Aristotle University of Thessaloniki

Faculty of Sciences

School of Biology

Department of Genetics,
Development and Molecular Biology

PhD dissertation

Comparative genomic analysis of halophilic organisms

Alexios Loukas

Thessaloniki, 2024

Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης

Σχολή Θετικών Επιστημών

Τμήμα Βιολογίας

Τομέας Γενετικής,
Ανάπτυξης και Μοριακής Βιολογίας

Διδακτορική Διατριβή

Συγκριτική γονιδιωματική ανάλυση αλόφιλων οργανισμών

Αλέξιος Λούκας

Θεσσαλονίκη, 2024

I hereby certify that I am the author of this dissertation and that I have cited or referenced, explicitly and specifically, all sources from which I have used data, ideas, suggestions, or words, whether they are precise quotes (in the original or translated) or paraphrased.

Βεβαιώνω ότι είμαι συγγραφέας της παρούσας εργασίας και ότι έχω αναφέρει ή παραπέμψει σε αυτήν, ρητά και συγκεκριμένα, όλες τις πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών, προτάσεων ή λέξεων, είτε αυτές μεταφέρονται επακριβώς (στο πρωτότυπο ή μεταφρασμένες) είτε παραφρασμένες.

The approval of this dissertation by the School of Biology, Faculty of Sciences, Aristotle University of Thessaloniki does not imply acceptance of the author's opinions (according to Law 5343/1932, article 202, paragraph 2).

Η έγκρισης της παρούσης διατριβής υπό του Τμήματος Βιολογίας της Σχολής Θετικών Επιστημών του Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης δεν υποδηλοί αποδοχή των γνωμών του συγγραφέως (Ν.5343/1932, άρθρ. 202, παρ. 2).

Cite this dissertation as follows:

Loukas, A., 2024. Comparative genomic analysis of halophilic organisms. Doctoral dissertation. Department of Genetics, Development and Molecular Biology, School of Biology, Aristotle University of Thessaloniki.

Λούκας, Α., 2024. Συγκριτική γονιδιωματική ανάλυση αλόφιλων οργανισμών. Διδακτορική διατριβή. Τομέας Γενετικής, Ανάπτυξης και Μοριακής Βιολογίας, Τμήμα Βιολογίας, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης.

Advisory Committee

Ilias Kappas, Supervisor
School of Biology, Aristotle University of Thessaloniki

Theodore Abatzopoulos, Member
School of Biology, Aristotle University of Thessaloniki

Christos Ouzounis, Member
Department of Informatics, Aristotle University of Thessaloniki

Examination Committee

Ilias Kappas (Assistant Professor)
School of Biology, Aristotle University of Thessaloniki

Theodore Abatzopoulos (Professor)
School of Biology, Aristotle University of Thessaloniki

Christos Ouzounis (Professor)
Department of Informatics, Aristotle University of Thessaloniki

Minas Yiangou (Professor)
School of Biology, Aristotle University of Thessaloniki

Alexandros Triantafyllidis (Professor)
School of Biology, Aristotle University of Thessaloniki

Spyros Gkelis (Associate Professor)
School of Biology, Aristotle University of Thessaloniki

Elena Drosopoulou (Associate Professor)
School of Biology, Aristotle University of Thessaloniki

Funding

Alexios Loukas was supported by a grant (number: 815) of the General Secretariat for Research and Technology (GSRT) and the Hellenic Foundation for Research and Innovation (HFRI).

Acknowledgements

The present PhD thesis was prepared in the Department of Genetics, Development and Molecular Biology, School of Biology, AUTh under the supervision of Assistant Professor Ilias Kappas, during the period of March 2016 - September 2022.

During the dissertation I gained important knowledge in the field of bioinformatics which will undoubtedly help in my further professional development. Bioinformatics research, as well as any research, requires collaboration and communication with people, who play a very important role in the outcome of each study. For this reason, I would like in the following paragraphs to thank all the people who contributed to the very positive development and completion of my doctoral thesis.

First, I would like to thank my supervisor Dr. Ilias Kappas whom I view as a mentor and during our collaboration conveyed to me his analytical skills and determination. Our meetings have been very productive, and we have always been able to find solutions to arising problems. I am grateful for the friendly and calm working environment he created so that I could perform to the best of my abilities. In this environment, my involvement with bioinformatics became for the first time a pleasant everyday routine and escaped the context of studies.

I also thank Professor Theodore Abatzopoulos who accepted me in his laboratory, for his advice and discussions on the important topics of the thesis, but also the help he offered regarding the procedural and funding aspects of some parts of the thesis.

I owe a big thank you to Professor Christos Ouzounis for his constant support and the communication channel he created so that we could interact with colleagues from all over the world. Also, both during my studies as a graduate student and now as a PhD candidate, our interaction laid a solid foundation that greatly helped to improve and develop my knowledge and skills in the exciting field of bioinformatics.

I also thank the other four members of the seven-member examination committee, for the time they devoted to the evaluation of the thesis but also for the questions and discussions that took place on the topic of the thesis, that helped view our work with different angles and spark a conversation.

I would also like to thank Dr. Athanasios Baxevanis for our casual but meaningful discussions on the thesis and on the effectiveness of some of the methods used. His contribution was particularly important and his practical as well as psychological support was invaluable.

I must also thank Dr. Anastasia Laggis who as a post-doc she was always able to prepare me for what comes next, what the requirements will be, and set an example to follow. I am very happy that we shared the laboratory for the most of this important time.

Additionally, I would like to thank a few colleagues that spent time working with me or sharing the laboratory. Ksenia Gorgia and Maria Alvanou, thank you for supporting me during stressful times, I hope you learned something from me, and I am sure you have a great career ahead of you. Also, thanks to my fellow PhD candidate Ilias Strachinis; I greatly enjoyed our talks about sea life and reptiles in particular.

I am grateful to the Hellenic Foundation for Research & Innovation (HFRI) for the financial support of my thesis. Also, to ELIXIR-GR for funding a short additional project and the Biomedical Sciences Research Center “Alexander Fleming” for facilitating this funding. Results presented in this thesis have been produced using the AUTh High Performance Computing Infrastructure “Aristotelis”.

I would like to thank my colleague and friend Ioannis Pantis for advising me to apply for a MSc in bioinformatics at AUTh; until today I did not regret this decision. Also, I would like to thank him for all the advice, help, and service he provided to me and the Greek scientific community.

I would also like to thank all my friends. We are a very diverse group of people, but I always look up to you, you push me to move forward, and I am always gaining knowledge when interacting with any of you. Also, a big thank you to my girlfriend Katerina Gkikopoulou who supported me in difficult times during my thesis and during a whole pandemic.

Finally, none of what I have achieved today would be possible without the emotional and financial support of my family. So, a big thank you to my mother and father for helping me get here, I am forever in your debt.

Στην οικογένειά μου

Publications in peer reviewed journals

Published

1. Loukas A., Kappas I. & Abatzopoulos T.J. 2018. HaloDom: a new database of halophiles across all life domains. Journal of Biological Research-Thessaloniki 25, 2. https://doi.org/10.1186/s40709-017-0072-0.

In preparation

- Loukas A., Baxevanis A.D., Abatzopoulos T.J. & Kappas I. Amino acid profile clustering of halophiles.

- Loukas A., Abatzopoulos T.J. & Kappas I. Pangenome analysis and phylogenetic distribution of protein families in Halobacteria.

CONTENTS

Abstract

Halophilic organisms are extremophiles that grow optimally in environments with high salt concentrations. They have representatives across all life domains showing considerable diversity in metabolic strategies and physiological responses, especially among microbes. Their study goes back decades and a considerable number of articles every year report new halophilic species. Research on halophiles has mainly focused on the specific adaptations and molecular mechanisms that enable them to maintain their osmotic balance under salt stress. A great deal of interest has also been channeled towards the investigation of their diversity and phylogenetic relationships as several representatives constitute ancient evolutionary lineages. On a different avenue, biotechnology has recently delved into the survival kits of extremophiles in the hunt of biocatalysts functioning in hostile environments. Hence, the interest on halophiles spans a wide range of topics, from their exotic biology to industrial and medical applications.

Despite the intensity of research in the field and the subsequent gradual increase in the reporting of new halophilic taxa every year, an extensive organized collection of these salt-loving organisms does not exist. In addition, the large volumes of data produced by next-generation sequencing have significantly enriched the molecular information of these lineages. In the first part of this work, we have aimed at cataloging all documented halophilic species in an online database repository, named "HaloDom". More than 1250 halophilic species mentioned in scientific papers were recorded. From these species, currently 22.3% belong to Archaea, 52.9% to Bacteria and 24.8% to Eukaryotes. Records contain basic information such as the salinity that a particular organism was isolated from, its taxonomy and genomic information availability.

Archaea possess a complex evolutionary history. Determining phylogeny with the use of single-marker phylogenies such as 16S rRNA has been proven sub-optimal, especially among Archaea, which exhibit great genetic diversity and variability even on typically conserved genetic sequences. Phylogenetic relationships between the class Halobacteria (Archaea) and other halophiles, such as the recently discovered Nanohaloarchaea, and halophilic Bacteria have not been fully established yet. A phylogenetic tree of 124 Halobacteria and 33 outgroup species including a diverse set of halophiles was assembled with the use of 242 previously published core protein archaeal markers. This multi-marker tree was used to gain more insight into the phylogenetic relationships of halophilic species, serving as an update of the microbial halophilic tree of life and providing evidence that salt adaptation has originated independently.

It has been previously shown that thermophilic and halophilic microbes bear specific signatures of amino acid compositions. In this dissertation, we took advantage of the large amount of microbial sequence data to confirm prior research, and also investigate for new distinguishing features that can potentially be uncovered by a large-scale protein sequence analysis. Hierarchical clustering and principal components analysis revealed a detailed picture of the amino acid profiles of several microbial taxonomic groups. In total 222 microbial species with available genome sequences (complete or partial) on NCBI were included. Among these proteomes, 134 were halophilic Archaea and 88 several other taxa including thermophilic, non-halophilic, acidophilic, alkaliphilic, and psychrophilic Archaea and Bacteria. Regarding the amino acid composition of halophiles, proteomic data revealed important differences between Halanaerobiales, Halobacteria, Methanomicrobia and Nanohaloarchaea. Also, several taxa from Bacteroidetes, Rhodothermaeota and Proteobacteria were observed to have similar amino acid profiles with the extremely halophilic class of Halobacteria, suggesting either horizontal gene transfer or convergent evolution. Several aspects of proteins and amino acids were investigated including protein function and family, different protein sizes and compositionally biased proteins both from halophilic and non-halophilic species. Our data dictate that different protein groups are subject to different levels of evolutionary pressure, depending on their thermodynamic efficiency, localization, and, yet unknown factors.

As mentioned, these protein sequence adaptations are detectable and can be used to distinguish adapted halophiles from other species. With the use of linear discriminant analysis of halophilic and non-halophilic amino acid profiles, a web tool was designed. This tool, named "HaloPredictor" can detect specific salinity adaptations based on protein sequences from the archaeal class of Halobacteria. The observed adaptations are registered as an altered amino acid profile in halophilic proteins, with certain amino acids like alanine (A) and aspartic acid (D) present in higher percentages, compared with non-halophiles. HaloPredictor is available as an online tool in HaloDom, but also as a local version for increased input capacity.

A pan-genome analysis of the archaeal class Halobacteria was conducted using complete genome data from 76 species, present in NCBI. Results indicate an open pan-genome and the existence of a plethora of novel genes to be discovered as new species of Halobacteria are isolated and sequenced. We identified 814 genes as core, 13752 as accessory, and 9010 as cloud for a total of 23576 identified gene clusters. From the 217783 protein sequences predicted from the 76 Halobacteria genomes, 182884 were annotated successfully and assigned to a functional category while for 34899 sequences the annotation was unsuccessful. Genes annotated as "Function unknown" were present in all gene groups (core, accessory, cloud, and multi-copy genes). Additionally, multi-copy genes were highly present in functional groups such as "Amino acid transport and metabolism", "Post-translational modification”, “Protein turnover and chaperones", and "Transcription". With the use of presence-absence matrices, phylogenetic profiles were created for the halophile pan-genome, revealing novel genetic traits.

The present work builds on past sporadic findings and extends significantly current knowledge on halophiles using large-scale data previously unavailable. Novel observations are presented about the amino acid profiles of extremophiles and their pan-genome. It also highlights the necessity for discovering and validating new protein families and the need for further research on the *mechanics* of salinity adaptations and cellular functions of halophiles. It thus provides a wide range of starting points for future analyses with a strong potential for advances in metagenomics, medicine, and the bio-industry.

Ελληνική περίληψη

Η ηλικία του πλανήτη μας είναι περίπου 4.5 δισεκατομμυρίων ετών. Ωστόσο οι αρχικές περιβαλλοντικές συνθήκες δεν ευνόησαν την ανάπτυξη ζωής για εκατομμύρια χρόνια. Τα πιο πρόσφατα στοιχεία δείχνουν ότι οι πρώτες μορφές ζωής εμφανίστηκαν 3.7 δισεκατομμύρια χρόνια πριν, πιθανώς και νωρίτερα. Αν και δεν είναι ακριβώς γνωστές οι επικρατούσες συνθήκες όπως οι θερμοκρασίες ή τα διάφορα αέρια στην ατμόσφαιρα του πλανήτη, μπορεί με ασφάλεια να θεωρηθεί ότι το περιβάλλον εκείνο ήταν εξαιρετικά ακραίο σύμφωνα με τις σημερινές προδιαγραφές της ζωής.

Παρά τις δύσκολες συνθήκες, η ζωή εμφανίστηκε κατά την πρώιμη αυτή εποχή, ενώ αργότερα μια από τις πρώτες οργανωμένες και πολύπλοκες μορφές ζωής, οι μικροοργανισμοί, ήταν ήδη σε θέση να επηρεάσουν το περιβάλλον του πλανήτη. Τα μικρόβια θεωρούνται από τις αρχαιότερες μορφές ζωής εποικίζοντας πρακτικά κάθε περιβάλλον. Πιστεύεται ότι οι ακραιόφιλοι (extremophiles) μικροοργανισμοί εμφανίστηκαν κατά την διάρκεια της πρώιμης περιόδου του πλανήτη. Η πλειοψηφία των στοιχείων για εκείνη την περίοδο προέρχεται από τα ακραία περιβάλλοντα που συναντάμε σήμερα, όπως είναι οι έρημοι, τα ηφαίστεια, οι παγετώνες, οι υδροθερμικές διέξοδοι (χερσαίες ή υδάτινες), οι αλμυρές λίμνες και οι υδατοσυλλογές, μερικά από τα οποία εκτιμάται ότι αποτελούσαν χαρακτηριστικά στοιχεία του πλανήτη στην πρώιμη περίοδο και για εκατομμύρια χρόνια.

Κάποιες ομάδες μικροοργανισμών κατάφεραν να προσαρμοστούν στα περιβάλλοντα εκείνα και διάφορα είδη ακραιόφιλων οργανισμών εδραιώθηκαν στον πλανήτη. Για παράδειγμα, τα θερμόφιλα (thermophiles) είδη, που συναντώνται στα Βακτήρια, τα Αρχαία και τους Μύκητες, και τα οποία είναι ικανά να επιβιώνουν και να αναπτύσσονται σε θερμοκρασίες μέχρι και 122°C. Στο άλλο άκρο συναντούμε τα ψυχρόφιλα (psychrophiles) είδη, κυρίως Βακτήρια και Αρχαία, που μπορούν να επιβιώσουν σε εξαιρετικά χαμηλές θερμοκρασίες καθώς έχουν απομονωθεί από περιβάλλοντα όπως οι πάγοι των πόλων και οι παγετώνες. Οι ψυχρόφιλες ιδιότητες έχουν παρατηρηθεί και σε φωτοσυνθετικούς ευκαρυωτικούς οργανισμούς όπως είναι τα Φύκη και τα Διάτομα.

Άλλες ομάδες ακραιόφιλων είναι τα οξεόφιλα (acidophiles) και τα αλκαλόφιλα (alkaliphiles), είδη τα οποία αναπτύσσονται σε περιβάλλοντα με ακραίες τιμές pH, είτε σε όξινες είτε σε βασικές συνθήκες. Επίσης τα πιεζόφιλα (piezophiles) είδη μπορούν να επιβιώσουν σε

ακραίες υδροστατικές πιέσεις που μπορεί να φτάσουν μέχρι και τα 18.855 psi. Συναντώνται και σε προκαρυωτικούς αλλά και ευκαρυωτικούς οργανισμούς.

Η παρούσα διατριβή επικεντρώνεται σε μια άλλη μεγάλη ομάδα οργανισμών, τα αλόφιλα (halophiles), τα οποία επιβιώνουν σε περιβάλλοντα με υψηλή αλατότητα, αλλά ακόμη και σε διαλύματα κορεσμένα με άλατα. Τέτοια περιβάλλοντα είναι οι αλμυρές και υπεράλμυρες λίμνες, αλυκές και πεδία άλατος, αλμυρά ποτάμια, υπεράλμυρες λίμνες στον πυθμένα της θάλασσας αλλά και διάφορα αλμυρά προϊόντα διατροφής. Υπάρχει ένα ευρύ φάσμα από αναφερόμενες αλατότητες επιβίωσης οι οποίες ξεκινούν από 5% (w/v) και φτάνουν μέχρι και 36%. Οι αλόφιλοι οργανισμοί παρατηρούνται και στις τρεις επικράτειες της ζωής, Αρχαία, Βακτήρια και Ευκαρυώτες. Ωστόσο δεν επιβιώνουν όλα τα είδη σε εξαιρετικά υψηλά επίπεδα αλατότητας καθώς μέχρι στιγμής έχουν παρατηρηθεί δύο διαφορετικοί μηχανισμοί προσαρμογής. Η πρώτη στρατηγική επιτρέπει στα άλατα να εισέρχονται στο κυτταρόπλασμα (salt-in strategy) και χρησιμοποιείται κυρίως από αντιπροσώπους της κλάσης Halobacteria (Αρχαία). Σε αυτή την στρατηγική οι πρωτεΐνες έχουν προσαρμοστεί ώστε να μπορούν να λειτουργούν σε υψηλές συγκεντρώσεις αλάτων. Η δεύτερη στρατηγική (salt-out) χρησιμοποιείται κυρίως από ομάδες βακτηρίων όπως τα Proteobacteria, Firmicutes και Cyanobacteria καθώς επίσης και σε διάφορους μύκητες (*Saccharomyces cerevisiae*). Πρέπει να τονιστεί ότι, πέρα από τις δύο παραπάνω βασικές στρατηγικές, η ποικιλία των προσαρμογών στην αλατότητα είναι πολύ μεγάλη και συνεχώς ανακαλύπτονται νέες.

Στα αλόφιλα παρατηρείται επίσης τεράστια μεταβολική ποικιλότητα όπως οξυγονικά, ανοξυγόνα φωτότροφα, αερόβια ετερότροφα, ζυμωτές, απονιτροποιητές, μειωτές θειικών και μεθανογόνα. Επιπλέον, έχουν απομονωθεί αλόφιλοι οξειδωτές θείου, μειωτές σιδήρου και ακετογόνα είδη. Επίσης, πολλά αλόφιλα βακτήρια έχουν βρεθεί ότι εκτελούν ζύμωση, ακετογένεση, αναγωγή θειικών και μεθανογένεση.

Οι αλόφιλοι και αρκετοί άλλοι ακραιόφιλοι οργανισμοί χρησιμοποιούνται σε πληθώρα ιατρικών, βιομηχανικών και εργαστηριακών εφαρμογών. Το χαρακτηριστικότερο παράδειγμα αποτελεί η θέρμο-σταθερή πολυμεράση "*Taq*" από το θερμόφιλο βακτήριο *Thermus aquaticus* η οποία χρησιμοποιείται ευρέως στο πρωτόκολλο της αλυσιδωτής αντίδρασης πολυμεράσης (PCR). Ομοίως, άλλα ένζυμα από ακραιόφιλους οργανισμούς έχουν βρει εφαρμογή σε διάφορους ερευνητικούς τομείς. Η εκμετάλλευση των ενζύμων ήταν από τις πρώτες ερευνητικές κατευθύνσεις που ακολούθησαν την ανακάλυψη των οργανισμών αυτών. Επιπλέον, με την πρόοδο της τεχνολογίας και της επεξεργασίας μακρομορίων, αυτά τα ένζυμα μπορούν να τροποποιηθούν ώστε να ταιριάζουν καλύτερα σε κάθε εφαρμογή.

Η παραγωγή βιοκαυσίμων είναι άλλη μια εφαρμογή, όπου διάφορα μεθανογόνα και θερμόφιλα στελέχη βακτηρίων χρησιμοποιούνται σε μεγάλες ποσότητες για την παραγωγή βουτανόλης, μεθανίου και βιο-ντίζελ (biodiesel). Οξεόφιλοι οργανισμοί χρησιμοποιούνται επίσης στην εξόρυξη μετάλλων για την απομάκρυνση αδιάλυτων σουλφιδίων και οξειδίων.

Η παραγωγή καροτενοειδών είναι δυνατή με τη χρήση ακραιόφιλων. Παρόλο που τα περισσότερα καροτενοειδή δεν μπορούν να παραχθούν σε υψηλούς ρυθμούς, κάτι τέτοιο είναι δυνατό με την βακτηριοδοψίνη, την κανθαξανθίνη και το β-καροτένιο. Η βακτηριοροδοψίνη συλλέγεται από το αλόφιλο αρχαιοβακτήριο *Halobacterium salinarum* και χρησιμοποιείται σε ένα ευρύ φάσμα εφαρμογών (ολογραφία, τεχνητοί αμφιβληστροειδείς, βαφές, ανανέωση βιοχημικής ενέργειας). Η κανθαξανθίνη από το είδος *Haloferax alexandrinus* χρησιμοποιείται ως βαφή και ως πρόσθετο τροφίμων. Το β-καροτένιο χρησιμοποιείται επίσης ως προσθετικό φαγητού και ως συμπλήρωμα διατροφής. Παράγεται από την αλόφιλη άλγη *Dunaliella salina*. Την τελευταία δεκαετία η έρευνα που αφορά τη χρήση πρωτεασών/λιπασών και γλυκοζυλ-υδρολασών από ακραιόφιλους οργανισμούς έχει ενταθεί. Αυτές οι ουσίες χρησιμοποιούνται ευρέως σε απορρυπαντικά ρούχων, στην παρασκευή τυριών και στη ζυθοποιία. Επιπλέον, μια σημαντική πιθανή εφαρμογή είναι η χρήση αλόφιλων στην παραγωγή βιο-διασπώμενων πλαστικών με στόχο την ολική αντικατάσταση των συμβατικών πλαστικών που βασίζονται στο πετρέλαιο.

Από την πλευρά των ιατρικών εφαρμογών, οι ακραιόφιλοι οργανισμοί χρησιμοποιούνται στην παραγωγή αντιβιοτικών, αντιμυκητιασικών και αντικαρκινικών μορίων. Όντας στην πλειοψηφία τους μικροοργανισμοί, αυτά τα είδη μπορούν και παράγουν ουσίες που αναστέλλουν βιολογικά συστήματα αλλά και την ανάπτυξη άλλων μικροβίων, γεγονός το οποίο μπορούν να εκμεταλλευτούν οι ερευνητές για να επιτύχουν ένα συγκεκριμένο στόχο ή θεραπεία. Μια από τις πιθανές εφαρμογές είναι η χρήση κυστιδίων αερίου (gas vesicles), τα οποία έχουν τροποποιηθεί στο αλόφιλο είδος *Halobacterium* NRC-1 για την χρήση τους ως εναλλακτικού συστήματος χορήγησης εμβολίων. Αυτού του είδους τα κυστίδια έχει αποδειχτεί ότι προκαλούν επαρκή ανοσοαπόκριση στα ποντίκια και τα αποτελέσματα είναι ενθαρρυντικά. Οι αντικαρκινικές ιδιότητες μεταβολιτών από το *Halobacterium salinarum* έχουν παρατηρηθεί ήδη *in vitro* και *in vivo* σε ποντίκια. Οι ιδιότητες αυτές των μεταβολιτών παρατηρήθηκαν πρόσφατα και εκτιμάται ότι περισσότερες τέτοιες ουσίες θα εντοπιστούν στο μέλλον από τους ερευνητές, καθώς ολοένα και περισσότερα ακραιόφιλα είδη ανακαλύπτονται και επιστρατεύονται στην έρευνα κατά του καρκίνου.

Τέλος, οι ακραιόφιλοι οργανισμοί έχουν πολύ μεγάλο ενδιαφέρον για το πεδίο της αστροβιολογίας όπου η μελέτη των συναρπαστικών προσαρμογών τους μπορεί να καθοδηγήσει σε μεγάλο βαθμό πρωτόκολλα διερεύνησης της πιθανότητας ζωής σε άλλους πλανήτες.

Στην παρούσα διατριβή επιχειρήθηκε μια συνολική συγκριτική γονιδιωματική ανάλυση των αλόφιλων οργανισμών, των φυλογενετικών τους σχέσεων και των προσαρμογών τους. Οι επιμέρους στόχοι ήταν οι εξής: **1)** Η καταγραφή των αποσπασματικών μέχρι σήμερα πληροφοριών των αλόφιλων ειδών από όλες τις επικράτειες του δέντρου της ζωής (Βακτήρια, Αρχαία, Ευκαρυώτες) και η οργάνωσή τους σε μια βάση δεδομένων με πληροφορίες εύρους αλατότητας, ταξινομικής θέσης και δεδομένων ακολουθιών ή γονιδιωμάτων, **2)** Η φυλογενωμική ανάλυση αλόφιλων αρχαιοβακτηρίων της κλάσης Halobacteria και η διαλεύκανση των γενεαλογικών τους σχέσεων, **3)** Η συγκριτική ανάλυση του προφίλ των αμινοξέων στα πρωτεώματα αλόφιλων και μη-αλόφιλων ειδών και η ανίχνευση μοριακών υπογραφών αλόφιλων σε ευρεία κλίμακα, **4)** Ο σχεδιασμός ενός λογισμικού εργαλείου για τον αυτόματο εντοπισμό μοριακών υπογραφών προσαρμογών στην αλατότητα σε πρωτεϊνικές ακολουθίες, και **5)** Η παν-γονιδιωματική ανάλυση της κλάσης Halobacteria, η διερεύνηση της γονιδιακής δεξαμενής και των χαρακτηριστικών της. Οι παραπάνω στόχοι οργανώνονται και περιγράφονται σε επτά κεφάλαια:

- **Κεφάλαιο 1:** Εισαγωγή στους ακραιόφιλους και αλόφιλους οργανισμούς.
- **Κεφάλαιο 2:** HaloDom: Μια βάση δεδομένων αλόφιλων οργανισμών από κάθε επικράτεια.
- **Κεφάλαιο 3:** Φυλογενωμική ανάλυση αλόφιλων αρχαιοβακτηρίων.
- **Κεφάλαιο 4:** Ανάλυση του προφίλ των αμινοξέων αλόφιλων ειδών.
- **Κεφάλαιο 5:** HaloPredictor: Ένα εργαλείο για τον εντοπισμό προσαρμογών στην αλατότητα.
- **Κεφάλαιο 6:** Παν-γονιδιωματική ανάλυση της κλάσης Halobacteria.
- **Κεφάλαιο 7:** Συμπεράσματα και προοπτικές.

Στο **Κεφάλαιο 1** γίνεται μια σύντομη εισαγωγή στους ακραιόφιλους οργανισμούς και στα διαφορετικά περιβάλλοντα στα οποία καταφέρνουν να επιβιώσουν με επιτυχία. Κατόπιν, δίνεται έμφαση σε λεπτομέρειες και ιδιαιτερότητες των αλόφιλων οργανισμών που είναι και το κύριο αντικείμενο της διατριβής. Αναφέρονται οι ιατρικές και βιομηχανικές εφαρμογές πολλών αλόφιλων και ακραιόφιλων οργανισμών και γίνεται μια εκτίμηση για τις μελλοντικές προοπτικές αξιοποίησής τους.

Στο **Κεφάλαιο 2** περιγράφεται η δημιουργία μιας βάσης δεδομένων αλόφιλων οργανισμών, με τίτλο "HaloDom". Πάνω από 1250 είδη αλόφιλων οργανισμών καταγράφηκαν από τη βιβλιογραφία. Οι εγγραφές της βάσης περιλαμβάνουν βασικές πληροφορίες όπως το επίπεδο αλατότητας στο οποίο βρέθηκε ο οργανισμός, η ταξινομική θέση του αλλά και δεδομένα ακολουθιών όπως η ύπαρξη ή όχι διαθέσιμου πλήρους γονιδιώματος, πρωτεϊνών κλπ. Περιγράφονται αναλυτικά οι διαδικασίες και τα πρωτόκολλα που χρησιμοποιήθηκαν για τον σχεδιασμό της βάσης δεδομένων και την ολοκλήρωσή της στην ιστοσελίδα http://www.halodom.bio.auth.gr.

Στο **Κεφάλαιο 3** πραγματοποιείται φυλογενωμική ανάλυση 124 μελών της κλάσης Halobacteria και 33 εξωομάδων αρχαιοβακτηρίων και βακτηρίων με τη χρήση 242 συντηρημένων πρωτεϊνών. Το φυλογενετικό δέντρο μέγιστης πιθανοφάνειας (maximum likelihood) που προέκυψε με τη χρήση του προγράμματος IQ-TREE διαφωτίζει και επικαιροποιεί προβληματικές μέχρι σήμερα γενεαλογικές σχέσεις διαφορετικών ομάδων της κλάσης Halobacteria.

Στο **Κεφάλαιο 4** γίνεται συγκριτική ανάλυση του προφίλ των αμινοξέων μεταξύ αλόφιλων και μη-αλόφιλων ειδών. Χρησιμοποιώντας μεθόδους ιεραρχικής ομαδοποίησης (hierarchical clustering) και ανάλυσης κυρίων συνιστωσών (principal components analysis) και αξιοποιώντας την πληθώρα των διαθέσιμων πρωτεϊνικών ακολουθιών επιτυγχάνεται η διακριτή ομαδοποίηση αλόφιλων και μη-αλόφιλων αντιπροσώπων με ξεχωριστή σύνθεση διαφορετικών ομάδων αμινοξέων (π.χ. όξινα, βασικά κλπ.). Οι διαφορές αυτές είναι εμφανείς και στις διαμεμβρανικές πρωτεΐνες των Halobacteria, εξαιρώντας όμως τα διαμεμβρανικά τους τμήματα τα οποία διατηρούν μη-αλόφιλη σύνθεση αμινοξέων.

Ως επέκταση του Κεφαλαίου 4, στο **Κεφάλαιο 5** τα προφίλ των αμινοξέων των αλόφιλων οργανισμών χρησιμοποιούνται για την ανάπτυξη μιας εφαρμογής η οποία ανιχνεύει προσαρμογές αλατότητας (μοριακές υπογραφές) σε πρωτεϊνικές ακολουθίες. Η εφαρμογή χρησιμοποιεί γραμμική διαφοροποιούσα ανάλυση (linear discriminant analysis) για τη σύγκριση των ποσοστιαίων αναλογιών των αμινοξέων σε πρωτεΐνες. Η εφαρμογή ονομάστηκε "HaloPredictor" και είναι ενσωματωμένη στην ιστοσελίδα της βάσης δεδομένων HaloDom. Αναπτύχθηκε επίσης αυτόνομη έκδοση για χρήση σε προσωπικό υπολογιστή, η οποία δέχεται πολλαπλές ακολουθίες πρωτεϊνών, αυξάνοντας έτσι την χρηστικότητά της.

Στο **Κεφάλαιο 6** πραγματοποιείται παν-γονιδιωματική ανάλυση της κλάσης Halobacteria με τη χρήση 76 γονιδιωμάτων. Τα αποτελέσματα έδειξαν ένα ανοιχτό παν-γονιδίωμα, με μικρό αριθμό συστατικών γονιδίων (core genes) και αρκετά μεγάλο αριθμό σπάνιων γονιδίων που εντοπίζονται σε μεμονωμένα είδη, γένη ή οικογένειες. Για τις 23.576 προβλεφθείσες γονιδιακές συστάδες (gene clusters) και τις 217.783 πρωτεΐνες που τις αποτελούν, έγινε σχολιασμός με το πρόγραμμα EggNOG. Ένα από τα κύρια ευρήματα που παρατηρήθηκαν μετά τον σχολιασμό είναι η ύπαρξη μεγάλου ποσοστού πρωτεϊνών με άγνωστη λειτουργία (“function unknown”). Κάτι τέτοιο δείχνει ότι στην κλάση Halobacteria υπάρχουν πιθανότατα νέες πρωτεϊνικές οικογένειες και βιολογικά συστήματα με άγνωστες λειτουργίες. Επίσης, ενδιαφέρον παρουσιάζει ο εντοπισμός γονιδίων τα οποία απαντώνται με παραπάνω από τρία αντίγραφα σε κάθε γονιδίωμα.

Η παρούσα διδακτορική διατριβή ολοκληρώνεται στο **Κεφάλαιο 7** με την παρουσίαση των συμπερασμάτων και μια προσπάθεια καταγραφής μελλοντικών προοπτικών σε αυτό το πολύ ενδιαφέρον και συναρπαστικό πεδίο έρευνας των αλόφιλων οργανισμών.

Chapter 1: Introduction to extremophiles and halophilic organisms

1.1 Extremophiles

Planet Earth is approximately 4.5 billion years old. However, the initial environmental conditions did not favor the emergence of life for millions of years. The latest evidence suggests that the first forms of life came to existence at 3.7 billion years ago and possibly earlier [1]. Although it is not yet clear about the exact temperatures, gases, and conditions that were dominant in Earth's atmosphere [2, 3], the conditions of the planet before the great oxygenation event [4, 5] can be considered extreme by today's life standards.

Despite the harsh conditions, life emerged during these early years and one of the first organized and complex life forms, microorganisms, were already influencing the planet's environment [6, 7]. Microbes are among the most ancient survivors on the planet, omnipresent, and adapted to a variety of extreme conditions. It is believed that microbial extremophiles emerged during that early period of the Earth. Most of the evidence of early life on Earth comes from extreme environments such as hydrothermal vents, arid expansions, and salt lakes which exist today and were more commonly present in the planet million years ago [8].

Several groups of microorganisms adapted to these early environments and many different types of extremophiles emerged [9]. **Thermophiles** are microbial species able to grow in temperatures of 45°C and survive at temperatures as high as 80-100°C. Most of the thermophiles discovered belong to Archaea, with fewer representatives from Bacteria. **Psychrophiles** on the other hand are able to survive in extremely low temperatures and have been isolated from places like the Antarctic Sea ice. Photosynthetic eukarya like algae and diatoms but also bacteria have been found bearing adaptations to low temperatures, although as more species are discovered, psychrophilic Archaea could also be found. Another group of extremophiles are **acidophiles** and **alkaliphiles**, which thrive in habitats with extremely low or high pH values. **Piezophiles** are organisms able to survive hydrostatic pressure equal or above certain psi levels**.** There are reports of piezophiles surviving and growing at 18,855 psi [10]. Piezophilic species can be found both in prokaryotic and eukaryotic domains.

The focus of this study however is on **halophilic** organisms. Halophiles live in saline environments sometimes saturated with salts, such as salt lakes, hypersaline lakes, salt flats, brine underwater pools, salt rivers, and even salty foods. There is a wide range of salinities reported starting from 5% (w/v) going up to 36% (seawater salinity is about 3.5%). Halophiles expand in all three domains of life, Archaea, Bacteria, and Eukarya [11], however, not all species are conditioned to survive in NaCl saturation as there are two different strategies for dealing with high salinity. The first strategy referred to as “salt-in” is used mainly by the archaeal class of Halobacteria. Species that use this strategy allow salt to enter the cytoplasm and contain salt-adapted proteins that can deal and function under salt-stress. In the second strategy called “salt-out”, salts are prevented from entering the cells, and instead compatible solutes are produced in the cytoplasm like glycine betaine and ectoine in order to combat the osmotic pressure [12-18]. The salt-out strategy is mostly used by several bacteria like Proteobacteria, Firmicutes, Cyanobacteria, and also fungi like *Saccharomyces cerevisiae*. However, several species of bacteria from the phylum Rhodothermaeota like *Salinivenus iranica* and *Salinivenus lutea*, but also *Salinibacter ruber* from Bacteroidetes, as it is supported by this study, seem to have salt-adapted proteins, probably transferred from Halobacteria [19-21]. Therefore, the diversity and dynamics of salt-adaptation appear to be high.

Halophilic metabolism is also very diverse, and halophiles are reported to include oxygenic and anoxygenic phototrophs, aerobic heterotrophs, fermenters, denitrifiers, sulfate reducers, and methanogens. Additionally, halophilic sulfide oxidizers, iron-reducers, and acetogens have also been isolated [22, 23]. Also, several halophilic bacteria have been found to perform fermentation, acetogenesis, sulfate reduction, phototrophy, and methanogenesis [24, 25].

1.2 Industrial and medical applications of extremophiles

Halophiles and many other extremophiles have a plethora of industrial, medical, and other applications. First, the several adaptations inside the cells of extremophiles produce capable enzymes for many different purposes. For example, the thermostable polymerase “Taq” from *Thermus aquaticus*, a thermophilic bacterium, is widely used in the Polymerase Chain Reaction (PCR) protocol. Other enzymes from extremophiles can be harvested and used in many research areas. Mining for, so-called, extremozymes is the first biotechnological application since these organisms were discovered. With recent advances in industrial enzyme technology, these enzymes can be modified to best suit each application [9, 26]. Biofuel production is another application where several methanogenic and thermophilic strains of bacteria are used in large quantities, where they are able to take in substances like cellulose, sugar, and waste products to produce butanol, methane, and biodiesel [27]. Additionally, some strains are engineered to handle higher concentrations of substances. The algal *Cyanidium caldarium* is also reported as a promising target for biofuel production [27]. Another application is biomining, where acidophiles are used to remove insoluble sulfides and oxides from various metals during mining [27]. Carotenoid production is possible with extremophiles, even though most carotenoids cannot be synthesized at high rates. In some species it is possible with the use of bacteriorhodopsin, canthaxanthin, and β-carotene [27, 28]. Bacteriorhodopsin can be acquired from *Halobacterium salinarum* and is broadly used from holography and artificial retinas to dyes and renewal of biochemical energy [29]. Canthaxanthin is used as a food dye and additive and is produced by halophilic Archaea and specifically *Haloferax alexandrinus [30]*. β-carotene is used as an additive in baking and as a food supplement. It is produced by the halophilic algal *Dunaliella salina.* In recent years, research on the use of extremophilic proteases/lipases and glycosyl hydrolases has intensified. These substances are used widely in laundry detergents, cheese making, brewing, and baking and are typically collected from mesophilic species [27]. Additionally, an important application is the use of halophilic extremophiles in the production of bio degradable plastics, in order to replace conventional oil-based plastics which will have a huge positive impact on the environment [31].

Extremophiles, from a medical applications perspective, are also used in the production of antibiotics, antifungals, and antitumor molecules. Extremophilic species are able to produce substances to inhibit other microbial species and systems, which in turn can be exploited by researchers for a specific purpose [27]. Another interesting application is the use of gas vesicles engineered in *Halobacterium* NRC-1 as an alternative vaccine delivery system. These specific type gas vesicles have been proven to elicit an adequate immune response in mice and show promising results [32]. The antitumor capabilities of supernatant metabolites from *Halobacterium salinarum* have also been demonstrated in-vitro and in-vivo in mice [33]. The antitumor effects of metabolites from Halobacteria are a recent discovery and more effective substances are expected to be discovered in the future, as more extremophilic species are utilized in cancer research.

Lastly, extremophiles are used as model organisms in the field of astrobiology. Relevant research topics range from evolution of life and mass extinctions to the future of life on earth and its expansion in outer space. Astrobiology could take advantage of these organisms to simulate environments from other planets and explore their habitability potential [34-36].

1.3 Research aims and objectives

From the above is it quite evident that the observation for the presence of life at the fringes of physical and chemical space unarguably raises the issue of the boundaries of the biological envelope. The investigation of the mechanisms and drivers of adaptive evolution in extreme environments has seen a tremendous interest in the last couple of decades. Part of this attention is related to fundamental questions as extremophiles seem to reframe the window of viability. Another part concerns potential applications of these organisms that could prove invaluable in biotechnology and medicine.

On this basis and with an emphasis on halophiles, the largest group of extremophiles, the aims of the present PhD dissertation focused on four major areas:

1) To perform an exhaustive literature search on halophilic species from all three domains of life, catalogue those with all relevant metadata, and design a database of halophiles.
2) Taking advantage of the plethora of available genomic sequences of halophilic taxa and non-halophilic relatives, to detect potential molecular signatures of salinity adaptation in the genomes and proteomes of halophiles.
3) To perform an extensive investigation of the phylogenetic relationships of halophiles with a focus on Halobacteria and clarify the topological placement of several unstable groups.
4) To investigate the evolutionary characteristics of the pangenome of halophiles and the diversity of protein families.

Chapter 2: HaloDom, a new database of halophiles across all life domains

2.1 Introduction

Halophilic organisms may thrive in or tolerate high salt concentrations. They have been studied for decades and a considerable number of papers reporting new halophilic species are being published every year. Halophiles are categorized as slight, moderate, and extreme, depending on their maximum salinity tolerance [24]. Halophilic species exist across all life domains [37, 38] showing considerable diversity in metabolic strategies and physiological responses, especially among microbes [24, 39-41]. Research on halophiles has mainly focused on the specific adaptations and molecular mechanisms that enable them to maintain their osmotic balance under salt-stress. A great deal of interest has also been channeled towards the investigation of their diversity and phylogenetic relationships as the majority of them constitute ancient evolutionary lineages [42, 43]. On a different avenue, biotechnology has recently decided to delve into the survival kits of extremophiles in the hunt for biocatalysts functioning in hostile environments. All this interest is reflected in the plethora of papers reporting new halophilic species every year [44-46], a trend which is expected to increase. Consequently, and due to the large quantities of data produced by next-generation sequencing, there is a need for a database repository of extremophiles which will be regularly updated.

So far, there are three halophilic databases available online: HaloWeb [47], HaloBase [48] and HProtDB [49]. HaloWeb focuses on genome information and provides complete genome sequences available for downloading. There are also features like blasting sequences against a genome and genomic maps. In total, 19 haloarchaeal species are registered in HaloWeb. HaloBase contains more general information in 23 halophilic archaeal and bacterial halophiles. GenBank sequence numbers, number of chromosomes and plasmids, gene/protein content, and cellular features are among the database entries. HaloBase provides user accounts, followed by the ability to add a new organism as a registered member. In HProtDB, the priority is protein content. The resource contains physical and biochemical properties of halophilic proteins for 21 strains of Archaea and Bacteria. It also allows users to register as members and enter their own halophilic data. All three databases are restricted to information about halophilic Archaea and Bacteria, their number of entries is limited to an average of 18 and are irregularly updated.

In this chapter, a new halophiles database covering more than 1250 halophilic species and spanning all three domains of life is presented. This new resource was named "HaloDom" and is available online at: http://halodom.bio.auth.gr.

2.2 Materials and Methods

An extensive literature search has been carried out through the Web of Science, Scopus, PubMed, and Google Scholar using appropriate keywords (i.e. haloph*, salt, saline, hypersaline, extremophile) as well as combinations of them. Ultimately, the Web of Science was chosen as the primary source of literature as it provided a sophisticated search/query engine that suited our methodology and was proven to contain most of the papers found in other literature databases. The keyword combination that returned most papers in Web of Science was "sp nov haloph*" (on title section), returning 610 papers reporting new halophilic species at the time of the search. The same keyword combination returned many results in Google Scholar (2410), but not all of these papers contained the desirable keywords in their titles making its search engine unsuitable for our purposes. Scopus returned 615 results, but the interface of Web of Science offered a more flexible environment. There was great overlap among all three literature databases. Google Scholar however also returned several unrelated papers. Finally, a small number of books and reports containing useful information about halophilic species (albeit with no salinity data) were also included.

The methodology followed for data extraction was the same for papers, books, reports, or other document types. Because of the query "sp nov haloph*" that was placed as a title search, every document result from Web of Science contained reports on novel species. The text was searched for salinity information about the new species, in particular minimum, maximum, and optimal salinity range reported either as weight-to-volume (w/v), parts per thousand (ppt) or molar concentration (M). All salinity information was converted to weight-to-volume scale for data homogeneity. Finally, the halophilic species were categorized as "slight", "moderate", or "extreme" according to Ollivier [24], and more information mentioned below were added to an excel spreadsheet.

The obtained list was initially refined by topic and document type and potential errors on the data were filtered out manually. The final dataset-spreadsheet was organized in several columns (i.e. full taxonomy of each species, salinity record or range, halotolerance classification, genome availability, bibliography, notes/other information). Several taxonomy databases were used for registering the taxonomy of halophilic organisms (Table 2.1). “Salinity recorded or range” column reports either a single salinity value or a range of salinities, or both depending on the available information from the scientific source. “Halotolerance classification” included three halophilic categories: “slight”, “moderate”, and “extreme”. We searched for full genomes for all our entries in the NCBI genome database. The column “Genome availability” contained five possible states: complete genome, shotgun, mitochondrial genome, chloroplast genome, and no (not available). “Bibliography” contained the scientific article/s from which the information was extracted. “Notes/other info” is a complementary column for any type of information or metadata gauged as necessary to be documented.

Table 2.1. Taxonomy databases that were used to record taxonomic information about halophiles in HaloDom.

Πίνακας 2.1. Οι βάσεις δεδομένων ταξινομίας που χρησιμοποιήθηκαν για την καταγραφή πληροφοριών στους αλόφιλους οργανισμούς της HaloDom.

Taxonomy Database	Number of species
NCBI taxonomy browser	942
algaeBASE	49
World Register of Marine Species	32
Encyclopedia of Life	32
Integraded Taxonomic Information System	17
Atlas of Living Australia	11
Catalogue of Life	1
Global Biodiversity Information Facility	1
Global species	1
INPN - Inventaire National du Patrimoine Naturel	1
Marine species identification portal	1
Sum	**1088**

The spreadsheet was converted to a comma separated values (csv) file, and uploaded to a local database with the use of XAMPP and apache server [50]. PhpMyAdmin was also used [51] in order to handle the administration of the MySQL database protocol locally or in a webpage. Additionally, NetBeans 8.1 IDE (Integrated Development Environment) [52] was installed for creating the website with the use of HTML (HyperText Markup Language) and the programming languages PHP and Javascript. The user interface was created and modified using HTML and cascading style sheets (CSS) for the visual parts, and both PHP and Javascript for all functional parts regarding interactions between users and the database.

After importing the spreadsheet to the database all data were converted from a csv file to a table called “halodb”. The table was assigned with a primary key column called “Species_ID”. A primary key in mySQL is a number for each individual row of a table and it is unique. In this case, every halophilic species has a unique primary key. This primary key, or “Species_ID” column, always contains an integer starting from 1 and set to “auto-increment”. As more species are added to the database, this number is automatically increased providing every species with its distinctive ID number.

The HaloDom data structure started as one table that contained all information. However, as data volume increased it was necessary to break down the database into several tables. This methodology improves the speed and efficiency of the database during user query. It is also a way of organizing data, so that administrators can easily check the data integrity, make changes, and reduce redundancy. The structure of the database was changed from the table called “halodb”, containing all recorded information, to three tables. The first information separated from “halodb” was the “Bibliography” column, which was moved to a table called “Bibliography”. “Bibliography” table was assigned a primary key called “Biblio_id” and four columns: “pub_title” which contains the title of the study, “authors” containing the study’s author/s, “journal” mentioning the name of the journal, and “biblio_link” providing a direct link to the study. The third table is called “genomes” and contains five columns: “Genome_id” which is the primary key, “Species_ID” which is a foreign key from “halodb” table, “Species” which is the species name, “Genome_type” which declares the type of genome, and the “ncbi_link” which contains the link to the genome details in the NCBI genome database. A graph of the relationships between all three tables can be found in Figure 2.1.

The website project in NetBeans 8.1, written mostly on HTML, CSS, and PHP, was named “HaloDB”. Several .php files were created in order to design the user interface and its database functions. Home page contains a welcoming text and a photo slide created with the use of a jQuery script. Moreover, pie charts were created with Google Charts [53] and the use of JavaScript. These charts were embedded to the webpage code and can be viewed through the user interface. Halophilic entries are presented in a new page when clicked, where all available information is listed. Additionally, if a full genome is available the user can be redirected to the corresponding NCBI genome page. Also, users can perform a nucleotide or protein search.

A short Perl script was written to utilize the ip counter incorporated into the website code. The script was named “ip2location.pl” (Perl script A1) and created a list of locations derived from the list of ip addresses of the website visitors. The script uses a Perl package from the Comprehensive Perl Archive Network [54], called “Geo::IP2Location”. The list of locations was saved in a .csv file format and inserted in Google maps [55] in order to visualize the website’s visitors . Also, the traffic monitoring of HaloDom was assigned to Google analytics [56] for extensive and more analytical information regarding the audience of the website.

