brown1, Egor Konyk1, Sergey Melnikov1

Affiliations

1 Biosciences Institute, Newcastle University, Newcastle upon Tyne, UK
2 These authors contributed equally to this work.

To whom correspondence should be addressed: sergey.melnikov@ncl.ac.uk

Abstract

Despite the rapidly increasing number of organisms with sequenced genomes, there is no existing resource that simultaneously contains information about genome sequences and the optimal growth conditions for a given species. In the absence of such a resource, we cannot immediately sort genomic sequences by growth conditions, making it difficult to study how organisms and biological molecules adapt to distinct environments. To address this problem, we have created a database called GSHC (Genome Sequences: Hot, Cold, and everything in between). This database, available at http://melnikovlab.com/gshc, brings together information about the genomic sequences and optimal growth temperatures for 25,324 species, including ~89% of the bacterial species with known genome sequences. Using this database, it is now possible to readily compare genomic sequences from thousands of species and correlate variations in genes and genomes with optimal growth temperatures, at the scale of the entire tree of life. The database interface allows users to retrieve protein sequences sorted by optimal growth temperature for their corresponding species, providing a tool to explore how organisms, genomes, and individual proteins and nucleic acids adapt to certain temperatures. We hope that this database will contribute to medicine and biotechnology by helping to create a better understanding of molecular adaptations to heat and cold, leading to new ways to preserve biological samples, engineer useful enzymes, and develop new biological materials and organisms with the desired tolerance to heat and cold.
Introduction

Despite significant research efforts to understand how biological molecules adapt to temperature change\textsuperscript{1-36}, we are still not able to accurately answer two fundamental questions: What are the most common strategies by which cellular proteins adapt to environmental conditions, such as heat and cold? And can we find a simple and robust approach to alter the thermal tolerance of natural proteins by introducing a minimal number of mutations to the protein sequence?

One challenge in answering these questions stems from the lack of a resource that stores easy-to-use information about the optimal growth conditions of living organisms, together with their genomic data. Currently, there are more than 14,400 genome sequences from representative bacterial species publicly available. In principle, we could use these sequences to study thousands of variants of a given protein, observing how its sequence and structure undergo changes upon transition from cold-adapted bacteria\textsuperscript{5,8,13,14,20,24,27,37-48} to heat-adapted bacteria\textsuperscript{12,18,28-31,33,49-69}. In practice, however, it is not immediately possible to organise thousands of organisms (and their genomic sequences) by their optimal growth temperatures, because the corresponding genomic sequences deposited in public repositories (such as NCBI Genomes) lack information about these organisms’ optimal growth conditions. Hence, although we have at our disposal genome sequences for thousands of distinct bacteria, eukaryotes, and archaea, we lack a simple tool to sort these organisms (and their genomic sequences) by optimal growth conditions, thereby hindering large-scale studies of molecular and organismal adaptations to temperature. Here, we develop such a resource for scientists and engineers interested in exploring and exploiting molecular adaptations to heat and cold. Using the NCBI database of sequenced genomes as a scaffold, we have created a database in which species with known genome sequences are annotated with information about these species’ optimal growth temperature. This database describes the optimal growth temperature of more than 25,000 microorganisms, including 89% of the representative bacteria whose genome sequences are deposited in the NCBI database.

This new resource makes it possible to retrieve up to 12,354 sequences of a given bacterial protein of interest, sort these sequences by the optimal growth temperature of its corresponding species, and explore how each residue in this protein varies in sequence and conservation upon transition from cold-adapted to heat-adapted organisms. Thus, we provide a tool for large-scale studies of the molecular and organismal adaptations to a specific temperature.
Database features

Downloadable lists of species with known genome sequence and together with their optimal growth temperature. Currently, the GSHC database contains information about the optimal growth temperature for 25,324 species. This information is continuously retrieved by web-scraping 23 public repositories of microorganisms (Table 1) and then added to the list of organisms deposited in the NCBI repository of organisms with sequenced genomes. The GSHC site contains the optimal growth temperatures for an organism; however, it does not yet include information about other growth conditions, such as oxygen requirement, pH, pressure, and salt concentration.

Lists of organisms with experimentally defined optimal growth temperatures and a reference to the corresponding genome data can be downloaded from the database. This includes the optimal growth temperature values for 12,265 representative bacteria, 414 representative archaea, and 973 representative eukaryotes. The datasets are updated monthly and are available in .csv format, enabling the species to be sorted by an organism’s name, phylogenetic group, genome size, genomic GC-content, number of protein coding genes, and optimal growth temperature. As shown in the example provided for bacterial species (Fig. 1), organisms contained in the datasets include thermophiles and psychrophiles from all major lineages of species, providing an opportunity to study molecular adaptation to heat and cold at the scale of the entire tree of life.