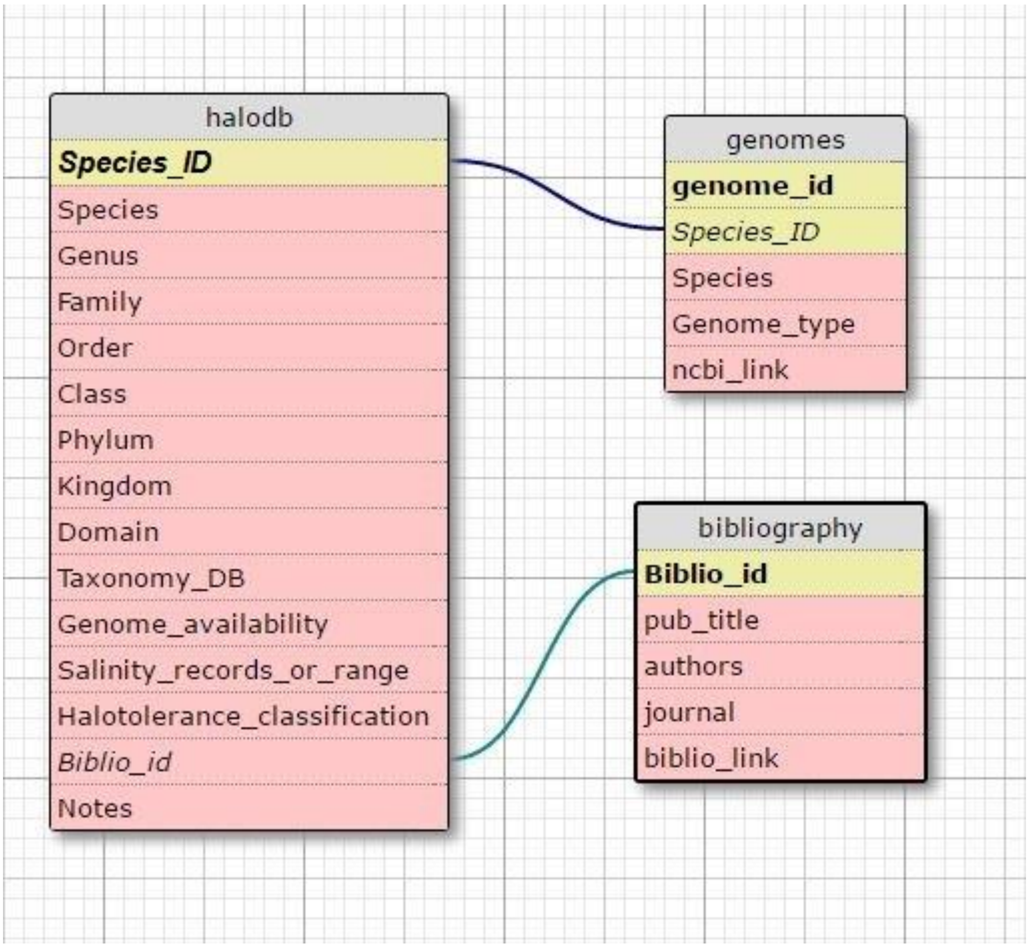


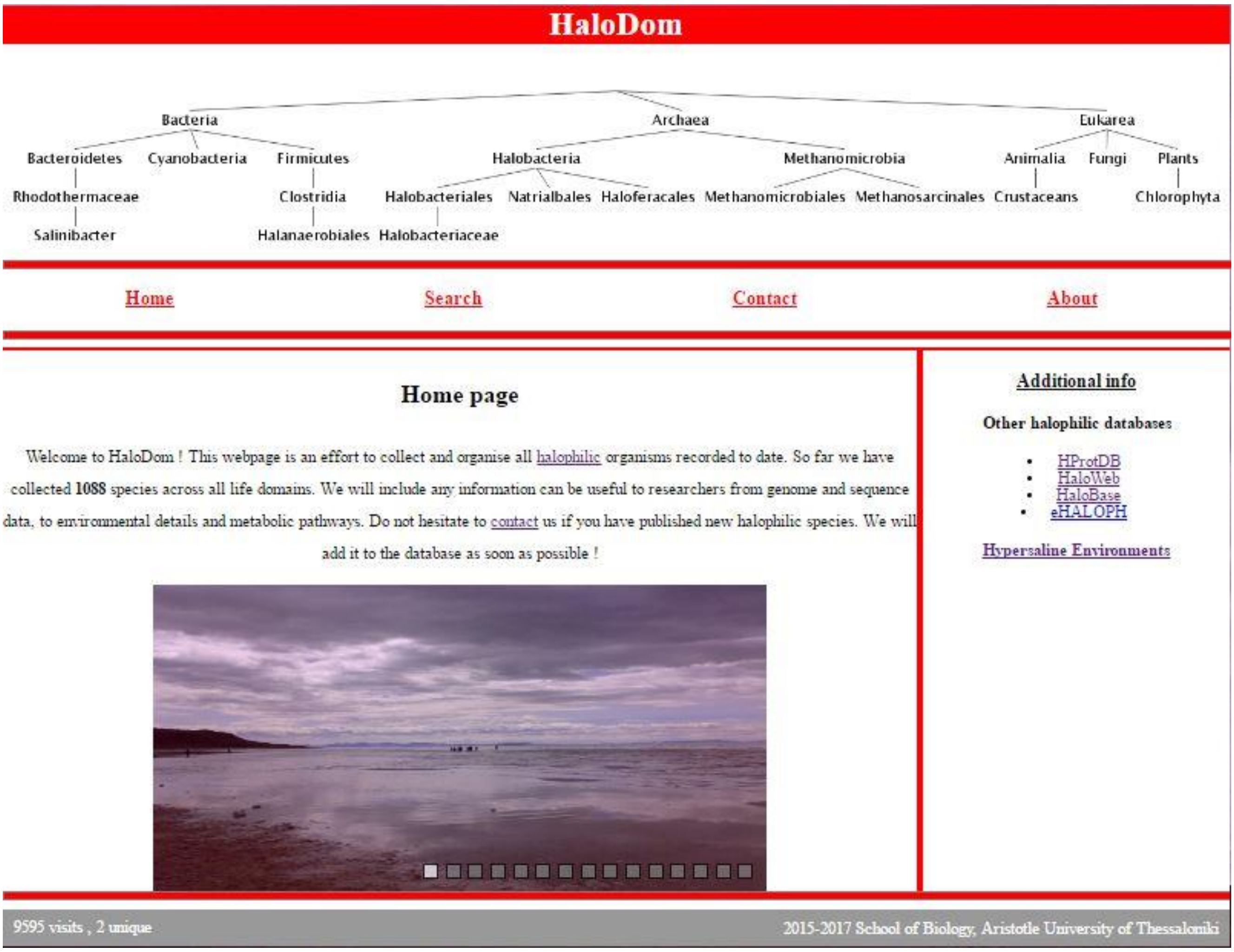
Figure 2.1. HaloDom consists of three tables shown in the picture. Foreign key relationships are shown with blue lines. Beige cells are integer numbers while purple cells are varchar, meaning mixed characters and numbers.

Εικόνα 2.1. Η Halodom αποτελείται από τρεις πίνακες δεδομένων οι οποίοι φαίνονται στην εικόνα. Οι μπλε γραμμές δηλώνουν τις σχέσεις ξένου κλειδιού. Τα κελιά χρώματος μπεζ περιέχουν ακέραιους αριθμούς ενώ τα μωβ περιέχουν χαρακτήρες αλλά και αριθμούς (αλφαριθμητικά).

2.3 Results

HaloDom is an online database containing more than 1200 halophilic species from all life domains. Users can perform a keyword search in all columns of the “halodb” table and retrieve all matching entries in numbered order. The homepage of HaloDom can be seen in Figure 2.2.

The main menu contains four options: “Home”, “Search”, “Contact”, and “About”. The search page, apart from retrieving data entries, can also show all recorded data and several pie charts created for a better visual interpretation of the listed halophilic data. The search page prompts the user to choose a column and perform a keyword search. When displaying the results, search always displays “Species_ID”, “Species”, and “Domain” columns. The column that the user has selected to perform the keyword search is shown in parentheses inside the “species” column. Exact or partial keyword matches are highlighted as light-colored text. The results are displayed in several pages, if necessary. Users can choose how many results per page should be displayed (10, 25, 50, 100). When a search is performed on “Bibliography” field, the results are shown on a different table that contains paper title, authors, journal, and corresponding species. Figure 2.3 shows the search results page for all fields except “Bibliography” while Figure 2.4 shows the results table for “Bibliography” searches. The species name is always clickable and leads to the corresponding entry. The entry page contains all available information and can lead the user to NCBI for more genomic information. Figure 2.5 displays an example entry page for *Artemia tibetiana*.



Home page

Welcome to HaloDom ! This webpage is an effort to collect and organise all halophilic organisms recorded to date. So far we have collected **1088** species across all life domains. We will include any information can be useful to researchers from genome and sequence data, to environmental details and metabolic pathways. Do not hesitate to contact us if you have published new halophilic species. We will add it to the database as soon as possible !

Additional info

Other halophilic databases

- HProtDB
- HaloWeb
- HaloBase
- eHALOPH

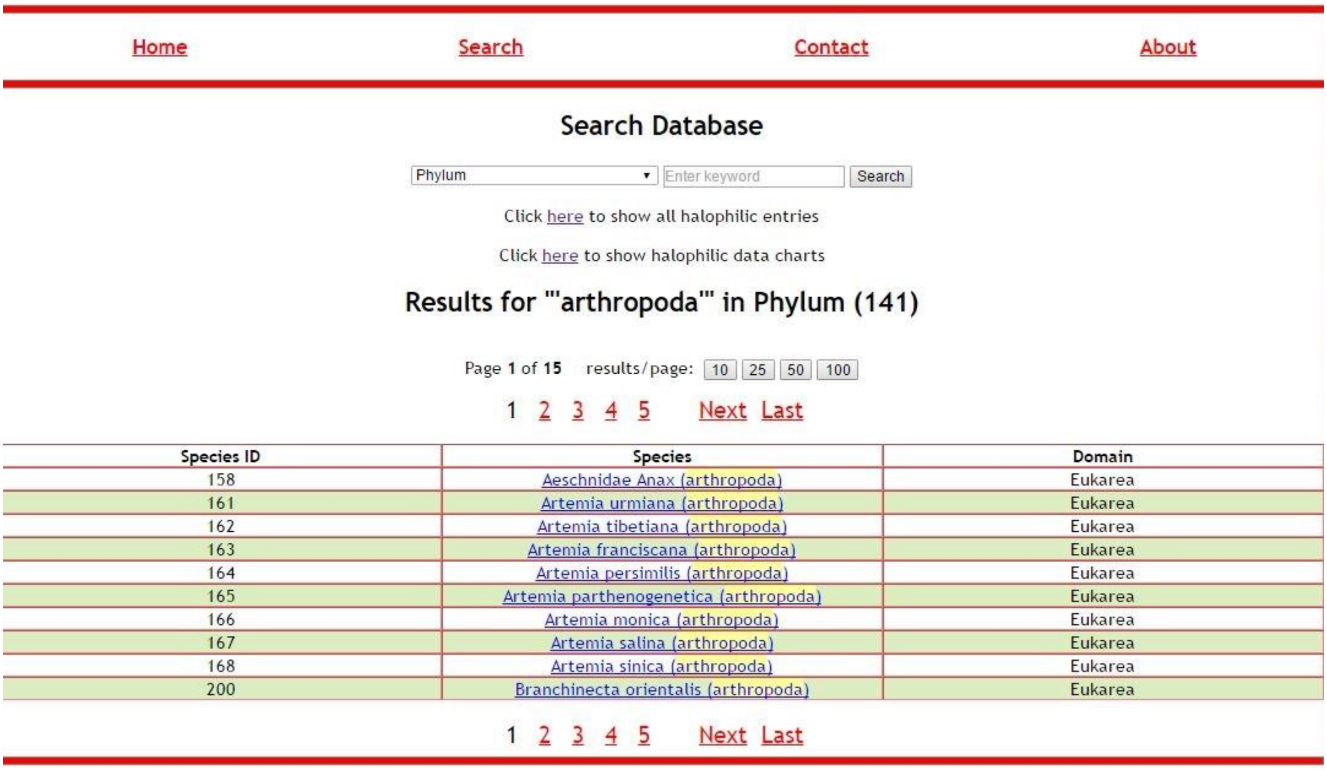
Hypersaline Environments

9595 visits , 2 unique

2015-2017 School of Biology, Aristotle University of Thessaloniki

Figure 2.2. Homepage of HaloDom contains mainly a welcoming text, a small tree graph, and a photo slide.

Εικόνα 2.2. Η κεντρική σελίδα της HaloDom περιέχει κυρίως ένα εισαγωγικό κείμενο, ένα μικρό δέντρο-γράφημα και μια σειρά από ολισθαίνουσες εικόνες.



Search Database

Click here to show all halophilic entries

Click here to show halophilic data charts

Results for "'arthropoda'" in Phylum (141)

Page **1** of **15** results/page: 10 25 50 100

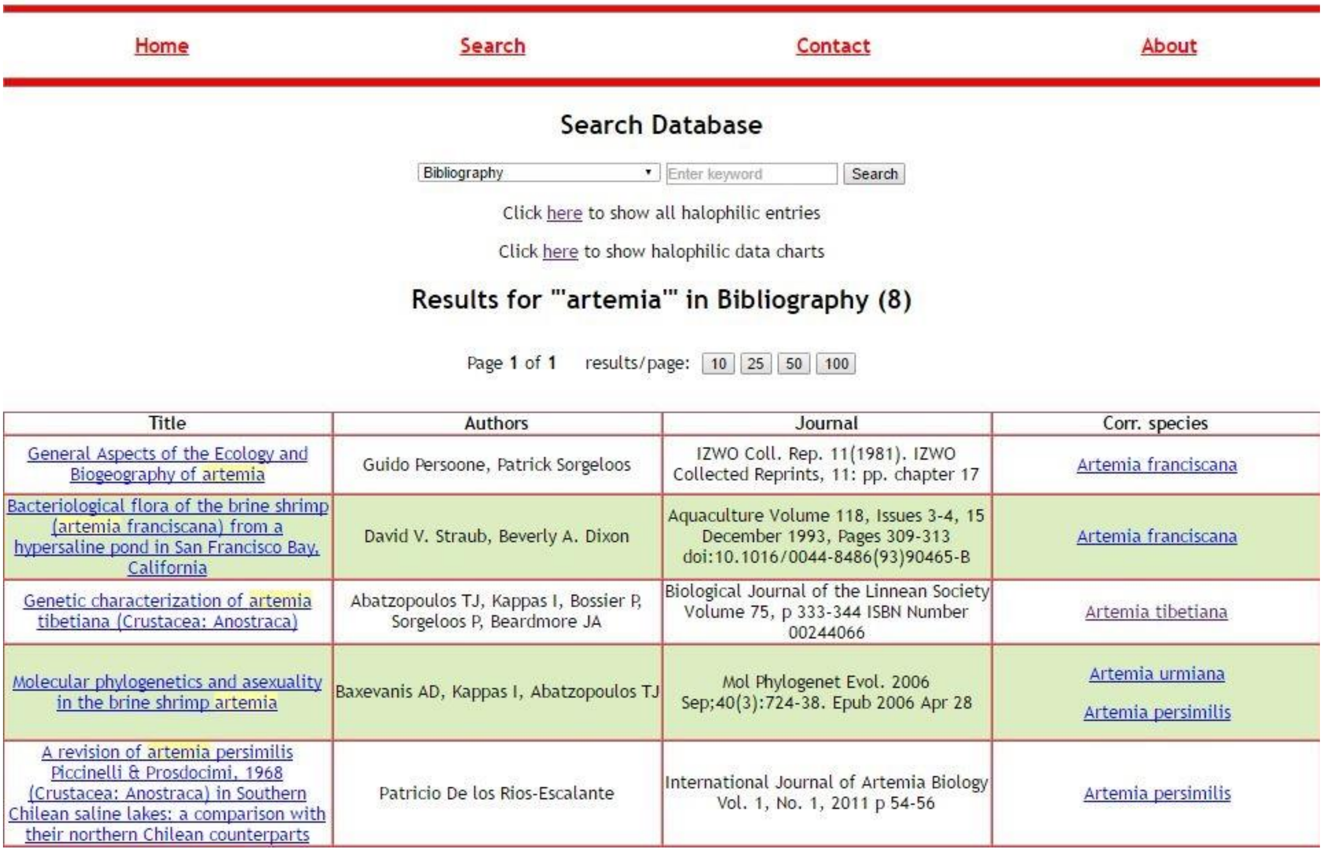
1 2 3 4 5 Next Last

Species ID	Species	Domain
158	Aeschnidae Anax (arthropoda)	Eukarea
161	Artemia urmiana (arthropoda)	Eukarea
162	Artemia tibetiana (arthropoda)	Eukarea
163	Artemia franciscana (arthropoda)	Eukarea
164	Artemia persimilis (arthropoda)	Eukarea
165	Artemia parthenogenetica (arthropoda)	Eukarea
166	Artemia monica (arthropoda)	Eukarea
167	Artemia salina (arthropoda)	Eukarea
168	Artemia sinica (arthropoda)	Eukarea
200	Branchinecta orientalis (arthropoda)	Eukarea

1 2 3 4 5 Next Last

Figure 2.3. The first ten results for keyword "arthropoda" in the "Phylum" column.

Εικόνα 2.3. Τα πρώτα δέκα αποτελέσματα αναζήτησης για τον όρο “arthropoda” στο πεδίο “Φύλο”.



Search Database

Click here to show all halophilic entries

Click here to show halophilic data charts

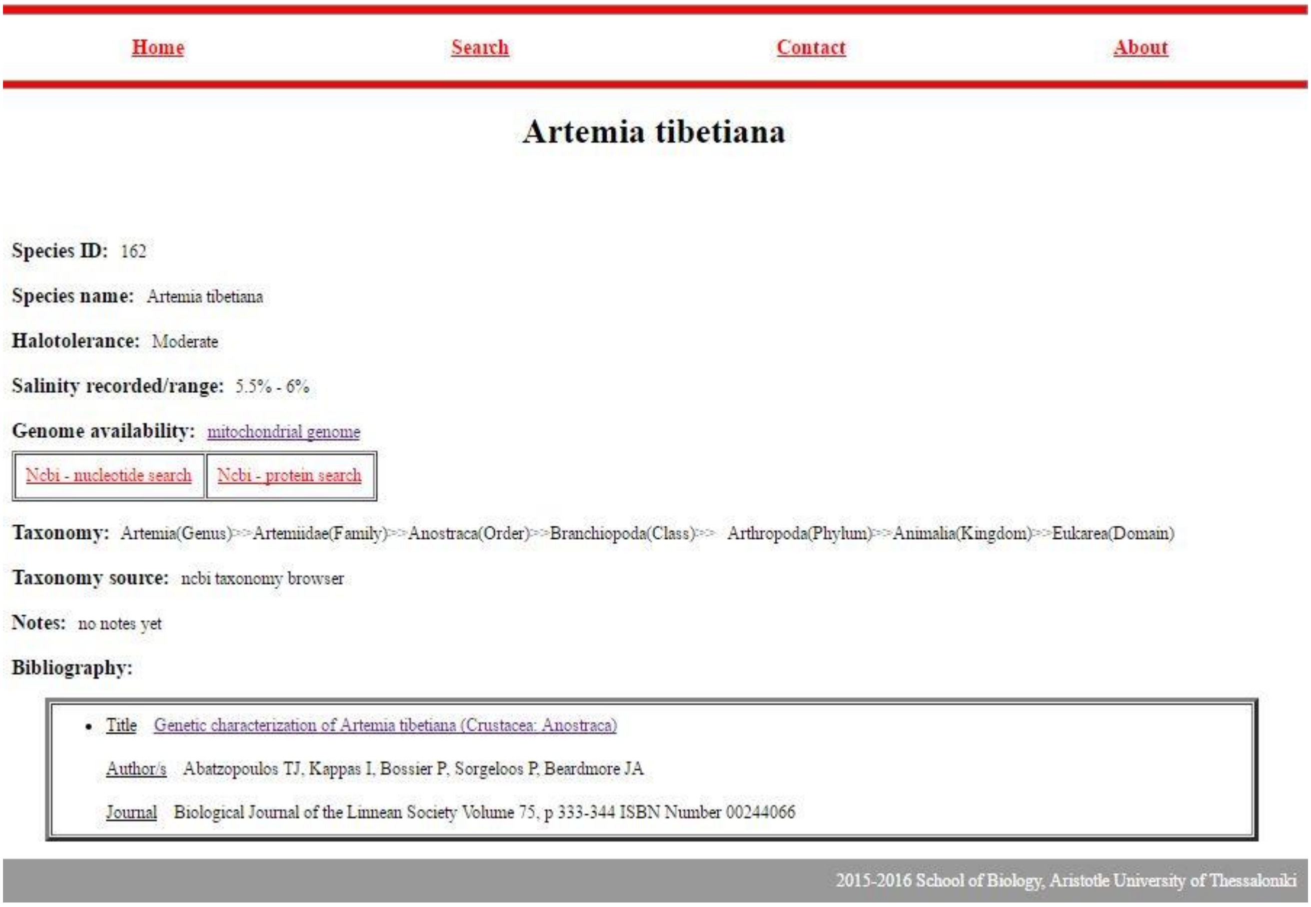
Results for "'artemia'" in Bibliography (8)

Page **1** of **1** results/page: 10 25 50 100

Title	Authors	Journal	Corr. species
General Aspects of the Ecology and Biogeography of artemia	Guido Persoone, Patrick Sorgeloos	IZWO Coll. Rep. 11(1981). IZWO Collected Reprints, 11: pp. chapter 17	Artemia franciscana
Bacteriological flora of the brine shrimp (artemia franciscana) from a hypersaline pond in San Francisco Bay, California	David V. Straub, Beverly A. Dixon	Aquaculture Volume 118, Issues 3-4, 15 December 1993, Pages 309-313 doi:10.1016/0044-8486(93)90465-B	Artemia franciscana
Genetic characterization of artemia tibetiana (Crustacea: Anostraca)	Abatzopoulos TJ, Kappas I, Bossier P, Sorgeloos P, Beardmore JA	Biological Journal of the Linnean Society Volume 75, p 333-344 ISBN Number 00244066	Artemia tibetiana
Molecular phylogenetics and asexuality in the brine shrimp artemia	Baxevanis AD, Kappas I, Abatzopoulos TJ	Mol Phylogenet Evol. 2006 Sep;40(3):724-38. Epub 2006 Apr 28	Artemia urmiana Artemia persimilis
A revision of artemia persimilis Piccinelli & Prosdocimi, 1968 (Crustacea: Anostraca) in Southern Chilean saline lakes: a comparison with their northern Chilean counterparts	Patricio De los Rios-Escalante	International Journal of Artemia Biology Vol. 1, No. 1, 2011 p 54-56	Artemia persimilis

Figure 2.4. The displaying format of "Bibliography" field results for the keyword "artemia".

Εικόνα 2.4. Ο τρόπος απεικόνισης αποτελεσμάτων αναζήτησης καταχωρημένης βιβλιογραφίας για τον όρο "artemia".



Artemia tibetiana

Species ID: 162

Species name: Artemia tibetiana

Halotolerance: Moderate

Salinity recorded/range: 5.5% - 6%

Genome availability: mitochondrial genome

Taxonomy: Artemia(Genus)=>Artemiidae(Family)=>Anostraca(Order)=>Branchiopoda(Class)=> Arthropoda(Phylum)=>Animalia(Kingdom)=>Eukarea(Domain)

Taxonomy source: ncbi taxonomy browser

Notes: no notes yet

Bibliography:

- Title Genetic characterization of Artemia tibetiana (Crustacea: Anostraca)

  Author/s Abatzopoulos TJ, Kappas I, Bossier P, Sorgeloos P, Beardmore JA

  Journal Biological Journal of the Linnean Society Volume 75, p 333-344 ISBN Number 00244066

2015-2016 School of Biology, Aristotle University of Thessaloniki

Figure 2.5. The entry page for *Artemia tibetiana*.

Εικόνα 2.5. Η σελίδα με την καταχώρηση του είδους *Artemia tibetiana*.

When showing all data from the search page, the user can select ascending or descending order with respect to a certain column. The pie charts visualize basic information about the data. For example, the first chart calculates the percentage of Archaea, Bacteria, and Eukarya in our database. When the user's mouse hovers above a certain piece, the frequency is shown first and then the corresponding percentage enclosed in parentheses. The first two pie charts are shown in Figure 2.6.

"Contact" section lists the administrators and contact information. "About" page shows the date of creation of HaloDom, current number of registered halophilic species, and the database version.

The results from traffic analysis of HaloDom show a wide range of visitors from across the globe (Fig. 2.7). Google analytics [56] further confirmed the results and provided extensive information about users, demographics, and locations. An overview about users per country as of March 2020 can be seen in Figure 2.8.

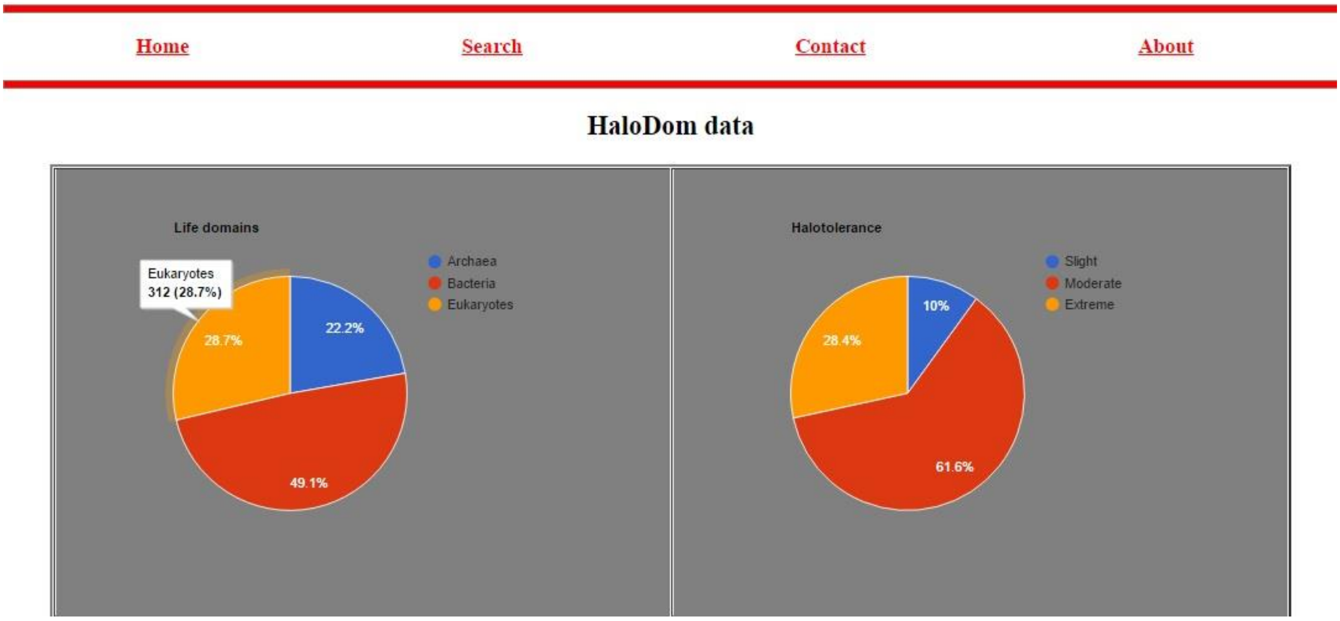


Figure 2.6. Halophilic data pie charts. Left: Frequency and percentage of Archaea, Bacteria, and Eukaryotes. Right: Frequency and percentage of Slight, Moderate, and Extreme halophiles in the database.

Εικόνα 2.6. Γραφήματα πίτας για τα δεδομένα αλόφιλων οργανισμών. Στην αριστερή πλευρά οι συχνότητες και τα ποσοστά για Αρχαία, Βακτήρια και Ευκαρυώτες. Στην δεξιά πλευρά οι συχνότητες και τα ποσοστά για ελαφρώς, μέτρια και ακραίως αλόφιλα είδη.

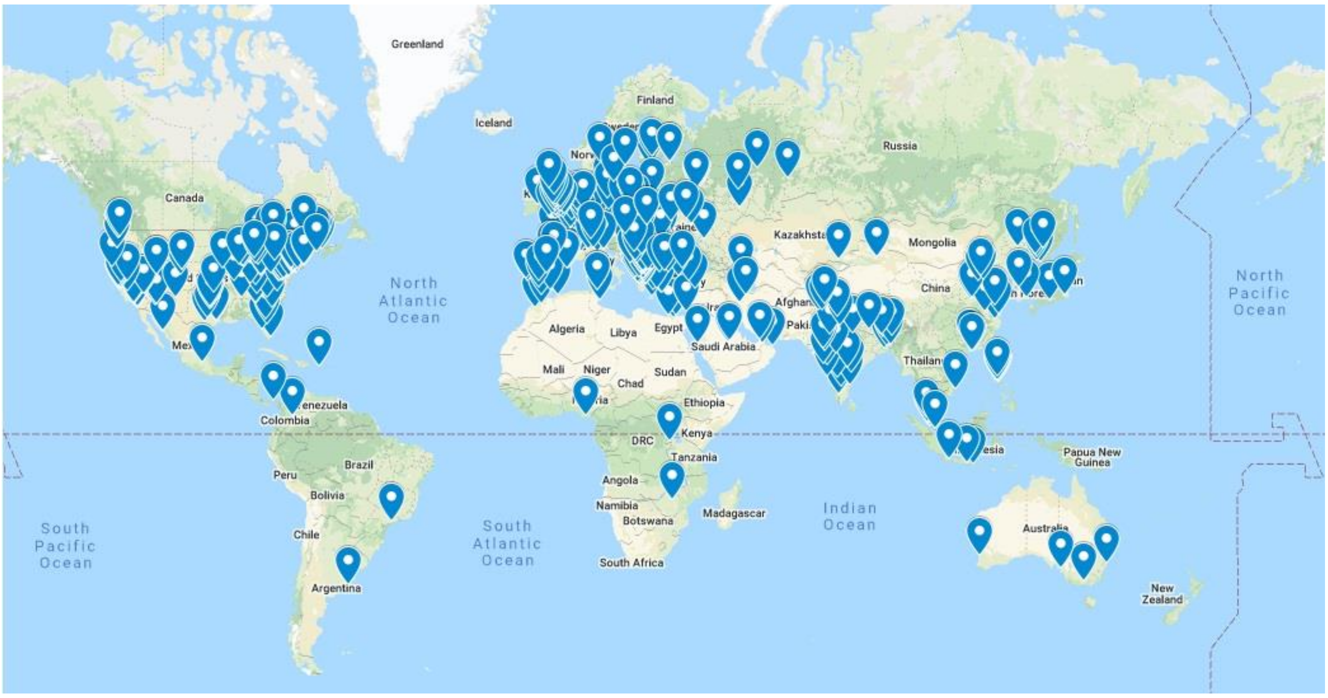
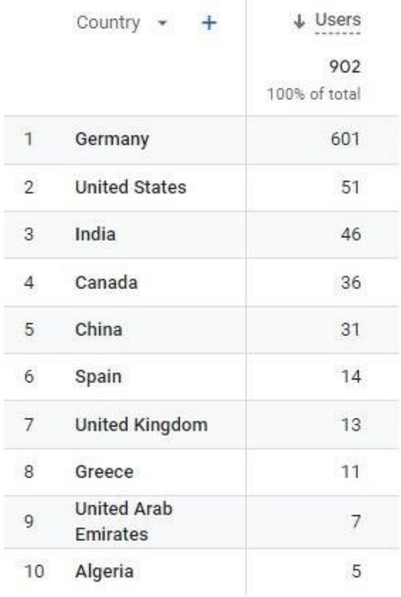


Figure 2.7. Unique ip addresses and their locations, collected from HaloDom's traffic monitor.

Εικόνα 2.7. Μοναδικές διευθύνσεις δικτύου με τις αντίστοιχες τοποθεσίες τους που συλλέγονται από τις καταγραφές επισκέψεων της HaloDom.



	Country	↓ Users
		902 100% of total
1	Germany	601
2	United States	51
3	India	46
4	Canada	36
5	China	31
6	Spain	14
7	United Kingdom	13
8	Greece	11
9	United Arab Emirates	7
10	Algeria	5

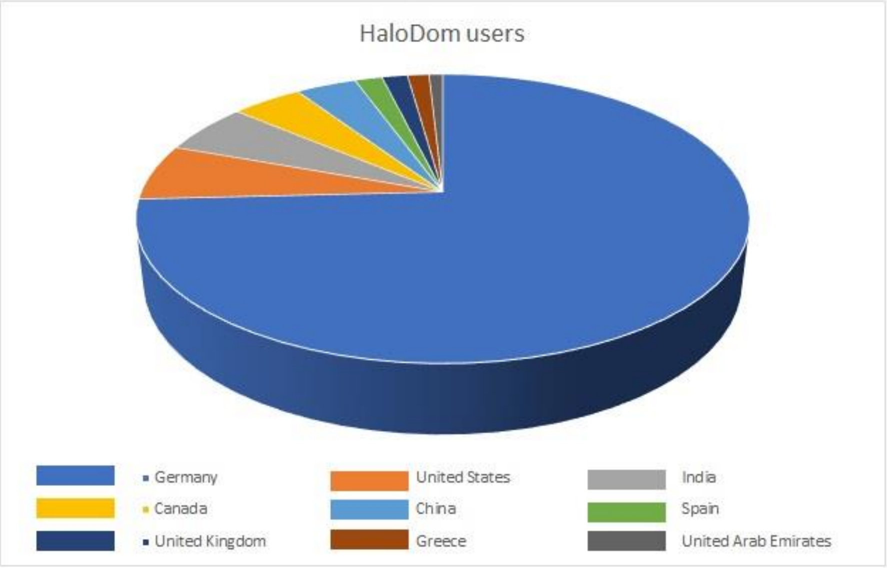


Figure 2.8. Users of HaloDom per country as of March 2024.

Εικόνα 2.8. Χρήστες της HaloDom ανά χώρα τον Μάρτιο του 2024.

2.4 Discussion

As a resource, HaloDom expands considerably compared with previous databases in terms of coverage (representatives from all life domains) and number of entries. Periodical updates are scheduled once every 2 months and as the database grows, additional metadata (e.g. geographic distribution, biochemical properties etc.) and analytical tools are planned to be incorporated.

Occasionally, during data curation and annotation, species nomenclature proved to be a challenge. This was especially true for Archaea and Bacteria given their notoriously difficult taxonomy and the fast discovery of new strains [57]. Considerable efforts were invested into resolving this issue by using several taxonomy databases (Table 2.1), but also user feedback is encouraged. A grey picture also exists in the literature regarding threshold values in halophile classification (slight/moderate/extreme). For example, in one study the copepod *Cletocamptus retrogressus* was found in 2-7.4% (w/v) salinity, and thus categorized as slight to moderate halophile, while in another study the recorded salinity range was 19.8-36% (w/v), characteristic of extreme halophiles. This probably reflects the limited knowledge on the biology of many species but as additional data are gathered more accurate annotations are expected. Also, in the light of idiosyncratic molecular mechanisms and signatures in extreme halophilic Archaea [58, 59], criteria for halophile classification could be refined.

As of January 2024, HaloDom contains 1268 entries of halophiles across all life domains. Novel halophilic species continue to be published in a steady rate and new data are incorporated to the database. Also, the simple ip address counter monitoring user traffic in HaloDom shows that more than 4000 unique ip addresses have visited the website. Moreover, further analysis and visualization dictates that users are visiting the website from several places across the globe (Fig. 2.7).

As of March 2020, HaloDom was registered in Google analytics, for better monitoring the website's user traffic. The first four countries that most users come from were United States, Canada, India, and Greece (Fig. 2.8). These results are to some extent a reflection of the intensity of research on halophiles in these countries (see also China included in the top 10 HaloDom visitors). Additionally, there are several regular users present in HaloDom each month which is probably a sign that HaloDom supports certain research projects in some locations.

The current database fills a gap in halophile research and can be used as a useful repository and starting point for a wide range of investigations. Over the last few years, research has focused on the mechanisms responsible for modulating survival in hypersaline settings [13, 38, 60, 61], on the biotechnological production of halophile macromolecules [9, 62], on the phylogenetic position of halophiles in the tree of life [63], on climate change [64, 65], and even on astrobiology [66]. It is therefore obvious that halophile research addresses appealing questions to several fields of biology, especially in combination with the diverse spectrum of extremophile organisms. The answer to the basic question whether sustaining life in physicochemical extremes is a matter of entire adaptation or due to the action of a few genes is crucial, multidisciplinary, and influential [67].

Chapter 3: Phylogenomic analysis of halophilic Archaea

3.1 Introduction

Archaea possess a complex evolutionary history. Being a diverse taxonomic group, it was initially categorized as Bacteria. However, the accumulation of gene sequences over the years has proven this arrangement obsolete [68]. Archaeal phylogeny has puzzled researchers and the discovery of new major phyla and species require a constant update on the phylogenetic status of the group [69]. Determining phylogeny with the use of single-marker phylogenetic techniques such as 16S rRNA has been proven sub-optimal, especially among Archaea which exhibit great genetic diversity and variability even on typically conserved genetic sequences [70].

The archaeal class of Halobacteria and other halophilic taxa within Euryarchaeota are a good example of such a diversity [71]. Importantly, with the constant growth of public sequence data, more complicated, multi-marker approaches to phylogeny can be applied [72]. The phylogenetic literature of Archaea as a whole or within certain clades is considerable, diverse, and broad. Examples include studies on the DNA replication machinery [73], on the metabolic abilities of Archaea [74, 75], but also on their evolutionary relationships with Eukarya [72, 76, 77].

All these studies show that the requirements in software tools, computational power, and data handling methods for large-scale phylogenomic analyses are now met. Although papers have investigated and kept track of the phylogenetic relationships of the major class of Halobacteria [78-80], the relationships between Halobacteria and other halophilic species such as the recently discovered Nanohaloarchaea or halophilic bacteria have not been fully established yet. In the present chapter, a phylogenetic tree of 124 Halobacteria and 33 outgroup species was assembled with the use of 242 previously published core protein archaeal markers [81, 82]. This multi-marker tree was used in order to gain more insights into the phylogenetic relationships of halophilic Archaea and, specifically, within the class of Halobacteria.

The topology of the tree partly confirmed existing placements but also provided extra information about the core proteins shared by Halobacteria. It also demonstrated that adaptation to salinity is not necessarily manifested by common ancestry but sharing a specific group of genetic traits that remain elusive in their functions.

3.2 Materials and Methods

Phylogenetic tree reconstruction from 242 core archaeal markers

For the construction of the multi-marker phylogenetic tree of Halobacteria, several protein markers were used in .hmm file format. The markers were extracted from the published PhyEco [81] group, a set of genes selected with focus on universality across Archaea and Bacteria, the ability to be used to produce robust phylogenetic trees that reflect as much as possible the evolution of the species from which the genes come, and also low variation in copy number across taxa. Markers were also collected from the Amphora2 package [82], which is an automated phylogenomic inference tool. Amphora2 offers a greatly expanded phylogenetic marker database and can analyze both Bacterial and Archaeal sequences. In total, 266 markers for Archaea and Bacteria were extracted. The PhyEco markers were 106 and already individual .hmm files. The markers from Amphora2 were 160, all in a single file. They were separated in individual files with a Perl script created for our purposes, called "separate_hmms.pl" (Perl script A2). In total 266 .hmm files were created and placed in a single folder.

With the help of HaloDom [11], 124 Halobacteria proteomes from complete genome records were downloaded in fasta format from NCBI and also 33 outgroup proteomes from both the archaeal and bacterial domains with completed genome records (Table A1). The outgroup species were selected for their known phylogenetic position as close relatives of Halobacteria. However additional species of interest like Nanohaloarchaea, thermophiles, and halophilic Bacteria were included in order to examine their relationship with Halobacteria. All proteomes were also placed in a single folder.

A search was conducted locally with HMMER [83] for the presence of these 266 markers in our downloaded proteomic fasta sequences. A Perl script was created ("hmmer_ex.pl") in order to execute HMMER multiple times for every marker in every genome (Perl script A3). The results were saved in another folder in the form of .out text files. These results contain a list of matches (if any) that HMMER found in our proteomes, along with statistical information like E-values and scores. From this list, the best hit was selected in the form of its unique GenBank ID. To extract GenBank IDs from more than 40,000 result files, a series of Perl scripts for text mining were developed in order to determine whether the files in question contained an HMMER hit, or if there were no matching sequences for each hmm profile. From this process 40,576 GenBank IDs were obtained.

With the GenBank IDs, the corresponding fasta sequences were retrieved from NCBI and a catalogue of our 157 species (124 Halobacteria and 33 outgroups) was created, along with their corresponding marker match sequence in fasta format. The fasta files were titled as their species name and the number of the corresponding marker (1-266), for example "Halobacterium_salinarum_157.fasta". From these fasta files, multiple alignments were created for every marker. In total 266 alignments were created with the use of the Perl script "create_alignments_w_mafft.pl" (Perl script A4) and the local, command line version of MAFFT [84].

All fasta files containing markers were inserted as input in the annotation tool HAMAP [85]. An excel spreadsheet was created with the results of the annotation process. Additionally, visual inspection of the 266 alignments also determined which markers consisted of highly conserved sequences. From the annotation, 198 markers scored high in HAMAP with the label "trusted". However, another 44 markers were additionally included in the final analysis, because of their highly conserved sequences after manual/visual inspection.

In total 242 alignments were concatenated with Geneious [86] into a single fasta file. This file was used an input for IQ-TREE [87, 88], a powerful software tool used for efficient maximum likelihood phylogenetic tree estimation, allowing for robust and accurate evolutionary analysis of genomic data. The software was used in its own server, but was also installed in the HPC of Aristotle University of Thessaloniki for testing purposes. Finally, the software ran with default parameters and 1000 ultrafast bootstrap cycles for reconstruction of the final phylogenetic tree.

3.3 Results

Phylogenetic tree from 242 core archaeal markers

A phylogenetic tree (Fig. 3.1) from 242 core archaeal markers was created for 124 Halobacteria and 33 outgroup species (Table A1). The topology separates all Halobacteria from outgroups. The node of Halobacteria contains all three known orders: Halobacteriales, Haloferacales, and Natrialbales. Haloferacales form a monophyletic group and are separated in two large branches. The same is true for Natrialbales which diverges from Halobacteriales. However, the order Halobacteriales is a paraphyletic assemblage: the node leading to Halobacteriales includes the common ancestor to the exclusion of four other members of Halobacteriales: *Halanaeroarchaeum sulfurireducens, Halarchaeum acidiphilum, Halobacterium jilantaiense*, and *Halobacterium salinarum.*

Outgroups are placed also in four monophyletic groups. The first group consists of 14 Archaea of class Methanomicrobia. The second group contains two members of class Archaeoglobi. The third group contains three members of the order Thermoplasmata, one member of candidate class Candidatus Poseidoniia, two members of proposed superphylum Asgardarchaeota, and five members of TACK superphylum. The fourth group contains three members of class Nanohaloarchaea and three species of bacteria. All species of the tree can be found in Table A1 in the appendix.

Throughout the tree, bootstrap values in the majority of nodes are higher than 90%. The Halobacteria node is supported by 100%. Within Halobacteria, the three orders are also well supported. In the outgroups, the clade of Thermoplasmata is at 66% while the clade of Methanomicrobia is at 63%. The complete tree topology with bootstrap values included can be seen in Figure 3.2.

Data assembly characteristics

Several datasets were created during reconstruction of Halobacteria phylogeny. First, a catalog of all downloaded protein markers in hmm format was created. These markers are reported to be abundant in Archaea and could be the starting point for further analyses. Also, 157 proteomes of the used species were assembled in fasta format. Several Perl scripts were produced in order to get the GenBank IDs of interest. These blocks of code can help in future projects regarding genomic data handling. The results of HMMER in .out format are also of considerable size including more than 40,000 lists of protein sequences that match a specific HMM protein model. The same is true for the top matches (protein sequences) extracted from these lists. Then, there are the fasta files containing all 242 protein markers for every species and the multiple alignment files derived from these files. Finally, the concatenated multiple alignments were created in several versions with different marker numbers included, for testing the runtimes of IQ-TREE on the HPC AUTh cluster of "Aristotelis" (10, 100, 200, and 242 markers). All files and folders created for the phylogenetic analysis alone occupied >2GB on hard drive, around 300,000 files, and 896 folders.

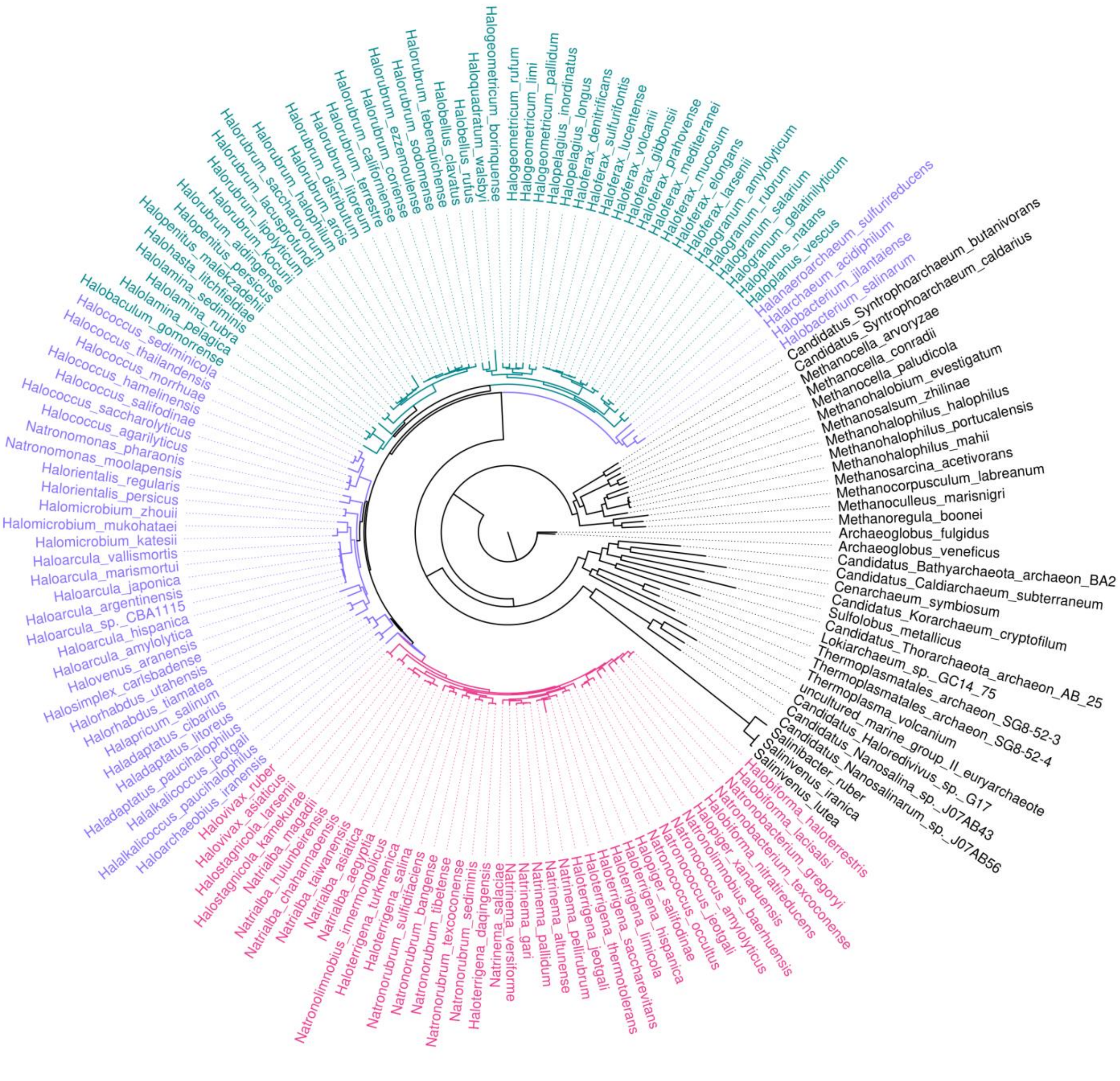
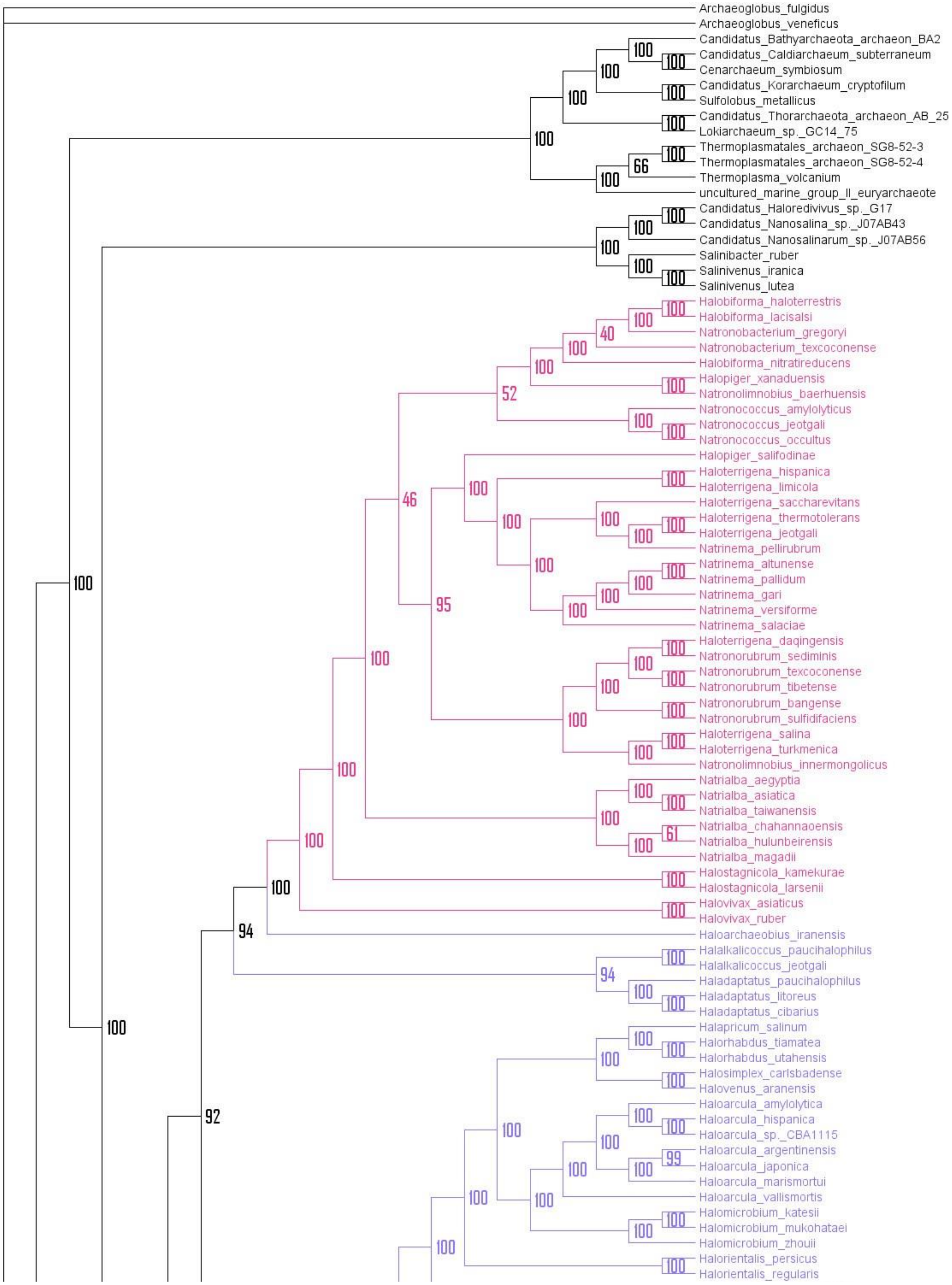


Figure 3.1. Radial maximum likelihood phylogeny from 242 conserved archaeal protein markers, produced with IQ-TREE software. The tree consists of 124 members of Halobacteria and 33 outgroups. Green: Order Haloferacales. Red: Order Natrialbales. Purple: Order Halobacteriales. Black: Outgroups.

Εικόνα 3.1. Φυλογενετικό δέντρο με τη μέθοδο μέγιστης πιθανοφάνειας από 242 συντηρημένους πρωτεϊνικούς δείκτες Αρχαιοβακτηρίων, κατασκευασμένο με το λογισμικό IQ-TREE. Το δέντρο απεικονίζει τις σχέσεις 124 μελών της κλάσης Halobacteria και 33 εξωομάδες. Πράσινο: Τάξη Haloferacales. Κόκκινο: Τάξη Natrialbales. Μωβ: Τάξη Halobacteriales. Μαύρο: Εξωομάδες.


Archaeoglobus_fulgidus
Archaeoglobus_veneficus
Candidatus_Bathyarchaeota_archaeon_BA2
Candidatus_Caldiarchaeum_subterraneum
Cenarchaeum_symbiosum
Candidatus_Korarchaeum_cryptofilum
Sulfolobus_metallicus
Candidatus_Thorarchaeota_archaeon_AB_25
Lokiarchaeum_sp._GC14_75
Thermoplasmatales_archaeon_SG8-52-3
Thermoplasmatales_archaeon_SG8-52-4
Thermoplasma_volcanium
uncultured_marine_group_II_euryarchaeote
Candidatus_Haloredivivus_sp._G17
Candidatus_Nanosalina_sp._J07AB43
Candidatus_Nanosalinarum_sp._J07AB56
Salinibacter_ruber
Salinivenus_iranica
Salinivenus_lutea
Halobiforma_haloterrestris
Halobiforma_lacisalsi
Natronobacterium_gregoryi
Natronobacterium_texcoconense
Halobiforma_nitratireducens
Halopiger_xanaduensis
Natronolimnobius_baerhuensis
Natronococcus_amylolyticus
Natronococcus_jeotgali
Natronococcus_occultus
Halopiger_salifodinae
Haloterrigena_hispanica
Haloterrigena_limicola
Haloterrigena_saccharevitans
Haloterrigena_thermotolerans
Haloterrigena_jeotgali
Natrinema_pellirubrum
Natrinema_altunense
Natrinema_pallidum
Natrinema_gari
Natrinema_versiforme
Natrinema_salaciae
Haloterrigena_daqingensis
Natronorubrum_sediminis
Natronorubrum_texcoconense
Natronorubrum_tibetense
Natronorubrum_bangense
Natronorubrum_sulfidifaciens
Haloterrigena_salina
Haloterrigena_turkmenica
Natronolimnobius_innermongolicus
Natrialba_aegyptia
Natrialba_asiatica
Natrialba_taiwanensis
Natrialba_chahannaoensis
Natrialba_hulunbeirensis
Natrialba_magadii
Halostagnicola_kamekurae
Halostagnicola_larsenii
Halovivax_asiaticus
Halovivax_ruber
Haloarchaeobius_iranensis
Halalkalicoccus_paucihalophilus
Halalkalicoccus_jeotgali
Haladaptatus_paucihalophilus
Haladaptatus_litoreus
Haladaptatus_cibarius
Halapricum_salinum
Halorhabdus_tiamatea
Halorhabdus_utahensis
Halosimplex_carlsbadense
Halovenus_aranensis
Haloarcula_amylolytica
Haloarcula_hispanica
Haloarcula_sp._CBA1115
Haloarcula_argentinensis
Haloarcula_japonica
Haloarcula_marismortui
Haloarcula_vallismortis
Halomicrobium_katesii
Halomicrobium_mukohataei
Halomicrobium_zhouii
Halorientalis_persicus
Halorientalis_regularis
100
100
100
100
100
100
100
100
66
100
100
100
100
100
100
100
100
40
100
100
100
52
100
100
100
46
100
100
100
100
100
100
100
100
100
95
100
100
100
100
100
100
100
100
100
100
100
100
61
100
100
100
100
94
94
100
100
100
100
100
92
100
100
100
100
100
100
99
100
100
100
100
100
100
100

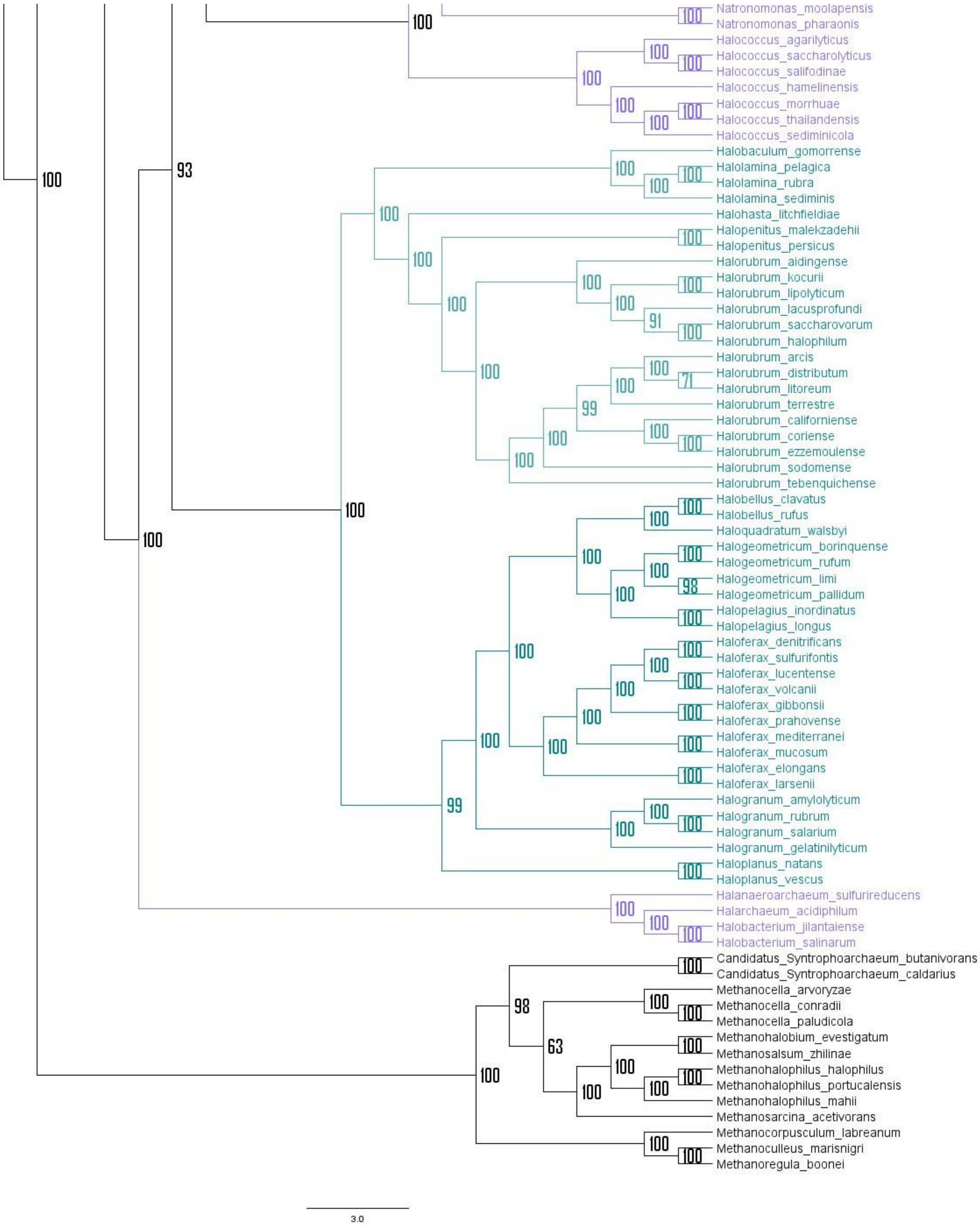

Natronomonas_moolapensis
Natronomonas_pharaonis
Halococcus_agarilyticus
Halococcus_saccharolyticus
Halococcus_salifodinae
Halococcus_hamelinensis
Halococcus_morrhuae
Halococcus_thailandensis
Halococcus_sediminicola
Halobaculum_gomorrense
Halolamina_pelagica
Halolamina_rubra
Halolamina_sediminis
Halohasta_litchfieldiae
Halopenitus_malekzadehii
Halopenitus_persicus
Halorubrum_aidingense
Halorubrum_kocurii
Halorubrum_lipolyticum
Halorubrum_lacusprofundi
Halorubrum_saccharovorum
Halorubrum_halophilum
Halorubrum_arcis
Halorubrum_distributum
Halorubrum_litoreum
Halorubrum_terrestre
Halorubrum_californiense
Halorubrum_coriense
Halorubrum_ezzemoulense
Halorubrum_sodomense
Halorubrum_tebenquichense
Halobellus_clavatus
Halobellus_rufus
Haloquadratum_walsbyi
Halogeometricum_borinquense
Halogeometricum_rufum
Halogeometricum_limi
Halogeometricum_pallidum
Halopelagius_inordinatus
Halopelagius_longus
Haloferax_denitrificans
Haloferax_sulfurifontis
Haloferax_lucentense
Haloferax_volcanii
Haloferax_gibbonsii
Haloferax_prahovense
Haloferax_mediterranei
Haloferax_mucosum
Haloferax_elongans
Haloferax_larsenii
Halogranum_amylolyticum
Halogranum_rubrum
Halogranum_salarium
Halogranum_gelatinilyticum
Haloplanus_natans
Haloplanus_vescus
Halanaeroarchaeum_sulfurireducens
Halarchaeum_acidiphilum
Halobacterium_jilantaiense
Halobacterium_salinarum
Candidatus_Syntrophoarchaeum_butanivorans
Candidatus_Syntrophoarchaeum_caldarius
Methanocella_arvoryzae
Methanocella_conradii
Methanocella_paludicola
Methanohalobium_evestigatum
Methanosalsum_zhilinae
Methanohalophilus_halophilus
Methanohalophilus_portucalensis
Methanohalophilus_mahii
Methanosarcina_acetivorans
Methanocorpusculum_labreanum
Methanoculleus_marisnigri
Methanoregula_boonei
3.0

Figure 3.2. The maximum likelihood phylogeny of figure 3.1 here shown in rectangular format. Green: Order Haloferacales. Red: Order Natrialbales. Purple: Order Halobacteriales. Black: Outgroups.

Εικόνα 3.2. Η φυλογένεση μέγιστης πιθανοφάνειας της εικόνας 3.1, εδώ σε ορθογώνια προβολή. Πράσινο: Τάξη Haloferacales. Κόκκινο: Τάξη Natrialbales. Μωβ: Τάξη Halobacteriales. Μαύρο: Εξωομάδες.

3.4 Discussion

The maximum likelihood phylogenetic tree (Fig. 3.1) created for 124 Halobacteria and 33 outgroups (Table A1), using 242 core archaeal protein markers, is mostly in accordance with the currently accepted Halobacteria phylogeny [79]. Both Haloferacales and Natrialbales are retrieved as monophyletic groups. In contrast to these two clades, the third order of Halobacteriales forms a paraphyletic assemblage. This is caused by the exclusion of the distant members *Halobacterium salinarum* and *Halarchaeum acidiphilum*, in line with previous research [79]. In the current phylogeny, two more members, *Halobacterium jilantaiense* and *Halanaeroarchaeum sulfurireducens*, are added to this distant group. The polyphyletic component of the inferred paraphyly of Halobacteriales strongly suggests that halophilicity probably originated repeatedly either through convergence or horizontal gene transfers. However, we should also take into account the frequent occurrence of strains with multiple extreme requirements or tolerances (poly-extremophiles, e.g. *Halarchaeum acidiphilum*). These situations may result in complex genome architectures producing a blurred evolutionary signal during phylogenetic investigations. Undoubtedly, this is a neglected topic in studies on extremophile taxa and more research is needed.

In previous phylogenies [78] using 80 genomes of Halobacteria and 40 PhyEco protein markers, also used in this chapter [81], the tree topology is similar. Haloferacales and Natrialbales are monophyletic, and Halobacteriales are separated in two distant groups. Bootstrap values are significantly lower in previous phylogenies [78] but the overall topology still confirmed the key phylogenetic relationships.

Outgroups (Table A1) are separated in four monophyletic clades. The first clade contains all members of the class Methanomicrobia. It appears that this methanogenic and halophilic group does differ significantly from Halobacteria in genetic traits. It is not yet clear exactly how Methanomicrobia deal with salinity, even though they are reported to be able to survive in high salt concentrations [89-91]. Another clade contains the recently discovered archaeal class Nanohaloarchaea [92] whose members are using Halobacteria as hosts [93, 94], are also extremely halophilic, even though without known mechanisms of adaptation. Nanohaloarchaea do not share most of the protein markers used here with the Halobacteria group. This clade also includes three bacterial outgroups. *Salinibacter ruber*, *Salinivenus iranica*, and *Salinivenus lutea* have been reported from saline environments, are extremely halophilic, and also closely related [95-97]. Though genetically they are considered to be closest to the thermophilic genus *Rhodothermus*, they are most comparable to the family Halobacteriaceae, because of similarity in protein structure. The salinity adaptation mechanisms of these species are in part attributed to gene sharing through lateral gene transfer from halophilic Archaea like Halobacteria [19], although it does not seem to be the case with the conserved protein markers used for this analysis. It will be interesting to compare genomic data between halophilic Archaea and Bacteria, to detect the core genomic machinery required for survival in extreme salinity. The third and most diverse clade contains members of class Thermoplasmata, Asgard Archaea, TACK group and one Marine Group II Euryarchaeota member which is reported as ubiquitous planktonic marine organism. The fact that it is placed in proximity with the thermophilic clade is an interesting observation because so far it is considered as “mesophilic”. The last outgroup clade is comprised by two members of class Archaeoglobi, *Archaeoglobus fulgidus* and *Archaeoglobus veneficus*. The latter is reported to be a polyextremophile. All outgroups are separated from the Halobacteria clade and the further investigation of their interesting -and in many cases unresolved- phylogenetic relationships is appropriate, however is not in the scope of this dissertation.

Chapter 4: Amino acid profile analysis of halophilic species

4.1 Introduction

It has been previously shown that thermophilic and halophilic microbes can be distinguished from the amino acid composition of their proteins [98]. In this chapter, we take advantage of the significantly increased microbial sequence data to evaluate how broad this finding is and also investigate for the presence of new traits that can potentially be uncovered by a large-scale proteomic analysis.

Previous studies on the amino acid profile of extremophilic proteins have helped researchers gain a better understanding of the molecular adaptations required for survival in extreme conditions [99], with potentially important applications [101-103, 109, 110]. Analyses of small-sized datasets have so far provided strong indications that the amino acid sequence alone is a reliable indicator of thermophilic or halophilic lifestyle, with the observed amino acid compositions aiming for better protein stabilization, correct folding, and thermodynamic stability under several kinds of stress [59, 98, 104-106, 111-114]. Similar but limited research has been performed on psychrophilic and piezophilic proteins [115-117]. Additionally, amino acid profile adaptations have also been implemented in AI algorithms in an effort to increase efficiency and speed of machine learning based predictions [116, 118, 119]. In spite of the extensive availability of genomic and proteomic data in biological databases, large-scale analyses from extremophiles are limited. Utilizing large datasets and through a systematic approach to amino acid profiles could yield a clearer picture of these adaptations and help uncover new details about life in extreme environments. Here, we have assembled, with the help of HaloDom [11], the largest dataset, up to date, of proteomes from halophilic, thermophilic, and mesophilic Archaea and Bacteria in order to investigate the generality of previous findings, detect additional signature traits of adaptation to extreme conditions, and explore how these adaptations are distributed in various protein families and sizes, a topic also overlooked by previous studies.

Hierarchical clustering and principal components analysis revealed a detailed picture of the amino acid profiles of several microbial taxonomic groups such as halophilic and thermophilic Archaea, halophilic Bacteria, Euryarchaeota and other major archaeal and bacterial lineages.

Regarding the amino acid composition of halophiles, proteomic data revealed important differences between Halanaerobiales, Halobacteria, Methanomicrobia, and Nanohaloarchaea.

Several protein characteristics, for the aforementioned groups and more, were investigated including protein function and family, different protein sizes, and compositionally biased proteins both from halophilic and non-halophilic species. Our data dictate that different protein groups are subject to different levels of evolutionary pressure, depending on thermodynamic efficiency and possibly yet unknown factors.

4.2 Materials and Methods

Data retrieval and organization

A large proteome dataset was assembled for the analysis. In total, 222 microbial species with available complete or partial genome sequences on NCBI were included (Table A2). Genome completeness does not affect the corresponding proteomic data, so we took advantage of as many data in the protein sequence level as possible. Among these proteomes, 134 were from halophilic Archaea and 88 from several other taxa including thermophilic, non-halophilic, acidophilic, alkaliphilic, and psychrophilic Archaea and Bacteria. A wide range of extremophiles and other species were included in the dataset in order to yield a better picture of the microbial world, regarding the amino acid profile. For every species, its RefSeq proteome was downloaded from NCBI in fasta format. Halophilic species were pinpointed from the database Halodom [11] and the relevant literature.

Data formatting and preparation

In order to calculate amino acid frequencies from the downloaded proteomes, ResidueFrequencySummarizer [120] was used which accepts fasta or text files with multiple protein sequences as input. To run ResidueFrequencySummarizer multiple times for all 222 proteomes, a Perl script was created (“residue_calculator.pl”) which would call ResidueFrequencySummarizer for every proteome and create result files in csv format (Perl script A5). In these csv files, the first column contained the raw frequencies of every amino acid for all sequences included in the fasta file in question while the second column contained the corresponding percentages. This process resulted in the creation of 222 csv files containing the amino acid statistics mentioned above. The raw frequencies of every species were integrated to a single csv file which was used as a basis for hierarchical clustering and principal components analysis (PCA), as described below.

Hierarchical clustering and principal components analysis

Our excel data sheet containing all 222 species and their amino acid frequencies was converted to a tab delimited text file, in order for hierarchical clustering to be performed with PermutMatrix [121]. Euclidean distance was used as dissimilarity measurement and McQuitty's method (WPGMA), complete linkage, and average linkage (UPGMA) for linkage rules. Rows (amino acid frequencies for every species) were normalized with Z-scores from the PermutMatrix interface. Visual comparisons were made between the different linkage rules. All methods delivered roughly the same data clusters in different order in the cladogram. The WPGMA results of clustering with PermutMatrix [121] are shown in Figures 4.1 and 4.2.

PCA was performed in RStudio with the R script "PCA_analysis.r", using the same dataset from hierarchical clustering, containing all species and amino acid frequencies (R script A1). Two separate R scripts were created, one for PCA analysis ("PCA_analysis.r") and one for plotting eigenvectors (R script A2, "PCA_analysis_EIGEN.r"). Excel data were converted to a csv file and normalized with z-scores before insertion in RStudio. The analysis produced several plots saved as jpg files. The parameters of the scripts (such as colors and legends) were tweaked several times for the desired plot to be created. Data refinement for obtaining protein sizes, protein functional categories, and transmembrane proteins was conducted with the use of several Perl scripts including the popular package BioPerl [122]. The protein length histogram (Fig. A2) was created by an R script called "protein_length_historgam.r" (R script A3). GC content data were also included in the analysis.

From the initial proteome dataset with 222 species, 34 Bacteria (halophilic and non-halophilic) and 31 halophilic Archaea that had a complete genome sequence available were chosen for GC content calculation and incorporation in hierarchical clustering and PCA. In total, 65 genomes were downloaded in fasta format from NCBI. A Perl script was used to calculate GC contents in these fasta files as a percentage ("Fasta_GC_counter.pl") of the total nucleotide count (Perl script A6). Two separate csv files were created for GC content analysis, one containing only the amino acid frequencies of the 65 species and another containing an extra column with the GC percentages of the species. The WPGMA was used as linkage rule and Euclidean distance as dissimilarity option. Data were normalized by rows (z-scores). The resulting topology can be seen in Figure A5 in the appendix.

A subgroup of the 65 species was selected for further investigation. The bacterial order Halanaerobiales was chosen along with all halophilic Archaea to be included in a PCA analysis conducted by the R script “PCA_GC_analysis.r” (R script A4).

4.3 Results

Hierarchical clustering distinguishes halophilic from non-halophilic protein groups

As mentioned, 222 archaeal and bacterial species were used for the analysis and the corresponding proteomes were assembled (Table A2) in order to obtain amino acid profiles for every species. The clustering algorithm placed all members of the class Halobacteria in the same large group with a few exceptions. It appears that halophilic Nanohaloarchaea and halophiles from the class Methanomicrobia have a different amino acid profile from their halophilic relatives in Halobacteria. Nanohaloarchaea are placed outside the Halobacteria group. Methanomicrobia form a separate group along with *Methanococcoides burtonii* (a psychrophile) and *Candidatus methanoperedens nitroreducens* (Fig. 4.1). *Methanohalobium evestigatum* was placed in a separate group with *Candidatus nitrocosmicus oleophilus,* closer to the cluster of Halanaerobiales. From Nanohaloarchaea, the strains *Candidatus haloredivivus* sp*.* G17 and *Candidatus nanosalina* were grouped with *Halarsenatibacter silvermanii* (Fig. 4.1). *Candidatus nanosalinarum* is placed in a single group, next to the large cluster of Halobacteria (Fig. 4.2). The Halophilic Archaea *Haloarcula salaria* and *Haloarcula* sp. CBA1115 are also placed on their own showing a slightly different amino acid pattern from the majority of Halobacteria. The former is placed near the large halophilic cluster and the latter is placed as a single group within Halobacteria (Fig. 4.2). The halophilic *Haloquadratum walsby* is grouped outside the large halophilic cluster along with *uncultured marine group II euryarchaeote*. Most species from our Halanaerobiales dataset (11/12) are clustered in one group (Fig. 4.1). The remaining Halanaerobiales member, *Acetohalobium arabaticum*, is placed along with the halophilic, alkalithermophilic bacterium *Natranaerobius thermophilus*. All members of halophilic Halanaerobiales are placed away from the large halophilic cluster of Halobacteria.

Halophilic Archaea of class Halobacteria and thermophiles form two separate clusters in PCA

Principal component analysis was used with the proteomic data described earlier to better visualize a small fraction of amino acid adaptations of the microbial world. As seen in Figure 4.3, halophilic

Archaea of the class Halobacteria form a very distinct cluster, suggesting a highly conserved amino acid profile. Thermophiles form a much larger cluster with thermophilic species being more scattered and sparser. Also, small subgroups seem to be forming within the thermophilic cluster as data increase, for example five *Methanocaldococcus* species in the upper left part of the thermophilic cluster (Figs 4.3 and 4.4). The blue colored eclipse marks the halophilic bacterial family of Halanaerobiales. Within the eclipse are also six halophilic methanogenic Archaea, two Nanohaloarchaea, two acidophilic species, one alkaliphilic, a psychrophilic Archaeon, a member of the Asgard Archaea, several thermophiles, Firmicutes, and Euryarchaeota (Figs 4.3 and 4.4). Therefore, this is the most diverse cluster in the analysis. On the contrary, the halophilic red colored cluster located to the right (Fig. 4.3) solidly contains only Archaea from the class Halobacteria. *Haloquadratum walsby* is the only member of Halobacteria placed away from their halophilic cluster. The closest species to the amino acid profile of Halobacteria is *Methanoculleus marisnigri*, a methanogenic archaeon [123, 124]. Several bacterial halophilic species are being situated below and close to the cluster of Halobacteria (Fig. 4.3). Among these bacteria is the extremely halophilic *Salinibacter ruber* [96].

PCA eigenvectors agree with previous estimations and reveal new traits of salinity adaptations

The eigenvalues calculated in our analysis are in agreement with previous studies [59, 111-113, 125]. As seen in Figure 4.5, the most positively contributing amino acids towards haloadaptation are alanine, aspartic acid, arginine, proline, and histidine. In addition, valine, threonine, glycine, tryptophan, and cysteine also have a moderate positive effect. On the contrary, the most negatively contributing amino acids are lysine, isoleucine, asparagine, and leucine. Also, moderate negative contribution is coming from tyrosine, phenylalanine, and serine. Glutamic acid, methionine, and glutamine do not affect the amino acid profile directly as they account for variation in y axis (Fig. 4.5).

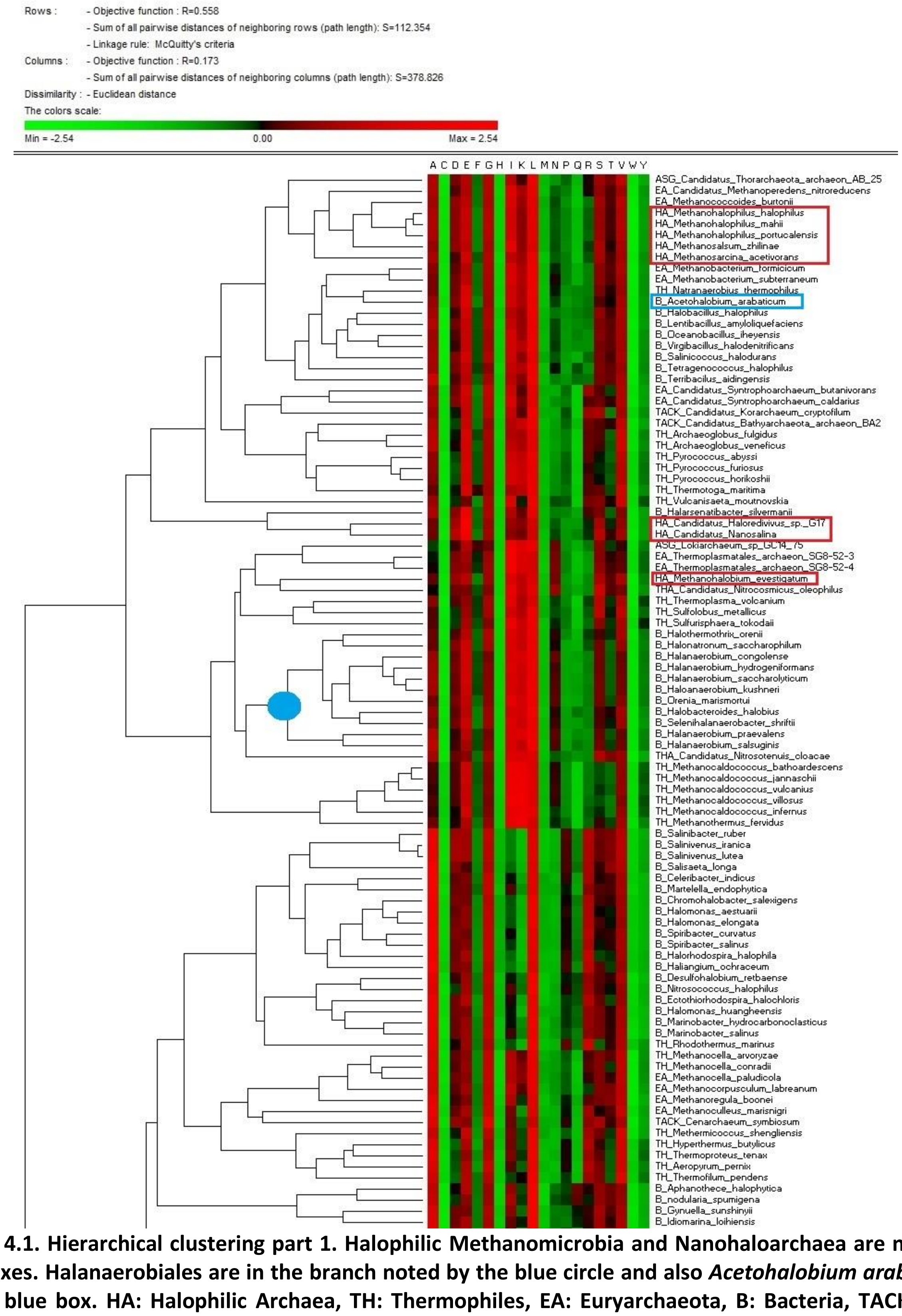


Figure 4.1. Hierarchical clustering part 1. Halophilic Methanomicrobia and Nanohaloarchaea are noted in red boxes. Halanaerobiales are in the branch noted by the blue circle and also *Acetohalobium arabaticum* in the blue box. HA: Halophilic Archaea, TH: Thermophiles, EA: Euryarchaeota, B: Bacteria, TACK: TACK superphylum Archaea, ASG: Asgard Archaea.

Εικόνα 4.1. Ιεραρχική ομαδοποίηση μέρος 1. Τα αλόφιλα Methanomicrobia και Nanohaloarchaea σημειώνονται στο σχήμα με κόκκινα πλαίσια. Τα Halanaerobiales είναι στον κλάδο που σηματοδοτείται

από τον μπλε κύκλο αλλά και στο μπλε κουτάκι (*Acetohalobium arabaticum*). HA: Αλόφιλα Αρχαία, TH: Θερμόφιλα, EA: Euryarchaeota, B: Βακτήρια, TACK: TACK superphylum Archaea, ASG: Asgard Αρχαία.

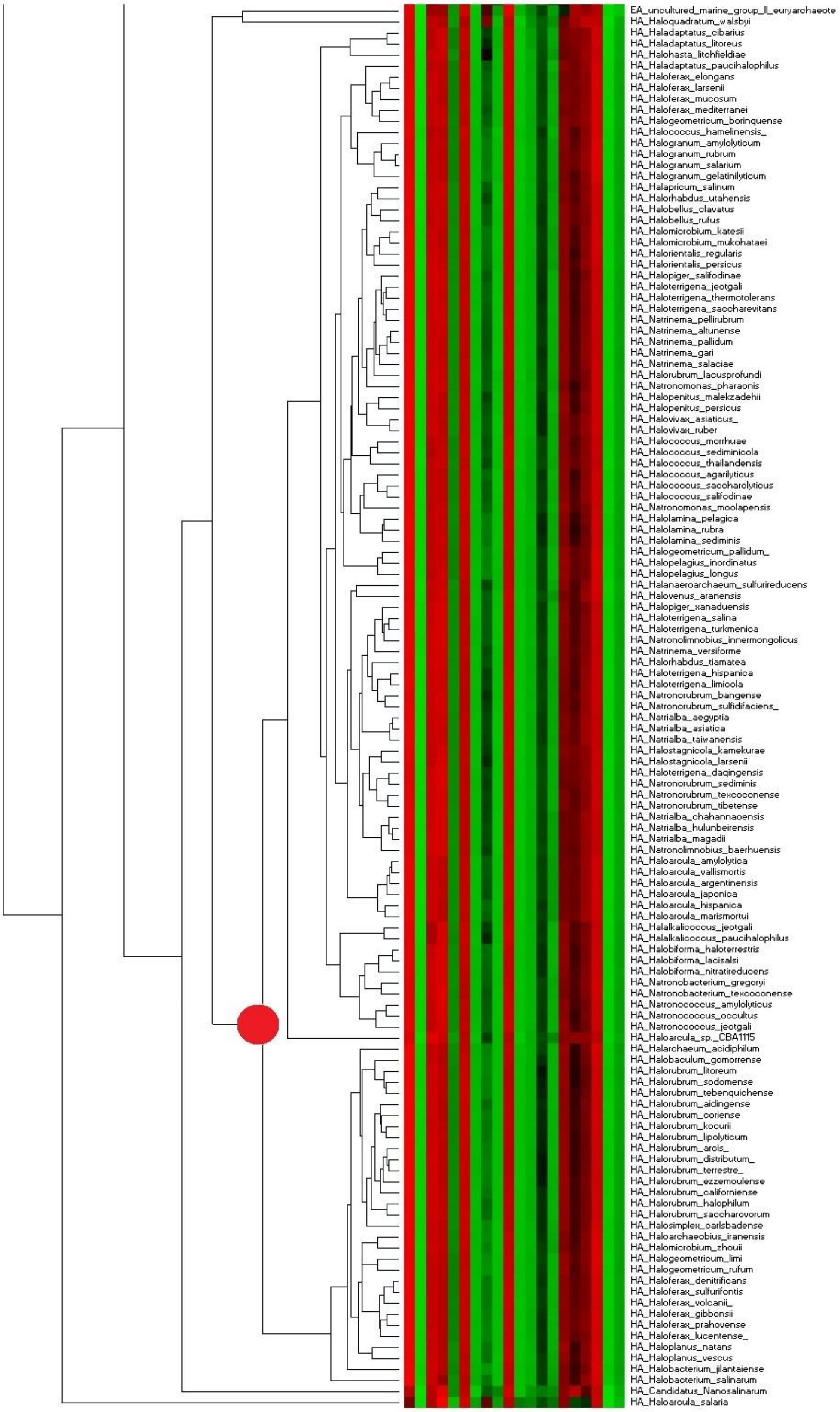


Figure 4.2. Hierarchical clustering part 2. The large cluster of Halobacteria is denoted with the red circle.

Εικόνα 4.2. Ιεραρχική ομαδοποίηση μέρος 2. Η μεγάλη ομάδα των Halobacteria συμβολίζεται με τον κόκκινο κύκλο.

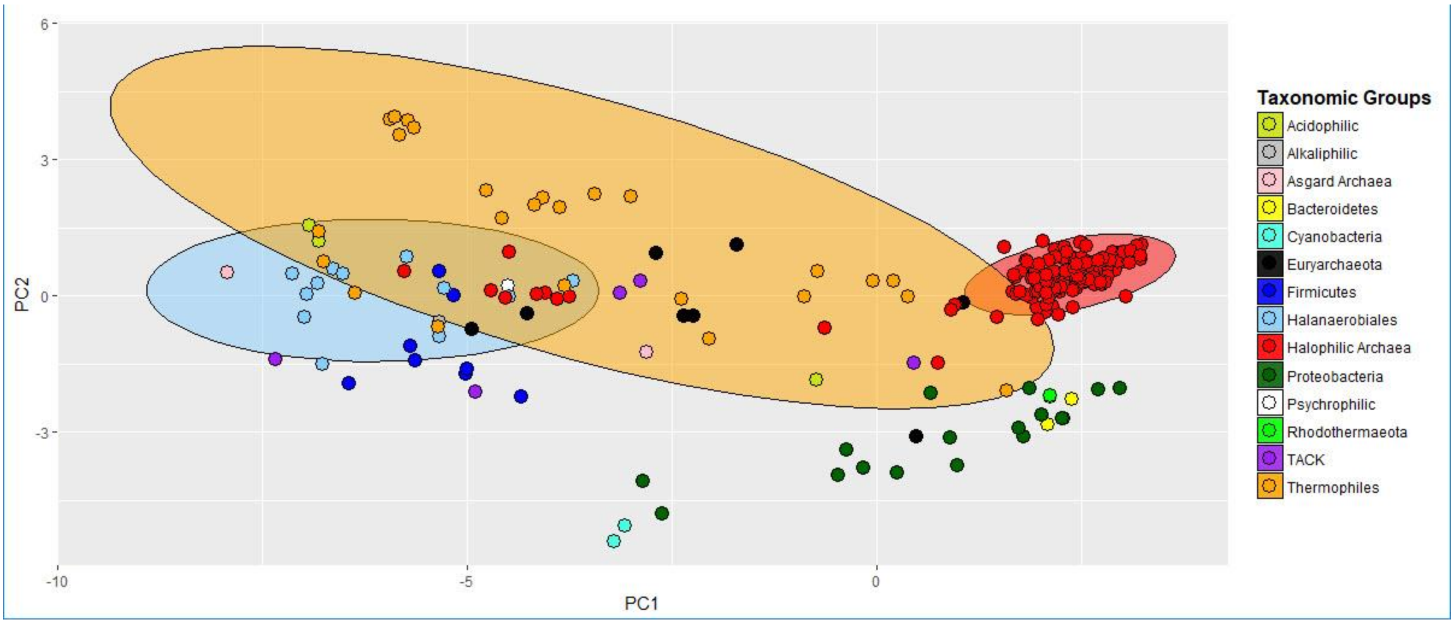


Figure 4.3. PCA analysis of amino acid profiles from 222 proteomes.

Εικόνα 4.3. Ανάλυση κυρίων συνιστωσών των προφίλ αμινοξέων για 222 πρωτεώματα.

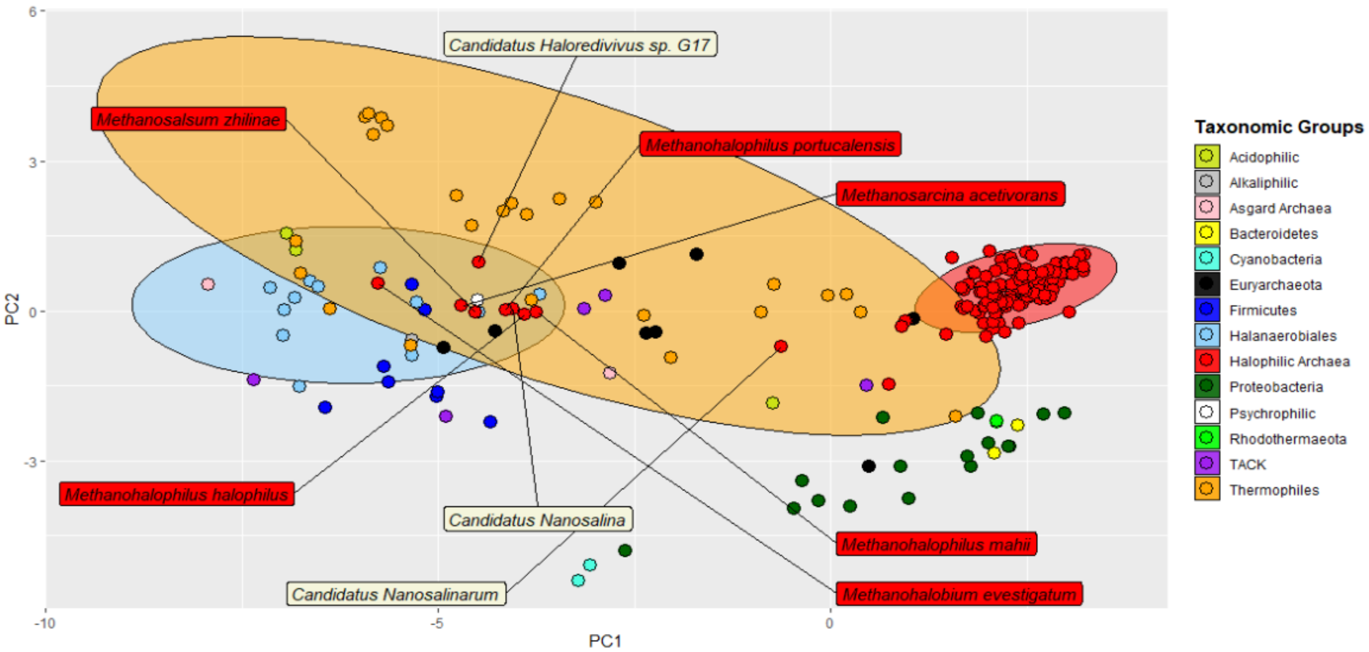


Figure 4.4. Halophilic Methanomicrobia (red labels) and Nanohaloarchaea (beige labels) have a different amino acid profile from the class Halobacteria (red cluster on the right).

Εικόνα 4.4. Τα αλόφιλα Methanomicrobia (κόκκινες ετικέτες) και Nanohaloarchaea (μπέζ ετικέτες) έχουν διαφορετικό προφίλ αμινοξέων από την κλάση Halobacteria (κόκκινη ομάδα στα δεξιά του σχήματος).

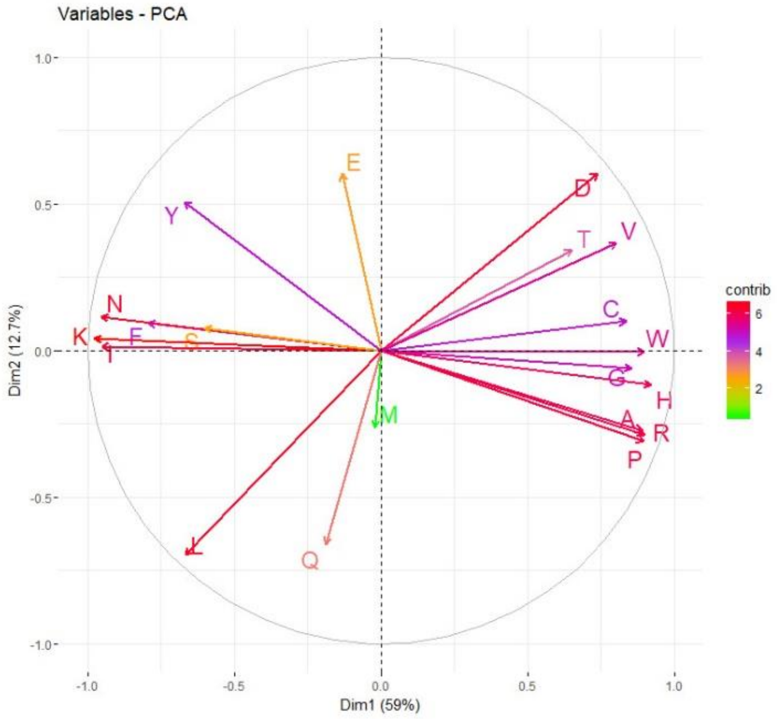


Figure 4.5. Eigenvectors of the PCA analysis shown in figure 4.3.

Εικόνα 4.5. Ιδιοδιανύσματα της ανάλυσης κυρίων συνιστωσών που απεικονίζεται στην εικόνα 4.3.

GC content, protein size, protein family, and composition bias in adapted halophiles

The incorporation of GC content in the analysis did not change the obtained topology for hierarchical clustering performed with PermutMatrix [121]. The topology of the clustering remained the same after the addition of GC data (Fig. A1).

A PCA analysis was conducted with data only from halophilic Archaea and Halanaerobiales, in order to investigate the effects of including GC content. Two PCA graphs were created, one with GC content included and one without. The resulting graphs do not have significant differences (Fig. A2).

In Figure 4.6 the amino acid profiles of small, medium, and large proteins were included in a PCA analysis containing halophilic Archaea, thermophiles, and Halanaerobiales. Protein sizes were chosen based on a protein length histogram of all halophilic proteins in our dataset (Fig. A3). Small proteins are situated outside the halophilic cluster of Halobacteria. Large proteins are situated in the lower section of the cluster, while medium-sized proteins are in the very center of the cluster. Similar results are shown in Figure 4.7 where several archaeal halophilic protein families are variably distributed. Translation factors and ribosomal proteins are situated outside and below the halophilic archaeal cluster, while other protein groups such as synthetases, polymerases, and many hypothetical proteins are situated in the core of the cluster. Transmembrane halophilic proteins are the most distant protein group from the halophilic cluster and present several differences in terms of amino acid profile and amino acid usage preferences. Additionally, in Figure 4.8, transmembrane proteins from different taxonomic groups, specifically from Halanaerobiales, thermophiles, and mesophiles also have a different amino acid profile in comparison with transmembrane proteins from halophilic Archaea of the class Halobacteria.

Finally, amino acid percentages from 123 proteomes of Halobacteria and 38 non-halophilic species from our data were plotted in a bar chart visualizing the differences between halophilic and non-halophilic amino acid profiles (Fig. 4.9).

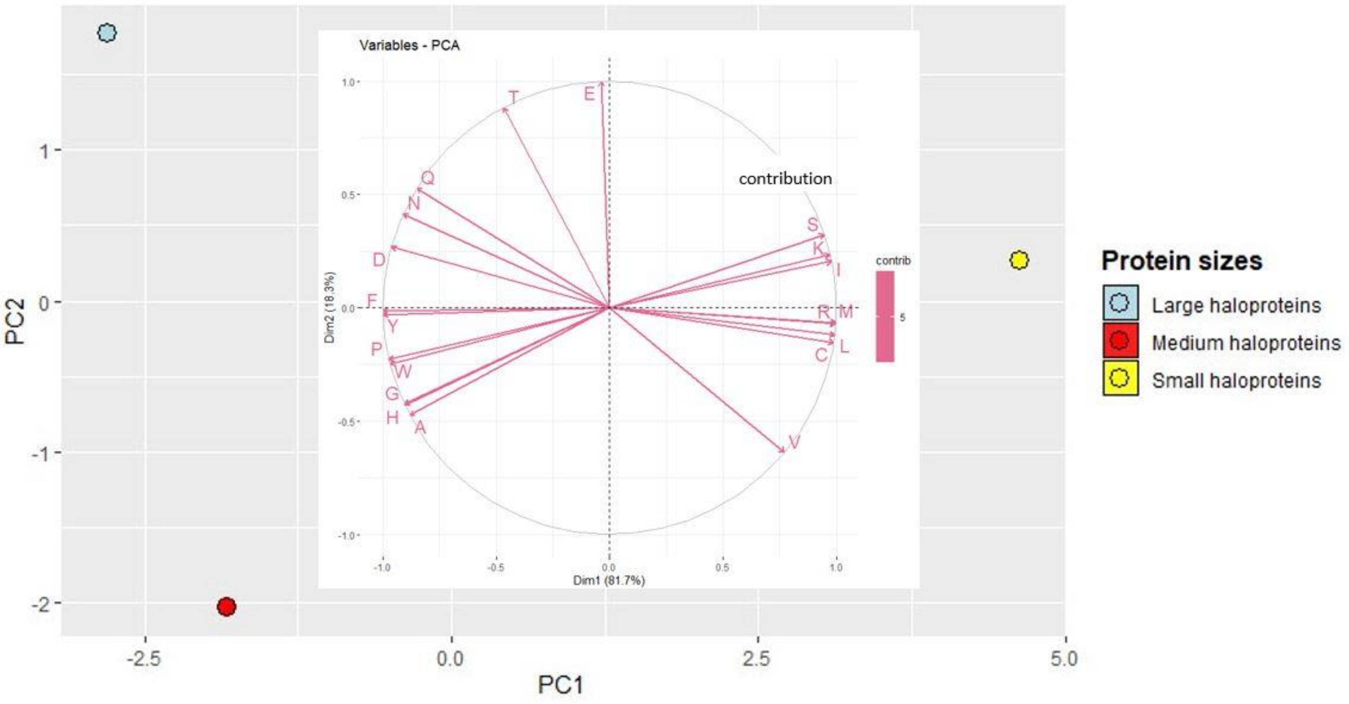


Figure 4.6. PCA analysis of amino acid profiles from proteins of various sizes. Small: 0-70 aa, medium: 71-500 aa, large: >501aa.

Εικόνα 4.6. Ανάλυση κυρίων συνιστωσών των προφίλ αμινοξέων από πρωτεΐνες διαφόρων μεγεθών. Μικρές: 0-70 αμινοξέα, μεσαίες: 71-500 αμινοξέα, μεγάλες: >501 αμινοξέα.

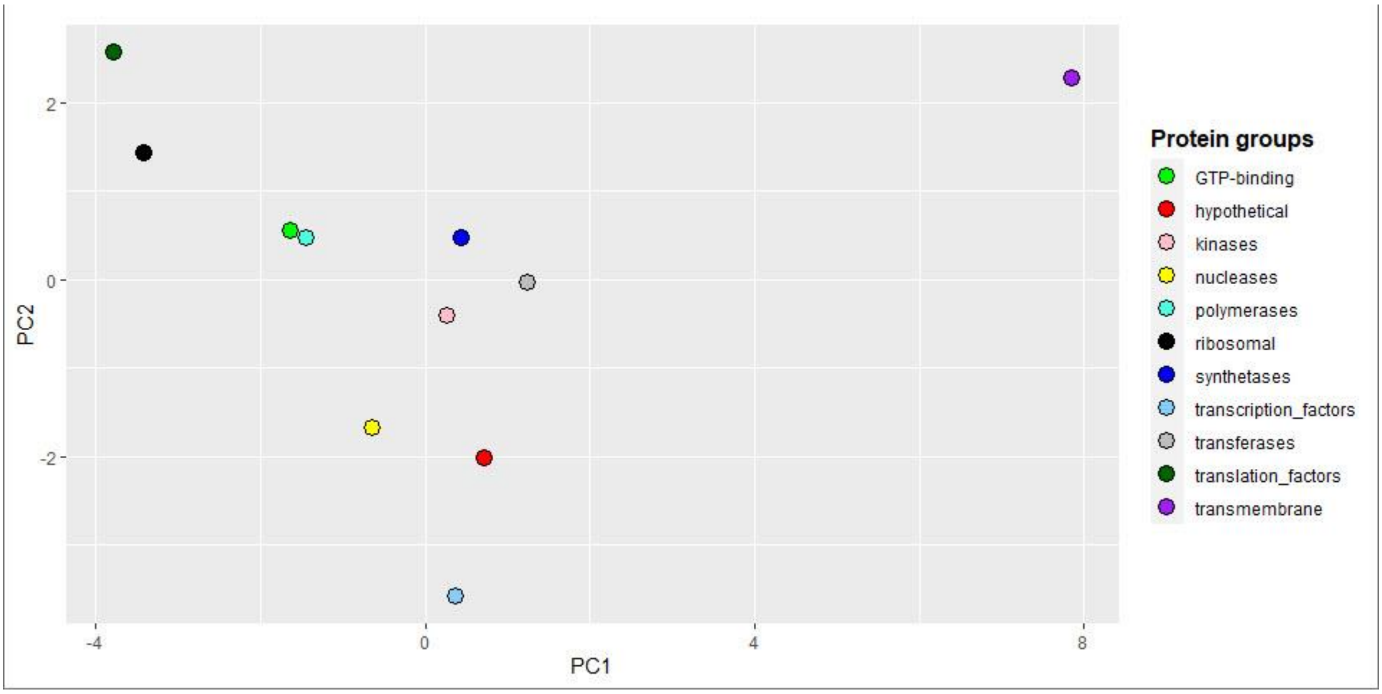


Figure 4.7. Distribution of amino acid profiles of various halophilic protein families. The graph includes all halophilic archaeal species for comparison.

Εικόνα 4.7. Κατανομή αμινοξικών προφίλ για διάφορες οικογένειες πρωτεϊνών. Το γράφημα περιλαμβάνει όλα τα αλόφιλα είδη Αρχαίων για λόγους σύγκρισης.

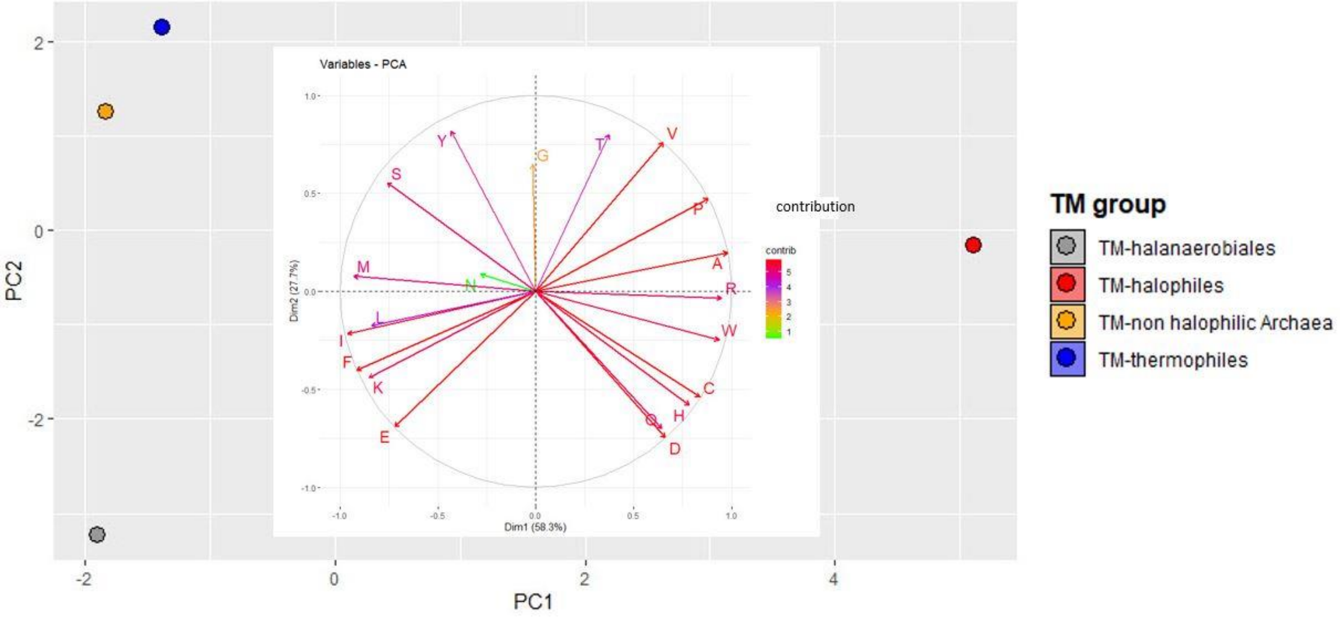


Figure 4.8. PCA analysis of amino acid profiles of all transmembrane proteins from Halanaerobiales, halophilic Archaea, thermophiles, and mesophilic species. TM: transmembrane proteins.