Optimal growth-temperature checker. In addition to the downloadable data, the database user interface allows searching for the optimal growth temperature for a given species. The user can enter a species name in a search window and retrieve the optimal growth temperature for the species of interest if it is present in the database.

Retrieval of protein sequences sorted by optimal growth temperature. In addition to the temperature checker, the database interface allows users to retrieve sequences of their protein of interest arranged by the optimal growth temperature of its corresponding species. The optimal growth temperature is automatically added to the sequence name which allows sequences to be easily aligned, enabling the exploration of how each residue in a protein of interest changes its identity and conservation upon transition from cold-adapted to heat-adapted organisms.

Error and request tracking. To document the rapidly expanding data, most entries in the database are automated, without inspection against published literature. We acknowledge that the optimal growth temperature values for some organisms may contain discrepancies or inaccuracies and therefore encourage users to provide feedback on any unaddressed discrepancies by submitting reports through our feedback tracking system. The submission form allows users to request changes in the database and monitor the progress of each inquiry.
Applications

How will this database help researchers to explore molecular mechanisms of thermal adaptation? Below, we propose some example applications:

- First, we can now sort genomes or homologous gene sequences by their optimal growth temperature (and not just by phylogenetic origin), making it possible to explore universal strategies of adaptations to heat and cold, as opposed to idiosyncratic adaptations within each lineage of species. Such studies would have far-reaching implications in biotechnology as they can simplify the rational design of biological molecules and organisms with a desired thermal stability 33,61,68,70-75.

- Second, this database simplifies studies of molecular adaptations to temperature within any given range of temperatures. This is important, because to date most studies have focused on extremophiles 49,50, leaving mostly unexplored how mesophilic organisms adapt to relatively subtle changes in the environment (e.g. temperature increases of a few degrees Celsius as a consequence of climate change 13,17,23,76).

- Third, we can monitor how the identity and conservation of each residue in a protein of interest gradually changes across a range of optimal growth temperatures from 2°C to more than 103°C, simplifying studies of structural constraints 23,46,63,75-80, and finding new ways to engineer useful proteins with a desired optimal thermal tolerance 25,70,72,74,75,80,81.

- Fourth, this database can help identify model organisms to observe “extremophiles in the making”. Currently, the database contains organisms from the same genus that have almost identical sequences for most of their cellular proteins but exhibit dramatically different optimal growth temperatures. For example, the genus Clostridium includes species with an optimal growth temperature ranging from just 5°C (Clostridium frigoris) 82, to 55°C (Clostridium thermobutyricum) 83. Comparing species in these genera can provide us with a rare opportunity to observe the natural transformation of mesophiles into extremophiles and gain an understanding of how organisms evolve the ability to tolerate heat and cold through minimal changes in their genomes 10,14,33,55,57,69,84. These studies are important, as they may help to simplify the design of economically useful microorganisms with the desired thermal tolerance.

Future directions

We are currently working to expand the database by including additional environmental parameters, such as optimal salt concentration, pH, pressure, and oxygen requirement. We are also testing a new data scraping algorithm to retrieve data not only from repositories of commercially available microorganisms but also from original research papers, to maximise the completeness of our datasets. Finally, we are testing scripts to allow the mapping of temperature-dependent variations in protein sequences to corresponding three-dimensional protein structures that are available in the Protein Data Bank. We hope that in the future our database or similar annotations will be integrated into centralised repositories of genome sequences, such as NCBI, making it possible to explore organismal adaptations to changing environments using the rapidly expanding collection of genomic sequences for both living and extinct species on our planet.
Figures

**Fig 1 | Bacterial tree of life colour-coded by optimal growth temperatures.** a The bacterial tree of life shows representative members of X bacterial genera, with each organism coloured by the optimal growth temperature. b Lineages of bacterial species in the database and their corresponding optimal growth temperature.
Tables

Table 1 | Websites used for web scraping to collect information about the optimal growth temperatures of microbial organisms.