Εικόνα 4.8. Ανάλυση κυρίων συνιστωσών αμινοξικών προφίλ όλων των διαμεμβρανικών πρωτεϊνών από Halanaerobiales, αλόφιλα Αρχαία, θερμόφιλα και μεσόφιλα είδη. ΤΜ: διαμεμβρανική πρωτεΐνη.

4.4 Discussion

Both hierarchical clustering and PCA analyses group all Archaea from class Halobacteria in a single cluster (Figs 4.2, 4.3), with the exception of *Haloquadratum walsby* and *Haloarcula salaria*. This suggests that Halobacteria are using a very distinct amino acid profile as an adaptation strategy, referred to as "salt-in". Part of the "salt-in" strategy is the influx of KCl ions inside the cytoplasm of the adapted cells [112]. Our analysis confirms that the amino acid profile for the majority of proteins within Halobacteria is different than the profile of mesophilic proteins and very specific, a clear sign of adaptation and evolutionary pressure. Halophilic Archaea from the class Methanomicrobia, but also recently discovered species from Nanohaloarchaea, have a different amino acid profile from Halobacteria, suggesting a different adaptation strategy (Fig. 4.4). It has recently been reported that Nanohaloarchaeota possibly use several taxa from Halobacteria as hosts, so this could partially explain the absence of salinity-adapted proteins in Nanohaloarchaeota [126]. *Candidatus nanosalinarum* appears to be in between a fully adapted amino acid profile and a non-adapted profile (Fig. 4.4).

The difference of the amino acid composition between Nanohaloarchaea and Halobacteria add to the recent discovery that these two taxonomic groups derive from two distinct methanogen Class II lineages [127]. The halophilic order Halanaerobiales is clustered away from Halobacteria in Figure 4.3, suggesting also a different amino acid profile and a different adaptation strategy, described as “salt-out”. In this strategy, salt is not entering the cells and compatible solutes are accumulated in the cytoplasm for osmoregulation, providing stability to proteins [15]. Our findings suggest that halophilic bacteria from Halanaerobiales do not possess adapted proteins to cope with high salinities and use the salt-out method. It has been suggested however, that *Haloanaerobium praevalens* and *Halobacteroides halobius* as well as several other species from Halanaerobiales do not produce compatible solutes and accumulate sodium and potassium ions in their cytoplasm [128-130]. There could be previously unknown and undetected molecular mechanisms for these species that enable their proteins to function in high intracellular salinity, since our data clearly place both species away from the adapted amino acid profile of Halobacteria (Fig. A4). More research is needed to gain a better understanding of compatible solutes and KCl accumulation as osmoregulation strategies.

Thermophiles can be distinguished from Halobacteria as they form a separate and considerably larger cluster (Fig. 4.3). This observation suggests that adaptation in high temperatures does not require a strict amino acid profile and can be achieved possibly with other molecular strategies, some of them already pinpointed [131]. Also, it can be observed that several thermophilic sub-groups are clustering together, such as the five *Methanocaldococcus* species, in the upper left part of the cluster and *Pyrococcus* species in the center of the thermophilic cluster (Fig. 4.3). There could be a process where thermophilic adaptations are acquired by closely related species through horizontal gene transfer or this clustering is attributed merely to close phylogenetic relationships between these taxa. On another note, the closest species to the cluster of Halobacteria is *Methanoculleus marisnigri*. The species was isolated from anoxic sediments of the Black Sea, it is known as a thermophilic methanogen, and it is categorized as slightly halophilic [123, 124]. Our data, however, show that it has the same adaptation mechanism and amino acid profile as Halobacteria. Interestingly, *Methanoculleus marisnigri* is the only methanogen from the class Methanomicrobia that shares the amino acid profile of Halobacteria while the rest of Methanomicrobia are clustering close to Halanaerobiales. There are three more thermophilic species close to the cluster of Halobacteria: *Methermicoccus shengliensis* (Methanomicrobia), *Hyperthermus butylicus*, and *Aeropyrum pernix* (both Thermoprotei). There may be more methanogens and other taxa that share this adaptation but additional genomic and experimental data are needed to clarify the picture and pinpoint the thresholds of halophilic adaptation.

Eigenvectors (Fig. 4.5) from PCA analysis in Figure 4.3 confirm previous studies regarding which amino acids are over and under-represented in Halobacteria and drive salinity adaptations. Additionally, the bar chart in Figure 4.9 is in accordance with previous research [59]. It is also worth mentioning that amino acids that drive halophilic adaptation in Halobacteria are also common in halophilic Bacteria situated in the bottom right corner of Figure 4.3. Several taxa from Bacteroidetes, Rhodothermaeota, and Proteobacteria were observed to have similar amino acid profiles with the extremely halophilic class of Halobacteria, suggesting either horizontal gene transfer or convergent evolution. It appears that the amino acids in the lower right quadrant of Figure 4.5 are preferred more for the bacterial species. Additional research is needed in order to compare amino acid profiles of the archaeal class against halophilic Proteobacteria, Rhodothermaeota, and Bacteroidetes. It has already been proposed that horizontal gene transfer has occurred between some of the taxonomic groups in question [19].

GC content was investigated as a factor affecting our PCA analysis. In both datasets containing only halophilic Archaea and Halanaerobiales (Fig. A2), no significant differences were found. Although halophilic Archaea have a slightly increased GC content in their genomes compared with Halanaerobiales, the results of PCA clustering are not affected. Several PCA analyses were conducted to determine how different groups of proteins are subject to evolutionary pressure and if there are significant differences between these proteins in terms of amino acid profile. In Figure 4.6, different protein sizes are not subject to the same evolutionary pressures. Smaller proteins do not enter the cluster of Halobacteria. We suggest that small molecules can fold and function correctly, requiring less compositional tunning. However, the amino acid profile of small proteins is still closer to Halobacteria than to non-adapted species. Medium-sized proteins correspond to the majority of proteins within Halobacteria. Their profile is situated in the center of the cluster. In total, 127,680 hypothetical proteins were of medium size in our analysis. Large proteins also have a different amino acid profile. It can be speculated that different residues and polymorphisms are preferred so that large proteins can fold properly without compromising thermodynamic efficiency.

A similar pattern can be observed in Figure 4.7, where 11 functional protein groups from halophilic Archaea were incorporated in PCA. There are differences in the amino acid profiles of these proteins. Ribosomal and translation factor proteins have amino acid profiles closer to halophilic Bacteria and are situated below the cluster of Halobacteria. Other groups like polymerases, nucleases, kinases, GTP-binding, and transcription factors are closer to the cluster but still show different profiles. Synthetases, transferases, and hypothetical proteins are in the core of Halobacteria, highlighting again the importance of isolation and characterization of several hypothetical proteins that are part of halophilic mechanisms of adaptation. Transmembrane proteins in halophilic Archaea are significantly different in terms of amino acid profile than the rest of the protein groups (Fig. 4.7). This is mostly due to composition bias and the necessary transmembrane regions of these molecules. However, the analysis in Figure 4.8 is demonstrating that halophilic transmembrane proteins are still subject to adaptations as they have a different profile in comparison with proteins from other taxa like Halanaerobiales, thermophiles, and mesophiles. It can be suggested that this scenario applies in other protein groups with composition bias and conserved regions, but also in extracellular proteins [132]. It has been shown in halophilic Actinobacteria that membrane proteins play an important role in salinity adaptation. They are responsible for the cell's primary reaction to hyperosmotic stress and trigger metabolic and possibly cell cycle processes [133]. The membrane proteomic machinery of Halobacteria could reveal more traits of haloadaptation in halophilic Archaea.

Our results shed light into the protein adaptation mechanisms used by extremophiles. It is worth mentioning that evolution creates a variety of solutions for every environment, but also specifically in every protein type of halophilic species. Several questions arise from the analysis. For example, why different molecules with the same adaptation mechanisms prefer slightly different residues? What molecular and environmental factors cause differences in amino acid profiles between Halobacteria and other adapted halophilic Bacteria from Bacteroidetes, Rhodothermaeota, and Proteobacteria? Are there specialized proteins for adapted halophiles that possibly assist compositionally biased proteins to fold, transport, and function correctly under osmotic pressure? Why some species from Halanaerobiales do not possess adapted proteins, yet KCl influx has been observed? Is there a common, baseline molecular signature among extremophilic groups? In Halobacteria many protein groups like transferases will probably not align correctly with mesophilic transferases and novel protein identification and annotation could be challenging if the amino acid profile is altered significantly and there is no reference point. Laboratory procedures may be necessary for proper functional characterization of certain proteins.

Looking closer at the cluster of Halanaerobiales in Figure 4.3, it is clear that Halanaerobiales, Methanomicrobia, Nanohaloarchaea, the psychrophilic *Methanococcoides burtonii*, two acidophilic species, one alkaliphilic, and several thermophiles share a rather similar, yet not so coherent amino acid profile as that of Halobacteria. This may be a sign that these groups of extremophiles share a range of adaptation mechanisms which manifests as a loose clustering in Figure 4.3. Methanogenic Archaea, being some of the oldest life forms on the planet and also the group from where halophilic Archaea originated from [134], could be a cradle group from where several extremophilic species began to differentiate and specialize in other extreme environments [135-137], with occasional lateral gene transfer contributions [67]. Detailed investigations on this issue are limited and more research is needed in order to shed light in the mechanics of extremophilicity.

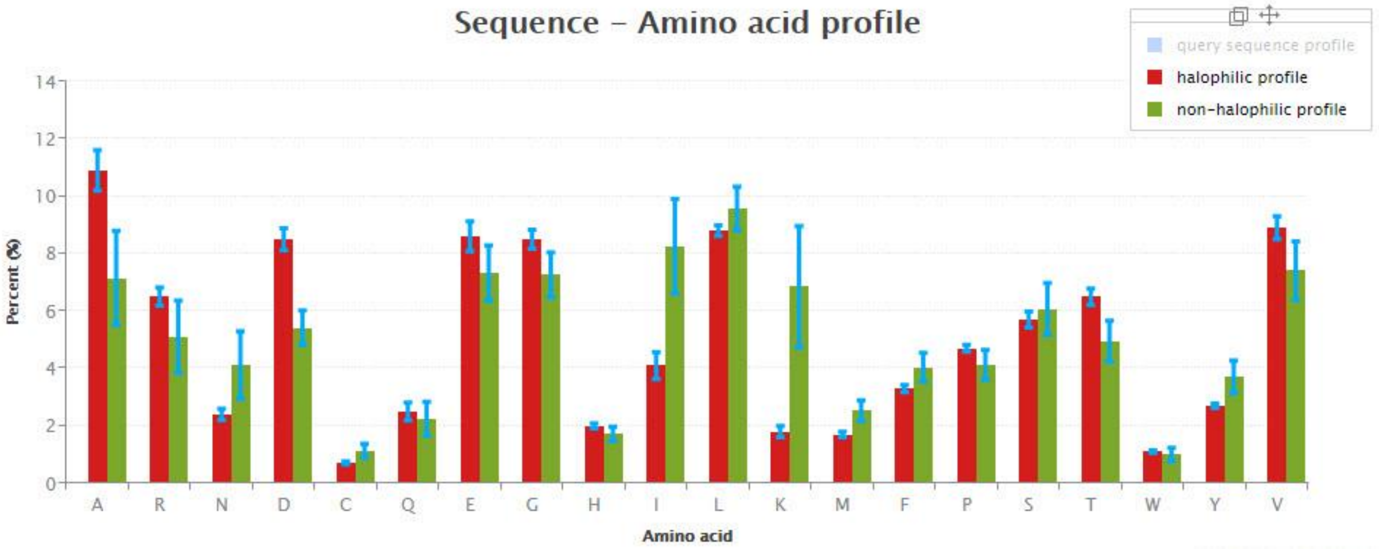


Figure 4.9. Halophilic vs non-halophilic amino acid profiles. Halophilic data come from 123 species of Halobacteria and non-halophilic data from 38 species from our dataset. Error bars: Standard deviations.

Εικόνα 4.9. Αλόφιλα έναντι μη-αλόφιλα προφίλ αμινοξέων. Τα δεδομένα αλόφιλων προέρχονται από 123 είδη Halobacteria και τα μη-αλόφιλα από 38 είδη μη-αλόφιλων καταγεγραμμένων οργανισμών. Γραμμές σφάλματος: Τυπικές αποκλίσεις.

Chapter 5: HaloPredictor, a tool for detecting halophilic adaptations

5.1 Introduction

Protein sequence adaptations reflect changes in genomes during evolution, in order for living organisms to adapt to certain environmental conditions [138-140]. The diversity of environments on Earth have driven species to adapt in conditions like extreme cold, acidic, hot, extreme pressure, but also increased salinity [116, 141, 142]. Protein adaptations have been detected sporadically in archaeal halophiles but also in other extremophilic species [59, 99, 113]. The observed trend is that changes in the amino acid sequence of proteins improve thermodynamic efficiency and help maintain protein fold with changes in residues that occupy important places in the protein's three-dimensional structure. For the case of halophilic proteins, such places, so far, are the protein surface and interior [111].

The significance of extremophile protein research has two parts. Firstly, extremophiles are among the most ancient life forms on Earth and occupy a diverse range of extreme environments with respect to salinity, acidity, temperature among others, and even polluted places due to anthropogenic activities. Therefore, they can provide clues about life and conditions on Earth millions of years ago, closer to the emergence of living organisms on the planet [100, 105, 134, 143]. Additionally, they are of ecological importance since their presence indicates environmental phenomena and changes such as desertification [144]. Secondly, extremophiles have numerous biotechnological and industrial uses, from thermostable proteins to stable membranes, compatible solutes, biomolecule production, and bioremediation [103, 144, 145]. Also, there is a tremendous unexplored diversity of extremophilic protein and enzyme families which could further enlighten basic research questions and the industry.

It is therefore important to increase the focus on extremophilic proteins. As mentioned above, protein adaptations have been observed in a plethora of extremophilic organisms, however to our knowledge, no distinct tools for detecting protein adaptations in amino acid sequences exist so far. As reported, adaptive changes in amino acid sequences include the over- or under-representation of certain residues, or changes/modifications in a protein's secondary structure.

In this chapter, we present a tool designed to detect salinity adaptations based on protein sequences from the Archaeal class of Halobacteria. The observed adaptation is a changed amino acid profile in halophilic proteins compared with that of non-halophilic relatives. The tool called

“HaloPredictor” can predict if a protein sequence is fully adapted to salinity, using simple statistical methods. HaloPredictor accepts a protein sequence as input and provides an indication about the protein's haloadaptation level. To our knowledge, HaloPredictor is the first user interface platform to help researchers determine the haloadaptation strategy of a new halophilic species. It is also the first tool for predicting any kind of extremophilic adaptations in protein sequences, which opens new possibilities and research opportunities on this exciting topic.

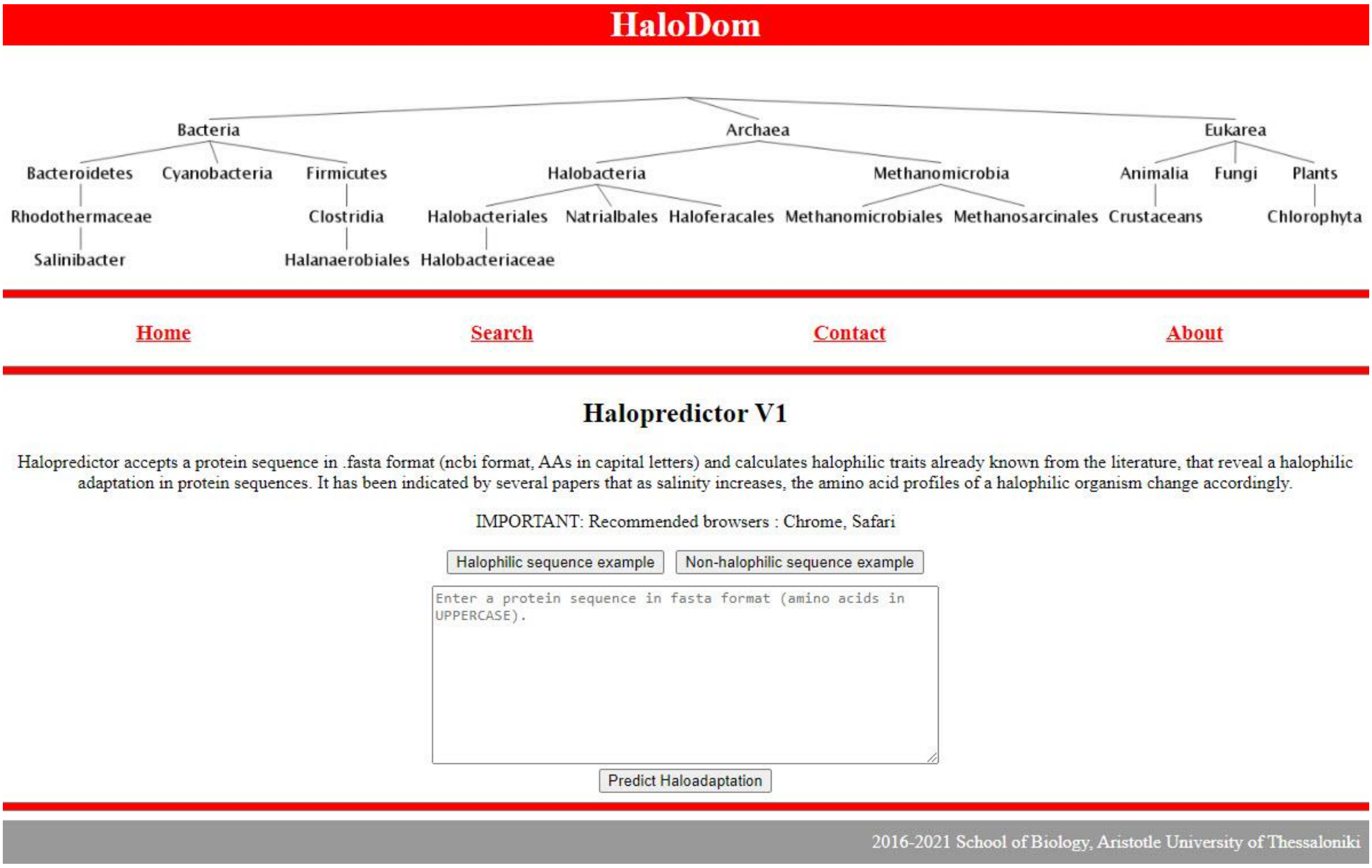
5.2 Materials and Methods

HaloPredictor development

HaloPredictor was designed with standard tools for website development such as PHP, HTML, javascript, and apache server for local and online testing. One protein sequence at a time is inserted to an online form (Fig. 5.1). From the sequence, basic amino acid statistics are calculated with PHP. Code from Google gauges [146] is used to highlight halophilic traits while the amino acid profile chart is written in java and is provided by ZingChart [147]. LDA analysis is performed with an R script called “LDA_with_unknown.r” which refers to the unknown sequence input coming from users (R script A5).

The script is executed to an external server and sends the results of LDA analysis back to the webpage. The results of LDA provide a data graph and a prediction, where the query protein is classified as halophilic or non-halophilic and its position is visualized among all the data. LDA analysis uses a csv file with halophilic and non-halophilic amino acid compositions in order to calculate a classification for the query protein. All results and graphs of HaloPredictor can be downloaded in png format.

The local version of HaloPredictor is executed by a Perl script called “halopredictor_local_V1.pl” (Perl script A7). The script is executed from the command line of Windows or Unix operating systems and utilizes the following three R scripts: “halopredictor_local_LDA.r” (R script A6) for performing the LDA analysis required for the predictions, ”stats.r” (R script A7) for calculating basic statistics about the input sequences and ”halopredictor_local_visualize.r” (R script A8) for visualizing the results. The input sequences are inserted in the form of a fasta file, containing the desired protein sequences. The file is named “input.fasta” in order for the software to recognize it as input. The results of the analysis are saved in a folder named “results” and contain a density plot (Fig. 5.2), a violin plot (Fig. 5.3) and two csv files, “halo_results.csv” and “statistics.csv”. In the latter file, the total number of inserted sequences is shown as well as the percentage of halophilic sequences found. In “halo_results.csv” all inserted sequences are presented along with their LDA scores and their classifications as either “Halophilic” or “Non-halophilic”.



Halopredictor V1

Halopredictor accepts a protein sequence in .fasta format (ncbi format, AAs in capital letters) and calculates halophilic traits already known from the literature, that reveal a halophilic adaptation in protein sequences. It has been indicated by several papers that as salinity increases, the amino acid profiles of a halophilic organism change accordingly.

IMPORTANT: Recommended browsers : Chrome, Safari

Halophilic sequence example
Non-halophilic sequence example
Enter a protein sequence in fasta format (amino acids in UPPERCASE).
Predict Haloadaptation

2016-2021 School of Biology, Aristotle University of Thessaloniki

Figure 5.1. Input form of the online version of HaloPredictor. The form accepts one protein sequence at a time, in fasta format.

Εικόνα 5.1. Φόρμα εισόδου δεδομένων της διαδικτυακής έκδοσης του HaloPredictor. Η φόρμα δέχεται μια πρωτεϊνική ακολουθία τη φορά, σε μορφοποίηση fasta.

5.3 Results

A new web-tool for detecting salinity adaptations in protein sequences

HaloPredictor is a web tool that accepts a protein sequence in fasta format and provides an indication about the protein's adaptation to salinity, based on the amino acid composition. The online version, given a query protein, calculates percentages for acidic, basic, and all residues. It also determines the acidic/basic amino acid ratio. Results are presented in gauges showing if a certain protein trait is closer to a “halophilic” (red) or to a “normal” (green) percentage, on average. In total, there are six haloadaptation traits for users to consider, the most reliable being the acidic/basic ratio. There is also a graph (Fig. 5.4) showing percentage differences (± std) in all amino acids between halophilic and non-halophilic profiles along with the current query protein sequence. A linear discriminant analysis (LDA) is also performed and viewed, in order to statistically classify the query sequence as “Halophilic” or “Non-halophilic”. All the results from the online version are displayed in Figure 5.4.

The results format of the local version of HaloPerdictor differs from the online version. The input of the local version accepts multiple fasta sequences and so the results are in the form of a list (Table 5.1). Also, descriptive graphs are produced. A density plot of LDA scores (Fig. 5.2) from the input data and a violin plot (Fig. 5.3). Both graphs describe the distribution of LDA scores of all inserted sequences.

Table 5.1. The first 10 results of 1972 protein sequences from *Haloarcula salaria* (RefSeq records from NCBI), extracted from the “halo_results.csv” file containing HaloPredictor predictions.

Πίνακας 5.1. Τα 10 πρώτα αποτελέσματα από 1972 πρωτεϊνικές ακολουθίες του *Haloarcula salaria* (καταχωρήσεις RefSeq του NCBI), προερχόμενα από το αρχείο «halo_results.csv» το οποίο περιέχει τις προβλέψεις του HaloPredictor.

Sequence	Group	Prediction	LDA
WP_188853512.1 hydroxyphenylacetyl-CoA thioesterase PaaI [Haloarcula salaria]	1	halophilic	-34.4936
WP_188853511.1 phenylacetate-CoA oxygenase subunit PaaC [Haloarcula salaria]	1	halophilic	-5.86227
WP_188851879.1 molecular chaperone DnaJ [Haloarcula salaria]	2	non-halophilic	5.672601
WP_050007479.1 hydroxyphenylacetyl-CoA thioesterase PaaI [Haloarcula salaria]	1	halophilic	-28.2375
WP_127015113.1 HIT domain-containing protein [Haloarcula salaria]	2	non-halophilic	18.48329
WP_206508226.1 hypothetical protein [Haloarcula salaria]	1	halophilic	-2.81539
WP_206508224.1 hypothetical protein [Haloarcula salaria]	1	halophilic	-21.1062
WP_206508222.1 S8 family serine peptidase [Haloarcula salaria]	1	halophilic	-23.2354
WP_206508221.1 MFS transporter [Haloarcula salaria]	1	halophilic	-7.79241
WP_206508218.1 SDR family oxidoreductase [Haloarcula salaria]	1	halophilic	-1.57784

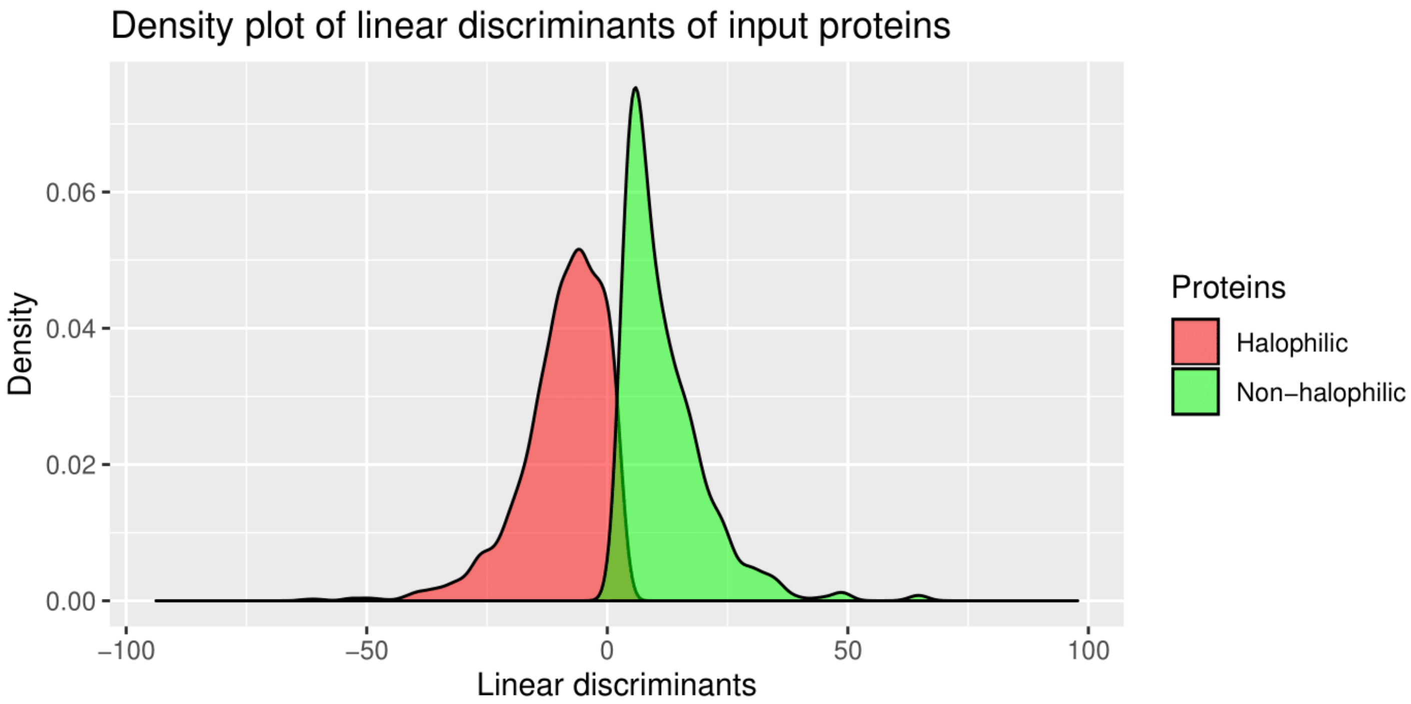


Figure 5.2. Density plot of linear discriminants of 1972 input protein sequences from *Haloarcula salaria* (RefSeq records from NCBI). The graph shows the distribution of halophilic and non-halophilic LDA scores.

Εικόνα 5.2. Διάγραμμα πυκνοτήτων των γραμμικών διαχωριστών των 1972 πρωτεϊνικών ακολουθιών από το *Haloarcula salaria* (καταχωρήσεις RefSeq του NCBI). Το γράφημα δείχνει την κατανομή των αλόφιλων και μη-αλόφιλων τιμών της ανάλυσης γραμμικής διάκρισης.

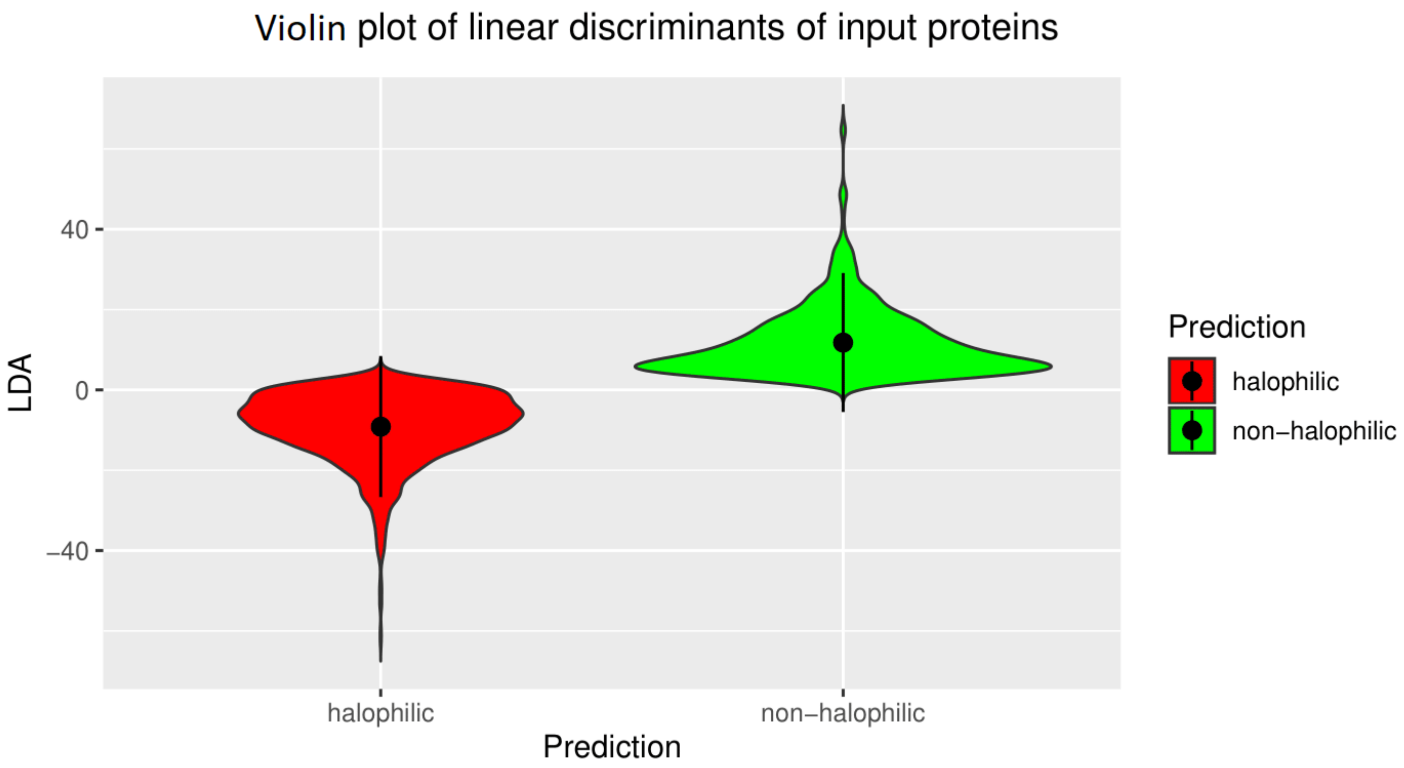


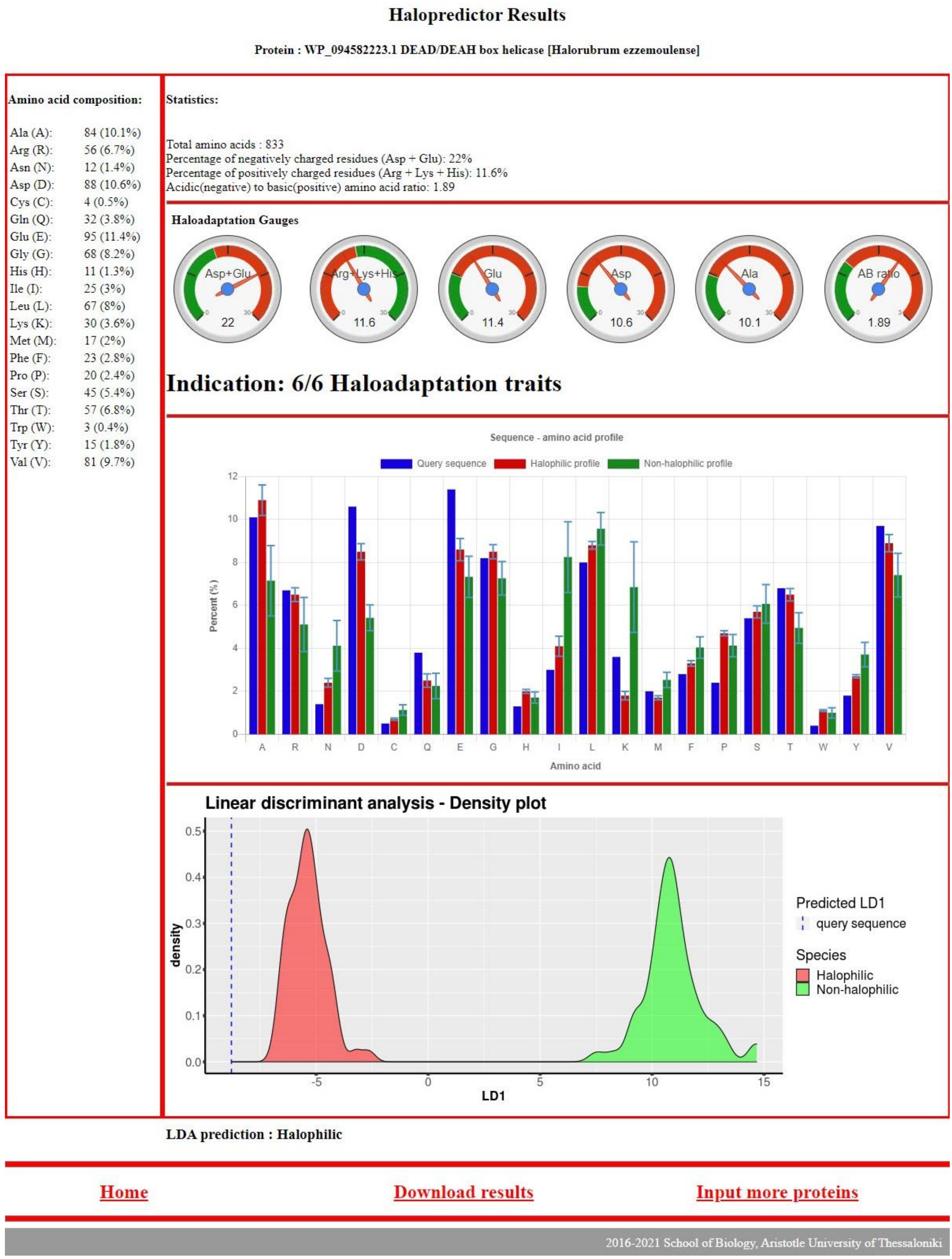
Figure 5.3. Violin plot of linear discriminants of 1972 input protein sequences from *Haloarcula salaria* (RefSeq records from NCBI). The graph shows the distribution of halophilic and non-halophilic LDA scores.

Εικόνα 5.3. Διάγραμμα τύπου βιολιού των γραμμικών διαχωριστών από 1972 πρωτεϊνικές ακολουθίες του *Haloarcula salaria* (καταχωρήσεις RefSeq του NCBI). Το γράφημα δείχνει την κατανομή αλόφιλων και μη-αλόφιλων τιμών της ανάλυσης γραμμικής διάκρισης.

5.4 Discussion

Currently, the most distinctive trait for a truly adapted halophile is the amino acid profile of its proteins. Most species that use the salt-in strategy do possess modified proteins in order to survive hypersalinity and osmotic pressure [59, 112, 148]. However, to our knowledge there are no computational tools to detect such adaptations, and when new halophilic species are isolated from metagenomic samples for instance, the only method for adaptation detection is to analyze the amino acid profile of their proteomes. HaloPredictor accepts a protein sequence in fasta format and provides information and predictions about the protein’s haloadaptation level. Halophilic traits, amino acid profile comparison, and LDA analysis (Fig. 5.4) can provide significant information about salinity adaptations. Currently there are 122 proteomes of adapted Halobacteria and 60 proteomes of non-halophilic control species in our LDA classification dataset. On this scale, the HaloPredictor analysis provides solid evidence for the universality of certain halophilic adaptations. Efficiency estimation needs to be addressed and more proteomic data need to be included in the LDA analysis for better prediction accuracy. Taking into account the position of acidic residues in a protein (surface or interior) will arguably improve sensitivity [111]. Additionally, more adaptation mechanisms, besides sequence composition, need to be explored. Salt-tolerant proteins also exist in non-halophiles [111], a very interesting fact that raises new questions regarding the role and evolutionary history of halophilic proteins. It is highly possible that lateral gene transfer of salt-adapted proteins has occurred between certain microbial taxa [19, 112] and HaloPredictor can assist in comparative genomic analyses trying to detect gene transfers between species.

Moreover, HaloPredictor can help researchers quickly determine the haloadaptation strategy (salt-in) of a new species, discover new molecular mechanisms conducive to brine life, and characterize paleoenvironments. Despite certain limitations of HaloPredictor it is worth noting that adaptation clues from previous research [59, 112, 113] have been confirmed. The presented tool could also provide a basis for uncovering novel protein and evolutionary traits in halophilic, extremophilic, and mesophilic microbial species. The software can be further developed both online and in its local version, in order to facilitate larger datasets and additional functions, thus contributing to innovative research on extremophiles.



Halopredictor Results

Protein : WP_094582223.1 DEAD/DEAH box helicase [Halorubrum ezzemoulense]

Amino acid composition:

Ala (A):	84 (10.1%)
Arg (R):	56 (6.7%)
Asn (N):	12 (1.4%)
Asp (D):	88 (10.6%)
Cys (C):	4 (0.5%)
Gln (Q):	32 (3.8%)
Glu (E):	95 (11.4%)
Gly (G):	68 (8.2%)
His (H):	11 (1.3%)
Ile (I):	25 (3%)
Leu (L):	67 (8%)
Lys (K):	30 (3.6%)
Met (M):	17 (2%)
Phe (F):	23 (2.8%)
Pro (P):	20 (2.4%)
Ser (S):	45 (5.4%)
Thr (T):	57 (6.8%)
Trp (W):	3 (0.4%)
Tyr (Y):	15 (1.8%)
Val (V):	81 (9.7%)

Statistics:

Total amino acids : 833
Percentage of negatively charged residues (Asp + Glu): 22%
Percentage of positively charged residues (Arg + Lys + His): 11.6%
Acidic(negative) to basic(positive) amino acid ratio: 1.89

Haloadaptation Gauges

Indication: 6/6 Haloadaptation traits

LDA prediction : Halophilic

Home
Download results
Input more proteins
2016-2021 School of Biology, Aristotle University of Thessaloniki

Figure 5.4. HaloPredictor results for a query helicase from the halophilic archaeon *Halorubrum ezzemoulense*, a clear case of a haloadapted protein. The amino acid composition of the query protein is shown on the left tab. Right tab displays basic amino acid statistics, haloadaptation traits in gauges, graphical representation of the query amino acid profile for comparison purposes, and the results of LDA analysis for classification of the query protein as "halophilic" or "non-halophilic".

Εικόνα 5.4. Αποτελέσματα του HaloPredictor για μία ελικάση από το αλόφιλο αρχαιοβακτήριο *Halorubrum ezzemoulense*, μια ξεκάθαρη περίπτωση προσαρμοσμένης στην αλατότητα πρωτεΐνης. Η σύνθεση των αμινοξέων της πρωτεΐνης φαίνεται στην αριστερή καρτέλα. Στην δεξιά καρτέλα παρουσιάζονται στατιστικά στοιχεία των αμινοξέων της ακολουθίας, χαρακτηριστικά της προσαρμογής στην αλατότητα (κυκλικοί μετρητές), γραφική αναπαράσταση του προφίλ των αμινοξέων και των ποσοστών παρουσίας τους αλλά και τα αποτελέσματα της ανάλυσης γραμμικής διάκρισης για την κατηγοριοποίηση της πρωτεΐνης σε αλόφιλη ή μη-αλόφιλη.

Chapter 6: Pangenome analysis of class Halobacteria

6.1 Introduction

The accumulation of public sequence data in the last few years, especially whole genome data, has enabled researchers to conduct large scale genomic analyses [149]. Pangenomics have introduced one more systemic approach for genetic research, by analyzing or predicting all genes in a taxonomic group [150-153] using several bioinformatics tools and pipelines [154-162]. Separating genes in categories as core, accessory, and cloud can provide useful information in terms of population dynamics, taxonomy, the evolution of species, and more. The pangenome literature in the last five years has increased considerably throughout Archaea, Bacteria, and Eukarya [163-167], but also in viruses (including SARS-CoV-2) [168-170]. Most often, the scale of pangenome analyses refers to the species or genus levels [151, 163, 164].

In this chapter, we use the pangenome approach for a wider taxonomic group, the archaeal class of Halobacteria. To our knowledge, a systemic genetic approach for this taxonomic group is not yet available. At the time of writing there were 76 complete genome sequences of Halobacteria available in NCBI (Table 6.1). Halobacteria, among other halophiles across the tree of life, have developed several cellular functions to cope with and thrive in high salinity settings [171]. Although it has been established that many of their proteins carry a specific amino acid profile that allows proper protein folding and function under osmotic stress [59, 111, 112], little is known about the intracellular functions of their cells and the complete genetic architecture of life under these conditions. Here, we aimed to explore the genetic potential of Halobacteria and investigate for the presence of novel genes and protein families. We also constructed phylogenetic profiles of all predicted protein clusters and made comparisons with the reconstructed phylogeny (see Chapter 3) of 124 Halobacteria and 33 outgroups from 242 core archaeal markers [81, 82].

Our pangenome analysis revealed an increasing rarefaction curve after all 76 genomes were accounted, indicative of an open pangenome and the existence of novel genes as new species of Halobacteria are isolated and sequenced. We identified 814 genes as “core”, 13,752 as “accessory”, and 9010 as “cloud” for a total of 23,576 identified gene clusters. The majority of 217,783 protein sequences predicted from the 76 genomes of Halobacteria were annotated with the use of EggNOG orthology database and assigned to a functional category. With the use of the presence-absence pangenome matrices, phylogenetic profiles were created revealing new genetic traits. Finally, with the use of HaloPredictor, all proteins of Halobacteria and several proteomes of model organisms were analyzed and novel proteomic patterns were found. Our analysis provides valuable information on the genetic repertoire of halophilic Archaea and highlights the need for discovering and validating new protein families relevant to salinity adaptation.

6.2 Materials and Methods

Data retrieval and pangenome construction

The 76 genome sequences (Table 6.1) of Halobacteria were obtained through a combination of NCBI's services and Micropan v2 [161], a tool written in R and executed locally. Micropan's pipeline offers a complete solution for pangenome analysis through a series of steps including downloading genome data, finding coding regions with Prodigal [172], creating gene alignments using BLAST, gene clustering, pan-matrix and dendrogram (both weighted or unweighted) construction.

The genome data were downloaded from NCBI's online genome repository (https://www.ncbi.nlm.nih.gov/genome/browse#!/prokaryotes/). The keyword "Halobacteria" was used as a search term and the results were further filtered. Only species with assembly level as "complete" were chosen for the analysis. The data were downloaded in csv format and saved as a txt file, in a specific folder in order to be recognized by Micropan v2 which downloads the corresponding genome sequences.

For the identification of coding genes in the genome sequences, Micropan v2 uses Prodigal [172] which is a popular software for prokaryotic gene finding. The pipeline prompts users to run Prodigal on a randomly generated DNA sequence in order to decide on a reliable score cutoff threshold for the predicted gene sequences.

The predicted genes are stored as fasta files and the pipeline proceeds with aligning all the sequences using BLAST. A BLAST all-against-all step will compute distances between all protein pairs. The all-against-all procedure was the most demanding in terms of CPU power and the corresponding part of the code was executed in the HPC AUTh cluster "Aristotelis".

Next, the distances between all protein sequences are computed by Micropan v2. Distance values range between 0 and 1. A value closer to 1 denotes that the protein pair is very distant, while the value of 0 means that the protein pair is identical. With the computed distances, a distance threshold can be chosen for clustering the protein sequences. We used a distance threshold of 0.75. The calculation of pairwise distances between all protein sequences was also executed in the HPC AUTh cluster.

The hierarchical clustering of our protein sequences was executed with the “complete linkage” method and resulted in 23,576 gene clusters. To proceed with the pangenome analysis, a pan-matrix is created, which is a matrix with all 76 complete genomes as rows and all 23,576 gene clusters as columns. The contents of the matrix are the observed frequencies of every predicted gene cluster in every genome.

With the use of the pan-matrix, several characteristics of the pangenome can be calculated including the openness of the pangenome according to Heaps’ law and a power law, the pangenome rarefaction curve, the distribution of gene clusters, and the detection probabilities of gene families.

With the use of the pan-matrix, the Manhattan distances between Halobacteria species were calculated and both simple and weighted dendrograms were created. Micropan v2 enables the construction of a weighted dendrogram, placing the weights on either “shell” genes (which include core and shell genes) or “cloud” genes, which places weights on the cloud genes of the pangenome. In this study, two weighted dendrograms were created, with weights on “shell” genes. One dendrogram was created with complete linkage clustering and one with average linkage clustering. The visualization of the dendrograms was modified and is different than the default dendrogram graphs of Micropan v2. An R script was created (“circular_weighted_tree.r”) that changed the format of the dendrogram in a circular tree, also adding coloring in the branches and the topology (R script A9).

Table 6.1. Genome sequences used for creating the pangenome of Halobacteria. Every genome has a unique genome ID (GID.tag) during the analysis. The assembly level of all genomes used is "complete".

No	Species	Strain	GID.tag	GenBank ID
1	*Halophilic archaeon* DL31	DL31	GID1	GCA_000224475.1
2	*Halovivax ruber* XH-70	XH-70	GID2	GCA_000328525.1
3	*Halorhabdus utahensis* DSM 12940	DSM 12940	GID3	GCA_000023945.1
4	*Halorhabdus tiamatea* SARL4B	SARL4B	GID4	GCA_000470655.1
5	*Halobiforma lacisalsi* AJ5	AJ5	GID5	GCA_000226975.3
6	*Natronomonas moolapensis* 8.8.11	8.8.11	GID6	GCA_000591055.1
7	*Natronobacterium gregoryi* SP2	SP2	GID7	GCA_000230715.3
8	*Halobacterium* sp. DL1	DL1	GID8	GCA_000230955.3
9	*Natrinema* sp. J7-2	J7-2	GID9	GCA_000281695.1
10	*Halorubrum* sp. BOL3-1	BOL3-1	GID10	GCA_004114375.1
11	*Natrinema pallidum*	BOL6-1	GID11	GCA_005890195.1
12	*Halorubrum ezzemoulense*	Fb21	GID12	GCA_004126515.1
13	*Salinarchaeum* sp. Harcht-Bsk1	Harcht-Bsk1	GID13	GCA_000403645.1
14	*Haloarchaeon* sp. 3A1-DGR	3A1-DGR	GID14	GCA_010092465.1
15	*Halapricum salinum*	CBA1105	GID15	GCA_004799665.1
16	*Halanaeroarchaeum sulfurireducens* HSR2	HSR2	GID16	GCA_001011115.1
17	*Halopenitus persicus*	CBA1233	GID17	GCA_002355635.1
18	*Halohasta litchfieldiae*	tADL	GID18	GCA_002788215.1
19	*Halodesulfurarchaeum formicicum* HSR6	HSR6	GID19	GCA_001886955.1
20	*Haloterrigena daqingensis*	JX313	GID20	GCA_001971705.1
21	*Halorientalis* sp. IM1011	IM1011	GID21	GCA_001989615.1
22	*Haloarcula taiwanensis*	Taiwanensis	GID22	GCA_002844335.1
23	*Halobellus limi*	CGMCC 1.10331	GID23	GCA_004799685.1
24	*Haloarculaceae archaeon* HArcel1	HArcel1	GID24	GCA_003058365.1
25	*Haloplanus* sp. CBA1112	CBA1112	GID25	GCA_003342675.1
26	*Halorussus* sp. ZS-3	ZS-3	GID26	GCA_008831545.1
27	*Haloprofundus* sp. MHR1	MHR1	GID27	GCA_005155585.1
28	*Halalkaliarchaeum desulfuricum*	AArc-Sl	GID28	GCA_002952775.1
29	*Natronolimnobius aegyptiacus*	JW/NM-HA 15	GID29	GCA_002156705.1
30	*Haloplanus aerogenes*	JCM 16430	GID30	GCA_003856835.1
31	*Halomicrobium* sp. LC1Hm	LC1Hm	GID31	GCA_009617995.1
32	*Halorhabdus* sp. CBA1104	CBA1104	GID32	GCA_009690625.1
33	*Salarchaeum* sp. JOR-1	JOR-1	GID33	GCA_007833275.1
34	*Natronolimnobius sulfurireducens* AArc1	AArc1	GID34	GCA_003430825.1
35	*Haloplanus rallus*	MBLA0036	GID35	GCA_009762275.1
36	*Halobacteriaceae archaeon* HD8-45	HD8-45	GID36	GCA_009789175.1
37	*Natronolimnobius sulfurireducens* AArc-Mg	AArc-Mg	GID37	GCA_003430805.1
38	*Halorussus* sp. RC-68	RC-68	GID38	GCA_004087835.1

39	*Haloplanus* sp. CBA1113	CBA1113	GID39	GCA_003342695.1
40	*Halodesulfurarchaeum formicicum* HTSR1	HTSR1	GID40	GCA_001767315.1
41	*Halanaeroarchaeum sulfurireducens* M27-SA2	M27-SA2	GID41	GCA_001305655.1
42	*Haloterrigena jeotgali*	A29	GID42	GCA_004799625.1
43	*Haloferax alexandrinus*	wsp1	GID43	GCA_010692905.1
44	*Halorubrum* sp. PV6	PV6	GID44	GCA_003990725.1
45	*Halomicrobium mukohataei* ZPS1	ZPS1	GID45	GCA_009217585.1
46	*Halogeometricum borinquense* wsp4	wsp4	GID46	GCA_010692885.1
47	*Haloquadratum walsbyi* DSM 16790	HBSQ001	GID47	GCA_000009185.1
48	*Halomicrobium mukohataei* JP60	JP60	GID48	GCA_004803735.1
49	*Haloferax volcanii* DS2	DS2	GID49	GCA_000025685.1
50	*Natronomonas pharaonis* DSM 2160	Gabara	GID50	GCA_000026045.1
51	*Haloquadratum walsbyi* C23	DSM 16854	GID51	GCA_000237865.1
52	*Natrialba magadii* ATCC 43099	ATCC 43099	GID52	GCA_000025625.1
53	*Haloarcula hispanica* ATCC 33960	CGMCC 1.2049	GID53	GCA_000223905.1
54	*Halopiger xanaduensis* SH-6	SH-6	GID54	GCA_000217715.1
55	*Halogeometricum borinquense* DSM 11551	PR 3	GID55	GCA_000172995.2
56	*Halomicrobium mukohataei* DSM 12286	DSM 12286	GID56	GCA_000023965.1
57	*Haloferax mediterranei* ATCC 33500	CGMCC 1.2087	GID57	GCA_000306765.2
58	*Natrinema pellirubrum* DSM 15624	DSM 15624	GID58	GCA_000230735.3
59	*Halostagnicola larsenii* XH-48	XH-48	GID59	GCA_000517625.1
60	*Natrinema versiforme*	BOL5-4	GID60	GCA_005576615.1
61	*Natronorubrum bangense*	JCM 10635	GID61	GCA_004799645.1
62	*Haloferax gibbonsii*	ARA6	GID62	GCA_001190965.1
63	*Haloarcula* sp. CBA1115	CBA1115	GID63	GCA_000827835.1
64	*Halobacterium hubeiense*	JI20-1	GID64	GCA_001488575.1
65	*Halorubrum trapanicum*	CBA1232	GID65	GCA_002355655.1
66	*Salinigranum rubrum*	GX10	GID66	GCA_002906575.1
67	*Halostella pelagica*	DL-M4	GID67	GCA_005954745.1
68	*Haloferax mediterranei* ATCC 33500	ATCC 33500	GID68	GCA_005406325.1
69	*Haloarcula hispanica* N601	N601	GID69	GCA_000504565.2
70	*Haloarcula marismortui* ATCC 43049_1	ATCC 43049	GID70	GCA_005310945.1
71	*Haloarcula marismortui* ATCC 43049_2	ATCC 43049	GID71	GCA_000011085.1
72	*Haloterrigena turkmenica* DSM 5511	DSM 5511	GID72	GCA_000025325.1
73	*Halalkalicoccus jeotgali* B3	B3	GID73	GCA_000196895.1
74	*Halobacterium salinarum* NRC-1	ATCC 700922	GID74	GCA_000006805.1
75	*Halobacterium salinarum* 91-R6	91-R6	GID75	GCA_004799605.1
76	*Halobacterium salinarum* R1	DSM 671	GID76	GCA_000069025.1

Data extraction from R package Micropan v2

Micropan's v2 pipeline is executed step by step in R. In this study, Rstudio [173] was used for the editing of the code. All graphs produced by Micropan v2 were saved in png and tiff formats. In order to handle fasta and csv files, create a catalogue of each gene cluster with its sequences, and create fasta files that contain sequences for "all", "core", "accessory", "cloud", and "multicopy" gene categories, several Perl [174] and BioPerl [122] scripts were used.

Functional annotation of predicted gene clusters

For the annotation of predicted genes the EggNOG database [175] and the eggnog-mapper v2 [176] were used. A series of five fasta files were placed as input in eggnog-mapper v2, each file containing one of the five gene categories: "core", "accessory", "cloud", "multicopy", and "all genes". The results were obtained in tsv format, from which annotation information were extracted with the use of Perl scripts. An excel sheet was created containing the frequencies of every annotation and gene category. A final graph was created with an R script ("category_bar_chart.r", R script A10), using the information from the excel sheets, showing the results of the annotation process.

Phylogenetic profiles created by the pangenome presence-absence matrices

During the pangenome analysis, Micropan v2 creates a presence-absence matrix called "the pan-matrix", which contains all predicted gene clusters and their presence in every genome used in the analysis. Absence is declared with the number zero, while presence with 1. In case there are more copies of the same gene present in the same genome, the number is larger than 1 and is the number of copies for this marker. Two types of matrices were used for phylogenetic profiling. The first is described above and we call it frequency pan-matrix. An example is shown in Table 6.2. The second is the same matrix converted to a binary presence-absence matrix with the use of an R script. We call it binary pan-matrix and an example is shown in Table 6.3.

Both the frequency pan-matrix and the binary pan-matrix were inserted into R, and with the use of an R script, three heatmaps were created. In these heatmaps, columns contain the predicted protein clusters and rows contain the 76 Halobacteria species of the analysis. With the use of hierarchical clustering the heatmaps were clustered with the complete distance method for columns (predicted protein clusters). Additionally, the same weighted tree calculated before with Micropan v2 was applied to all three heatmap rows (species). The data from pan-matrices were not normalized or scaled. In the case of the binary heatmap data are already scaled, while in the case of the frequency heatmaps scaling would hinder visualization of the presence of predicted multi-copy protein clusters and their distribution in the Halobacteria pangenome.

Table 6.2. Example of a frequency pan-matrix. Every number greater than 0 represents the copy number of a specific protein cluster in certain species.

Πίνακας 6.2. Παράδειγμα παν-μήτρας συχνότητας. Κάθε αριθμός μεγαλύτερος από 0 αναπαριστά τον αριθμό αντιγράφων μιας συγκεκριμένης πρωτεϊνικής συστάδας σε συγκεκριμένο είδος.

	Cluster481	Cluster482	Cluster483	Cluster484	Cluster485	Cluster486
Halophilic archaeon **DL31**	5	1	2	2	1	1
Halovivax ruber **XH-70**	0	0	0	0	0	0
Halorhabdus utahensis **DSM 12940**	2	0	2	0	2	2
Halorhabdus tiamatea **SARL4B**	2	1	0	0	1	1
Halobiforma lacisalsi **AJ5**	0	0	0	0	0	0
Natronomonas moolapensis **8.8.11**	0	0	0	0	0	0
Natronobacterium gregoryi **SP2**	0	0	0	1	2	0
Halobacterium **sp. DL1**	3	0	1	0	0	0
Natrinema **sp. J7-2**	0	0	0	0	0	0
Halorubrum **sp. BOL3-1**	1	0	1	2	2	0
Natrinema pallidum	0	0	0	0	0	0
Halorubrum ezzemoulense	1	1	0	3	2	0

Table 6.3. Example of a binary pan-matrix. Number 0 declares absence of a protein cluster in a species, while number 1 declares presence.

Πίνακας 6.3. Παράδειγμα δυαδικής παν-μήτρας. Ο αριθμός 0 δηλώνει απουσία μιας πρωτεϊνικής συστάδας σε ένα είδος, ενώ ο αριθμός 1 δηλώνει την παρουσία της.

	Cluster481	Cluster482	Cluster483	Cluster484	Cluster485	Cluster486
Halophilic archaeon **DL31**	1	1	1	1	1	1
Halovivax ruber **XH-70**	0	0	0	0	0	0
Halorhabdus utahensis **DSM 12940**	1	0	1	0	1	1
Halorhabdus tiamatea **SARL4B**	1	1	0	0	1	1
Halobiforma lacisalsi **AJ5**	0	0	0	0	0	0
Natronomonas moolapensis **8.8.11**	0	0	0	0	0	0
Natronobacterium gregoryi **SP2**	0	0	0	1	1	0
Halobacterium **sp. DL1**	1	0	1	0	0	0
Natrinema **sp. J7-2**	0	0	0	0	0	0
Halorubrum **sp. BOL3-1**	1	0	1	1	1	0
Natrinema pallidum	0	0	0	0	0	0
Halorubrum ezzemoulense	1	1	0	1	1	0

Several color and visualization combinations from R were used to reduce the complexity of the data and make the figures easily comprehensible. For the frequency heatmap, absence or zero is declared with white color. Presence with one copy is declared with yellow color. Presence with two copies is declared with red color. Three or more copies are declared with black color.

For the binary heatmap, absence is declared by dark yellow and presence with red color. Row and column names are omitted, as the data scale is too large to have distinctive and readable labels. All heatmaps are the same size which is 76 rows and 23,576 columns.

The third heatmap was created using the frequency heatmap. The data are the same as in frequency heatmap, but the only color shown is black, highlighting all the predicted gene clusters with more than two copies. The rest of the data are left with white color.

Protein sequence analysis with HaloPredictor

All predicted gene clusters such as core, accessory, cloud, and multi-copy genes of Halobacteria pangenome, but also several proteomes of model organisms downloaded from UniProt [177] in fasta format were uploaded into HaloPredictor. The details of the proteomes used can be found in Table 6.4. Apart from the csv file containing the results and predictions for every individual protein sequence, a density plot for all LDA scores was created. The plot highlights the distribution of LDA scores both in halophilic and non-halophilic predicted protein sequences, for all input sequences.

Finally, in the group of multi-copy genes derived from the Halobacteria pangenome, two sequences from the same genetic cluster (CL2172) were chosen and aligned with protein BLAST. This was done in order to visualize at the sequence level the effect of salinity adaptation (residue substitutions) on two homologous protein sequences.

Table 6.4. List of proteomes of model organisms used in HaloPredictor salinity adaptation analysis.

Πίνακας 6.4. Λίστα των πρωτεωμάτων των οργανισμών-μοντέλων που χρησιμοποιήθηκαν στην ανάλυση προσαρμογών αλατότητας με το HaloPredictor.

No	Species	Protein count	Domain	UniProt Proteome ID
1	*Synechococcus elongatus*	2657	Bacteria	UP000002717
2	*Salinibacter ruber*	2812	Bacteria	UP000008674
3	*Methanobacterium formicicum*	2519	Archaea	UP000007360
4	*Lokiarchaeum GC14_75*	5378	Archaea	UP000034722
5	*Helicobacter pylori*	1552	Bacteria	UP000000429
6	*Candidatus halorediviνus* sp.	2152	Archaea	UP000003484
7	*Halobacteroides halobius*	2452	Bacteria	UP000010880
8	*Halanaerobium praevalens*	2055	Bacteria	UP000006866
9	*Escherichia coli*	5062	Bacteria	UP000000558
10	All *Artemia* species	2687	Eukarya	data from NCBI-protein
11	*Atherina boyeri*	435	Eukarya	data from NCBI-protein
12	*Caenorhabditis elegans*	26625	Eukarya	UP000001940
13	*Danio rerio*	46849	Eukarya	UP000000437
14	*Drosophila melanogaster*	22114	Eukarya	UP000000803
15	*Homo sapiens*	77027	Eukarya	UP000005640
16	*Mus musculus*	55470	Eukarya	UP000000589
17	*Saccharomyces cerevisiae*	6050	Eukarya	UP000002311
18	*Xenopus laevis*	44571	Eukarya	UP000186698

6.3 Results

The pangenome from 76 Halobacteria is open and diverse

The rarefaction curve of Figure 6.1 reveals an open pangenome with 23,576 unique genetic clusters. Application of a power law analogue to Heaps' law on our data yields an intercept of 2028.9962312 and an alpha of 0.6263267. The open property of a given pangenome can be estimated by this model according to the formula $\Delta n = \kappa N^{-\alpha}$, where Δn is the number of newly added genes, N is the number of genomes used, and κ and α are the fitting parameters. For $\alpha > 1$, the pangenome is closed, and for $\alpha < 1$, the pangenome is open. Heaps' law and power law are mathematically similar.

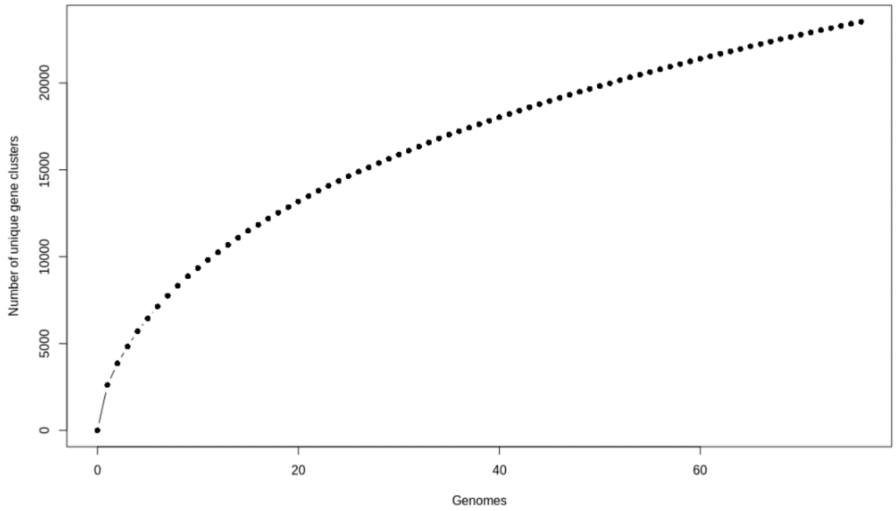


Figure 6.1. Rarefaction curve of the Halobacteria pangenome.

Εικόνα 6.1. Καμπύλη αραίωσης για το παν-γονιδίωμα της κλάσης Halobacteria.

As expected, very few gene clusters are present in all 76 genomes. The distribution of gene clusters among Halobacteria genomes reveals that 9391 out of 23,576 gene clusters are present in only one genome, while 458 out of 23,576 gene clusters are present in all 76 genomes. The remaining 13,727 clusters are ordered as shown in Fig. 6.2.

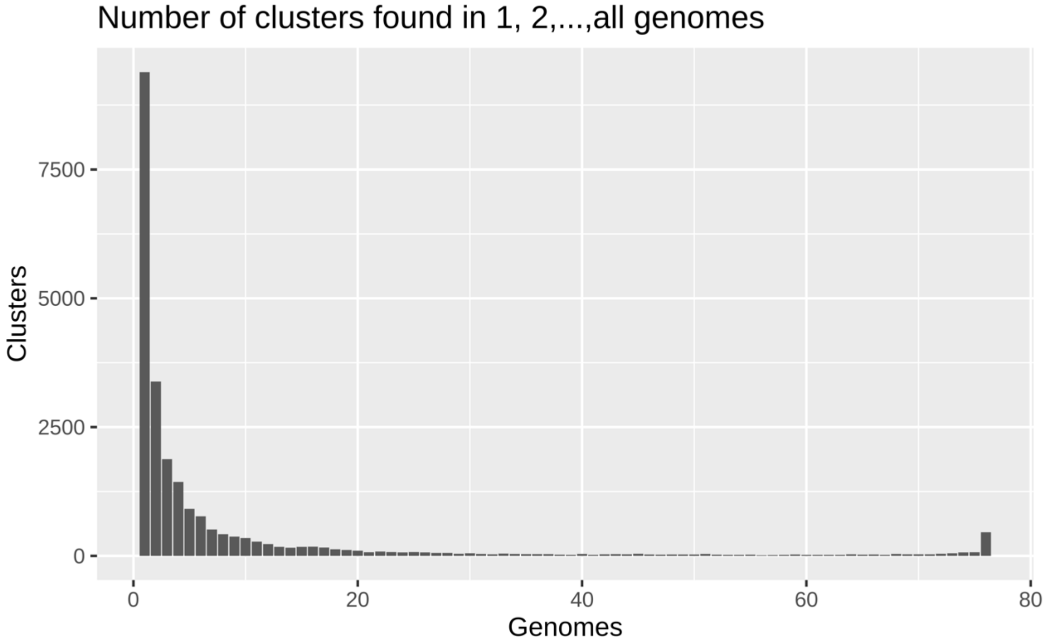


Figure 6.2. Distribution of predicted gene clusters in Halobacteria genomes. 458 out of 23,576 predicted genes are present in all 76 genomes.

Εικόνα 6.2. Κατανομή των προβλεφθέντων συστάδων γονιδίων στα γονιδιώματα των Halobacteria. 458 γονίδια από τα 23.576 που προβλέφθηκαν εντοπίζονται σε όλα τα 76 γονιδιώματα.

The detection probability of predicted gene clusters in Halobacteria pangenome (Fig. 6.3a) was also estimated. For comparison, probabilities of the pangenomes from *Halobacterium salinarium*, *Haloferax*, and *Halorubrum* were also estimated (Fig. 6.3b, c, d). For *Halobacterium salinarium* (Fig. 6.3b) gene clusters with very low detection probability are close to 12.5%. For *Haloferax* (Fig. 6.3c) they are close to 40% and for *Halorubrum* (Fig. 6.3d) more than 50% of genes. Regarding the whole pangenome of Halobacteria, more than 75% of gene clusters have low detection probability. The pangenomes of *Haloferax* and *Halorubrum* are estimated as “closed” and “nearly closed” with an alpha factor of 1.50 and 0.95, respectively. Weighted dendrograms were also created considering all predicted gene clusters for our 76 genomes and giving more weight to core and accessory genes and less to cloud and rare genes. The dendrograms were created with both complete and average Manhattan distances and can be seen in Figure 6.4 and Supplementary Figure A5, respectively.

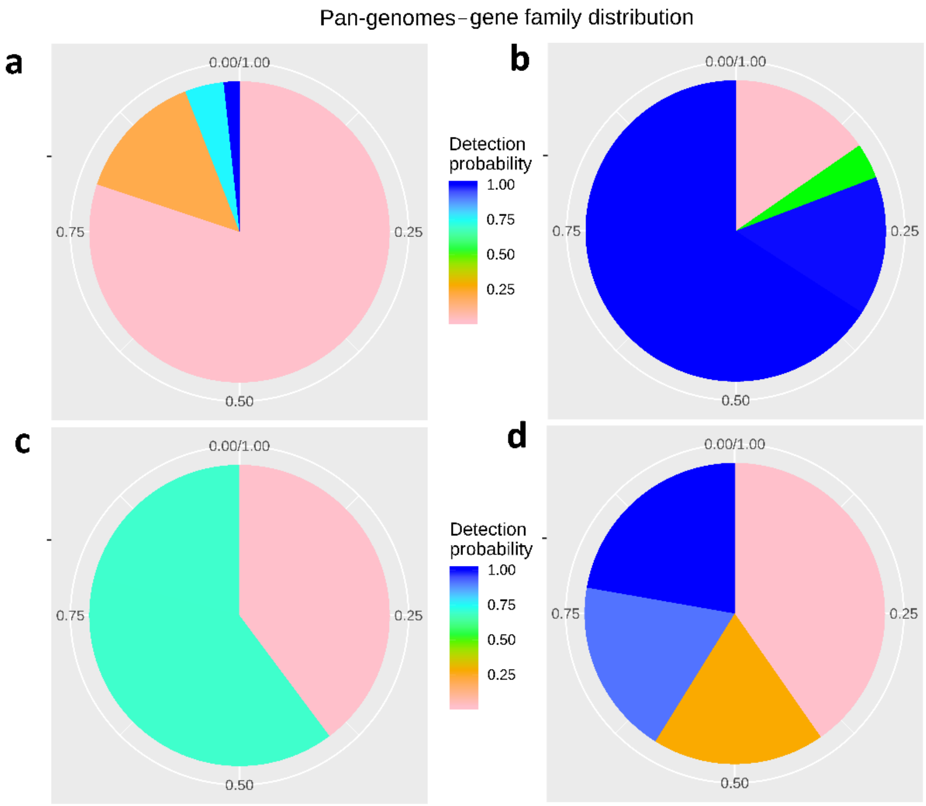


Figure 6.3. Detection probabilities of predicted gene clusters in pangenomes of Halobacteria (a), *Halobacterium salinarium* (b), *Haloferax* (c), and *Halorubrum* (d).

Εικόνα 6.3. Πιθανότητες εντοπισμού των προβλεφθέντων συστάδων γονιδίων στα παν-γονιδιώματα των Halobacteria (a), *Halobacterium salinarium* (b), *Haloferax* (c) και *Halorubrum* (d).

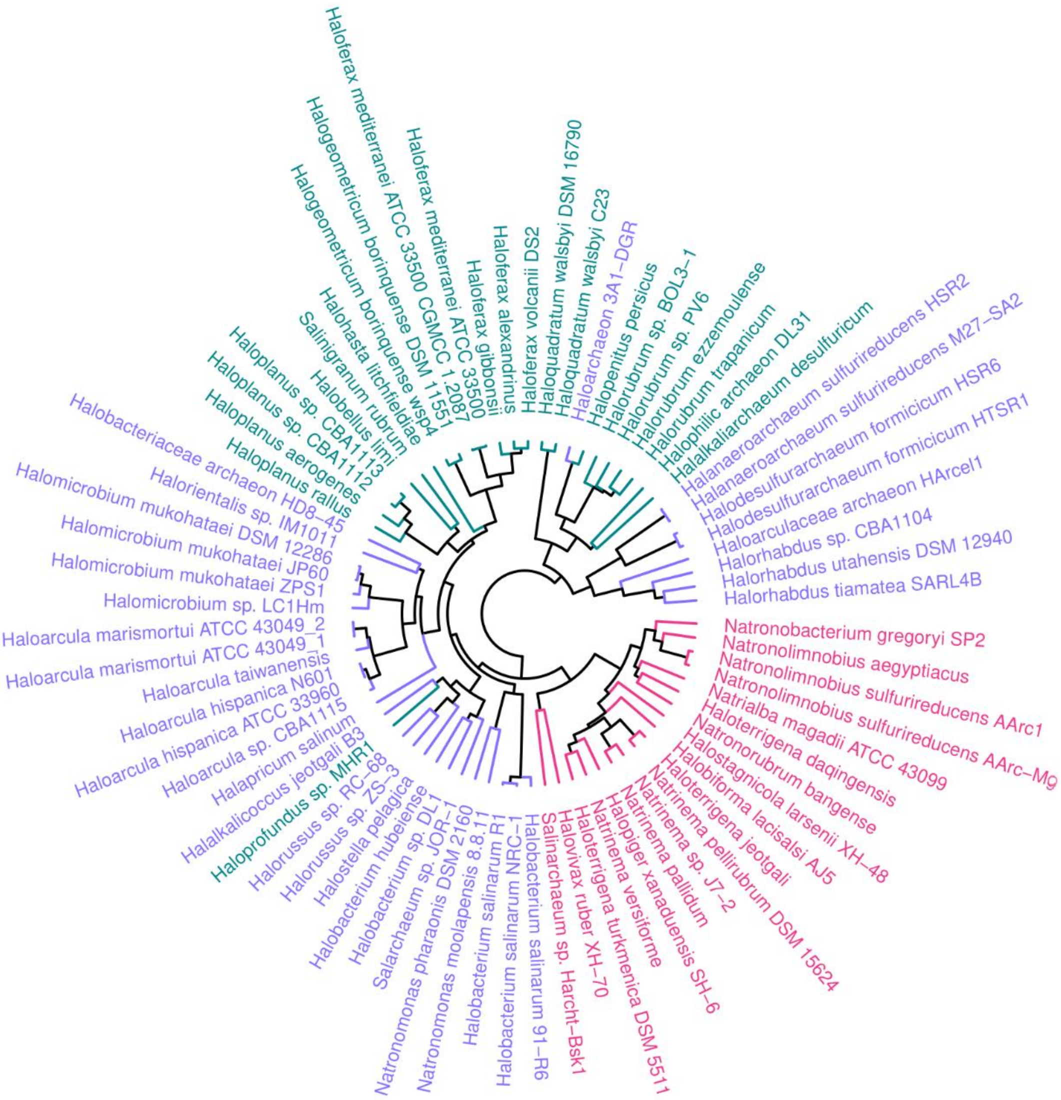

Haloplanus rallus
Haloplanus aerogenes
Haloplanus sp. CBA1112
Haloplanus sp. CBA1113
Halobellus limi
Salinigranum rubrum
Halohasta litchfieldiae
Halogeometricum borinquense wsp4
Halogeometricum borinquense DSM 11551
Haloferax mediterranei ATCC 33500 CGMCC 1.2087
Haloferax mediterranei ATCC 33500
Haloferax gibbonsii
Haloferax alexandrinus
Haloferax volcanii DS2
Haloquadratum walsbyi DSM 16790
Haloquadratum walsbyi C23
Haloarchaeon 3A1–DGR
Halopenitus persicus
Halorubrum sp. BOL3–1
Halorubrum sp. PV6
Halorubrum ezzemoulense
Halorubrum trapanicum
Halophilic archaeon DL31
Halalkaliarchaeum desulfuricum
Halanaeroarchaeum sulfurireducens HSR2
Halanaeroarchaeum sulfurireducens M27–SA2
Halodesulfurarchaeum formicicum HSR6
Halodesulfurarchaeum formicicum HTSR1
Haloarculaceae archaeon HArcel1
Halorhabdus sp. CBA1104
Halorhabdus utahensis DSM 12940
Halorhabdus tiamatea SARL4B
Natronobacterium gregoryi SP2
Natronolimnobius aegyptiacus
Natronolimnobius sulfurireducens AArc1
Natronolimnobius sulfurireducens AArc–Mg
Natrialba magadii ATCC 43099
Haloterrigena daqingensis
Natronorubrum bangense
Halostagnicola larsenii XH–48
Halobiforma lacisalsi AJ5
Haloterrigena jeotgali
Natrinema pellirubrum DSM 15624
Natrinema sp. J7–2
Natrinema pallidum
Halopiger xanaduensis SH–6
Natrinema versiforme
Haloterrigena turkmenica DSM 5511
Halovivax ruber XH–70
Salinarchaeum sp. Harcht–Bsk1
Halobacterium salinarum 91–R6
Halobacterium salinarum NRC–1
Halobacterium salinarum R1
Natronomonas moolapensis 8.8.11
Natronomonas pharaonis DSM 2160
Salarchaeum sp. JOR–1
Halobacterium sp. DL1
Halobacterium hubeiense
Halostella pelagica
Halorussus sp. ZS–3
Halorussus sp. RC–68
Haloprofundus sp. MHR1
Halalkalicoccus jeotgali B3
Halapricum salinum
Haloarcula sp. CBA1115
Haloarcula hispanica ATCC 33960
Haloarcula hispanica N601
Haloarcula taiwanensis
Haloarcula marismortui ATCC 43049_1
Haloarcula marismortui ATCC 43049_2
Halomicrobium sp. LC1Hm
Halomicrobium mukohataei ZPS1
Halomicrobium mukohataei JP60
Halomicrobium mukohataei DSM 12286
Halorientalis sp. IM1011
Halobacteriaceae archaeon HD8–45

Figure 6.4. Complete distance weighted dendrogram of 76 Halobacteria. Order Haloferacales in green, order Natrialbales in red, and order Halobacteriales in purple.

Εικόνα 6.4. Σταθμισμένο δενδρόγραμμα με τη μέθοδο complete distance για τα 76 μέλη των Halobacteria. Η τάξη Haloferacales με πράσινο, η τάξη Natrialbales με κόκκινο και η τάξη Halobacteriales με μωβ.

Functional annotation of genes predicted from Halobacteria pangenome

The annotation of predicted genes provided a wide view of the genetic repertoire of Halobacteria (Fig. 6.5). In the group of all predicted gene clusters, “Function unknown” is at 22%. “Amino acid transport and metabolism” is at 10.6% and “Energy production and conversion” at 7%. Also, “Inorganic ion transport and metabolism” functions are at 6.6%. In accessory clusters, “Inorganic ion transport and metabolism” is at 7.7%.

For cloud genes, “Function unknown” is at 30.8%. The next higher percentage categories for cloud genes, unlike core and accessory genes, are “Signal transduction mechanisms” and “Replication, recombination and repair”. Also, “Cell wall/membrane/envelope biogenesis” and “Inorganic ion transport and metabolism” have high percentages.

For multi-copy genes, “Amino acid transport and metabolism” is the first category at 28%, followed by “Post-translational modification, protein turnover, and chaperones” at 19%. “Transcription” and “Function unknown” are at 13%. From the group of all genes, for the annotation categories “Extracellular structures”, “General function prediction only”, and “Nuclear structure” no gene clusters are found. The same applies to core genes and additionally “Cytoskeleton” and “RNA processing and modification” also luck clusters. In accessory genes, no clusters are found in the above categories as well as for “Chromatin structure and dynamics”.

During sequence analysis, 415 multi-copy genes were detected that were present in more than one copy per genome, or several copies in some genomes. From those, the 14 genes with most-copies were present with 2-5 copies per genome (Table 6.5). Multi-copy genes were more present in the annotation categories “Amino acid transport and metabolism” (28.2%), “Post-translational modification, protein turnover, and chaperones” (19.5%), “Transcription” (13.4%), and “Function unknown” (13.3%). Multi-copy gene clusters could not be found in 16 annotation categories as it can be seen in Figure 6.5.

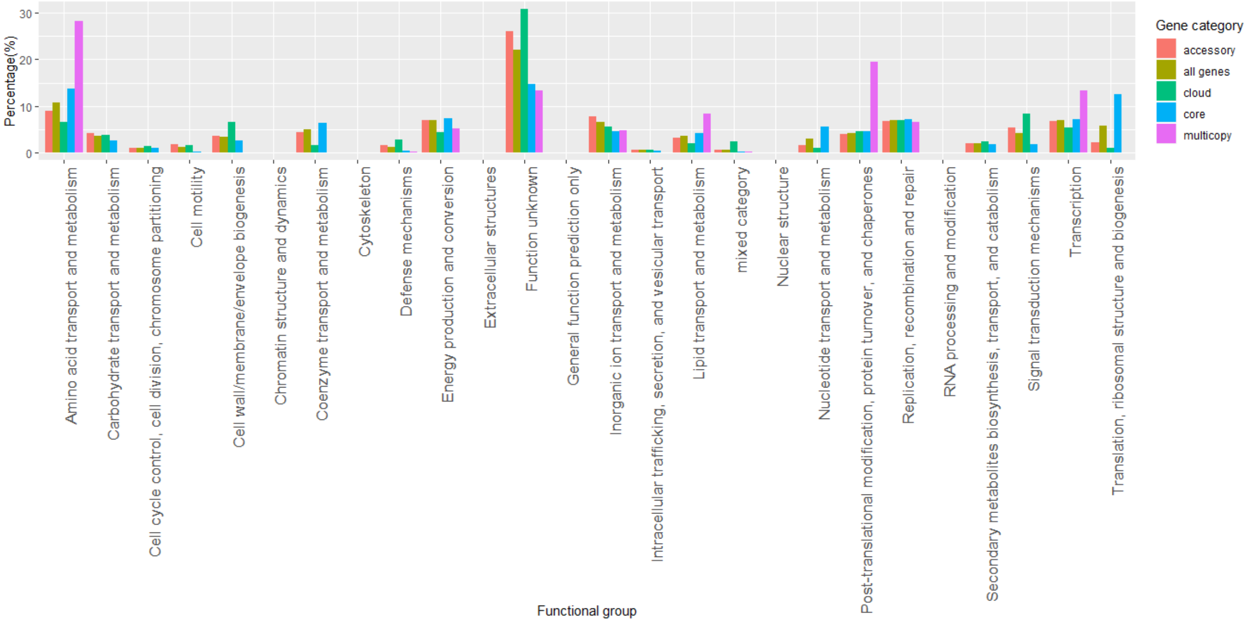


Figure 6.5. Functional annotation of Halobacteria pangenome for predicted gene clusters in 5 categories: all genes, core genes, accessory genes, cloud genes, and multi-copy genes.

Εικόνα 6.5. Λειτουργικός σχολιασμός του παν-γονιδιώματος των Halobacteria για τις προβλεφθείσες γονιδιακές συστάδες σε 5 κατηγορίες: όλα τα γονίδια, συστατικά γονίδια, συμπληρωματικά γονίδια, σπάνια γονίδια και γονίδια με πολλά αντίγραφα.

Table 6.5. The 14 most present multi-copy predicted gene clusters of Halobacteria pangenome.

Πίνακας 6.5. Οι 14 προβλεφθείσες συστάδες γονιδίων με τα περισσότερα αντίγραφα στο παν-γονιδίωμα των Halobacteria.

	Predicted gene cluster	**Gene presence in % of 76 genomes**	**No of sequences**	**Total genomes**	**COG**	**arCOG**	**EggNOG annotation result**
1	CL2172	550%	418	76	COG1405	arCOG01981	DNA-templated transcriptional preinitiation complex assembly
2	CL980	343%	261	76	COG1960	arCOG01707	Acyl-CoA dehydrogenases
3	CL5630	326%	248	76	COG0714	arCOG00434	ATPase associated with various cellular activities
4	CL1002	312%	237	76	COG0464	arCOG01308	ATPases of the AAA class
5	CL3468	287%	218	76	COG0459	arCOG01257	Belongs to the TCP-1 chaperonin family
6	CL619	275%	209	76	COG1131	arCOG00194	ABC-type multidrug transport system, ATPase component
7	CL2899	270%	205	76	COG3385	arCOG06160	FOG Transposase and inactivated derivatives
8	CL2181	254%	193	76	COG0334	arCOG01352	Belongs to the Glu Leu Phe Val dehydrogenases family
9	CL612	241%	183	76	COG4608	arCOG00184	ABC-type oligopeptide transport system, ATPase component
10	CL1443	218%	166	76	COG0656	arCOG01619	Aldo keto reductases, related to

							diketogulonate reductase
11	CL1718	217%	165	76	COG2309	arCOG01889	Leucyl aminopeptidase (Aminopeptidase T)
12	CL1353	205%	156	76	COG1013	arCOG01599	Pyruvate ferredoxin oxidoreductase and related 2-oxoacid ferredoxin oxidoreductases, beta subunit
13	CL613	201%	153	76	COG0444	arCOG00181	ABC-type dipeptide oligopeptide nickel transport system ATPase component
14	CL1003	200%	152	76	COG1222	arCOG01306	ATPase

Phylogenetic profiles created from the pangenome analysis

In total, three phylogenetic profiles/heatmaps were created from the pangenome data (Figs 6.6, 6.7, 6.8). Clustering the columns containing the predicted protein clusters of the pangenome reveals a core group of genes, present in most of our 76 species. The core group is present in all three graphs. On the frequency heatmap (Fig. 6.6) some genes from the core cluster are present in multiple copies, denoted with red and black color. This is also clear in Figure 6.8, where only genes with high frequencies are shown in black color.

On the left side of the frequency heatmap (Fig. 6.6, red arrow), a gene group present exclusively in the order Natrialbales (red colored branches) can be observed. Also, other distinctive protein groups can be found in some members of Natrialbales*,* present in multiple copies as well. The orders Halobacteriales and Haloferacales share a cluster of genes visible on the far-left edge of the graph (Fig. 6.6). A large genetic cluster is present only in six *Haloarcula* species (Fig. 6.6, purple arrow). Another cluster is present in both *Haloarcula* and *Halomicrobium* (Fig. 6.6, purple arrow). Several other large protein clusters can be observed on the graph (Fig. 6.6, green arrow), but also smaller and more specific to certain strains (Fig. 6.6, blue arrow).

Similar clustering can be observed in the binary heatmap (Fig. 6.7) which is expected since most of the data in frequency and binary matrices are identical. In the binary heatmap, the core cluster with genes present in all 76 species can be easily distinguished on the right side of the graph.

In the multi-copy frequency heatmap (Fig. 6.8), a core cluster of multi-copy genes can be seen on the right side of the graph. Several individual gene groups or single genes can be observed in the graph, for individual species or a small group of species. Additionally, on the left side of the graph, a cluster of multi-copy genes is present in all 76 Halobacteria species.

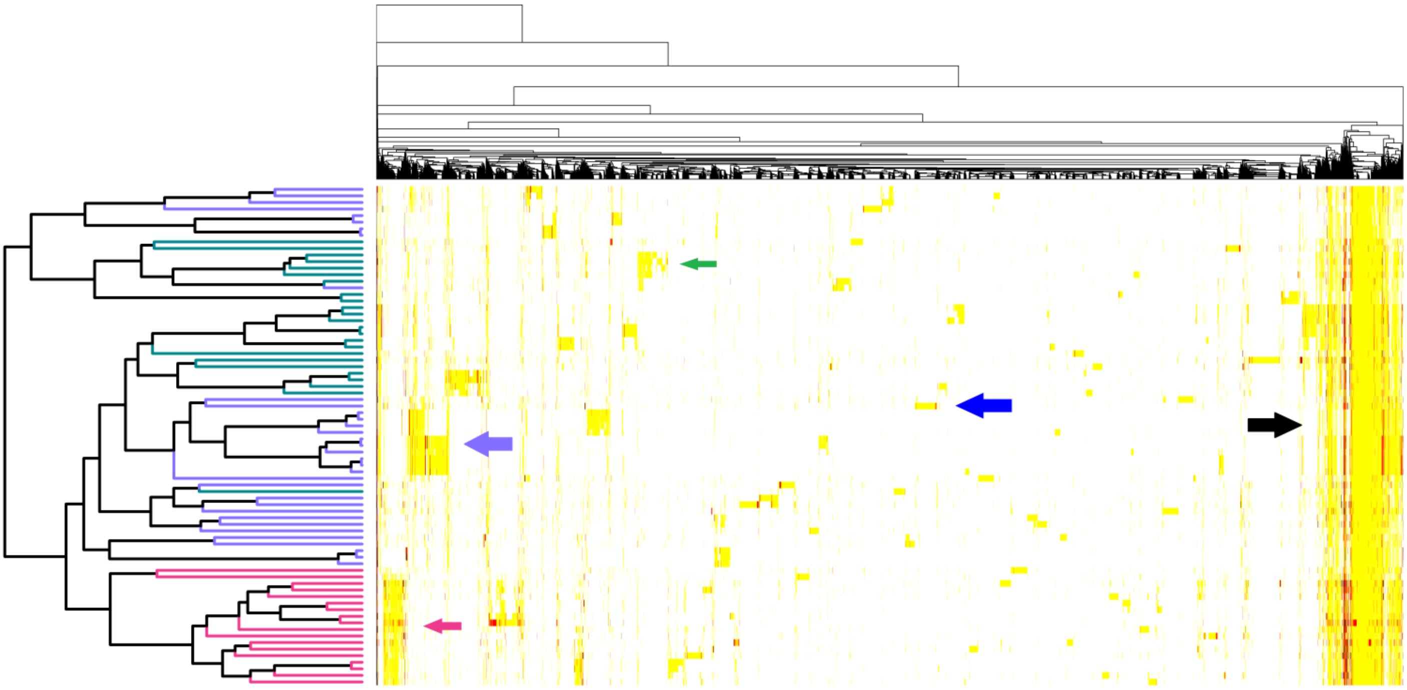


Figure 6.6. Frequency heatmap of the pan-matrix. Absence of a protein cluster is declared with white color. Presence with one copy in yellow, two copies in red, and more than two copies in black. Arrows indicate gene clusters of interest. Clustering of rows is structured according to the dendrogram of figure 6.4.

Εικόνα 6.6. Θερμικός χάρτης συχνοτήτων της παν-μήτρας. Η απουσία μιας συστάδας γονιδίων δηλώνεται με λευκό χρώμα. Η παρουσία με ένα αντίγραφο σε κίτρινο, δύο αντίγραφα σε κόκκινο και παραπάνω από δύο αντίγραφα σε μαύρο. Τα βέλη υποδεικνύουν συστάδες ενδιαφέροντος. Η ομαδοποίηση των γραμμών της μήτρας γίνεται σύμφωνα με το δενδρόγραμμα της εικόνας 6.4.

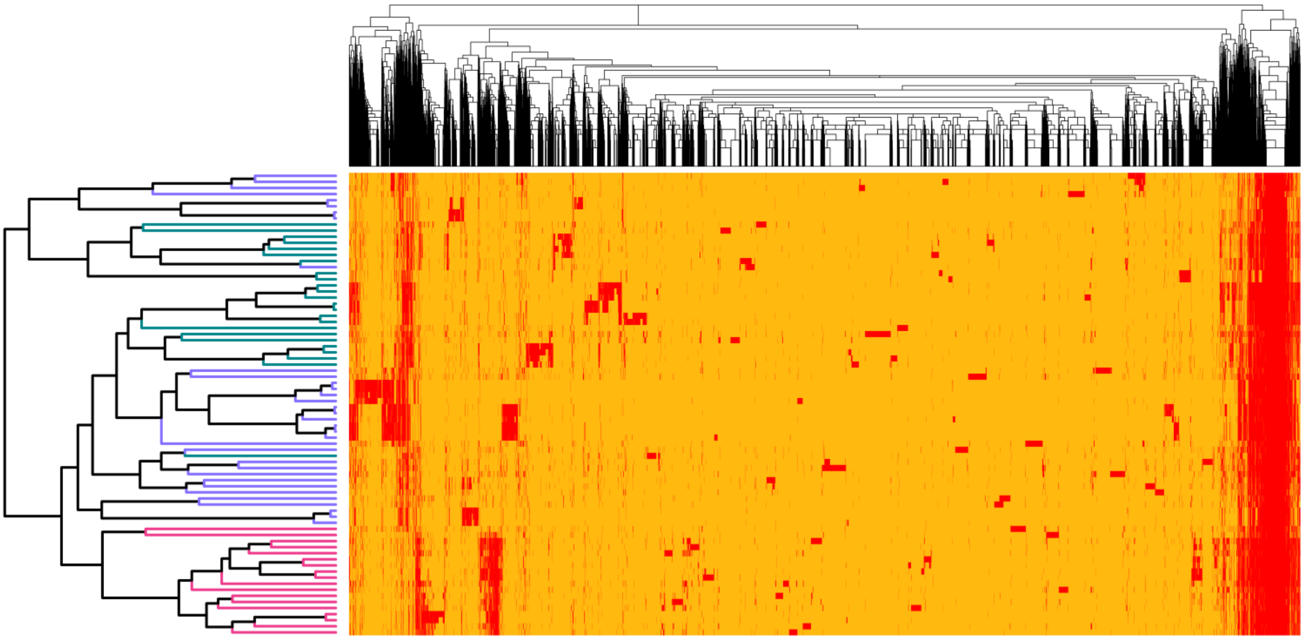


Figure 6.7. Binary heatmap of the pan-matrix. Dark yellow declares absence and red declares presence of predicted protein clusters. Clustering of rows is structured according to the dendrogram of figure 6.4.

Εικόνα 6.7. Δυαδικός θερμικός χάρτης της παν-μήτρας. Το σκούρο πορτοκαλί δηλώνει απουσία και το κόκκινο δηλώνει παρουσία των προβλεφθέντων γονιδιακών συστάδων. Η ομαδοποίηση των γραμμών της μήτρας γίνεται σύμφωνα με το δενδρόγραμμα της εικόνας 6.4.

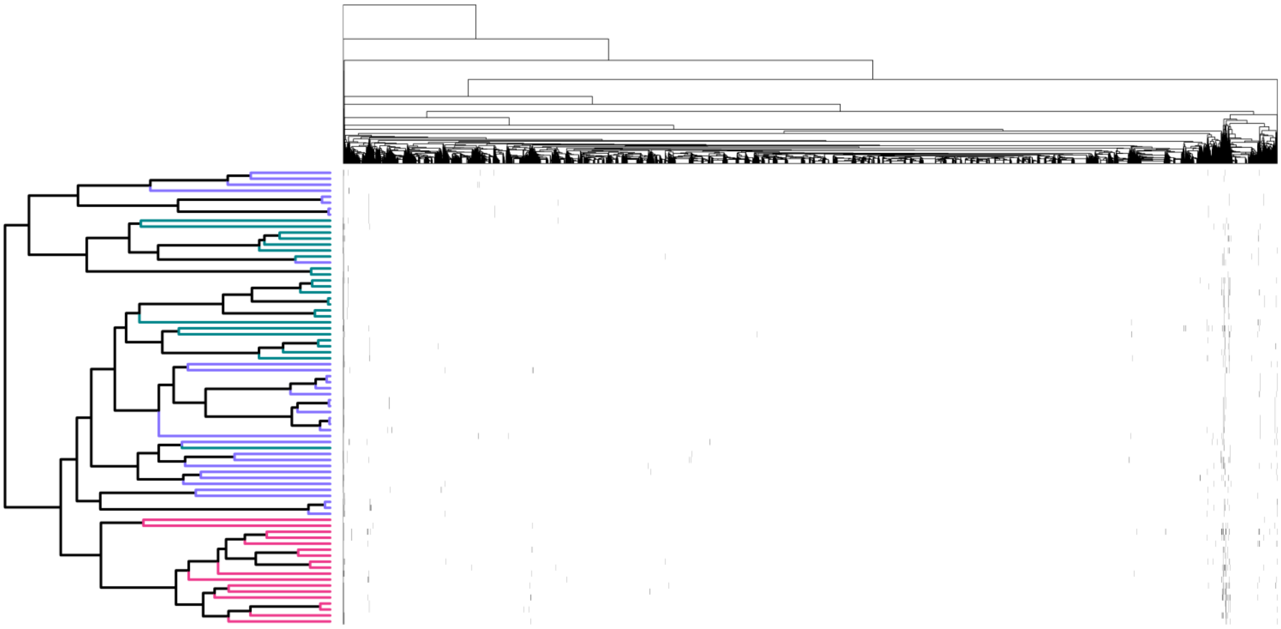


Figure 6.8. Frequency heatmap (same as figure 6.6) of protein clusters with more than two copies per genome. Multi-copy clusters declared with black color. All other data in white. Clustering of rows is structured according to the dendrogram of figure 6.4.

Εικόνα 6.8. Θερμικός χάρτης συχνοτήτων (όμοιος με την εικόνα 6.6) των πρωτεϊνικών συστάδων με παραπάνω από δύο αντίγραφα ανά γονιδίωμα. Οι συστάδες με πολλαπλά αντίγραφα δηλώνονται με μαύρο χρώμα. Όλα τα υπόλοιπα δεδομένα με λευκό. Η ομαδοποίηση των γραμμών της μήτρας γίνεται σύμφωνα με το δενδρόγραμμα της εικόνας 6.4.

Salinity adaptation analysis with HaloPredictor

The predicted protein sequences from the Halobacteria pangenome were tested for salinity adaptations with HaloPredictor [11]. In total, 217,783 Halobacteria proteins were tested and 77.3% were predicted as halophilic (Fig. 6.9.1). In a similar analysis, the proteomes of several model-organisms revealed a consistent percentage of halophilic proteins (12.6%) present in species that are non-halophilic (Fig. 6.10).

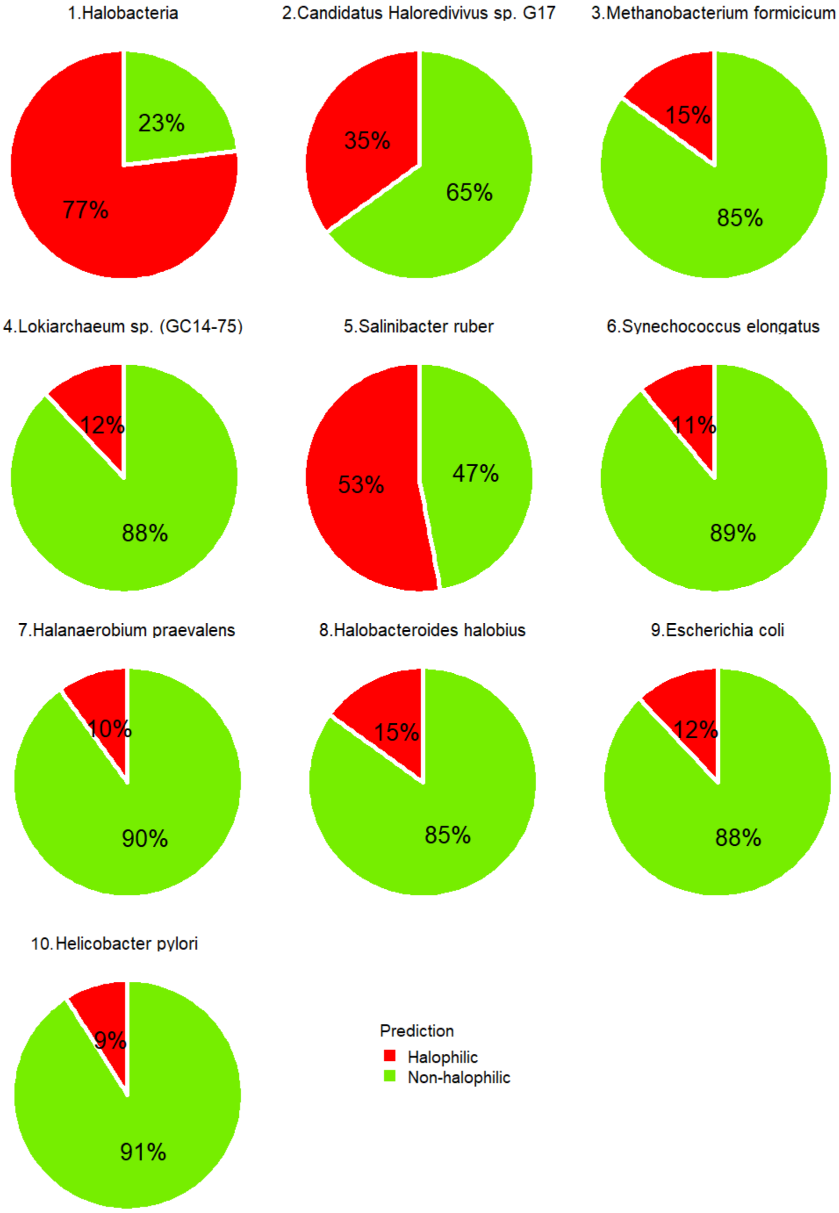


Figure 6.9. Percentages of halophilic and non-halophilic proteins in the class Halobacteria and closely or distantly related prokaryotic species as predicted by HaloPredictor.

Εικόνα 6.9. Ποσοστά αλόφιλων και μη-αλόφιλων πρωτεϊνών στην κλάση Halobacteria και σε συγγενικά (κοντινά ή μακρινά) προκαρυωτικά είδη όπως προβλέφθηκαν από το HaloPredictor.