<table>
<thead>
<tr>
<th>Database</th>
<th>Database Full Name</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARS (NRRL)</td>
<td>Agricultural Research Service Culture Collection, National Center for Agricultural Utilization Research, USA</td>
<td><a href="https://nrrl.ncaur.usda.gov/">https://nrrl.ncaur.usda.gov/</a></td>
</tr>
<tr>
<td>ATCC</td>
<td>American Type Culture Collection, USA</td>
<td><a href="https://www.atcc.org/">https://www.atcc.org/</a></td>
</tr>
<tr>
<td>BCCM</td>
<td>Belgian Coordinated Collections of Microorganisms, Belgium</td>
<td><a href="https://bccm.belspo.be/">https://bccm.belspo.be/</a></td>
</tr>
<tr>
<td>BCRC</td>
<td>Bioresources Collection and Research Center, Taiwan</td>
<td><a href="https://catalog.bcrc.firdi.org.tw">https://catalog.bcrc.firdi.org.tw</a></td>
</tr>
<tr>
<td>CABRI</td>
<td>Common Access to Biological Resources and Information</td>
<td><a href="http://www.cabri.org/">http://www.cabri.org/</a></td>
</tr>
<tr>
<td>CCARM</td>
<td>Culture Collection of Antibiotics Resistant Microbes, UK</td>
<td><a href="https://www.phe-culturecollections.org.uk">https://www.phe-culturecollections.org.uk</a></td>
</tr>
<tr>
<td>CCM</td>
<td>Czech Collection of Microorganisms, the Czech Republic</td>
<td><a href="https://www.sci.muni.cz/ccm/">https://www.sci.muni.cz/ccm/</a></td>
</tr>
<tr>
<td>CCUG</td>
<td>Culture Collection, University of Goteborg, Sweden</td>
<td><a href="https://www.ccug.se/">https://www.ccug.se/</a></td>
</tr>
<tr>
<td>CECT</td>
<td>Spanish Type Culture Collection, Spain</td>
<td><a href="https://www.uy.es/cect/">https://www.uy.es/cect/</a></td>
</tr>
<tr>
<td>CICC</td>
<td>China Center of Industrial Culture Collection</td>
<td><a href="http://english.china-cicc.org">http://english.china-cicc.org</a></td>
</tr>
<tr>
<td>CIP</td>
<td>Center for Biological Resources of the Institute Pasteur, France</td>
<td><a href="https://catalogue-erbp.pasteur.fr">https://catalogue-erbp.pasteur.fr</a></td>
</tr>
<tr>
<td>DSMZ</td>
<td>German Collection of Microorganisms and Cell Cultures GmbH, Germany</td>
<td><a href="https://www.dsmz.de">https://www.dsmz.de</a></td>
</tr>
<tr>
<td>GRIN</td>
<td>Agricultural Genetic Resources Information Center, USA</td>
<td><a href="https://www.ars-grin.gov">https://www.ars-grin.gov</a></td>
</tr>
<tr>
<td>HAMBI</td>
<td>Culture Collection of Department of Applied Chemistry and Microbiology, University of Helsinki, Finland</td>
<td><a href="https://kotka.luomus.fi/culture/bac">https://kotka.luomus.fi/culture/bac</a></td>
</tr>
<tr>
<td>JCM</td>
<td>Japan Collection of Microorganisms (RIKEN Bioresource Center), Japan</td>
<td><a href="https://jcm.brc.riken.jp/en/ordering_e">https://jcm.brc.riken.jp/en/ordering_e</a></td>
</tr>
<tr>
<td>KEMB</td>
<td>Korea Environmental Microorganisms Bank, South Korea</td>
<td><a href="https://kemb.or.kr">https://kemb.or.kr</a></td>
</tr>
<tr>
<td>NBIMCC</td>
<td>National Bank for Industrial Microorganisms and Cell Cultures, Bulgaria</td>
<td><a href="https://www.nbimcc.org/en/about.htm">https://www.nbimcc.org/en/about.htm</a></td>
</tr>
<tr>
<td>NBRC</td>
<td>NITE Biological Resource Center, Japan</td>
<td><a href="https://www.nite.go.jp">https://www.nite.go.jp</a></td>
</tr>
<tr>
<td>NCIMB</td>
<td>National Collections of Industrial, Food and Marine Bacteria, UK</td>
<td><a href="https://www.ncimb.com/">https://www.ncimb.com/</a></td>
</tr>
<tr>
<td>NCMA</td>
<td>National Center for Marine Algae and Microbiota (NCMA), USA</td>
<td><a href="https://ncma.bigelow.org/">https://ncma.bigelow.org/</a></td>
</tr>
<tr>
<td>NCTC</td>
<td>National Collection of Type Cultures, UK</td>
<td><a href="https://www.phe-culturecollections.org.uk/">https://www.phe-culturecollections.org.uk/</a></td>
</tr>
<tr>
<td>NIES</td>
<td>National Institute for Environmental Studies, Japan</td>
<td><a href="https://mcc.nies.go.jp/">https://mcc.nies.go.jp/</a></td>
</tr>
<tr>
<td>VKM</td>
<td>All-Russian Collection of Microorganisms, Russia</td>
<td><a href="http://www.vkm.ru">http://www.vkm.ru</a></td>
</tr>
</tbody>
</table>
Acknowledgements

This work was supported by Newcastle University (NUORS 2021 award to K.H-B.), BBSRC UK (4-year PhD studentship BB/T008695/1. to C.R.B.) and the Royal Society UK (RGS\R2 \ 202003 to S.M.).

References