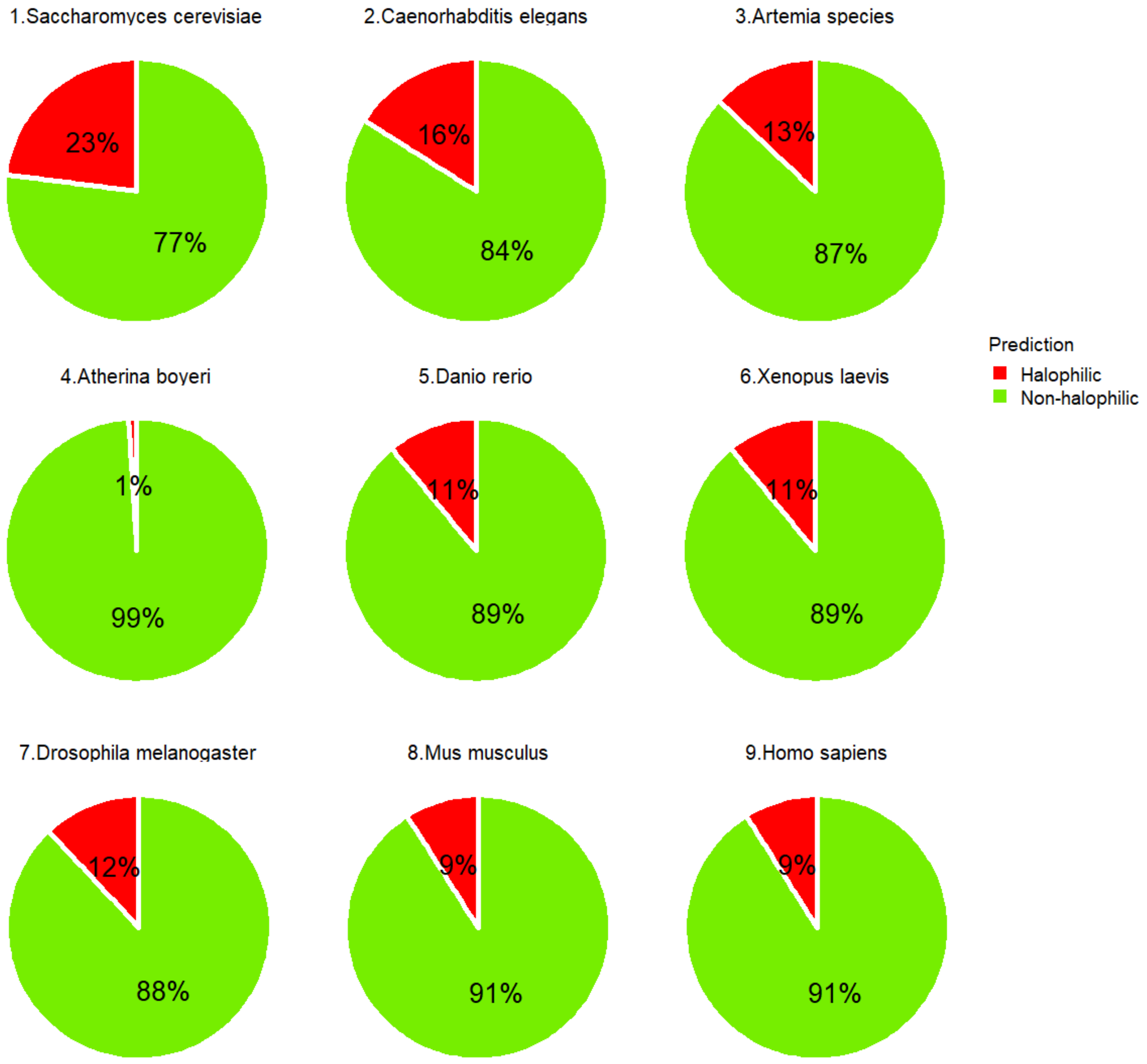


Figure 6.10. Percentages of halophilic and non-halophilic proteins in proteomes of several eukaryotes as predicted by HaloPredictor.

Εικόνα 6.10. Ποσοστά αλόφιλων και μη-αλόφιλων πρωτεϊνών στα πρωτεώματα ευκαριωτικών οργανισμών-μοντέλων όπως προβλέφθηκαν από το HaloPredictor.

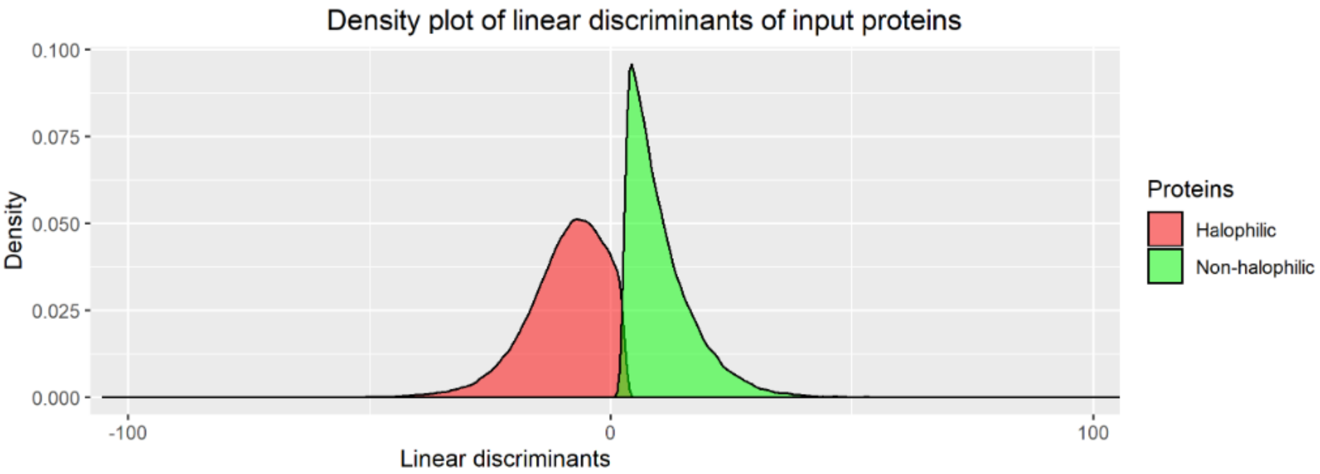
All gene groups were also investigated individually (core, accessory, cloud, and multi-copy genes) and the results are shown in Table 6.6. Additionally, the density plot of LDA scores of all Halobacteria proteins (Fig. 6.11) revealed a normal distribution of halophilic LDA scores with an average score of -9.44. Non-halophilic LDA scores are concentrated in a narrower range with an average score of 10.49. The maximum LDA score in the graph is 71.75 and the minimum is -100.7.

In the group of multi-copy genes, a specific trend was observed where, in the same gene clusters with multiple homologs in the same genome, the majority of proteins are halophilic, but there is also a non-halophilic homolog. A distinct example of this pattern can be found in the genome of *Halorubrum ezzemoulense* with genome id: GID12. There are five protein members in GID12 regarding predicted cluster 2172, which is the most abundant cluster in Halobacteria pangenome (Table 6.5). Four members are predicted as halophilic, while one member as non-halophilic (Table 6.7). As mentioned earlier, an alignment was performed between the most halophilic member of *Halorubrum ezzemoulense* (GID12_seq1490_CL2172) and the only non-halophilic member of CL2172 (GID12_seq1477_CL2172) according to their LDA scores (Table 6.7). The results showed significant differences between the two proteins, but the alignment score is still high (Fig. 6.12a). The amino acid profiles of the two proteins (Fig. 6.12b) have different percentages of key residues: halophilic proteins possess increased V, D, R, G, T, A and decreased L, N, Y. This is influencing adaptation in high salinity and increased osmotic pressure [59, 111, 112, 148].

Table 6.6. Percentages of halophilic sequences in several gene groups of Halobacteria, as predicted by HaloPredictor.

Πίνακας 6.6. Ποσοστά αλόφιλων ακολουθιών σε διάφορες ομάδες γονιδίων των Halobacteria, όπως προβλέφθηκαν από το HaloPredictor.

Group	Total sequences	Percentage of halophilic sequences
Core genes	68301	77.80%
Multi-copy genes	3115	76.90%
Accessory genes	140472	77.00%
Cloud genes	9010	78.20%

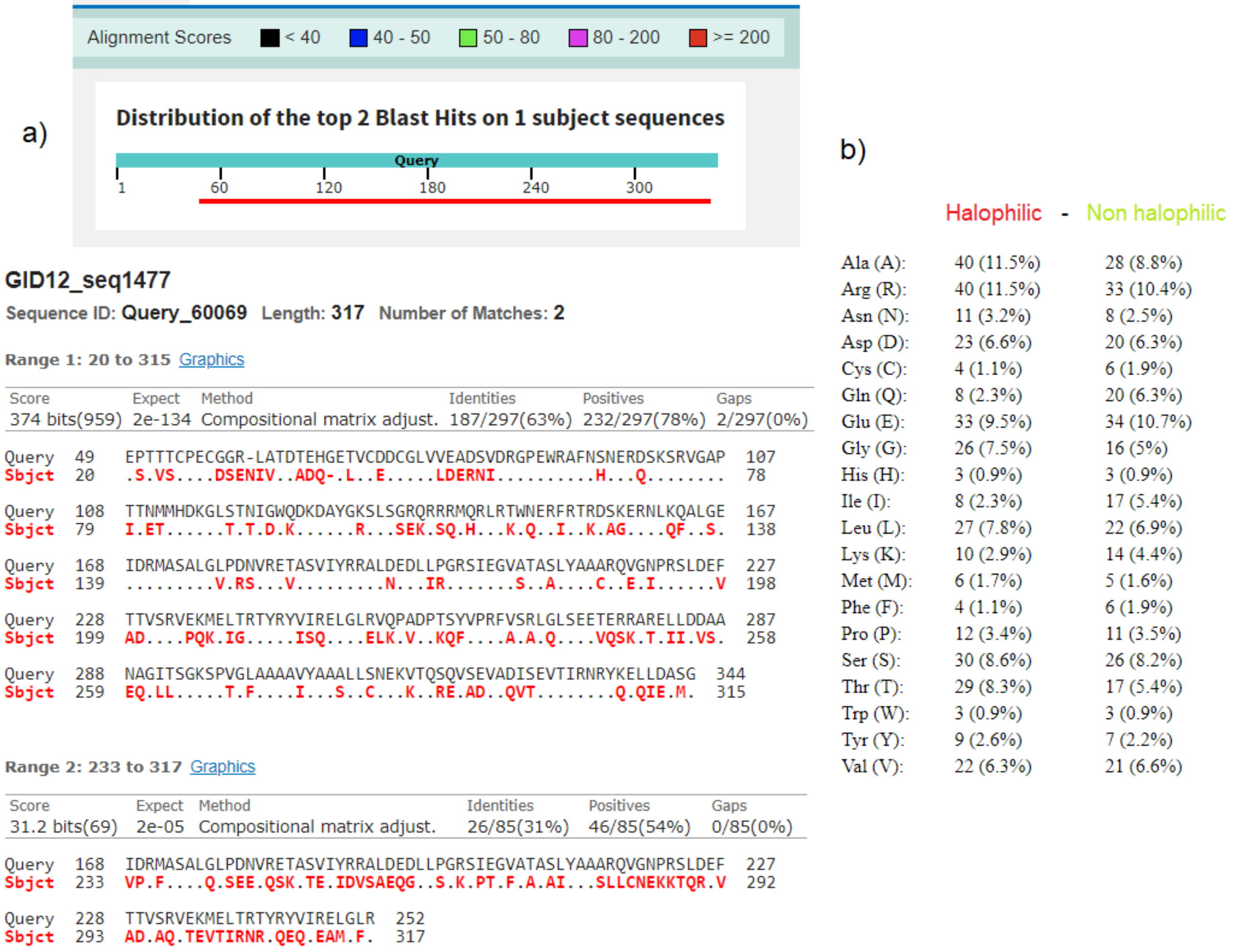


Εικόνα 6.11. Διάγραμμα πυκνοτήτων των γραμμικών διαχωριστών για όλες τις προβλεφθείσες πρωτεΐνες της κλάσης Halobacteria. Το μέγιστο LDA σκορ είναι 71.75 και το ελάχιστο -100.7.

Table 6.7. Predicted proteins of cluster 2172 from *Halorubrum ezzemoulense* (GID12) and their salinity adaptation prediction.

Πίνακας 6.7. Προβλεφθείσες πρωτεΐνες της συστάδας 2172 από το *Halorubrum ezzemoulense* (GID12) και οι αντίστοιχες προβλέψεις τους όσον αφορά προσαρμογές στην αλατότητα.

Sequence	**Group**	**Prediction**	**LDA**
GID12_seq1228_CL2172	1	halophilic	-6.675695113
GID12_seq1477_CL2172	2	non-halophilic	6.726886536
GID12_seq1490_CL2172	1	halophilic	-14.67131745
GID12_seq360_CL2172	1	halophilic	-0.223510102
GID12_seq626_CL2172	1	halophilic	-5.484858986



GID12_seq1477

Sequence ID: **Query_60069** Length: **317** Number of Matches: **2**

Range 1: 20 to 315 Graphics

Score	Expect	Method	Identities	Positives	Gaps
374 bits(959)	2e-134	Compositional matrix adjust.	187/297(63%)	232/297(78%)	2/297(0%)

```
Query  49   EPTTTCPECGGR-LATDTEHGETVCDDCGLVVEADSVDRGPEWRAFNSNERDSKSRVGAP  107
Sbjct  20   .S.VS....DSENIV..ADQ-.L..E.....LDERNI..........H...Q........  78

Query  108  TTNMMHDKGLSTNIGWQDKDAYGKSLSGRQRRRMQRLRTWNERFRTRDSKERNLKQALGE  167
Sbjct  79   I.ET......T.T.D.K......R...SEK.SQ.H...K.Q..I..K.AG....QF..S.  138

Query  168  IDRMASALGLPDNVRETASVIYRRALDEDLLPGRSIEGVATASLYAAARQVGNPRSLDEF  227
Sbjct  139  .........V.RS...V.........N...IR.......S..A....C..E.I......V  198

Query  228  TTVSRVEKMELTRTYRYVIRELGLRVQPADPTSYVPRFVSRLGLSEETERRARELLDDAA  287
Sbjct  199  AD....PQK.IG.....ISQ....ELK.V..KQF....A.A.Q....VQSK.T.II.VS.  258

Query  288  NAGITSGKSPVGLAAAAVYAAALLSNEKVTQSQVSEVADISEVTIRNRYKELLDASG  344
Sbjct  259  EQ.LL.....T.F....I...S..C...K..RE.AD..QVT........Q.QIE.M.  315
```

Range 2: 233 to 317 Graphics

Score	Expect	Method	Identities	Positives	Gaps
31.2 bits(69)	2e-05	Compositional matrix adjust.	26/85(31%)	46/85(54%)	0/85(0%)

```
Query  168  IDRMASALGLPDNVRETASVIYRRALDEDLLPGRSIEGVATASLYAAARQVGNPRSLDEF  227
Sbjct  233  VP.F....Q.SEE.QSK.TE.IDVSAEQG..S.K.PT.F.A.AI...SLLCNEKKTQR.V  292

Query  228  TTVSRVEKMELTRTYRYVIRELGLR  252
Sbjct  293  AD.AQ.TEVTIRNR.QEQ.EAM.F.  317
```

	Halophilic -	Non halophilic
Ala (A):	40 (11.5%)	28 (8.8%)
Arg (R):	40 (11.5%)	33 (10.4%)
Asn (N):	11 (3.2%)	8 (2.5%)
Asp (D):	23 (6.6%)	20 (6.3%)
Cys (C):	4 (1.1%)	6 (1.9%)
Gln (Q):	8 (2.3%)	20 (6.3%)
Glu (E):	33 (9.5%)	34 (10.7%)
Gly (G):	26 (7.5%)	16 (5%)
His (H):	3 (0.9%)	3 (0.9%)
Ile (I):	8 (2.3%)	17 (5.4%)
Leu (L):	27 (7.8%)	22 (6.9%)
Lys (K):	10 (2.9%)	14 (4.4%)
Met (M):	6 (1.7%)	5 (1.6%)
Phe (F):	4 (1.1%)	6 (1.9%)
Pro (P):	12 (3.4%)	11 (3.5%)
Ser (S):	30 (8.6%)	26 (8.2%)
Thr (T):	29 (8.3%)	17 (5.4%)
Trp (W):	3 (0.9%)	3 (0.9%)
Tyr (Y):	9 (2.6%)	7 (2.2%)
Val (V):	22 (6.3%)	21 (6.6%)

Figure 6.12. Alignment of two homologs from genetic cluster 2172. One is predicted as halophilic, while the other as non-halophilic. a) Details about the alignment. b) Amino acid profile of both homologs in frequency and percentage.

Εικόνα 6.12. Στοίχιση των δύο ομόλογων πρωτεϊνών από τη συστάδα 2172. Η μια ακολουθία προβλέπεται ως αλόφιλη ενώ η άλλη ως μη-αλόφιλη. a) Λεπτομέρειες που αφορούν τη στοίχιση. b) Προφίλ αμινοξέων των δύο ομόλογων πρωτεϊνών σε συχνότητες και ποσοστά.

6.4 Discussion

Pangenome analyses have revealed a great amount of genetic diversity in the microbial world, both in Bacteria [164] and Archaea [166]. However, to our knowledge, the current analysis in Halobacteria is the largest so far. The open pangenome of Halobacteria testifies to a genetic potential including new molecular pathways, protein families, and adaptation mechanisms awaiting exploration.

We organized the predicted gene clusters in three main groups, core, accessory, and cloud genes. Core genes were present in more than 95% of species, accessory genes from 95% to more than 1%, and cloud genes were present in 1% of the species. No strict standards exist for these thresholds, so these groups can be slightly variable. For example, we could account cloud genes to be present in 10% of species or less, which would increase the number of cloud genes and decrease the accessory genome. We oriented cloud genes to be the group referred to as “strain-specific” or “dispensable” in the literature, however the latter has been questioned because some of these genes have been shown to be important for survival and adaptation [178]. Core genes that are present in all 76 genomes amount to 660, including those with multiple copies per genome (415), while core genes with exactly one copy per genome amount to 245. The rest of the core genes are measured from 99 to above 95% presence (154). Therefore, there are 814 core genes, starting from 95% presence.

Detection probabilities shown in Figure 6.3 correspond to different groups of genes from the pangenome. Genes with detection probability closer to 1 correspond to core genes, while genes with low probability correspond to cloud and rare genes. At the species level, pangenomes of *Halomicrobium mukohataei* (data not shown) and *Halobacterium salinarium* (Fig. 6.3b) have a different distribution of low detection probability gene clusters. For the same number of three strains in their pangenome, *Halomicrobium mukohataei* has low detection probability clusters at 3% while *Halobacterium salinarium* is closer to 12.5%, more than fourfold increase. The pangenome of *Halomicrobium mukohataei* is closed with alpha factor 2.0 and the pangenome of *Halobacterium salinarium* is nearly closed with alpha factor of 0.98. For *Haloferax* and *Halorubrum*, using five and four genomes respectively, low detection probability clusters are close to 40%, while for *Halanaerobium* (not shown) they are close to 68%. This is not the case for the pangenome of *Haloplanus*, where strain-specific and rare genes are present in lower percentages [167]. These data suggest that some taxonomic groups within Halobacteria could have more conserved functions, while others may show large differences in terms of genetic composition and diversity. Nevertheless, more genomes are needed to draw a clearer picture.

As mentioned above, the amount of genetic diversity revealed in microbial species has occasionally been rather unexpected. Halobacteria survive and thrive in very specific environments, which raises several questions. Is this diversity essential for survival? Are there more adaptation strategies in salinity and extreme environments that we don't know about? It has been shown that horizontal gene transfer may cross domains [19, 166]. So, is this relevant to Halobacteria?

The dendrogram of the 76 Halobacteria (Fig. 6.4) with complete distance placed most of the species in their corresponding genera, on a monophyletic group when applicable. Many species like *Halostagnicola larsenii* XH-48 without other representatives in the genus were placed on a single branch. Even though the dendrogram is not a complete phylogenetic method, the genome data in our analysis confirm most previously known phylogenetic relationships. However, both on average (Supplementary Figure A5) and complete linkage (Fig. 6.4) trees, *Haloarchaeon* 3A1-DGR which according to NCBI taxonomy belongs to Halobacteriales was clustered within Haloferacales, and *Haloprofundus* sp. MHR1 which belongs to Haloferacales was clustered with Halobacteriales. *Haloarchaeon* 3A1-DGR is not yet phylogenetically resolved [179] and could share many core and accessory genes with Haloferacales*,* hence being clustered with this group on our weighted dendrogram. *Haloprofundus* sp. MHR1 may be clustering with Halobacteriales for the same reason*.* Also, *Salinarchaeum* sp. Harcht-Bsk1 which belongs to Natrialbales [180] is grouped with Haloferacales in the average distance graph (Supplementary Figure A5) while in the complete linkage graph it is on the last clade of Natrialbales, next to Halobacteriales. A more extensive analysis on how accessory genes are shared among Halobacteria may be required to explain the clustering results as this is a weighted dendrogram. Extensive phylogenomic analyses in Halobacteria are generally rare [78, 79] and information about these clades is still poor.

The functional annotation pinpointed interesting details about the genetic machinery of Halobacteria (Fig 6.5). The first observation concerns the large number of predicted proteins with unknown function. From a technical aspect, most of these predictions could be considered reliable as they involve large amounts of data coming from blast alignments, validation, cut-offs, and clustering. However, further analyses are needed to determine and experimentally validate the existence of these proteins and their roles in Halobacteria, as it occurred previously with bacteriorhodopsin [181-183] and carotenoids [184].

Amino acid transport, metabolism, and energy production-conversion is a large part of the predicted pangenome. It is already known that many functions necessary for survival in osmotic pressure require specific metabolic capacities and energy efficiency [185]. Having a large portion of genes responsible for the functions described above is therefore expected and our data provide strong confirmation for the case of Halobacteria.

Inorganic ion transport and metabolism takes up a considerable share of the pangenome. Halophiles use several strategies that include membrane ion transports and biogenesis of several compatible solutes to combat osmotic pressure [18, 78, 148, 186, 187]. It is not yet clear exactly which taxa have the capacity to use one or both strategies and it would be interesting for future studies to delve into this annotation category.

Regarding cloud genes, the largest annotation group of predicted cloud clusters belongs to unknown functions. The second group is “Signal transduction mechanisms”. There is plenty of information in the literature about signal transduction in Halobacteria [188-191], however having considerable percentage of signal transduction genes in the cloud group of the pangenome is observed for the first time. Additionally, great functional diversity has been observed in the cell wall and membrane composition of Halobacteria [192-194], which could be partially explained by the presence of “Cell wall/membrane/envelope biogenesis” annotation category in cloud genes of our pangenome. Annotation data from cloud genes mentioned above suggest there are many unknown and highly specific genes of various functions in Halobacteria.

Multi-copy genes highlight the gene expression priorities of the cells of Halobacteria. The most frequent gene cluster (CL2172) with five copies per genome is a transcription initiation factor IIB protein which stabilizes TBP binding to the archaeal box-A promoter (Table 6.5). It is also responsible for recruiting RNA polymerase II to the pre-initiation complex (DNA-TBP-TFIIB - arCOG01981) [175]. This result suggests the need for fast and efficient transcription. It has already been proposed that archaeal transcription and translation can happen simultaneously on the same mRNA [195-197], which combined with the use of operons [198] and the fact that highly expressed genes are already reported present in Archaea [199], points out faster gene expression in Halobacteria as highly probable.

The second gene cluster (CL980) from Table 6.5, with three copies per genome, belongs to Acyl-CoA dehydrogenases. These proteins are known to be involved in fatty acid metabolism [200]. Many processes within Halobacteria involve the cell wall, membrane proteins, and reading signals from the external environment. It is therefore essential for these cells to organize the function of their cell walls, lipid structure, and fatty acid processing. Although it was widely believed that Archaea did not possess a fatty acid processing system, recent papers report otherwise [201, 202]. The presence of Acyl-CoA dehydrogenases further supports these claims. Cell walls of Archaea have different structures and are diverse [192, 193, 203] so it could be likely for some extremophilic Archaea to possess fatty acids in their cell wall structure, despite the “lipid divide” [204]. However, more research on novel species of extremophilic Archaea is needed to confirm this hypothesis.

The next two gene clusters (CL5630, CL1002) from Table 6.5 are ATPases with three copies per genome. The presence of ATPases in multiple copies is expected in Archaea, since ATP synthases are highly present as part of an energy and mobility management mechanism [205, 206].

The remaining gene clusters in Table 6.5 are present in two copies per genome and range from chaperones to ATPases.

In total, 217,783 sequences were assigned as input for annotation with 182,884 returning an annotation result. That means 34,899 sequences still contain unknown information from the pangenome, hypothetical proteins, non-functional peptides, and disordered proteins. In other words, a considerable fraction of 16% in the Halobacteria pangenome involves undetermined and possibly novel functional elements. Also, despite the large number of proteins inserted into the annotation process, many annotation categories from EggNOG did have an extremely low percentage or no matches at all (Fig. 6.5). These categories for all predicted genes included "General function prediction only", "Extracellular structures", "Nuclear structure", "Cytoskeleton", "RNA processing and modification", and "Chromatin structure and dynamics".

The phylogenetic profiles created from the pangenome data matrices (Figs 6.6, 6.7, 6.8) revealed interesting information about Halobacteria. The frequency pan-matrix contained several predicted gene clusters with multiple copies per species. It is already known for some species of Halobacteria that several genes are present in multiple copies [207-210], however the current pangenome data show for the first time that this is a widespread phenomenon within the class. From our data, the predicted gene cluster numbered 2899, for strain *Haloferax volcanii* DS2, contains 37 copies of the archaeal orthologous group arCOG06160, which is predicted as a transposase of the ISH3 family according to EggNOG annotation. The presence of transposable elements in Halobacteria has not been studied in detail until now. A few papers exist [211-215] and our data could provide a basis for more research on this topic.

In Figures 6.6 and 6.7, on the right side, a large cluster of genes can be seen (Fig. 6.6, black arrow). During the pangenome analysis, we identified 814 core gene clusters, present in all 76 Halobacteria species. These clusters are visible in the graphs as a large solid colored block, both in frequency (Fig. 6.6) and binary (Fig. 6.7) heatmaps. These 814 core genes along with the rest of the gene clusters present in most species suggest the presence of about 1100 conserved genes that most Halobacteria possess.

The order Natrialbales in Figures 6.6 and 6.7 have an exclusive set of genes, not present in Halobacteriales or Haloferacales. To the best of our knowledge, apart from phylogenetic analyses on Halobacteria [78, 79], this is the first time Natrialbales are separated from the rest of Halobacteria by a specific trait. The presence of these genes indicates differences in several cellular functions in Natrialbales. Halobacteria members are known for their metabolic diversity [216-218], so this gene cluster could be linked to metabolic functions. No other distinct characteristics have been observed in Natrialbales [219] so far.

Additionally, there are two more distinct gene clusters in Figure 6.6. The first is present only in *Haloarcula* and the second is present both in *Haloarcula* and *Halomicrobium* (Fig. 6.6, purple arrow). *Haloarcula* species are known to differ from other Halobacteria by possessing the triglycosyl glycolipid TGD-2 [220]. *Halomicrobium* and *Haloarcula* strains are closely related [221, 222] and are reported having intraspecific polymorphisms in their 16S rRNA genes [223], as well as multiple copies of the gene as also reported in other members of the class Halobacteria [70]. The cluster with predicted genes belonging only to *Haloarcula* and *Halomicrobium* will hopefully reveal more clues regarding the genetic potential, origin, and evolutionary history of these closely related genera.

In comparison with the Halobacteria phylogeny in Chapter 3, the weighted dendrogram (Fig. 6.4) derived from the pangenome, successfully separated Haloferacales and Natrialbales. Halobacteriales were also divided in two groups, however the distant small group contains only *Halanaeroarchaeum sulfurireducens*. *Halobacterium salinarum* strains are placed in the main group, but in a separate sub-cluster.

The majority (77%) of the 217,783 predicted protein sequences from the Halobacteria pangenome were assigned by HaloPredictor as halophilic (Fig. 6.9.1). All gene categories follow a constant percentage of halophilic proteins at 77.5% on average (Table 6.6). This suggests that most Halobacteria require these adaptations to a great extent in order to survive in high salinity, and even gene groups that are not highly conserved or abundant have to be adapted in their majority. On a further reading, this indicates that, at least in Halobacteria, haloadaptation does not emerge by the action of a handful of genes but requires an extensive genetic rewiring throughout the genome. It would be interesting to see the distribution of the percentages of halophilic and non-halophilic proteins in individual members of Halobacteria to determine whether there are some species with higher proportions of adapted proteins than others, or the percentages are equally distributed in the whole class. It is also very interesting to see what protein families are usually predicted as non-halophilic and whether they are common in members of Halobacteria.

The non-halophilic proteins of Halobacteria account for a fraction of 22.5%. It is also very interesting that even though Halobacteria use the salt-in strategy and salts are entering their cytoplasm, it is unknown how the non-adapted proteins function. The non-halophilic predictions of HaloPredictor contain mostly functional proteins. However, there are also proteins, both of halophilic and non-halophilic nature, with compositionally biased regions such as transmembrane proteins. The conserved parts of these proteins do not allow for changes in the usage of amino acids for haloadaptation. We have noticed, specifically in transmembrane proteins, that they do carry an adaptation signature in their amino acid profile, when excluding the compositionally biased regions (Chapter 4). It is still unknown what amino acid profile adaptations other compositionally biased proteins or proteins with low complexity regions possess and what their functions are.

Several other prokaryotes have a high percentage of halophilic proteins in their proteomes. Our analysis confirms previous reports on high salinity tolerance for *Candidatus haloredivivus* sp. G17, *Salinibacter ruber*, and *Saccharomyces cerevisiae* (Figs 6.9.2, 6.9.5, 6.10.1) [92, 96, 224, 225]. Nanohaloarchaea such as *Candidatus haloredivivus* sp. G17 are reported using Halobacteria as hosts [93, 94]. It is possible that many of the adapted genes of Halobacteria have been transferred to several species of Nanohaloarchaea, although convergent evolution could also be taking place [93]. This is also the case for *Salinibacter ruber* regarding gene transfers [19]. *Saccharomyces cerevisiae* has been reported having salt-tolerance abilities but mostly through the basic stress response mechanisms, activation of biological pathways, and also compatible solute accumulation [225]. To our knowledge, this is the first study that proteins with adapted amino acid profiles are found in Fungi. Additionally, more species of Fungi are also reported with halophilic capabilities [226, 227]. More research is needed to trace the origin of these adapted halophilic proteins.

Most model-organisms used in this study (Figs 6.9, 6.10) present a consistent pattern regarding the presence of adapted halophilic proteins in their proteomes. It has been reported that non-halophilic species do contain halophilic proteins, however this subject has not been thoroughly studied [111]. From our proteome data it is suggested that most non-halophilic species have a percentage of 12.6% in halophilic proteins (Figs 6.9, 6.10). It would be interesting to see what protein families have conserved their amino acid profiles which gives them halophilic properties even in higher eukaryotes such as *Homo sapiens*.

The alignment between the halophilic and non-halophilic homologs of CL2172 from *Halorubrum ezzemoulense* confirmed the differences in their amino acid compositions. However, these data raise the question of the presence of non-halophilic homologs in the multi-copy genes of Halobacteria. Ancestral gene content reconstructions through phylogenetic profiling are needed to resolve these questions.

Regarding the density plot of LDA scores (Fig. 6.11), the typical LDA scores for a halophilic or a non-halophilic protein are -9.44 and 10.49, respectively. However, there are also more extreme values present in our data (min: -100.7, max: 71.75). Such outlier scores are caused either by short proteins (<150 amino acids) or by proteins containing low complexity or compositionally biased regions. Most of these proteins in Halobacteria are annotated as hypothetical and their functions remain unknown.

The open pangenome of Halobacteria revealed a small core of about 1100 genes. Cloud genes are present in more than ¾ of the pangenome, revealing a large genetic reservoir consisting of many unknown and possibly novel genes that need to be accounted for, validated, and annotated. These data will hopefully lead to the discovery of new molecular pathways and mechanisms in Archaea. Multi-copy genes will reveal more information regarding the genetic mechanisms that require fast expression in cells adapted to high salinity. The novel genetic traits discovered in *Haloarcula* and *Halomicrobium* along with the rest of the pangenome data call for further and extensive research on Halobacteria. Additionally, the presence of halophilic proteins in non-halophilic species across the tree of life raises new questions and research topics.

Chapter 7: Conclusions and prospects

Over the last few years, research on halophiles, and extremophiles in general, has entered an accelerating phase owing to an exciting combination of basic mechanisms for adaptation and potential applications these organisms bear. Databases like HaloDom can serve as a useful resource and starting point for researchers interested in studying the exotic biology of halophiles and their adaptations. HaloDom can be further expanded to include other types of extremophiles and aid in the devise of novel approaches and questions in this field. The accompanying online version of HaloPredictor is the only software, by the time of writing, that can determine a protein's halophilic or non-halophilic nature. Future improvements of the tool may include the detection of additional signatures of haloadaptation as well as appropriate adjustments for other extremophiles.

The lineage of Halobacteria forms a monophyletic group in phylogenetic trees constructed with conserved archaeal markers. From the three separate Halobacteria orders distinguished so far, Haloferacales and Natrialbales are also monophyletic while Halobacteriales have a more complex topology (Figs 3.1, 3.2). Despite the extensive battery of markers used (242 in total), the phylogenetic position of Halobacteriales remains largely unresolved. This is an indication that additional events, like horizontal gene transfers, may have played a role in the evolutionary history of certain groups and, consequently, advanced methods of topology estimation need to be implemented.

The majority of proteins in Halobacteria (77%) have the distinct amino acid profile of salt adaptation. It was shown that important and conserved regions in proteins are not affected by amino acid changes during the adaptation process. This pattern was confirmed in transmembrane proteins of Halobacteria but needs to be validated with the use of more data from other protein families and experimental proof.

Protein family and size also affects the amino acid adaptations. Large and small proteins have slightly less adaptations in their amino acids while hypothetical and known proteins of medium size have the most observed adaptations. Other halophilic species from Bacteroidetes, Proteobacteria, and Rhodothermaeota have a slightly different adapted amino acid profile.

The pangenome analysis of the class Halobacteria revealed an open pangenome with a restricted core genome but large accessory and cloud genomes. Considering that the annotation process revealed large numbers of proteins with "function unknown" in accessory and cloud categories, the pangenome of Halobacteria can be a source of novel protein families and functions.

It was also observed that pangenomes of different taxa within Halobacteria can diverge significantly in terms of the fractions of the different gene categories and their pangenomes' openness. For example, the pangenome of *Halomicrobium mukohataei* has 3% of its genome as cloud genes, while *Halobacterium salinarium* is closer to 12.5%. Additionally, for *Haloferax* its cloud genome is at 40%, while for *Halanaerobium* at 68%. More research is needed to determine which species have less conserved pangenomes and are more likely to contain novel and rare genes.

As mentioned above, the majority of predicted proteins from Halobacteria show salinity adaptations. In predicted genes with multiple copies per genome, there are also non-halophilic homologues present. This has important implications regarding the exact genealogy and evolutionary coalescence of these homologues.

The halophilic bacterium *Salinibacter ruber* has more than half of its proteome adapted to salinity (53%) and it is possible that more halophilic Bacteria follow the same pattern. It is widely believed that genes from Halobacteria have been horizontally transferred to bacterial lineages although the extent of this transfer is not known. Interestingly, several model, non-halophilic organisms that were examined in this work have a small percentage of their proteins adapted to salinity, which currently defies explanation.

The pangenome analysis of the class Halobacteria opens the way for novel research and the exploration of new genes, biological pathways, protein to protein interactions, metabolic systems, and undiscovered biological mechanisms of adaptation. The current dissertation provides novel data and tools regarding specific molecular signatures of haloadaptation and presents, for the first time, an extended view of the genetic repertoire of halophilicity.

References

1. Pearce, B.K.D., et al., *Constraining the Time Interval for the Origin of Life on Earth.* Astrobiology, 2018. **18**(3): p. 343-364.
2. Kasting, J.F., *Earth's early atmosphere.* Science, 1993. **259**(5097): p. 920-6.
3. Zahnle, K., L. Schaefer, and B. Fegley, *Earth's earliest atmospheres.* Cold Spring Harb Perspect Biol, 2010. **2**(10): p. a004895.
4. Falkowski, P.G., *The biological and geological contingencies for the rise of oxygen on Earth.* Photosynth Res, 2011. **107**(1): p. 7-10.
5. Lyons, T.W., C.T. Reinhard, and N.J. Planavsky, *The rise of oxygen in Earth's early ocean and atmosphere.* Nature, 2014. **506**(7488): p. 307-15.
6. Falkowski, P.G., T. Fenchel, and E.F. Delong, *The microbial engines that drive Earth's biogeochemical cycles.* Science, 2008. **320**(5879): p. 1034-9.
7. Pavlov, A.A., et al., *Greenhouse warming by CH4 in the atmosphere of early Earth.* J Geophys Res, 2000. **105**(E5): p. 11981-90.
8. Westall, F., et al., *Early Earth and early life: An extreme environment and extremophiles - Application to the search for life on Mars.* Proceedings of the Second European Workshop on Exo-Astrobiology, 2002. **518**: p. 131-136.
9. Madigan, M.T. and B.L. Marrs, *Extremophiles (vol 276, pg 82, 1997).* Scientific American, 1997. **277**(1): p. 8-8.
10. Dalmasso, C., et al., *Thermococcus piezophilus sp. nov., a novel hyperthermophilic and piezophilic archaeon with a broad pressure range for growth, isolated from a deepest hydrothermal vent at the Mid-Cayman Rise.* Syst Appl Microbiol, 2016. **39**(7): p. 440-444.
11. Loukas, A., I. Kappas, and T.J. Abatzopoulos, *HaloDom: a new database of halophiles across all life domains.* J Biol Res (Thessalon), 2018. **25**: p. 2.
12. Roberts, M.F., *Characterization of organic compatible solutes of halotolerant and halophilic microorganisms*, in *Methods in Microbiology, Extremophiles*, F.A. Rainey and A. Oren, Editors. 2006, Elsevier-Academic Press: Amsterdam.
13. Santos, H. and M.S. da Costa, *Compatible solutes of organisms that live in hot saline environments.* Environ Microbiol, 2002. **4**(9): p. 501-9.
14. Lai, M., et al., *Distribution of compatible solutes in the halophilic methanogenic archaebacteria.* J Bacteriol, 1991. **173**.
15. Shivanand, P. and G. Mugeraya, *Halophilic bacteria and their compatible solutes - osmoregulation and potential applications.* Current Science, 2011. **100**(10): p. 1516-1521.
16. Roberts, M.F., *Organic compatible solutes of halotolerant and halophilic microorganisms.* Saline Syst, 2005. **1**.
17. Empadinhas, N. and M.S. da Costa, *Osmoadaptation mechanisms in prokaryotes: distribution of compatible solutes.* Int Microbiol, 2008. **11**(3): p. 151-61.
18. da Costa, M.S., H. Santos, and E.A. Galinski, *An overview of the role and diversity of compatible solutes in Bacteria and Archaea.* Adv Biochem Eng Biotechnol, 1998. **61**: p. 117-53.
19. Mongodin, E.F., et al., *The genome of Salinibacter ruber: convergence and gene exchange among hyperhalophilic bacteria and archaea.* Proc Natl Acad Sci U S A, 2005. **102**(50): p. 18147-52.
20. Rhodes, M.E., et al., *Differences in lateral gene transfer in hypersaline versus thermal environments.* BMC Evol Biol, 2011. **11**: p. 199.
21. Fuchsman, C.A., et al., *Effect of the environment on horizontal gene transfer between bacteria and archaea.* Peerj, 2017. **5**.
22. Vannini, C., et al., *Sulphide oxidation to elemental sulphur in a membrane bioreactor: performance and characterization of the selected microbial sulphur-oxidizing community.* Syst Appl Microbiol, 2008. **31**(6-8): p. 461-73.
23. Oren, A., *Diversity of halophilic microorganisms: environments, phylogeny, physiology, and applications.* J Ind Microbiol Biotechnol, 2002. **28**.

24. Ollivier, B., et al., *Anaerobic-Bacteria from Hypersaline Environments.* Microbiological Reviews, 1994. **58**(1): p. 27-38.

25. Oren, A., *Bioenergetic aspects of halophilism.* Microbiol Mol Biol Rev, 1999. **63**.

26. Dumorne, K., et al., *Extremozymes: A Potential Source for Industrial Applications.* J Microbiol Biotechnol, 2017. **27**(4): p. 649-659.

27. Coker, J.A., *Extremophiles and biotechnology: current uses and prospects.* F1000Res, 2016. **5**.

28. Chandi, G.K. and B.S. Gill, *Production and Characterization of Microbial Carotenoids as an Alternative to Synthetic Colors: a Review.* International Journal of Food Properties, 2011. **14**(3): p. 503-513.

29. Charlesworth, J.C. and B.P. Burns, *Untapped Resources: Biotechnological Potential of Peptides and Secondary Metabolites in Archaea.* Archaea, 2015. **2015**: p. 282035.

30. Asker, D. and Y. Ohta, *Production of canthaxanthin by Haloferax alexandrinus under non-aseptic conditions and a simple, rapid method for its extraction.* Appl Microbiol Biotechnol, 2002. **58**(6): p. 743-50.

31. Eichler, J., *Biotechnological uses of archaeal extremozymes.* Biotechnol Adv, 2001. **19**(4): p. 261-78.

32. Stuart, E.S., et al., *Cassette-based presentation of SIV epitopes with recombinant gas vesicles from halophilic archaea.* J Biotechnol, 2004. **114**(3): p. 225-37.

33. Safarpour, A., et al., *Supernatant Metabolites from Halophilic Archaea to Reduce Tumorigenesis in Prostate Cancer In-vitro and In-vivo.* Iran J Pharm Res, 2019. **18**(1): p. 241-253.

34. Dassarma, S., *Extreme Halophiles Are Models for Astrobiology.* Microbe, 2006. **1**.

35. Merino, N., et al., *Living at the Extremes: Extremophiles and the Limits of Life in a Planetary Context (vol 10, pg 780, 2019).* Frontiers in Microbiology, 2019. **10**.

36. Thombre, R.S., P.A. Vaishampayan, and F. Gomez, *Chapter 7 - Applications of extremophiles in astrobiology*, in *Physiological and Biotechnological Aspects of Extremophiles*, R. Salwan and V. Sharma, Editors. 2020, Academic Press. p. 89-104.

37. Gunde-Cimerman, N., A. Oren, and A. Plemenitas, *Adaptation to life at high salt concentrations in Archaea, Bacteria, and Eukarya - Introduction.* Adaptation to Life at High Salt Concentrations in Archaea, Bacteria, and Eukarya, 2005. **9**: p. 1-6.

38. Oren, A., *Microbial life at high salt concentrations: phylogenetic and metabolic diversity.* Saline Systems, 2008. **4**(1): p. 1-13.

39. Banciu, H., et al., *Thialkalivibrio halophilussp. nov., a novel obligately chemolithoautotrophic, facultatively alkaliphilic, and extremely salt-tolerant, sulfur-oxidizing bacterium from a hypersaline alkaline lake.* Extremophiles, 2004. **8**.

40. Cho, B.C., *Heterotrophic flagellates in hypersaline waters*, in *Adaptation to Life at High Salt Concentrations in Archaea, Bacteria, and Eukarya*, N. Gunde-Cimerman, A. Oren, and A. Plemenitaš, Editors. 2005, Springer: Dordrecht.

41. Hauer, G. and A. Rogerson, *Heterotrophic protozoa from hypersaline environments*, in *Adaptation to Life at High Salt Concentrations in Archaea, Bacteria, and Eukarya*, N. Gunde-Cimerman, A. Oren, and A. Plemenitaš, Editors. 2005, Springer: Dordrecht.

42. Ali, I., et al., *Identification, phylogenetic analysis and characterization of obligate halophilic fungi isolated from a man-made solar saltern in Phetchaburi province, Thailand.* Annals of Microbiology, 2013. **63**(3): p. 887-895.

43. Roohi, A., et al., *Isolation and Phylogenetic Identification of Halotolerant/Halophilic Bacteria from the Salt Mines of Karak, Pakistan.* International Journal of Agriculture and Biology, 2014. **16**(3): p. 564-570.

44. Luo, X.X., et al., *Paraglycomyces xinjiangensis gen. nov., sp nov., a halophilic actinomycete.* International Journal of Systematic and Evolutionary Microbiology, 2015. **65**: p. 4263-4269.

45. Kim, S.J., et al., *Halobacillus salicampi sp nov., a moderately halophilic bacterium isolated from a solar saltern sediment.* Antonie Van Leeuwenhoek International Journal of General and Molecular Microbiology, 2016. **109**(5): p. 713-720.

46. Albuquerque, L., et al., *Halorhabdus rudnickae sp nov., a halophilic archaeon isolated from a salt mine borehole in Poland.* Systematic and Applied Microbiology, 2016. **39**(2): p. 100-105.

47. Dassarma, S.L., et al., *HaloWeb: the haloarchaeal genomes database.* Saline Systems, 2010. **6**: p. 12.

48. Ukani, H., et al., *HaloBase: development of database system for halophilic bacteria and archaea with respect to proteomics, genomics & other molecular traits.* Journal of Scientific & Industrial Research, 2011. **70**(11): p. 976-981.

49. Sharma, N., et al., *The Halophile protein database.* Database (Oxford), 2014. **2014**: p. bau114.

50. *XAMPP official webpage*. Available from: https://www.apachefriends.org/index.html.

51. *Phpmyadmin official webpage*. Available from: https://www.phpmyadmin.net/.

52. *NetBeans official webpage.*

53. Google. Available from: https://developers.google.com/chart/.

54. *Comprehensive Perl Archive Network*. Available from: https://www.cpan.org/.

55. *Google maps*. Available from: https://www.google.com/maps.

56. *Google analytics*. Available from: https://analytics.google.com/.

57. Kamekura, M., *Diversity of extremely halophilic bacteria.* Extremophiles, 1998. **2**(3): p. 289-295.

58. Paul, S., et al., *Molecular signature of hypersaline adaptation: insights from genome and proteome composition of halophilic prokaryotes.* Genome Biology, 2008. **9**(4).

59. Kastritis, P.L., N.C. Papandreou, and S.J. Hamodrakas, *Haloadaptation: insights from comparative modeling studies of halophilic archaeal DHFRs.* Int J Biol Macromol, 2007. **41**(4): p. 447-53.

60. Empadinhas, N. and M.S. da Costa, *Osmoadaptation mechanisms in prokaryotes: distribution of compatible solutes.* International Microbiology, 2008. **11**(3): p. 151-161.

61. Empadinhas, N. and M.S. da Costa, *To be or not to be a compatible solute: Bioversatility of mannosylglycerate and glucosylglycerate.* Systematic and Applied Microbiology, 2008. **31**(3): p. 159-168.

62. Ding, J.Y. and M.C. Lai, *The biotechnological potential of the extreme halophilic archaea Haloterrigena sp. H13 in xenobiotic metabolism using a comparative genomics approach.* Environ Technol, 2010. **31**(8-9): p. 905-14.

63. Munoz, R., et al., *Release LTPs104 of the All-Species Living Tree.* Syst Appl Microbiol, 2011. **34**(3): p. 169-70.

64. Clarke, C.J., et al., *Dryland salinity in south-western Australia: its origins, remedies, and future research directions.* Australian Journal of Soil Research, 2002. **40**(1): p. 93-113.

65. Bielanska-Grajner, I. and A. Cudak, *Effects of Salinity on Species Diversity of Rotifers in Anthropogenic Water Bodies.* Polish Journal of Environmental Studies, 2014. **23**(1): p. 27-34.

66. Cavicchioli, R., *Extremophiles and the search for extraterrestrial life.* Astrobiology, 2002. **2**(3): p. 281-292.

67. de Lorenzo, V., *Genes that move the window of viability of life: Lessons from bacteria thriving at the cold extreme Mesophiles can be turned into extremophiles by substituting essential genes.* Bioessays, 2011. **33**(1): p. 38-42.

68. Pace, N.R., *Time for a change.* Nature, 2006. **441**(7091): p. 289.

69. Brochier-Armanet, C., P. Forterre, and S. Gribaldo, *Phylogeny and evolution of the Archaea: one hundred genomes later.* Curr Opin Microbiol, 2011. **14**(3): p. 274-81.

70. Barreteau, H., et al., *Haloarcula sebkhae sp. nov., an extremely halophilic archaeon from Algerian hypersaline environment.* Int J Syst Evol Microbiol, 2019. **69**(3): p. 732-738.

71. Andrei, A.S., H.L. Banciu, and A. Oren, *Living with salt: metabolic and phylogenetic diversity of archaea inhabiting saline ecosystems.* FEMS Microbiol Lett, 2012. **330**(1): p. 1-9.

72. Zhu, Q., et al., *Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea.* Nat Commun, 2019. **10**(1): p. 5477.

73. Raymann, K., et al., *Global phylogenomic analysis disentangles the complex evolutionary history of DNA replication in archaea.* Genome Biol Evol, 2014. **6**(1): p. 192-212.

74. Borrel, G., et al., *Wide diversity of methane and short-chain alkane metabolisms in uncultured archaea.* Nat Microbiol, 2019. **4**(4): p. 603-613.

75. Dong, X., et al., *Metabolic potential of uncultured bacteria and archaea associated with petroleum seepage in deep-sea sediments.* Nat Commun, 2019. **10**(1): p. 1816.

76. Raymann, K., C. Brochier-Armanet, and S. Gribaldo, *The two-domain tree of life is linked to a new root for the Archaea.* Proc Natl Acad Sci U S A, 2015. **112**(21): p. 6670-5.

77. Villanueva, L., S. Schouten, and J.S.S. Damste, *Phylogenomic analysis of lipid biosynthetic genes of Archaea shed light on the "lipid divide'.* Environmental Microbiology, 2017. **19**(1): p. 54-69.

78. Becker, E.A., et al., *Phylogenetically driven sequencing of extremely halophilic archaea reveals strategies for static and dynamic osmo-response.* PLoS Genet, 2014. **10**(11): p. e1004784.

79. Gupta, R.S., et al., *A phylogenomic reappraisal of family-level divisions within the class Halobacteria: proposal to divide the order Halobacteriales into the families Halobacteriaceae, Haloarculaceae fam. nov., and Halococcaceae fam. nov., and the order Haloferacales into the families, Haloferacaceae and Halorubraceae fam nov.* Antonie Van Leeuwenhoek International Journal of General and Molecular Microbiology, 2016. **109**(4): p. 565-587.

80. Gupta, R.S., S. Naushad, and S. Baker, *Phylogenomic analyses and molecular signatures for the class Halobacteria and its two major clades: a proposal for division of the class Halobacteria into an emended order Halobacteriales and two new orders, Haloferacales ord. nov. and Natrialbales ord. nov., containing the novel families Haloferacaceae fam. nov. and Natrialbaceae fam. nov.* Int J Syst Evol Microbiol, 2015. **65**(Pt 3): p. 1050-1069.

81. Wu, D., G. Jospin, and J.A. Eisen, *Systematic identification of gene families for use as "markers" for phylogenetic and phylogeny-driven ecological studies of bacteria and archaea and their major subgroups.* PLoS One, 2013. **8**(10): p. e77033.

82. Wu, M. and A.J. Scott, *Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2.* Bioinformatics, 2012. **28**(7): p. 1033-4.

83. *HMMER official website*. Available from: http://www.hmmer.org/.

84. Katoh, K. and D.M. Standley, *MAFFT multiple sequence alignment software version 7: improvements in performance and usability.* Mol Biol Evol, 2013. **30**(4): p. 772-80.

85. Pedruzzi, I., et al., *HAMAP in 2015: updates to the protein family classification and annotation system.* Nucleic Acids Res, 2015. **43**(Database issue): p. D1064-70.

86. *Geneious*. Available from: https://www.geneious.com.

87. Trifinopoulos, J., et al., *W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis.* Nucleic Acids Res, 2016. **44**(W1): p. W232-5.

88. Hoang, D.T., et al., *UFBoot2: Improving the Ultrafast Bootstrap Approximation.* Mol Biol Evol, 2018. **35**(2): p. 518-522.

89. Mathrani, I.M., et al., *Methanohalophilus zhilinae sp. nov., an alkaliphilic, halophilic, methylotrophic methanogen.* Int J Syst Bacteriol, 1988. **38**(2): p. 139-42.

90. Paterek, J.R. and P.H. Smith, *Methanohalophilus-Mahii Gen-Nov, Sp-Nov, a Methylotrophic Halophilic Methanogen.* International Journal of Systematic Bacteriology, 1988. **38**(1): p. 122-123.

91. Yu, I.K. and F. Kawamura, *Halomethanococcus-Doii Gen-Nov Sp-Nov - an Obligately Halophilic Methanogenic Bacterium from Solar Salt Ponds.* Journal of General and Applied Microbiology, 1987. **33**(4): p. 303-310.

92. Narasingarao, P., et al., *De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities.* ISME J, 2012. **6**(1): p. 81-93.

93. Hamm, J.N., et al., *Unexpected host dependency of Antarctic Nanohaloarchaeota.* Proceedings of the National Academy of Sciences of the United States of America, 2019. **116**(29): p. 14661-14670.

94. La Cono, V., et al., *Symbiosis between nanohaloarchaeon and haloarchaeon is based on utilization of different polysaccharides.* Proceedings of the National Academy of Sciences of the United States of America, 2020. **117**(33): p. 20223-20234.

95. Makhdoumi-Kakhki, A., M.A. Amoozegar, and A. Ventosa, *Salinibacter iranicus sp. nov. and Salinibacter luteus sp. nov., isolated from a salt lake, and emended descriptions of the genus Salinibacter and of Salinibacter ruber.* Int J Syst Evol Microbiol, 2012. **62**(Pt 7): p. 1521-1527.

96. Anton, J., et al., *Salinibacter ruber gen. nov., sp. nov., a novel, extremely halophilic member of the Bacteria from saltern crystallizer ponds.* Int J Syst Evol Microbiol, 2002. **52**(Pt 2): p. 485-91.

97. Munoz, R., R. Rossello-Mora, and R. Amann, *Revised phylogeny of Bacteroidetes and proposal of sixteen new taxa and two new combinations including Rhodothermaeota phyl. nov.* Syst Appl Microbiol, 2016. **39**(5): p. 281-96.

98. Kreil, D.P. and C.A. Ouzounis, *Identification of thermophilic species by the amino acid compositions deduced from their genomes.* Nucleic Acids Res, 2001. **29**(7): p. 1608-15.

99. Reed, C.J., et al., *Protein adaptations in archaeal extremophiles.* Archaea, 2013. **2013**: p. 373275.

100. Gribaldo, S. and C. Brochier-Armanet, *The origin and evolution of Archaea: a state of the art.* Philos Trans R Soc Lond B Biol Sci, 2006. **361**(1470): p. 1007-22.

101. Trent, J.D., *Extremophiles in astrobiology: per Ardua ad Astra.* Gravit Space Biol Bull, 2000. **13**(2): p. 5-11.

102. Schiraldi, C. and M. De Rosa, *The production of biocatalysts and biomolecules from extremophiles.* Trends Biotechnol, 2002. **20**(12): p. 515-21.

103. Yin, J., et al., *Halophiles, coming stars for industrial biotechnology.* Biotechnol Adv, 2015. **33**(7): p. 1433-42.

104. Gudhka, R.K., B.A. Neilan, and B.P. Burns, *Adaptation, ecology, and evolution of the halophilic stromatolite archaeon Halococcus hamelinensis inferred through genome analyses.* Archaea, 2015. **2015**: p. 241608.

105. Albokari, M., et al., *Niche for high abundant extremophilic microbial communities in an ancient crater.* International Journal of Astrobiology, 2018. **17**(4): p. 345-355.

106. Orange, F., et al., *Experimental silicification of the extremophilic Archaea Pyrococcus abyssi and Methanocaldococcus jannaschii: applications in the search for evidence of life in early Earth and extraterrestrial rocks.* Geobiology, 2009. **7**(4): p. 403-18.

107. Baltscheffsky, H. and B. Persson, *On an early gene for membrane-integral inorganic pyrophosphatase in the genome of an apparently pre-luca extremophile, the archaeon Candidatus Korarchaeum cryptofilum.* J Mol Evol, 2014. **78**(2): p. 140-7.

108. Martin, W.F., S. Garg, and V. Zimorski, *Endosymbiotic theories for eukaryote origin.* Philos Trans R Soc Lond B Biol Sci, 2015. **370**(1678): p. 20140330.

109. Irwin, J.A., *Extremophiles and their application to veterinary medicine.* Environ Technol, 2010. **31**(8-9): p. 857-69.

110. Kruger, A., et al., *Towards a sustainable biobased industry - Highlighting the impact of extremophiles.* N Biotechnol, 2018. **40**(Pt A): p. 144-153.

111. Fukuchi, S., et al., *Unique amino acid composition of proteins in halophilic bacteria.* J Mol Biol, 2003. **327**(2): p. 347-57.

112. Rhodes, M.E., et al., *Amino acid signatures of salinity on an environmental scale with a focus on the Dead Sea.* Environ Microbiol, 2010. **12**(9): p. 2613-23.

113. Nath, A., *Insights into the sequence parameters for halophilic adaptation.* Amino Acids, 2016. **48**(3): p. 751-762.

114. Dassarma, S., et al., *Amino acid substitutions in cold-adapted proteins from Halorubrum lacusprofundi, an extremely halophilic microbe from antarctica.* PLoS One, 2013. **8**(3): p. e58587.

115. Metpally, R.P. and B.V. Reddy, *Comparative proteome analysis of psychrophilic versus mesophilic bacterial species: Insights into the molecular basis of cold adaptation of proteins.* BMC Genomics, 2009. **10**: p. 11.

116. Nath, A., R. Chaube, and S. Karthikeyan, *Discrimination of Psychrophilic and Mesophilic Proteins Using Random Forest Algorithm.* 2012 International Conference on Biomedical Engineering and Biotechnology, 2012: p. 179-182.

117. Michoud, G. and M. Jebbar, *High hydrostatic pressure adaptive strategies in an obligate piezophile Pyrococcus yayanosii.* Sci Rep, 2016. **6**: p. 27289.

118. Zhang, G. and H. Ge, *Protein hypersaline adaptation: insight from amino acids with machine learning algorithms.* Protein J, 2013. **32**(4): p. 239-45.

119. Zhu, W., A. Lomsadze, and M. Borodovsky, *Ab initio gene identification in metagenomic sequences.* Nucleic Acids Res, 2010. **38**(12): p. e132.

120. Grant, W.D., et al., *Class III Halobacteriaclass. nov*, in *Bergey's Manual of Systematic Bacteriology*, D.R. Boone, R.W. Castenholz, and G.M. Garrity, Editors. 2001, Springer: New York.

121. Caraux, G. and S. Pinloche, *PermutMatrix: a graphical environment to arrange gene expression profiles in optimal linear order.* Bioinformatics, 2005. **21**(7): p. 1280-1.

122. Stajich, J.E., et al., *The Bioperl toolkit: Perl modules for the life sciences.* Genome Res, 2002. **12**(10): p. 1611-8.

123. Maestrojuan, G.M., et al., *Transfer of Methanogenium-Bourgense, Methanogenium-Marisnigri, Methanogenium-Olentangyi, and Methanogenium-Thermophilicum to the Genus Methanoculleus Gen-Nov, Emendation of Methanoculleus-Marisnigri and Methanogenium, and Description of New Strains of Methanoculleus-Bourgense and Methanoculleus-Marisnigri.* International Journal of Systematic Bacteriology, 1990. **40**(2): p. 117-122.

124. Anderson, I.J., et al., *Complete genome sequence of Methanoculleus marisnigri Romesser et al. 1981 type strain JR1.* Stand Genomic Sci, 2009. **1**(2): p. 189-96.

125. Elevi Bardavid, R. and A. Oren, *The amino acid composition of proteins from anaerobic halophilic bacteria of the order Halanaerobiales.* Extremophiles, 2012. **16**(3): p. 567-72.

126. Hamm, J.N., et al., *Unexpected host dependency of Antarctic Nanohaloarchaeota.* Proc Natl Acad Sci U S A, 2019. **116**(29): p. 14661-14670.

127. Aouad, M., et al., *Extreme halophilic archaea derive from two distinct methanogen Class II lineages.* Mol Phylogenet Evol, 2018. **127**: p. 46-54.

128. Oren, A., *Intracellular Salt Concentrations of the Anaerobic Halophilic Eubacteria Haloanaerobium-Praevalens and Halobacteroides-Halobius.* Canadian Journal of Microbiology, 1986. **32**(1): p. 4-9.

129. Oren, A., M. Heldal, and S. Norland, *X-ray microanalysis of intracellular ions in the anaerobic halophilic eubacterium Haloanaerobium praevalens.* Canadian Journal of Microbiology, 1997. **43**(6): p. 588-592.

130. Oren, A., *Life at high salt concentrations, intracellular KCl concentrations, and acidic proteomes.* Front Microbiol, 2013. **4**: p. 315.

131. Klipcan, L., et al., *Optimal growth temperature of prokaryotes correlates with class II amino acid composition.* FEBS Lett, 2006. **580**(6): p. 1672-6.

132. Nakashima, H. and K. Nishikawa, *Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies.* J Mol Biol, 1994. **238**(1): p. 54-61.

133. Zhang, Y., et al., *Quantitative Proteomics Reveals Membrane Protein-Mediated Hypersaline Sensitivity and Adaptation in Halophilic Nocardiopsis xinjiangensis.* J Proteome Res, 2016. **15**(1): p. 68-85.

134. Rascovan, N., et al., *Metagenomic study of red biofilms from Diamante Lake reveals ancient arsenic bioenergetics in haloarchaea.* ISME J, 2016. **10**(2): p. 299-309.

135. Spang, A. and T.J.G. Ettema, *The methanogenic roots of Archaea.* Nature Microbiology, 2017. **2**(8).

136. Kelly, S., B. Wickstead, and K. Gull, *Archaeal phylogenomics provides evidence in support of a methanogenic origin of the Archaea and a thaumarchaeal origin for the eukaryotes.* Proc Biol Sci, 2011. **278**(1708): p. 1009-18.

137. Bapteste, E., C. Brochier, and Y. Boucher, *Higher-level classification of the Archaea: evolution of methanogenesis and methanogens.* Archaea, 2005. **1**(5): p. 353-63.

138. Collevatti, R.G., et al., *A genome-wide scan shows evidence for local adaptation in a widespread keystone Neotropical forest tree.* Heredity, 2019. **123**(2): p. 117-137.

139. Kozma, R., P. Rodin-Morch, and J. Hoglund, *Genomic regions of speciation and adaptation among three species of grouse.* Scientific Reports, 2019. **9**.

140. Fu, W. and J.M. Akey, *Selection and adaptation in the human genome.* Annu Rev Genomics Hum Genet, 2013. **14**: p. 467-89.

141. Zeldovich, K.B., I.N. Berezovsky, and E.I. Shakhnovich, *Protein and DNA sequence determinants of thermophilic adaptation.* PLoS Comput Biol, 2007. **3**(1): p. e5.

142. Gianese, G., P. Argos, and S. Pascarella, *Structural adaptation of enzymes to low temperatures.* Protein Eng, 2001. **14**(3): p. 141-8.

143. Schopf, J.W., *Fossil evidence of Archaean life.* Philos Trans R Soc Lond B Biol Sci, 2006. **361**(1470): p. 869-85.

144. Orellana, R., et al., *Living at the Frontiers of Life: Extremophiles in Chile and Their Potential for Bioremediation.* Frontiers in Microbiology, 2018. **9**.

145. Oren, A., *Industrial and environmental applications of halophilic microorganisms.* Environ Technol, 2010. **31**(8-9): p. 825-34.

146. Elshahed, M.S., et al., *Haloferax sulfurifontissp. nov., a halophilic archaeon isolated from a sulfide and sulfur-rich spring.* Int J Syst Evol Microbiol, 2004. **54**.

147. Savage, K.N., et al., *Haladaptatus paucihalophilusgen. nov., sp. nov., a halophilic archaeon isolated from a low-salt, high-sulfide spring.* Int J Syst Evol Microbiol, 2007. **57**.

148. Paul, S., et al., *Molecular signature of hypersaline adaptation: insights from genome and proteome composition of halophilic prokaryotes.* Genome Biol, 2008. **9**(4): p. R70.

149. *GenBank and WGS Statistics - NCBI.* Available from: https://www.ncbi.nlm.nih.gov/genbank/statistics/.

150. Morgante, M., E. De Paoli, and S. Radovic, *Transposable elements and the plant pan-genomes.* Curr Opin Plant Biol, 2007. **10**(2): p. 149-55.

151. Wu, Y., N. Zaiden, and B. Cao, *The Core- and Pan-Genomic Analyses of the Genus Comamonas: From Environmental Adaptation to Potential Virulence.* Front Microbiol, 2018. **9**: p. 3096.

152. Gao, L., et al., *The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor.* Nat Genet, 2019. **51**(6): p. 1044-1051.

153. Psomopoulos, F.E., et al., *The chlamydiales pangenome revisited: structural stability and functional coherence.* Genes (Basel), 2012. **3**(2): p. 291-319.

154. Contreras-Moreira, B., et al., *Analysis of Plant Pan-Genomes and Transcriptomes with GET_HOMOLOGUES-EST, a Clustering Solution for Sequences of the Same Species.* Front Plant Sci, 2017. **8**: p. 184.

155. Fouts, D.E., et al., *PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species.* Nucleic Acids Res, 2012. **40**(22): p. e172.

156. Chaudhari, N.M., V.K. Gupta, and C. Dutta, *BPGA- an ultra-fast pan-genome analysis pipeline.* Sci Rep, 2016. **6**: p. 24373.

157. Pantoja, Y., et al., *PanWeb: A web interface for pan-genomic analysis.* PLoS One, 2017. **12**(5): p. e0178154.

158. Page, A.J., et al., *Roary: rapid large-scale prokaryote pan genome analysis.* Bioinformatics, 2015. **31**(22): p. 3691-3.

159. Xiao, J., et al., *A brief review of software tools for pangenomics.* Genomics Proteomics Bioinformatics, 2015. **13**(1): p. 73-6.

160. Clarke, T.H., et al., *PanACEA: a bioinformatics tool for the exploration and visualization of bacterial pan-chromosomes.* BMC Bioinformatics, 2018. **19**(1): p. 246.

161. Snipen, L. and K.H. Liland, *micropan: an R-package for microbial pan-genomics.* BMC Bioinformatics, 2015. **16**: p. 79.

162. Vernikos, G.S., *A Review of Pangenome Tools and Recent Studies*, in *The Pangenome: Diversity, Dynamics and Evolution of Genomes*, H. Tettelin and D. Medini, Editors. 2020: Cham (CH). p. 89-112.

163. Gordon, S.P., et al., *Extensive gene content variation in the Brachypodium distachyon pan-genome correlates with population structure.* Nat Commun, 2017. **8**(1): p. 2184.

164. Livingstone, P.G., R.M. Morphew, and D.E. Whitworth, *Genome Sequencing and Pan-Genome Analysis of 23 Corallococcus spp. Strains Reveal Unexpected Diversity, With Particular Plasticity of Predatory Gene Sets.* Front Microbiol, 2018. **9**: p. 3187.

165. Udaondo, Z., E. Duque, and J.L. Ramos, *The pangenome of the genus Clostridium.* Environ Microbiol, 2017. **19**(7): p. 2588-2603.

166. Deschamps, P., et al., *Pangenome evidence for extensive interdomain horizontal transfer affecting lineage core and shell genes in uncultured planktonic thaumarchaeota and euryarchaeota.* Genome Biol Evol, 2014. **6**(7): p. 1549-63.
167. Kim, Y.B., et al., *Haloplanus rubicundus sp. nov., an extremely halophilic archaeon isolated from solar salt.* Syst Appl Microbiol, 2020. **43**(3): p. 126085.
168. Aherfi, S., et al., *A Large Open Pangenome and a Small Core Genome for Giant Pandoraviruses.* Front Microbiol, 2018. **9**: p. 1486.
169. Wang, L., et al., *Comparative genomic analysis reveals an 'open' pan-genome of African swine fever virus.* Transbound Emerg Dis, 2020. **67**(4): p. 1553-1562.
170. Parlikar, A., et al., *Understanding genomic diversity, pan-genome, and evolution of SARS-CoV-2.* PeerJ, 2020. **8**: p. e9576.
171. Ma, Y., et al., *Halophiles 2010: life in saline environments.* Appl Environ Microbiol, 2010. **76**(21): p. 6971-81.
172. Hyatt, D., et al., *Prodigal: prokaryotic gene recognition and translation initiation site identification.* BMC Bioinformatics, 2010. **11**: p. 119.
173. *RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL* Available from: http://www.rstudio.com/.
174. *Perl programming language*. Available from: https://www.perl.org/.
175. Huerta-Cepas, J., et al., *eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses.* Nucleic Acids Research, 2019. **47**(D1): p. D309-D314.
176. Huerta-Cepas, J., et al., *Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper.* Molecular Biology and Evolution, 2017. **34**(8): p. 2115-2122.
177. UniProt, C., *UniProt: the universal protein knowledgebase in 2021.* Nucleic Acids Res, 2021. **49**(D1): p. D480-D489.
178. Marroni, F., S. Pinosio, and M. Morgante, *Structural variation and genome complexity: is dispensable really dispensable?* Curr Opin Plant Biol, 2014. **18**: p. 31-6.
179. Pal, K.K., et al., *Draft Genome Sequence of an Extreme Haloarchaeon 3A1-DGR Isolated from a Saltern Crystallizer of the Little Rann of Kutch, India.* Indian J Microbiol, 2014. **54**(4): p. 471-3.
180. Dominova, I.N., et al., *Complete Genome Sequence of Salinarchaeum sp. Strain HArcht-Bsk1T, Isolated from Hypersaline Lake Baskunchak, Russia.* Genome Announc, 2013. **1**(4).
181. Kushwaha, S.C. and M. Kates, *Isolation and Identification of Bacteriorhodopsin and Minor-C40-Carotenoids in Halobacterium-Cutirubrum.* Biochimica Et Biophysica Acta, 1973. **316**(2): p. 235-243.
182. Bibikov, S.I., et al., *Bacteriorhodopsin is involved in halobacterial photoreception.* Proc Natl Acad Sci U S A, 1993. **90**(20): p. 9446-50.
183. Bibikov, S.I., et al., *The proton pump bacteriorhodopsin is a photoreceptor for signal transduction in Halobacterium halobium.* FEBS Lett, 1991. **295**(1-3): p. 223-6.
184. Rodrigo-Banos, M., et al., *Carotenoids from Haloarchaea and Their Potential in Biotechnology.* Mar Drugs, 2015. **13**(9): p. 5508-32.
185. Gunde-Cimerman, N., A. Plemenitas, and A. Oren, *Strategies of adaptation of microorganisms of the three domains of life to high salt concentrations.* FEMS Microbiol Rev, 2018. **42**(3): p. 353-375.
186. Argandona, M., et al., *Interplay between Iron Homeostasis and the Osmotic Stress Response in the Halophilic Bacterium Chromohalobacter salexigens.* Applied and Environmental Microbiology, 2010. **76**(11): p. 3575-3589.
187. Muller, V., R. Spanheimer, and H. Santos, *Stress response by solute accumulation in archaea.* Curr Opin Microbiol, 2005. **8**(6): p. 729-36.
188. Makarova, K.S., M.Y. Galperin, and E.V. Koonin, *Proposed Role for KaiC-Like ATPases as Major Signal Transduction Hubs in Archaea.* mBio, 2017. **8**(6).
189. Rudolph, J., et al., *Phosphorylation in Halobacterial Signal-Transduction.* Embo Journal, 1995. **14**(17): p. 4249-4257.

190. Zhang, W., et al., *Signal transduction in the archaeon Halobacterium salinarium is processed through three subfamilies of 13 soluble and membrane-bound transducer proteins.* Proc Natl Acad Sci U S A, 1996. **93**(10): p. 4649-54.

191. Schlesner, M., et al., *The protein interaction network of a taxis signal transduction system in a Halophilic Archaeon.* Bmc Microbiology, 2012. **12**.

192. Martin-Cuadrado, A.B., L. Pasic, and F. Rodriguez-Valera, *Diversity of the cell-wall associated genomic island of the archaeon Haloquadratum walsbyi.* BMC Genomics, 2015. **16**: p. 603.

193. Yoon, J.H., et al., *Halobacillus campisalis sp. nov., containing meso-diaminopimelic acid in the cell-wall peptidoglycan, and emended description of the genus Halobacillus.* Int J Syst Evol Microbiol, 2007. **57**(Pt 9): p. 2021-2025.

194. Niemetz, R., et al., *The cell wall polymer of the extremely halophilic archaeon Natronococcus occultus.* Eur J Biochem, 1997. **249**(3): p. 905-11.

195. French, S.L., et al., *Transcription and translation are coupled in Archaea.* Mol Biol Evol, 2007. **24**(4): p. 893-5.

196. Burmann, B.M., et al., *A NusE:NusG complex links transcription and translation.* Science, 2010. **328**(5977): p. 501-4.

197. Robinson, N.P., et al., *The cutting edge of archaeal transcription.* Emerging Topics in Life Sciences, 2018. **2**(4): p. 517-533.

198. Santangelo, T.J., et al., *Polarity in archaeal operon transcription in Thermococcus kodakaraensis.* J Bacteriol, 2008. **190**(6): p. 2244-8.

199. Karlin, S., et al., *Predicted highly expressed genes in archaeal genomes.* Proc Natl Acad Sci U S A, 2005. **102**(20): p. 7303-8.

200. Carbone, V., et al., *Structure and Evolution of the Archaeal Lipid Synthesis Enzyme sn-Glycerol-1-phosphate Dehydrogenase.* J Biol Chem, 2015. **290**(35): p. 21690-704.

201. Dibrova, D.V., M.Y. Galperin, and A.Y. Mulkidjanian, *Phylogenomic reconstruction of archaeal fatty acid metabolism.* Environ Microbiol, 2014. **16**(4): p. 907-18.

202. Hamerly, T., et al., *Characterization of Fatty Acids in Crenarchaeota by GC-MS and NMR.* Archaea, 2015. **2015**: p. 472726.

203. Klingl, A., *S-layer and cytoplasmic membrane - exceptions from the typical archaeal cell wall with a focus on double membranes.* Front Microbiol, 2014. **5**: p. 624.

204. Jordan, S.F., E. Nee, and N. Lane, *Isoprenoids enhance the stability of fatty acid membranes at the emergence of life potentially leading to an early lipid divide.* Interface Focus, 2019. **9**(6).

205. Gruber, G., et al., *ATP synthases from archaea: the beauty of a molecular motor.* Biochim Biophys Acta, 2014. **1837**(6): p. 940-52.

206. Ghosh, A., et al., *Archaeal flagellar ATPase motor shows ATP-dependent hexameric assembly and activity stimulation by specific lipid binding.* Biochem J, 2011. **437**(1): p. 43-52.

207. Ng, W.V., et al., *Genome sequence of Halobacterium species NRC-1.* Proc Natl Acad Sci U S A, 2000. **97**(22): p. 12176-81.

208. Bleiholder, A., et al., *Expression of multiple tfb genes in different Halobacterium salinarum strains and interaction of TFB with transcriptional activator GvpE.* Arch Microbiol, 2012. **194**(4): p. 269-79.

209. Ghanmi, F., et al., *The extremely halophilic archaeon Halobacterium salinarum ETD5 from the solar saltern of Sfax (Tunisia) produces multiple halocins.* Res Microbiol, 2020. **171**(2): p. 80-90.

210. Coker, J.A. and S. DasSarma, *Genetic and transcriptomic analysis of transcription factor genes in the model halophilic Archaeon: coordinate action of TbpD and TfbA.* BMC Genet, 2007. **8**: p. 61.

211. Pfeifer, F., et al., *Transposable Elements of Halobacterium-Halobium.* Molecular and General Genetics, 1983. **191**(2): p. 182-188.

212. DasSarma, S., U.L. RajBhandary, and H.G. Khorana, *High-frequency spontaneous mutation in the bacterio-opsin gene in Halobacterium halobium is mediated by transposable elements.* Proc Natl Acad Sci U S A, 1983. **80**(8): p. 2201-5.

213. Chen, Y., et al., *MUST: A system for identification of miniature inverted-repeat transposable elements and applications to Anabaena variabilis and Haloquadratum walsbyi.* Gene, 2009. **436**(1-2): p. 1-7.

214. Kiljunen, S., et al., *Generation of comprehensive transposon insertion mutant library for the model archaeon, Haloferax volcanii, and its use for gene discovery.* Bmc Biology, 2014. **12**.

215. Woods, W.G., K. Ngui, and M.L. Dyall-Smith, *An improved transposon for the halophilic archaeon Haloarcula hispanica.* Journal of Bacteriology, 1999. **181**(22): p. 7140-7142.

216. Falb, M., et al., *Metabolism of halophilic archaea.* Extremophiles, 2008. **12**(2): p. 177-96.

217. Brasen, C., et al., *Carbohydrate Metabolism in Archaea: Current Insights into Unusual Enzymes and Pathways and Their Regulation.* Microbiology and Molecular Biology Reviews, 2014. **78**(1): p. 89-175.

218. Oren, A. and P. Gurevich, *Diversity of lactate metabolism in halophilic archaea.* Can J Microbiol, 1995. **41**(3): p. 302-7.

219. Oren, A. and A. Ventosa, *Natrialbales*, in *Bergey's Manual of Systematics of Archaea and Bacteria*. 2017. p. 1-2.

220. Oren, A., et al., *Haloarcula quadrata sp. nov., a square, motile archaeon isolated from a brine pool in Sinai (Egypt).* Int J Syst Bacteriol, 1999. **49 Pt 3**: p. 1149-55.

221. Ihara, K., S. Watanabe, and T. Tamura, *Haloarcula argentinensis sp. nov. and Haloarcula mukohataei sp. nov., two new extremely halophilic archaea collected in Argentina.* Int J Syst Bacteriol, 1997. **47**(1): p. 73-7.

222. Oren, A., et al., *Halomicrobium mukohataei gen. nov., comb. nov., and emended description of Halomicrobium mukohataei.* Int J Syst Evol Microbiol, 2002. **52**(Pt 5): p. 1831-1835.

223. Cui, H.L., et al., *Intraspecific polymorphism of 16S rRNA genes in two halophilic archaeal genera, Haloarcula and Halomicrobium.* Extremophiles, 2009. **13**(1): p. 31-7.

224. Ghai, R., et al., *New abundant microbial groups in aquatic hypersaline environments.* Sci Rep, 2011. **1**: p. 135.

225. Arachchige, M.S.A., S. Yoshida, and H. Toyama, *Thermo-and salt-tolerant Saccharomyces cerevisiae strains isolated from fermenting coconut toddy from Sri Lanka.* Biotechnology & Biotechnological Equipment, 2019. **33**(1): p. 937-944.

226. Gostincar, C., et al., *Fungal adaptation to extremely high salt concentrations.* Adv Appl Microbiol, 2011. **77**: p. 71-96.

227. Plemenitas, A., et al., *Adaptation to high salt concentrations in halotolerant/halophilic fungi: a molecular perspective.* Front Microbiol, 2014. **5**: p. 199.

Appendix

Table A1. List of the species and proteomes used for constructing the phylogenetic trees of Figures 3.1 and 3.2. Ingroups in bold.

Πίνακας A1. Λίστα των ειδών των οποίων τα πρωτεώματα χρησιμοποιήθηκαν για την κατασκευή των φυλογενετικών δέντρων των Εικόνων 3.1 και 3.2. Οι εσωομάδες δίνονται με έντονη γραμματοσειρά.

No	Species	Kingdom	Phylum	Class
1	*Archaeoglobus fulgidus*	Archaea	Euryarchaeota	Archaeoglobi
2	*Archaeoglobus veneficus*	Archaea	Euryarchaeota	Archaeoglobi
3	*Candidatus Bathyarchaeota archaeon BA2*	Archaea	Candidatus Bathyarchaeota	undetermined
4	*Candidatus Caldiarchaeum subterraneum*	Archaea	Thaumarchaeota	undetermined
5	*Candidatus Haloredivivus sp. G17*	Archaea	Euryarchaeota	Nanohaloarchaea
6	*Candidatus Korarchaeum cryptofilum*	Archaea	Candidatus Korarchaeota	undetermined
7	*Candidatus Nanosalina sp. J07AB43*	Archaea	Euryarchaeota	Nanohaloarchaea
8	*Candidatus Nanosalinarum sp. J07AB56*	Archaea	Euryarchaeota	Nanohaloarchaea
9	*Candidatus Syntrophoarchaeum butanivorans*	Archaea	Euryarchaeota	Methanomicrobia
10	*Candidatus Syntrophoarchaeum caldarius*	Archaea	Euryarchaeota	Methanomicrobia
11	*Candidatus Thorarchaeota archaeon AB 25*	Archaea	Candidatus Thorarchaeota	undetermined
12	*Cenarchaeum symbiosum*	Archaea	Thaumarchaeota	undetermined
13	***Haladaptatus cibarius***	Archaea	Euryarchaeota	Halobacteria
14	***Haladaptatus litoreus***	Archaea	Euryarchaeota	Halobacteria
15	***Haladaptatus paucihalophilus***	Archaea	Euryarchaeota	Halobacteria
16	***Halalkalicoccus jeotgali***	Archaea	Euryarchaeota	Halobacteria
17	***Halalkalicoccus paucihalophilus***	Archaea	Euryarchaeota	Halobacteria
18	***Halanaeroarchaeum sulfurireducens***	Archaea	Euryarchaeota	Halobacteria
19	***Halapricum salinum***	Archaea	Euryarchaeota	Halobacteria
20	***Halarchaeum acidiphilum***	Archaea	Euryarchaeota	Halobacteria
21	***Haloarchaeobius iranensis***	Archaea	Euryarchaeota	Halobacteria
22	***Haloarcula amylolytica***	Archaea	Euryarchaeota	Halobacteria
23	***Haloarcula argentinensis***	Archaea	Euryarchaeota	Halobacteria
24	***Haloarcula hispanica***	Archaea	Euryarchaeota	Halobacteria
25	***Haloarcula japonica***	Archaea	Euryarchaeota	Halobacteria
26	***Haloarcula marismortui***	Archaea	Euryarchaeota	Halobacteria
27	***Haloarcula sp. CBA1115***	Archaea	Euryarchaeota	Halobacteria
28	***Haloarcula vallismortis***	Archaea	Euryarchaeota	Halobacteria
29	***Halobacterium jilantaiense***	Archaea	Euryarchaeota	Halobacteria
30	***Halobacterium salinarum***	Archaea	Euryarchaeota	Halobacteria
31	***Halobaculum gomorrense***	Archaea	Euryarchaeota	Halobacteria
32	***Halobellus clavatus***	Archaea	Euryarchaeota	Halobacteria
33	***Halobellus rufus***	Archaea	Euryarchaeota	Halobacteria
34	***Halobiforma haloterrestris***	Archaea	Euryarchaeota	Halobacteria
35	***Halobiforma lacisalsi***	Archaea	Euryarchaeota	Halobacteria

36	***Halobiforma nitratireducens***	Archaea	Euryarchaeota	Halobacteria
37	***Halococcus agarilyticus***	Archaea	Euryarchaeota	Halobacteria
38	***Halococcus hamelinensis***	Archaea	Euryarchaeota	Halobacteria
39	***Halococcus morrhuae***	Archaea	Euryarchaeota	Halobacteria
40	***Halococcus saccharolyticus***	Archaea	Euryarchaeota	Halobacteria
41	***Halococcus salifodinae***	Archaea	Euryarchaeota	Halobacteria
42	***Halococcus sediminicola***	Archaea	Euryarchaeota	Halobacteria
43	***Halococcus thailandensis***	Archaea	Euryarchaeota	Halobacteria
44	***Haloferax denitrificans***	Archaea	Euryarchaeota	Halobacteria
45	***Haloferax elongans***	Archaea	Euryarchaeota	Halobacteria
46	***Haloferax gibbonsii***	Archaea	Euryarchaeota	Halobacteria
47	***Haloferax larsenii***	Archaea	Euryarchaeota	Halobacteria
48	***Haloferax lucentense***	Archaea	Euryarchaeota	Halobacteria
49	***Haloferax mediterranei***	Archaea	Euryarchaeota	Halobacteria
50	***Haloferax mucosum***	Archaea	Euryarchaeota	Halobacteria
51	***Haloferax prahovense***	Archaea	Euryarchaeota	Halobacteria
52	***Haloferax sulfurifontis***	Archaea	Euryarchaeota	Halobacteria
53	***Haloferax volcanii***	Archaea	Euryarchaeota	Halobacteria
54	***Halogeometricum borinquense***	Archaea	Euryarchaeota	Halobacteria
55	***Halogeometricum limi***	Archaea	Euryarchaeota	Halobacteria
56	***Halogeometricum pallidum***	Archaea	Euryarchaeota	Halobacteria
57	***Halogeometricum rufum***	Archaea	Euryarchaeota	Halobacteria
58	***Halogranum amylolyticum***	Archaea	Euryarchaeota	Halobacteria
59	***Halogranum gelatinilyticum***	Archaea	Euryarchaeota	Halobacteria
60	***Halogranum rubrum***	Archaea	Euryarchaeota	Halobacteria
61	***Halogranum salarium***	Archaea	Euryarchaeota	Halobacteria
62	***Halohasta litchfieldiae***	Archaea	Euryarchaeota	Halobacteria
63	***Halolamina pelagica***	Archaea	Euryarchaeota	Halobacteria
64	***Halolamina rubra***	Archaea	Euryarchaeota	Halobacteria
65	***Halolamina sediminis***	Archaea	Euryarchaeota	Halobacteria
66	***Halomicrobium katesii***	Archaea	Euryarchaeota	Halobacteria
67	***Halomicrobium mukohataei***	Archaea	Euryarchaeota	Halobacteria
68	***Halomicrobium zhouii***	Archaea	Euryarchaeota	Halobacteria
69	***Halopelagius inordinatus***	Archaea	Euryarchaeota	Halobacteria
70	***Halopelagius longus***	Archaea	Euryarchaeota	Halobacteria
71	***Halopenitus malekzadehii***	Archaea	Euryarchaeota	Halobacteria
72	***Halopenitus persicus***	Archaea	Euryarchaeota	Halobacteria
73	***Halopiger salifodinae***	Archaea	Euryarchaeota	Halobacteria
74	***Halopiger xanaduensis***	Archaea	Euryarchaeota	Halobacteria
75	***Haloplanus natans***	Archaea	Euryarchaeota	Halobacteria
76	***Haloplanus vescus***	Archaea	Euryarchaeota	Halobacteria
77	***Haloquadratum walsbyi***	Archaea	Euryarchaeota	Halobacteria
78	***Halorhabdus tiamatea***	Archaea	Euryarchaeota	Halobacteria
79	***Halorhabdus utahensis***	Archaea	Euryarchaeota	Halobacteria

80	***Halorientalis persicus***	Archaea	Euryarchaeota	Halobacteria
81	***Halorientalis regularis***	Archaea	Euryarchaeota	Halobacteria
82	***Halorubrum aidingense***	Archaea	Euryarchaeota	Halobacteria
83	***Halorubrum arcis***	Archaea	Euryarchaeota	Halobacteria
84	***Halorubrum californiense***	Archaea	Euryarchaeota	Halobacteria
85	***Halorubrum coriense***	Archaea	Euryarchaeota	Halobacteria
86	***Halorubrum distributum***	Archaea	Euryarchaeota	Halobacteria
87	***Halorubrum ezzemoulense***	Archaea	Euryarchaeota	Halobacteria
88	***Halorubrum halophilum***	Archaea	Euryarchaeota	Halobacteria
89	***Halorubrum kocurii***	Archaea	Euryarchaeota	Halobacteria
90	***Halorubrum lacusprofundi***	Archaea	Euryarchaeota	Halobacteria
91	***Halorubrum lipolyticum***	Archaea	Euryarchaeota	Halobacteria
92	***Halorubrum litoreum***	Archaea	Euryarchaeota	Halobacteria
93	***Halorubrum saccharovorum***	Archaea	Euryarchaeota	Halobacteria
94	***Halorubrum sodomense***	Archaea	Euryarchaeota	Halobacteria
95	***Halorubrum tebenquichense***	Archaea	Euryarchaeota	Halobacteria
96	***Halorubrum terrestre***	Archaea	Euryarchaeota	Halobacteria
97	***Halosimplex carlsbadense***	Archaea	Euryarchaeota	Halobacteria
98	***Halostagnicola kamekurae***	Archaea	Euryarchaeota	Halobacteria
99	***Halostagnicola larsenii***	Archaea	Euryarchaeota	Halobacteria
100	***Haloterrigena daqingensis***	Archaea	Euryarchaeota	Halobacteria
101	***Haloterrigena hispanica***	Archaea	Euryarchaeota	Halobacteria
102	***Haloterrigena jeotgali***	Archaea	Euryarchaeota	Halobacteria
103	***Haloterrigena limicola***	Archaea	Euryarchaeota	Halobacteria
104	***Haloterrigena saccharevitans***	Archaea	Euryarchaeota	Halobacteria
105	***Haloterrigena salina***	Archaea	Euryarchaeota	Halobacteria
106	***Haloterrigena thermotolerans***	Archaea	Euryarchaeota	Halobacteria
107	***Haloterrigena turkmenica***	Archaea	Euryarchaeota	Halobacteria
108	***Halovenus aranensis***	Archaea	Euryarchaeota	Halobacteria
109	***Halovivax asiaticus***	Archaea	Euryarchaeota	Halobacteria
110	***Halovivax ruber***	Archaea	Euryarchaeota	Halobacteria
111	*Lokiarchaeum sp. GC14 75*	Archaea	Candidatus Lokiarchaeota	undetermined
112	*Methanocella arvoryzae*	Archaea	Euryarchaeota	Methanomicrobia
113	*Methanocella conradii*	Archaea	Euryarchaeota	Methanomicrobia
114	*Methanocella paludicola*	Archaea	Euryarchaeota	Methanomicrobia
115	*Methanocorpusculum labreanum*	Archaea	Euryarchaeota	Methanomicrobia
116	*Methanoculleus marisnigri*	Archaea	Euryarchaeota	Methanomicrobia
117	*Methanohalobium evestigatum*	Archaea	Euryarchaeota	Methanomicrobia
118	*Methanohalophilus halophilus*	Archaea	Euryarchaeota	Methanomicrobia
119	*Methanohalophilus mahii*	Archaea	Euryarchaeota	Methanomicrobia
120	*Methanohalophilus portucalensis*	Archaea	Euryarchaeota	Methanomicrobia
121	*Methanoregula boonei*	Archaea	Euryarchaeota	Methanomicrobia
122	*Methanosalsum zhilinae*	Archaea	Euryarchaeota	Methanomicrobia
123	*Methanosarcina acetivorans*	Archaea	Euryarchaeota	Methanomicrobia

124	***Natrialba aegyptia***	Archaea	Euryarchaeota	Halobacteria
125	***Natrialba asiatica***	Archaea	Euryarchaeota	Halobacteria
126	***Natrialba chahannaoensis***	Archaea	Euryarchaeota	Halobacteria
127	***Natrialba hulunbeirensis***	Archaea	Euryarchaeota	Halobacteria
128	***Natrialba magadii***	Archaea	Euryarchaeota	Halobacteria
129	***Natrialba taiwanensis***	Archaea	Euryarchaeota	Halobacteria
130	***Natrinema altunense***	Archaea	Euryarchaeota	Halobacteria
131	***Natrinema gari***	Archaea	Euryarchaeota	Halobacteria
132	***Natrinema pallidum***	Archaea	Euryarchaeota	Halobacteria
133	***Natrinema pellirubrum***	Archaea	Euryarchaeota	Halobacteria
134	***Natrinema salaciae***	Archaea	Euryarchaeota	Halobacteria
135	***Natrinema versiforme***	Archaea	Euryarchaeota	Halobacteria
136	***Natronobacterium gregoryi***	Archaea	Euryarchaeota	Halobacteria
137	***Natronobacterium texcoconense***	Archaea	Euryarchaeota	Halobacteria
138	***Natronococcus amylolyticus***	Archaea	Euryarchaeota	Halobacteria
139	***Natronococcus jeotgali***	Archaea	Euryarchaeota	Halobacteria
140	***Natronococcus occultus***	Archaea	Euryarchaeota	Halobacteria
141	***Natronolimnobius baerhuensis***	Archaea	Euryarchaeota	Halobacteria
142	***Natronolimnobius innermongolicus***	Archaea	Euryarchaeota	Halobacteria
143	***Natronomonas moolapensis***	Archaea	Euryarchaeota	Halobacteria
144	***Natronomonas pharaonis***	Archaea	Euryarchaeota	Halobacteria
145	***Natronorubrum bangense***	Archaea	Euryarchaeota	Halobacteria
146	***Natronorubrum sediminis***	Archaea	Euryarchaeota	Halobacteria
147	***Natronorubrum sulfidifaciens***	Archaea	Euryarchaeota	Halobacteria
148	***Natronorubrum texcoconense***	Archaea	Euryarchaeota	Halobacteria
149	***Natronorubrum tibetense***	Archaea	Euryarchaeota	Halobacteria
150	*Salinibacter ruber*	Bacteria	Bacteroidetes	Bacteroidia
151	*Salinivenus iranica*	Bacteria	Rhodothermaeota	Rhodothermia
152	*Salinivenus lutea*	Bacteria	Rhodothermaeota	Rhodothermia
153	*Sulfolobus metallicus*	Archaea	Crenarchaeota	Thermoprotei
154	*Thermoplasma volcanium*	Archaea	Euryarchaeota	Thermoplasmata
155	*Thermoplasmatales archaeon SG8-52-3*	Archaea	Euryarchaeota	Thermoplasmata
156	*Thermoplasmatales archaeon SG8-52-4*	Archaea	Euryarchaeota	Thermoplasmata
157	*uncultured marine group II euryarchaeote*	Archaea	Euryarchaeota	Candidatus Poseidoniia

Table A2. All 222 species whose proteomes were used in Chapter 4.

Πίνακας Α2. Όλα τα 222 είδη των οποίων τα πρωτεώματα χρησιμοποιήθηκαν στο Κεφάλαιο 4.

No	Species	Taxonomic group
1	*Lokiarchaeum sp GC14 75*	Asgard Archaea
2	*Candidatus Thorarchaeota archaeon A25*	Asgard Archaea
3	*Salisaeta longa*	Bacteroidetes
4	*Salinibacter ruber*	Bacteroidetes
5	*nodularia spumigena*	Cyanobacteria
6	*Aphanothece halophytica*	Cyanobacteria
7	*Thermoplasmatales archaeon SG8-52-3*	Euryarchaeota
8	*Thermoplasmatales archaeon SG8-52-4*	Euryarchaeota
9	*Methanobacterium subterraneum*	Euryarchaeota
10	*Methanobacterium formicicum*	Euryarchaeota
11	*Methanococcoides burtonii*	Euryarchaeota
12	*Candidatus Methanoperedens nitroreducens*	Euryarchaeota
13	*Candidatus Syntrophoarchaeum caldarius*	Euryarchaeota
14	*Methanocorpusculum labreanum*	Euryarchaeota
15	*Candidatus Syntrophoarchaeum butanivorans*	Euryarchaeota
16	*Methanoregula boonei*	Euryarchaeota
17	*uncultured marine group II euryarchaeote*	Euryarchaeota
18	*Methanoculleus marisnigri*	Euryarchaeota
19	*Tetragenococcus halophilus*	Firmicutes
20	*Virgibacillus halodenitrificans*	Firmicutes
21	*Oceanobacillus iheyensis*	Firmicutes
22	*Halothermothrix orenii*	Firmicutes
23	*Salinicoccus halodurans*	Firmicutes
24	*Halobacillus halophilus*	Firmicutes
25	*Lentibacillus amyloliquefaciens*	Firmicutes
26	*Terribacilus aidingensis*	Firmicutes
27	*Halanaerobium congolense*	Halanaerobiales
28	*Halanaerobium praevalens*	Halanaerobiales
29	*Haloanaerobium kushneri*	Halanaerobiales
30	*Halanaerobium saccharolyticum*	Halanaerobiales
31	*Halanaerobium salsuginis*	Halanaerobiales
32	*Halanaerobium hydrogeniformans*	Halanaerobiales
33	*Orenia marismortui*	Halanaerobiales
34	*Halonatronum saccharophilum*	Halanaerobiales
35	*Halobacteroides halobius*	Halanaerobiales
36	*Selenihalanaerobacter shriftii*	Halanaerobiales
37	*Acetohalobium arabaticum*	Halanaerobiales
38	*Halarsenatibacter silvermanii*	Halanaerobiales
39	*Methanohalobium evestigatum*	Halophilic Archaea
40	*Methanosarcina acetivorans*	Halophilic Archaea

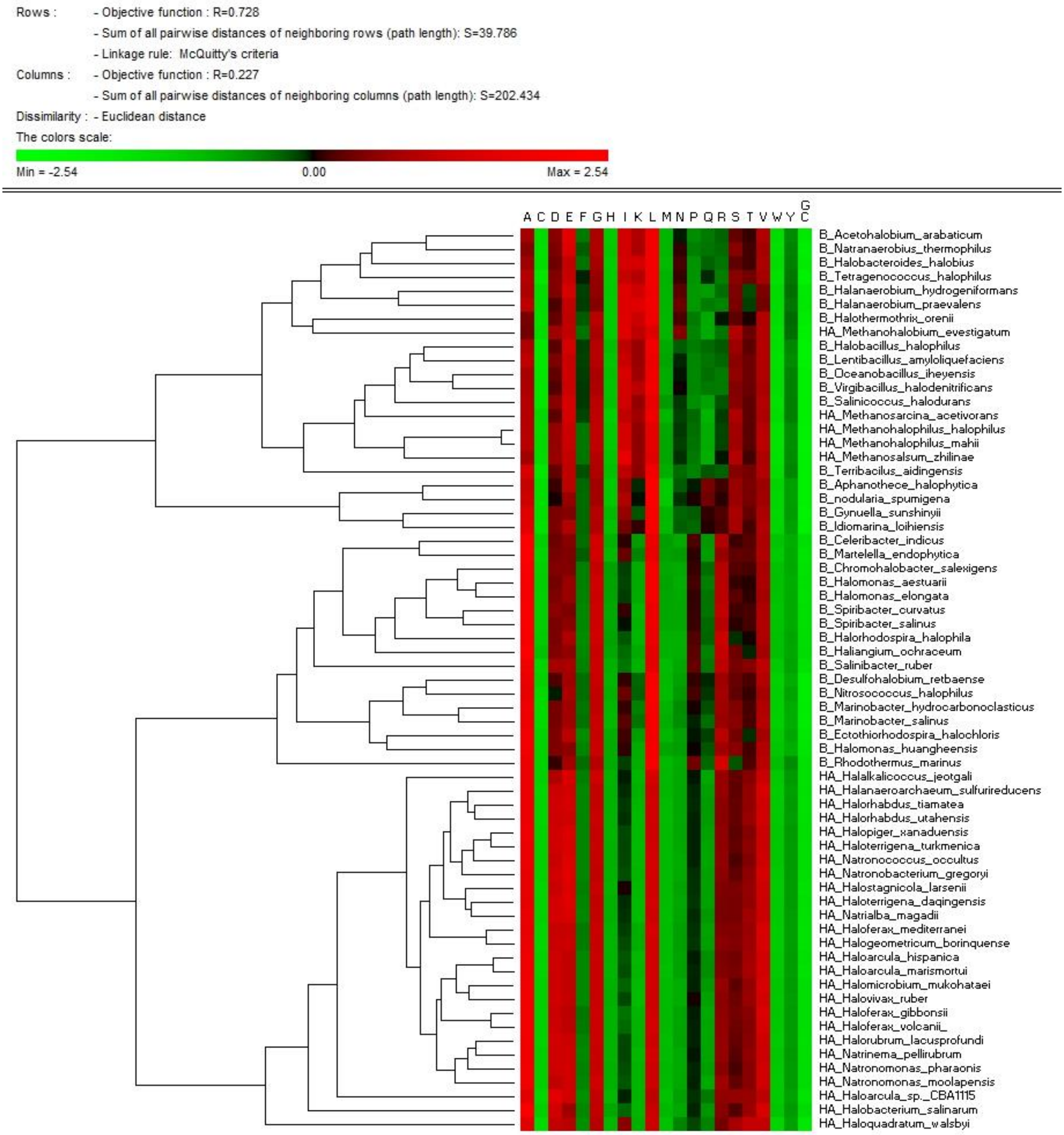
41	*Methanosalsum zhilinae*	Halophilic Archaea
42	*Candidatus Haloredivivus sp. G17*	Halophilic Archaea
43	*Methanohalophilus portucalensis*	Halophilic Archaea
44	*Candidatus Nanosalina*	Halophilic Archaea
45	*Methanohalophilus halophilus*	Halophilic Archaea
46	*Methanohalophilus mahii*	Halophilic Archaea
47	*Candidatus Nanosalinarum*	Halophilic Archaea
48	*Haloquadratum walsbyi*	Halophilic Archaea
49	*Haladaptatus litoreus*	Halophilic Archaea
50	*Haladaptatus cibarius*	Halophilic Archaea
51	*Halohasta litchfieldiae*	Halophilic Archaea
52	*Haloarcula salaria*	Halophilic Archaea
53	*Haladaptatus paucihalophilus*	Halophilic Archaea
54	*Halostagnicola larsenii*	Halophilic Archaea
55	*Halostagnicola kamekurae*	Halophilic Archaea
56	*Haloferax mediterranei*	Halophilic Archaea
57	*Halogeometricum borinquense*	Halophilic Archaea
58	*Halalkalicoccus paucihalophilus*	Halophilic Archaea
59	*Haloterrigena daqingensis*	Halophilic Archaea
60	*Natronorubrum sediminis*	Halophilic Archaea
61	*Haloferax elongans*	Halophilic Archaea
62	*Halococcus thailandensis*	Halophilic Archaea
63	*Natronolimnobius baerhuensis*	Halophilic Archaea
64	*Halalkalicoccus jeotgali*	Halophilic Archaea
65	*Natronorubrum bangense*	Halophilic Archaea
66	*Haloferax larsenii*	Halophilic Archaea
67	*Haloarcula japonica*	Halophilic Archaea
68	*Haloferax mucosum*	Halophilic Archaea
69	*Haloarcula sp. CBA1115*	Halophilic Archaea
70	*Natronorubrum tibetense*	Halophilic Archaea
71	*Halococcus sediminicola*	Halophilic Archaea
72	*Natrialba chahannaoensis*	Halophilic Archaea
73	*Haloterrigena hispanica*	Halophilic Archaea
74	*Natronobacterium texcoconense*	Halophilic Archaea
75	*Halovenus aranensis*	Halophilic Archaea
76	*Haloarcula argentinensis*	Halophilic Archaea
77	*Haloarcula vallismortis*	Halophilic Archaea
78	*Natronorubrum sulfidifaciens*	Halophilic Archaea
79	*Natrialba magadii*	Halophilic Archaea
80	*Natrialba taiwanensis*	Halophilic Archaea
81	*Halorhabdus tiamatea*	Halophilic Archaea
82	*Halobellus rufus*	Halophilic Archaea
83	*Natrialba hulunbeirensis*	Halophilic Archaea
84	*Halogranum salarium*	Halophilic Archaea
85	*Natronorubrum texcoconense*	Halophilic Archaea

86	*Halogranum rubrum*	Halophilic Archaea
87	*Natrialba aegyptia*	Halophilic Archaea
88	*Haloterrigena limicola*	Halophilic Archaea
89	*Haloarcula amylolytica*	Halophilic Archaea
90	*Natrinema versiforme*	Halophilic Archaea
91	*Halapricum salinum*	Halophilic Archaea
92	*Halopiger xanaduensis*	Halophilic Archaea
93	*Natrialba asiatica*	Halophilic Archaea
94	*Halanaeroarchaeum sulfurireducens*	Halophilic Archaea
95	*Natronococcus amylolyticus*	Halophilic Archaea
96	*Natronobacterium gregoryi*	Halophilic Archaea
97	*Halococcus morrhuae*	Halophilic Archaea
98	*Natronococcus jeotgali*	Halophilic Archaea
99	*Halogranum amylolyticum*	Halophilic Archaea
100	*Haloarcula marismortui*	Halophilic Archaea
101	*Halobellus clavatus*	Halophilic Archaea
102	*Halorhabdus utahensis*	Halophilic Archaea
103	*Halopelagius longus*	Halophilic Archaea
104	*Haloterrigena salina*	Halophilic Archaea
105	*Haloterrigena turkmenica*	Halophilic Archaea
106	*Haloferax volcanii*	Halophilic Archaea
107	*Halogranum gelatinilyticum*	Halophilic Archaea
108	*Natronococcus occultus*	Halophilic Archaea
109	*Halorubrum lacusprofundi*	Halophilic Archaea
110	*Natronolimnobius innermongolicus*	Halophilic Archaea
111	*Halogeometricum pallidum*	Halophilic Archaea
112	*Haloarcula hispanica*	Halophilic Archaea
113	*Haloferax sulfurifontis*	Halophilic Archaea
114	*Halorientalis persicus*	Halophilic Archaea
115	*Halopelagius inordinatus*	Halophilic Archaea
116	*Natrinema pellirubrum*	Halophilic Archaea
117	*Halopiger salifodinae*	Halophilic Archaea
118	*Haloferax denitrificans*	Halophilic Archaea
119	*Haloferax gibbonsii*	Halophilic Archaea
120	*Halococcus salifodinae*	Halophilic Archaea
121	*Natronomonas pharaonis*	Halophilic Archaea
122	*Halococcus hamelinensis*	Halophilic Archaea
123	*Natrinema altunense*	Halophilic Archaea
124	*Natrinema pallidum*	Halophilic Archaea
125	*Natrinema gari*	Halophilic Archaea
126	*Halobiforma lacisalsi*	Halophilic Archaea
127	*Natrinema salaciae*	Halophilic Archaea
128	*Haloterrigena saccharevitans*	Halophilic Archaea
129	*Halobiforma haloterrestris*	Halophilic Archaea
130	*Haloferax prahovense*	Halophilic Archaea

131	*Halogeometricum limi*	Halophilic Archaea
132	*Halococcus saccharolyticus*	Halophilic Archaea
133	*Natronomonas moolapensis*	Halophilic Archaea
134	*Halobiforma nitratireducens*	Halophilic Archaea
135	*Haloterrigena jeotgali*	Halophilic Archaea
136	*Haloferax lucentense*	Halophilic Archaea
137	*Halorientalis regularis*	Halophilic Archaea
138	*Halolamina sediminis*	Halophilic Archaea
139	*Haloplanus vescus*	Halophilic Archaea
140	*Halomicrobium katesii*	Halophilic Archaea
141	*Haloterrigena thermotolerans*	Halophilic Archaea
142	*Halomicrobium zhouii*	Halophilic Archaea
143	*Halovivax ruber*	Halophilic Archaea
144	*Halolamina pelagica*	Halophilic Archaea
145	*Halorubrum halophilum*	Halophilic Archaea
146	*Halomicrobium mukohataei*	Halophilic Archaea
147	*Halorubrum saccharovorum*	Halophilic Archaea
148	*Halogeometricum rufum*	Halophilic Archaea
149	*Halolamina rubra*	Halophilic Archaea
150	*Halococcus agarilyticus*	Halophilic Archaea
151	*Halovivax asiaticus*	Halophilic Archaea
152	*Halopenitus malekzadehii*	Halophilic Archaea
153	*Halopenitus persicus*	Halophilic Archaea
154	*Haloplanus natans*	Halophilic Archaea
155	*Halorubrum californiense*	Halophilic Archaea
156	*Halorubrum kocurii*	Halophilic Archaea
157	*Halarchaeum acidiphilum*	Halophilic Archaea
158	*Haloarchaeobius iranensis*	Halophilic Archaea
159	*Halorubrum coriense*	Halophilic Archaea
160	*Halorubrum aidingense*	Halophilic Archaea
161	*Halorubrum arcis*	Halophilic Archaea
162	*Halobacterium jilantaiense*	Halophilic Archaea
163	*Halorubrum lipolyticum*	Halophilic Archaea
164	*Halobacterium salinarum*	Halophilic Archaea
165	*Halorubrum ezzemoulense*	Halophilic Archaea
166	*Halosimplex carlsbadense*	Halophilic Archaea
167	*Halorubrum tebenquichense*	Halophilic Archaea
168	*Halorubrum distributum*	Halophilic Archaea
169	*Halorubrum sodomense*	Halophilic Archaea
170	*Halobaculum gomorrense*	Halophilic Archaea
171	*Halorubrum terrestre*	Halophilic Archaea
172	*Halorubrum litoreum*	Halophilic Archaea
173	*Idiomarina loihiensis*	Proteobacteria
174	*Gynuella sunshinyii*	Proteobacteria
175	*Nitrosococcus halophilus*	Proteobacteria

176	*Marinobacter salinus*	Proteobacteria
177	*Marinobacter hydrocarbonoclasticus*	Proteobacteria
178	*Desulfohalobium retbaense*	Proteobacteria
179	*Martelella endophytica*	Proteobacteria
180	*Ectothiorhodospira halochloris*	Proteobacteria
181	*Halomonas huangheensis*	Proteobacteria
182	*Halomonas elongata*	Proteobacteria
183	*Chromohalobacter salexigens*	Proteobacteria
184	*Celeribacter indicus*	Proteobacteria
185	*Halomonas aestuarii*	Proteobacteria
186	*Spiribacter salinus*	Proteobacteria
187	*Spiribacter curvatus*	Proteobacteria
188	*Haliangium ochraceum*	Proteobacteria
189	*Halorhodospira halophila*	Proteobacteria
190	*Salinivenus iranica*	Rhodothermaeota
191	*Salinivenus lutea*	Rhodothermaeota
192	*Candidatus Nitrocosmicus oleophilus*	TACK
193	*Candidatus Nitrosotenuis cloacae*	TACK
194	*Candidatus Korarchaeum cryptofilum*	TACK
195	*Candidatus Bathyarchaeota archaeon BA2*	TACK
196	*Cenarchaeum symbiosum*	TACK
197	*Sulfurisphaera tokodaii*	Thermophiles
198	*Thermoplasma volcanium*	Thermophiles
199	*Sulfolobus metallicus*	Thermophiles
200	*Methanocaldococcus jannaschii*	Thermophiles
201	*Methanocaldococcus villosus*	Thermophiles
202	*Methanocaldococcus vulcanius*	Thermophiles
203	*Methanocaldococcus bathoardescens*	Thermophiles
204	*Methanocaldococcus infernus*	Thermophiles
205	*Natranaerobius thermophilus*	Thermophiles
206	*Methanothermus fervidus*	Thermophiles
207	*Thermotoga maritima*	Thermophiles
208	*Pyrococcus horikoshii*	Thermophiles
209	*Pyrococcus furiosus*	Thermophiles
210	*Pyrococcus abyssi*	Thermophiles
211	*Vulcanisaeta moutnovskia*	Thermophiles
212	*Archaeoglobus fulgidus*	Thermophiles
213	*Archaeoglobus veneficus*	Thermophiles
214	*Methanocella conradii*	Thermophiles
215	*Methanocella paludicola*	Thermophiles
216	*Methanocella arvoryzae*	Thermophiles
217	*Thermoproteus tenax*	Thermophiles
218	*Thermofilum pendens*	Thermophiles
219	*Methermicoccus shengliensis*	Thermophiles
220	*Hyperthermus butylicus*	Thermophiles

221	*Aeropyrum pernix*	Thermophiles
222	*Rhodothermus marinus*	Thermophiles



```
Rows :          - Objective function : R=0.728
                - Sum of all pairwise distances of neighboring rows (path length): S=39.786
                - Linkage rule:  McQuitty's criteria
Columns :       - Objective function : R=0.227
                - Sum of all pairwise distances of neighboring columns (path length): S=202.434
Dissimilarity : - Euclidean distance
The colors scale:
Min = -2.54                          0.00                          Max = 2.54
```

Figure A1. The topology of hierarchical clustering of 65 species from Bacteria (prefix B_) and halophilic Archaea (prefix HA_). GC content (last column) does not influence the clustering.

Εικόνα Α1. Η τοπολογία της ιεραρχικής ομαδοποίησης των 65 ειδών βακτηρίων (πρόθεμα B_) και αλόφιλων αρχαίων (πρόθεμα HA_). Το περιεχόμενο GC (τελευταία στήλη) δεν επηρεάζει την ομαδοποίηση.

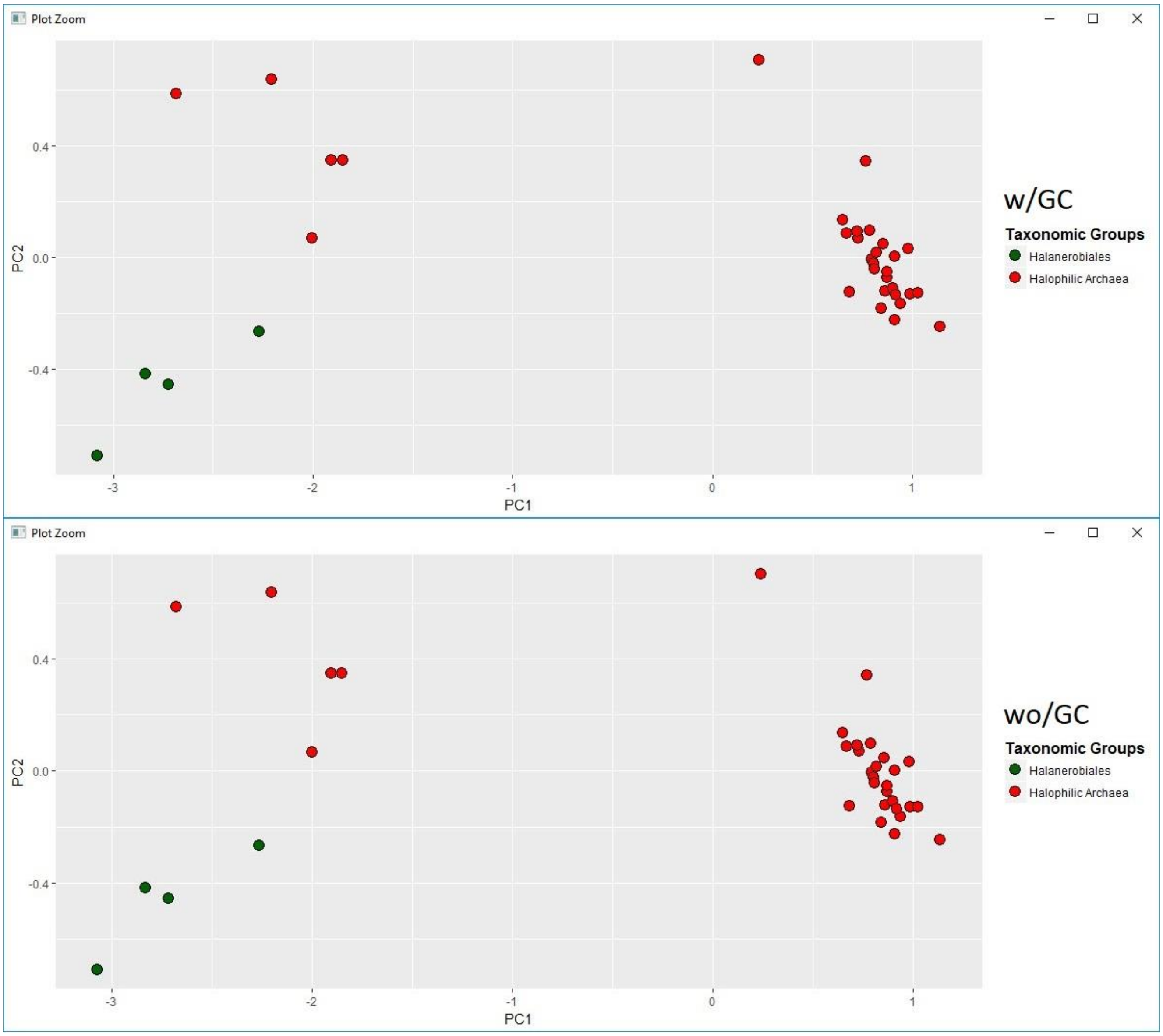


Figure A2. PCA analysis of Halanaerobiales and halophilic Archaea with GC included and GC excluded. Results are almost identical.

Εικόνα A2. Ανάλυση κυρίων συνιστωσών (PCA) των Halanaerobiales και των αλόφιλων αρχαιοβακτηρίων, με το περιεχόμενο GC να λαμβάνεται υπόψιν αλλά και να μη λαμβάνεται. Τα αποτελέσματα είναι σχεδόν όμοια.

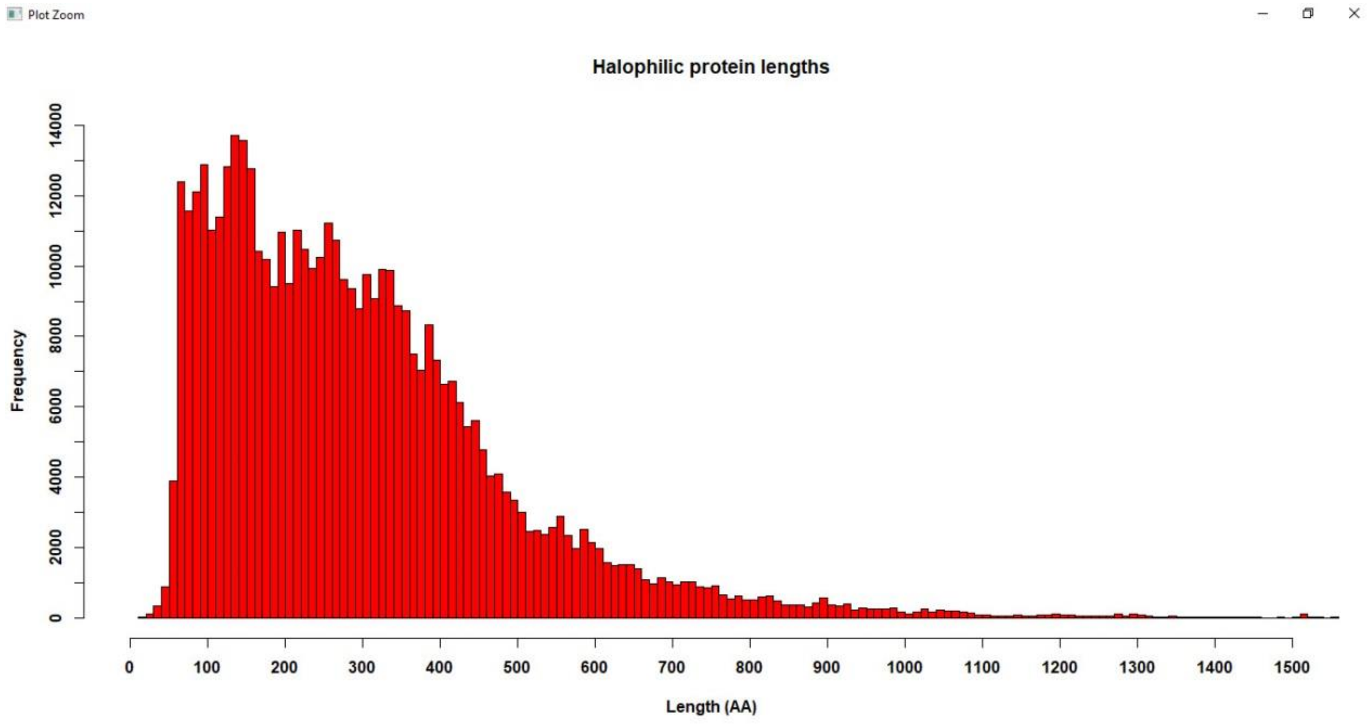


Figure A3. Length histogram of all halophilic proteins from halophilic Archaea.

Εικόνα A3. Ιστόγραμμα μηκών όλων των πρωτεϊνών από τα αλόφιλα Αρχαία.

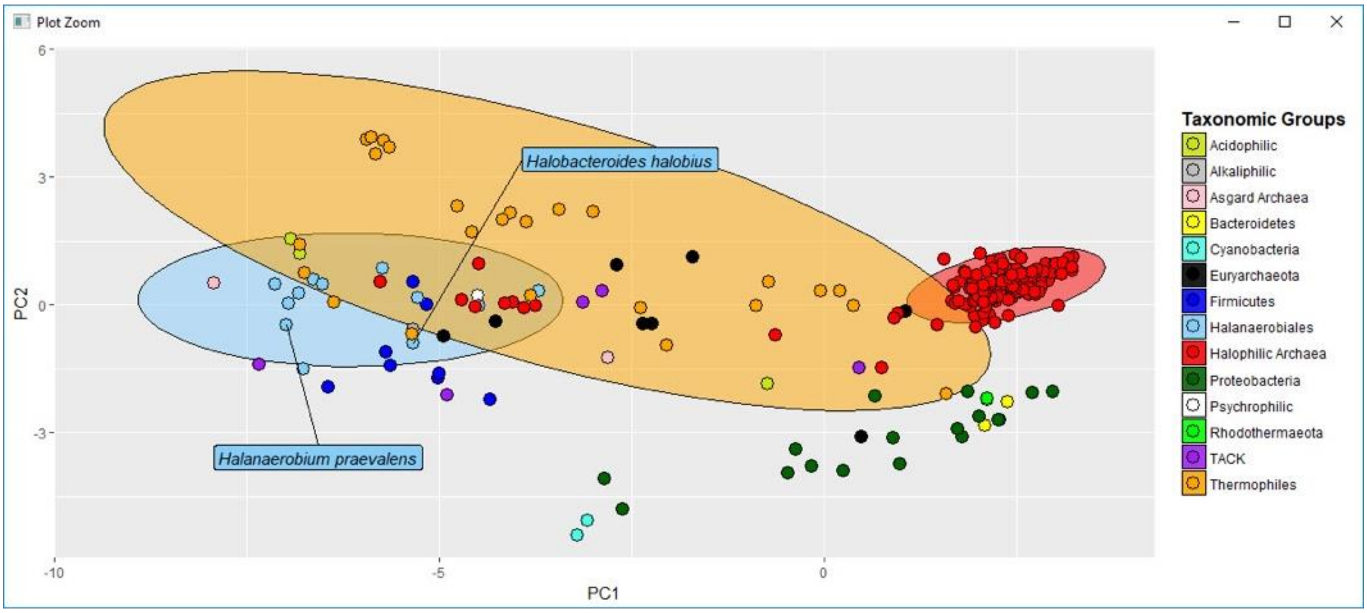


Figure A4. *Haloanaerobium praevalens* and *Halobacteroides halobius* are placed away from the halophilic cluster of Halobacteria (on the right).

Εικόνα A4. Τα είδη *Haloanaerobium praevalens* και *Halobacteroides halobius* τοποθετούνται μακριά από την ομάδα των αλόφιλων Αρχαίων (στα δεξιά).

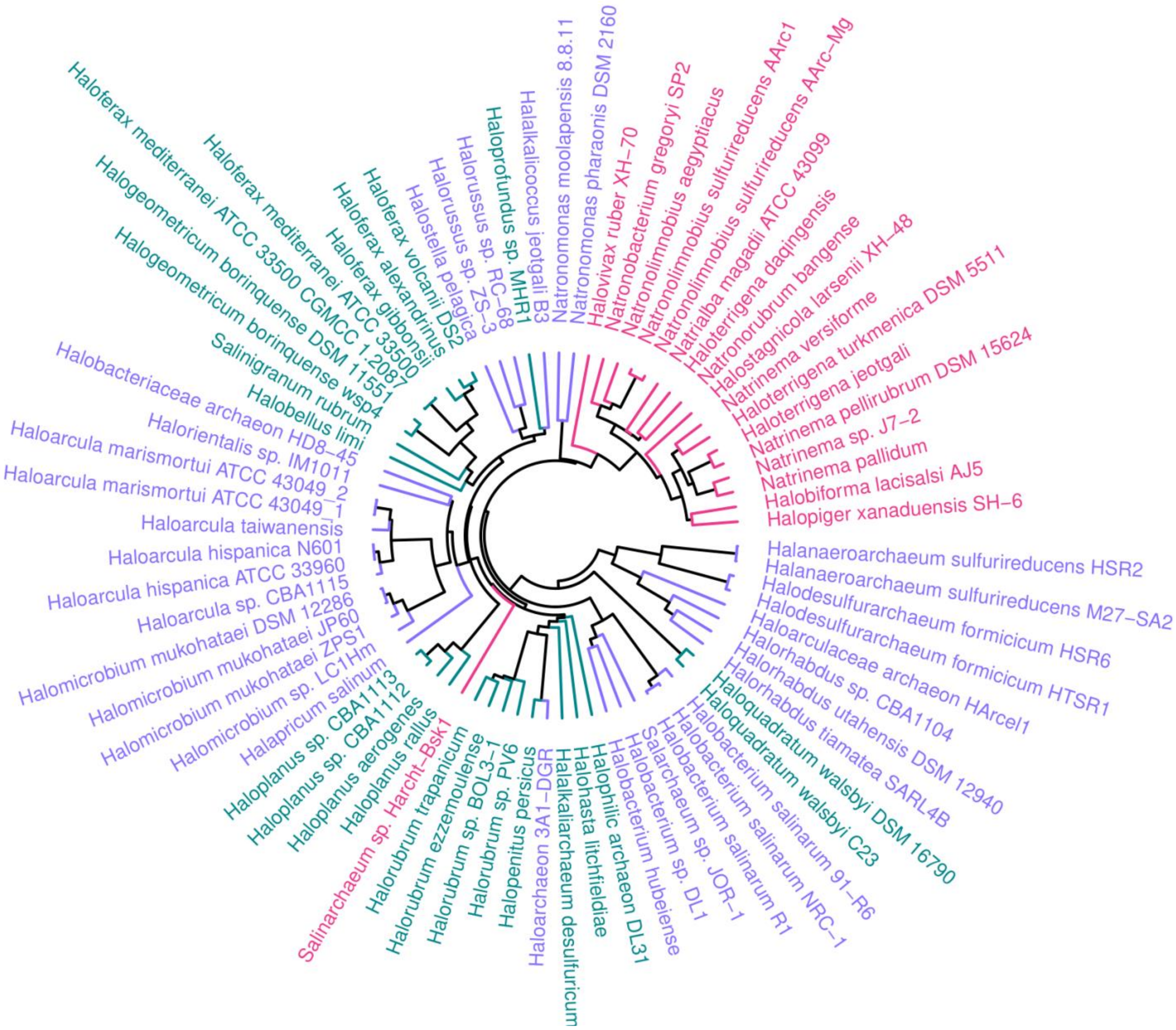

Haloferax mediterranei ATCC 33500 CGMCC 1.2087
Haloferax mediterranei ATCC 33500
Halogeometricum borinquense DSM 11551
Halogeometricum borinquense wsp4
Haloferax gibbonsii
Haloferax alexandrinus
Haloferax volcanii DS2
Halostella pelagica
Halorussus sp. ZS−3
Halorussus sp. RC−68
Haloprofundus sp. MHR1
Halalkalicoccus jeotgali B3
Natronomonas moolapensis 8.8.11
Natronomonas pharaonis DSM 2160
Halovivax ruber XH−70
Natronobacterium gregoryi SP2
Natronolimnobius aegyptiacus
Natronolimnobius sulfurireducens AArc1
Natronolimnobius sulfurireducens AArc−Mg
Natrialba magadii ATCC 43099
Haloterrigena daqingensis
Natronorubrum bangense
Halostagnicola larsenii XH−48
Natrinema versiforme
Haloterrigena turkmenica DSM 5511
Haloterrigena jeotgali
Natrinema pellirubrum DSM 15624
Natrinema sp. J7−2
Natrinema pallidum
Halobiforma lacisalsi AJ5
Halopiger xanaduensis SH−6
Halanaeroarchaeum sulfurireducens HSR2
Halanaeroarchaeum sulfurireducens M27−SA2
Halodesulfurarchaeum formicicum HSR6
Halodesulfurarchaeum formicicum HTSR1
Haloarculaceae archaeon HArcel1
Halorhabdus sp. CBA1104
Halorhabdus utahensis DSM 12940
Halorhabdus tiamatea SARL4B
Haloquadratum walsbyi DSM 16790
Haloquadratum walsbyi C23
Halobacterium salinarum 91−R6
Halobacterium salinarum NRC−1
Halobacterium salinarum R1
Halobacterium sp. JOR−1
Salarchaeum sp. DL1
Halobacterium hubeiense
Halophilic archaeon DL31
Halohasta litchfieldiae
Halalkaliarchaeum desulfuricum
Haloarchaeon 3A1−DGR
Halopenitus persicus
Halorubrum sp. PV6
Halorubrum sp. BOL3−1
Halorubrum ezzemoulense
Halorubrum trapanicum
Salinarchaeum sp. Harcht−Bsk1
Haloplanus rallus
Haloplanus aerogenes
Haloplanus sp. CBA1112
Haloplanus sp. CBA1113
Halapricum salinum
Halomicrobium sp. LC1Hm
Halomicrobium mukohataei ZPS1
Halomicrobium mukohataei JP60
Halomicrobium mukohataei DSM 12286
Haloarcula sp. CBA1115
Haloarcula hispanica ATCC 33960
Haloarcula hispanica N601
Haloarcula taiwanensis
Haloarcula marismortui ATCC 43049_1
Haloarcula marismortui ATCC 43049_2
Halorientalis sp. IM1011
Halobacteriaceae archaeon HD8−45
Halobellus limi
Salinigranum rubrum

Figure A5. Average distance, weighted dendrogram for 76 Halobacteria. Order Haloferacales in green, order Natrialbales in red, and order Halobacteriales in purple.

Εικόνα Α5. Σταθμισμένο δενδρόγραμμα με την μέθοδο average distance για τα 76 μέλη της κλάσης Halobacteria. Η τάξη Haloferacales με πράσινο, η τάξη Natrialbales με κόκκινο και η τάξη Halobacteriales με μωβ.

Perl script A1. “ip2location.pl”

```
use strict;
use warnings;
use Tie::File;

use Geo::IP2Location;

my $obj = Geo::IP2Location->open('IPV6.BIN');

open my $fh, '>', "ip_results.txt" or die "Cannot open ip_results.txt: $!";

tie my @ips, 'Tie::File', "ip_v6.txt";

foreach my $ips (@ips){

my $ip = $ips;
=pod
my $countryshort = $obj->get_country_short($ip);
print $countryshort,"\n";
my $countrylong = $obj->get_country_long($ip);
my $region = $obj->get_region($ip);
print $region,"\n";
my $city = $obj->get_city($ip);
print $city,"\n";
my $latitude = $obj->get_latitude($ip);
my $longitude = $obj->get_longitude($ip);
my $isp = $obj->get_isp($ip);
my $domain = $obj->get_domain($ip);
my $zipcode = $obj->get_zipcode($ip);
my $timezone = $obj->get_timezone($ip);
my $netspeed = $obj->get_netspeed($ip);
my $iddcode = $obj->get_iddcode($ip);
my $areacode = $obj->get_areacode($ip);
my $weatherstationcode = $obj->get_weatherstationcode($ip);
my $weatherstationname = $obj->get_weatherstationname($ip);
my $mcc = $obj->get_mcc($ip);
my $mnc = $obj->get_mnc($ip);
my $mobilebrand = $obj->get_mobilebrand($ip);
my $elevation = $obj->get_elevation($ip);
my $usagetype = $obj->get_usagetype($ip);
=cut



my ($cos, $col, $reg, $cit, $lat, $lon, $zip, $tmz, $isp, $dom, $ns, $idd, $area, $wcode, $wname,
$mcc, $mnc, $brand, $elevation, $usagetype) = $obj->get_all($ip);


print $fh $col.','.$reg.','.$cit.','.$lat.','.$lon,"\n";


}
```

Perl script A2. “separate_hmms.pl”

```
use strict;
  use warnings;


 my $content;
 my $filename = "archaea_markers.hmm";
    open(my $fh, '<', $filename) or die "cannot open file $filename";
    {
        local $/;
        $content = <$fh>;
    }
    close($fh);



#print $content."\n";


  my @values = split('//', $content);


my $count=0;
my $filecode=1;
  foreach my $val (@values){
    #print "$val\n";
    #writing values to separate hmm files START
    my $filename2 = 'amphora_marker'.$filecode.'.hmm';
    print $filename2,"\n";
             open(my $fh, '>', $filename2) or die "Could not open file '$filename2' $!";
             print $fh $val;
             print $fh "//";
             close $fh;
    #writing values to separate hmm files END
    $count++;
    $filecode++;
    print $count."\n";
  }

  exit 0;
```

Perl script A3. “hmmer_ex.pl”

```
use strict;
use warnings;



my $dir = '/home/loukalexis/Desktop/markers';
my $dir2 = '/home/loukalexis/Desktop/proteins';



opendir(DIR,$dir);
my @markers = readdir(DIR);
closedir(DIR);
#foreach(@markers){
#  print $_,"\n";
#}

opendir(DIR,$dir2);
my @species = readdir(DIR);
closedir(DIR);
#foreach(@species){
#  print $_,"\n";
#}



#print "Array size is ", scalar(@markers), "\n";
#print "Array size is ", scalar(@species), "\n";




#print "files array before: ",@markers,"\n";
#checking array for invalid files
```

```
my $inv_exists=1;
while ($inv_exists) {

my $index=0;
my $occ=0;
foreach my $marker (@markers) {
	if ($marker !~ m/.hmm/) {
		#print "Will be deleted: ",$file,"\n\n";
		splice(@markers,$index,1);
		$occ++;
	}
	if ($occ==0) {
		$inv_exists=0;
	}
	$index++;
}


}

#checking array for invalid files
my $inv_exists2=1;
while ($inv_exists2) {

my $index2=0;
my $occ2=0;
foreach my $specie (@species) {
	if ($specie !~ m/.fasta/) {
		#print "Will be deleted: ",$file,"\n\n";
		splice(@species,$index2,1);
		$occ2++;
	}
```

```
        if ($occ2==0) {
                $inv_exists2=0;
        }
        $index2++;
}


}

#save species names into an array

#print "files array after: ",@markers,"\n";
#print "Array size now is ", scalar(@markers), "\n";
#print "last element: ",$markers[265],"\n\n";

#print "files array after: ",@species,"\n";
#print "Array size now is ", scalar(@species), "\n";
#print "last element: ",$species[133],"\n\n";



#hmmer search
my $nr=1;
for (my $i=0; $i<scalar(@markers); $i++){
        for (my $j=0; $j<scalar(@species); $j++){
                my $tmp_title=$species[$j];
                $tmp_title=~ s/.fasta//;
                print $tmp_title;
#system ('hmmsearch /home/loukalexis/Desktop/markers/'.$markers[$i].'
'/home/loukalexis/Desktop/proteins/'.$species[$j] >
/home/loukalexis/Desktop/results/results'.$nr.'.out');
```

```
print 'hmmsearch
/home/loukalexis/Desktop/markers/'.$markers[$i].'/home/loukalexis/Desktop/proteins/'.$species[
$j].' > /home/loukalexis/Desktop/results/results'.$nr.'.out';
$nr++;
}
}
```

Perl script A4. “create_alignments_w_mafft.pl”

```
use strict;
use warnings;



my $dir = 'C:\Users\Alex\Desktop\Perl works\Phylogenomics\Unified data from Archaea,
Outgroups and Salinibacters\Halobacteria and some Methanogens
markers\concatenated_markers\corrected';


opendir(DIR,$dir);
my @filenames = readdir(DIR);
closedir(DIR);
#foreach(@filenames){
# print $_,"\n";
#}


my $inv_exists=1;
while ($inv_exists) {

my $index=0;
my $occ=0;
foreach my $fas (@filenames) {
        if ($fas =~/^.fasta$/) {
                #print "Will be deleted: ",$file,"\n\n";
                splice(@filenames,$index,1);
                $occ++;
        }
        if ($occ==0) {
```

```
                    $inv_exists=0;
          }
          $index++;
}
}



foreach my $file (@filenames){
          print $file,"\n";
          #$input = '';
          #$output= '';
          #system('mafft $input > $output');
}
```

Perl script A5 “residue_calculator.pl”

```
use strict;
use warnings;


my $dir = 'C:\Users\Alex\Desktop\Perl works\Phylogenomics\Amino acid compositions\complete genomes tmp\proteomes';


opendir(DIR,$dir);
my @files = readdir(DIR);
closedir(DIR);


my $inv_exists=1;
while ($inv_exists) {

my $index=0;
my $occ=0;
foreach my $out (@files) {
	if ($out !~ m/.fasta/) {
		#print "Will be deleted: ",$file,"\n\n";
		splice(@files,$index,1);
		$occ++;
	}
	if ($occ==0) {
		$inv_exists=0;
	}
	$index++;
}
}
```

```
my $input = ".fasta";
my $output = ".csv";

foreach my $files (@files){
        my $n = $files;
        $n =~ s/fasta/csv/;
system('"C:\Users\Alex\Desktop\Perl works\Phylogenomics\Amino acid
compositions\ResidueFrequencySummarizer\ResidueFrequencySummarizer.exe"
"C:\Users\Alex\Desktop\Perl works\Phylogenomics\Amino acid compositions\complete genomes
tmp\proteomes\\'.$files.'" "/O:C:\Users\Alex\Desktop\Perl works\Phylogenomics\Amino acid
compositions\complete genomes tmp\proteomes\rest_csv\\'.$n.'"');
}
```

Perl script A6. “Fasta_GC_counter.pl”

```
use strict;
use warnings;



my $dir = 'C:\Users\Alex\Desktop\Perl works\Phylogenomics\Amino acid compositions\complete genomes data\genomes';



opendir(DIR,$dir);
my @files = readdir(DIR);
closedir(DIR);



my $inv_exists=1;
while ($inv_exists) {

my $index=0;
my $occ=0;
foreach my $out (@files) {
	if ($out !~ m/.fasta/) {
		#print "Will be deleted: ",$file,"\n\n";
		splice(@files,$index,1);
		$occ++;
	}
	if ($occ==0) {
		$inv_exists=0;
	}
	$index++;
}
}
```

```
open (my $gcs, '>>', 'C:\Users\Alex\Desktop\Perl works\Phylogenomics\Amino acid
compositions\complete genomes data\genomes\GC_contents\GC_conents.csv') || die "Can't open
<file:$>!";

foreach my $fastas (@files){



my $file = 'C:\Users\Alex\Desktop\Perl works\Phylogenomics\Amino acid compositions\complete
genomes data\genomes\\'.$fastas;
open my $info, $file or die "Could not open $file: $!";

my $counter = 0;
my $GC_count=0;
while( my $line = <$info>)  {
	if ($line !~ />/){
		my $G = $line;
		my $C = $line;
		my $A = $line;
		my $T = $line;
		my $count_G = () = $G =~ /G/g;
		my $count_C = () = $C =~ /C/g;
		my $count_A = () = $A =~ /A/g;
		my $count_T = () = $T =~ /T/g;
		my $count_final = $count_G + $count_C + $count_A + $count_T;
		$GC_count = $GC_count + $count_G + $count_C;
		$counter = $counter + $count_final;
	}
}

my $GC_content = $GC_count/$counter;
close $info;

print $gcs $fastas.",".$GC_content."\n";


}
```

Perl script A7. “halopredictor_local_V1.pl”

```
#!/bin/perl -w


use strict;
use warnings;
use Bio::SeqIO;
use File::Path;
use File::Path qw(make_path);

my $dir = 'results';
my $halo_results = 'C:\Halopredictor_Local_V1\results\halo_results.csv';
my $halo_un = "halo_un.csv";
my $r_out = "r_out.csv";
my $log = "logfile.txt";

#check if halo_results.csv exists
if (-e $halo_results){
	#print "The file exists\n";
	unlink $halo_results or warn "Could not unlink $halo_results: $!";
}

#check if folder "results" exists
if (-e $dir and -d $dir){
	rmtree($dir);
}

#creating the folder results again
mkdir 'results';

if ((-e $halo_un) || (-e $r_out)|| (-e $log) ) {
unlink $halo_un,$r_out,;
}



my $seqio_obj = Bio::SeqIO->new(-file => "input.fasta",
                      -format => "fasta" );


my @amino_acids = ('A','C','D','E','F','G','H','I','K','L','M','N','P','Q','R','S','T','V','W','Y');
my $tmp_prof;
my $firstline = 'type,A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y';
my $cmd = 'Rscript.exe C:/halopredictor_Local_v1/halopredictor_local_LDA.r';
my $prediction="";
my $hal_counter=0;
my $nonhal_counter=0;

#inserting title to the results csv file
make_path("results");
my $res_t = 'results/halo_results.csv';
	open my $th, '>>', $res_t or die "Could not open file '$res_t' $!";
	print $th "Sequence,Group,Prediction,LDA\n";
	close $th;





#loop for finding the number of sequences
my $seq_nr=0;
while ( my $seq_obj = $seqio_obj->next_seq ) {
	$seq_nr++;
}
```

```
my $seqio_obj2 = Bio::SeqIO->new(-file => "input.fasta",
                       -format => "fasta" );

my $progress=0;
print "\n\n";
while ( my $seq_obj2 = $seqio_obj2->next_seq ) {
        #empty the temp amino acid profile array
        $tmp_prof = ",";
        #measuring the amino acid percentages and creating CSV file for R input
        #print "-----------------------------------------------------------------------------\n";
        #print $seq_obj2->display_id." ".$seq_obj2->desc,"\n";
        my $string = $seq_obj2->seq;
        foreach my $aa (@amino_acids){
   #print $seq_obj->seq,"\n";
        my $c = () = $string =~ /$aa/g;
        #print "The number of $aa in this sequence is :", $c,"\n";
        my $length = length($string);
        #print "The length of the sequence is :", $length,"\n";
        my $prc = ($c/$length)*100;
        my $prc_r = sprintf("%.1f", $prc);
        #print "The percentage of $aa is : ", $prc_r,"\n";
        $tmp_prof = $tmp_prof.$prc_r.",";
        }
                $progress++;
        print "       ".$progress."/".$seq_nr," sequences completed\r";
        chop($tmp_prof);
        #print "The temp profile from variable is : ",$tmp_prof,"\n";
        #now the amino acid profile of the sequence is writen to a csv file for the R script input
        my $filename = 'halo_un.csv';
        open(my $fh, '>', $filename) or die "Could not open file '$filename' $!";
        print $fh $firstline,"\n";
        print $fh $tmp_prof,"\n";
```

```
	close $fh;
	#call the R script with system
	system(qq{cd C:/Program Files/R/R-3.6.1/bin && $cmd}[[12gPhylPhyP);
	#get results of the R script from csv
	open my $file, '<', "r_out.csv";
	my $csv = <$file>;
	close $file;
	my @csv_element = split /,/, $csv;
	#print $csv_element[0],"\n";
	#now writing results of LDA in a file
	my $res = 'results/halo_results.csv';
	open(my $fh2, '>>', $res) or die "Could not open file '$res' $!";
	my $desc_correct = $seq_obj2->desc;
	$desc_correct =~ s/,//g;
	if ($csv_element[0] == 1){
		$prediction = "halophilic";
		$hal_counter++;
	}
	else{
		$prediction = "non-halophilic";
		$nonhal_counter++;
	}
	print $fh2 $seq_obj2->display_id."
".$desc_correct.",".$csv_element[0].",".$prediction.",".$csv_element[1];
	close $fh2;
}
print "\n\n\n		-Process completed-\n";



#print "Halophilic counter: ",$hal_counter,"\n";
#print "Non-halophilic counter: ",$nonhal_counter,"\n";
```

```
#calculate some simple statistics about inserted sequences
my $cmd_stats = 'Rscript.exe C:/halopredictor_Local_v1/stats.r';
system(qq{cd C:/Program Files/R/R-3.6.1/bin && $cmd_stats}7171[);
#now asking the user if the script should proceed with visualizing the results
my $control=0;
while ($control==0){
print "\n\nWould you like to visualize results? [type \"yes\" or \"no\"]";
my $choice = <STDIN>;
chomp $choice;
 if (($choice eq "yes")||($choice eq "no")){
        $control=1;
 }
 if ($choice eq "no"){
        print "\n\nResults can be found in halo_results.csv in \"results\" folder\n";
        print "\n\nThank you for using Halopredictor_Local_V1\n\n";
 }
 if($choice eq "yes"){
        if (($hal_counter>=2)&&($nonhal_counter>=2)){
        my $cmd2 = 'Rscript.exe C:/halopredictor_Local_v1/halopredictor_local_visualize.r';
        system(qq{cd C:/Program Files/R/R-3.6.1/bin && $cmd2});
        print "\nThe results can be found in \"results\" folder\n";
        print "\nThank you for using Halopredictor_Local_V1\n";
        }else{
                print "Sorry there must be minimum 2 halophilic and 2 non-halophilic sequences in
order to visualise results, check halo_results.csv\n";
                print "\nThank you for using Halopredictor_Local_V1\n";
                $control=1;
        }
 }
}
```

R script A1. “PCA_analysis.r”

```
#PCA analysis for amino acid profiles of all data
HaloDom = read.csv("C:\\Users\\Alex\\Desktop\\Perl works\\Phylogenomics\\Amino acid
compositions\\Clustering of all Halophiles\\PCA\\PCA_norm.csv", header = TRUE)
data(HaloDom)
head(HaloDom)
summary(HaloDom)
#View(HaloDom)
myPr <- prcomp(HaloDom[3:22], scale=TRUE)
summary(myPr)
plot(myPr)
plot(myPr, type='l')
biplot(myPr, scale=0)
str(myPr)
myPr$x
HaloDom2 <- cbind(HaloDom, myPr$x[,1:2])
head(HaloDom2)
library(ggplot2)
require("ggrepel")
ggplot(HaloDom2, aes(PC1, PC2, label=Residue, fill=category))+
guides(fill=guide_legend(title="Taxonomic Groups", title.theme = element_text(colour = "black",
size = 12, angle = 0, face = "bold"), title.hjust = -3))+
scale_fill_manual(values=c("#CBE315","grey", "pink", "yellow",
"#4DFFE1","black","blue","lightskyblue","red","darkgreen","white","green", "purple","orange"))+
stat_ellipse(data = subset(HaloDom2, category=="Halanaerobiales"), geom="polygon", col="black",
alpha="0.5")+
stat_ellipse(data = subset(HaloDom2, category=="Halophilic Archaea"), geom="polygon",
col="black", alpha="0.5")+
stat_ellipse(data = subset(HaloDom2, category=="Thermophiles"), geom="polygon", col="black",
alpha="0.5")+
geom_point(shape=21,color="black",size=4)+
#geom_label_repel(data = subset(HaloDom2, Residue=="HA_Haloarcula_argentinensis"),
show.legend = FALSE, inherit.aes = TRUE, colour="black", ylim =(-5.7), xlim = (-2) , force=10,
fontface="italic")+
#geom_label_repel(data = subset(HaloDom2, Residue=="HA_Haloarcula_hispanica"), show.legend
= FALSE, inherit.aes = TRUE, colour="black", ylim =(-5.7), xlim=10, force=10, fontface="italic")+
#geom_label_repel(data = subset(HaloDom2, Residue=="Haloarcula sp. CBA1115"), show.legend =
FALSE, inherit.aes = TRUE, colour="black", ylim =(-5.7), xlim=(-8.9), force=10, fontface="italic")+
#geom_label_repel(data = subset(HaloDom2, Residue=="Candidatus Nanosalinarum"), show.legend
= FALSE, inherit.aes = TRUE, colour="black", ylim =(-6), xlim=(-7), force=10, fontface="italic")+
#geom_label_repel(data = subset(HaloDom2, Residue=="Candidatus Haloredivivus sp. G17"),
show.legend = FALSE, inherit.aes = TRUE, colour="black", ylim =5, xlim=(-5), force=10,
fontface="italic")+
#geom_label_repel(data = subset(HaloDom2, Residue=="Candidatus Nanosalina"), show.legend =
FALSE, inherit.aes = TRUE, colour="black", ylim=(-4.5) ,xlim=(-5), force=10, fontface="italic")+
#geom_label_repel(data = subset(HaloDom2, Residue=="Haloarcula salaria"), show.legend = FALSE,
inherit.aes = TRUE, colour="black", ylim=3 ,xlim=0, force=10, fontface="italic")
#geom_label_repel(data = subset(HaloDom2, Residue=="Halobacteroides halobius"), show.legend
= FALSE, inherit.aes = TRUE, colour="black", ylim=3 ,xlim=(-4), force=10, fontface="italic")+
#geom_label_repel(data = subset(HaloDom2, Residue=="Methanosalsum zhilinae"), show.legend =
FALSE, inherit.aes = TRUE, colour="black", ylim=(3.5) ,xlim=(-9.5), force=10, fontface="italic")+
#geom_label_repel(data = subset(HaloDom2, Residue=="Methanohalophilus portucalensis"),
show.legend = FALSE, inherit.aes = TRUE, colour="black", ylim=3 ,xlim=(-2.5), force=10,
fontface="italic")+
#geom_label_repel(data = subset(HaloDom2, Residue=="Methanohalophilus halophilus"),
show.legend = FALSE, inherit.aes = TRUE, colour="black", ylim=(-4) ,xlim=(-9.9), force=10,
fontface="italic")+
#geom_label_repel(data = subset(HaloDom2, Residue=="Methanohalobium evestigatum"),
show.legend = FALSE, inherit.aes = TRUE, colour="black", ylim=(-6) ,xlim=(0), force=10,
fontface="italic")+
#geom_label_repel(data = subset(HaloDom2, Residue=="Methanosarcina acetivorans"),
show.legend = FALSE, inherit.aes = TRUE, colour="black", ylim=(2) ,xlim=(0), force=10,
fontface="italic")+
#geom_label_repel(data = subset(HaloDom2, Residue=="Methanohalophilus mahii"), show.legend
= FALSE, inherit.aes = TRUE, colour="black", ylim=(-5) ,xlim=(0), force=10, fontface="italic")
#geom_label_repel(data = subset(HaloDom2, Residue=="Halanaerobium praevalens"), show.legend
= FALSE, inherit.aes = TRUE, colour="black", ylim=(-4) ,xlim=(-8), force=10, fontface="italic")
#geom_label_repel(data = subset(HaloDom2, Residue=="HA_Haloarcula_salaria"), show.legend =
FALSE, inherit.aes = TRUE, colour="black", ylim=(5.7) ,xlim=(-6), force=10, fontface="italic")+
#geom_label_repel(data = subset(HaloDom2, Residue=="Halococcus hamelinensis"), show.legend =
FALSE, inherit.aes = TRUE, colour="black", ylim=(-4) ,xlim=(-6), force=10, fontface="italic")
#geom_label_repel(data = subset(HaloDom2, Residue=="Rhodothermus marinus"), show.legend =
FALSE, inherit.aes = TRUE, colour="black", ylim=(-5.8) ,xlim=0, force=10, fontface="italic")+
#geom_label_repel(data = subset(HaloDom2, Residue=="Salinibacter ruber"), show.legend = FALSE,
inherit.aes = TRUE, colour="black", ylim=(-5) ,xlim=2, force=10, fontface="italic")+
#geom_label_repel(data = subset(HaloDom2, Residue=="Salisaeta longa"), show.legend = FALSE,
inherit.aes = TRUE, colour="black", ylim=(-5.5) ,xlim=(-2), force=10, fontface="italic")+
#geom_label_repel(data = subset(HaloDom2, Residue=="Salinivenus iranica"), show.legend =
FALSE, inherit.aes = TRUE, colour="black", ylim=(-5) ,xlim=(-8), force=10, fontface="italic")+
#geom_label_repel(data = subset(HaloDom2, Residue=="Salinivenus lutea"), show.legend = FALSE,
inherit.aes = TRUE, colour="black", ylim=(-4) ,xlim=(-9.5), force=10, fontface="italic")+
#geom_label_repel(data = subset(HaloDom2, Residue=="Halorhodospira halophila"), show.legend =
FALSE, inherit.aes = TRUE, ylim=(3.5) ,xlim=(0), force=10, fontface="italic", colour="white",
segment.colour = "black")+
#geom_label_repel(data = subset(HaloDom2, Residue=="Haliangium ochraceum"), show.legend =
FALSE, inherit.aes = TRUE, ylim=(3.5) ,xlim=(-3), force=10, fontface="italic", colour="white",
segment.colour = "black")+
#geom_label_repel(data = subset(HaloDom2, Residue=="Haloquadratum walsbyi"), show.legend =
FALSE, inherit.aes = TRUE, ylim=(3.5) ,xlim=(-9.5), force=10, fontface="italic", colour="black")+
#geom_label_repel(data = subset(HaloDom2, Residue=="Martelella endophytica"), show.legend =
FALSE, inherit.aes = TRUE, ylim=(2.5) ,xlim=(-9.5), force=10, fontface="italic", colour="white",
segment.colour = "black")
#geom_label_repel(data = subset(HaloDom2, Residue=="Spiribacter salinus"), show.legend =
FALSE, inherit.aes = TRUE, ylim=(-4) ,xlim=(2.3), force=10, fontface="italic", colour="white",
segment.colour = "black")+
#geom_label_repel(data = subset(HaloDom2, Residue=="Spiribacter curvatus"), show.legend =
FALSE, inherit.aes = TRUE, ylim=(-5.5) ,xlim=(-2), force=10, fontface="italic", colour="white",
segment.colour = "black")+
#geom_label_repel(data = subset(HaloDom2, Residue=="Halomonas aestuarii"), show.legend =
FALSE, inherit.aes = TRUE, ylim=(-5) ,xlim=(1), force=10, fontface="italic", colour="white",
segment.colour = "black")+
#geom_label_repel(data = subset(HaloDom2, Residue=="Celeribacter indicus"), show.legend =
FALSE, inherit.aes = TRUE, ylim=(3.5) ,xlim=(2), force=10, fontface="italic", colour="white",
segment.colour = "black")+
#geom_label_repel(data = subset(HaloDom2, Residue=="Chromohalobacter salexigens"),
show.legend = FALSE, inherit.aes = TRUE, ylim=(-3) ,xlim=(-4), force=10, fontface="italic",
colour="white", segment.colour = "black")+
#geom_label_repel(data = subset(HaloDom2, Residue=="Halomonas elongata"), show.legend =
FALSE, inherit.aes = TRUE, ylim=3 ,xlim=(-1), force=10, fontface="italic", colour="white",
segment.colour = "black")
#geom_label_repel(data = subset(HaloDom2, Residue=="Methanoculleus marisnigri"), show.legend
= FALSE, inherit.aes = TRUE, ylim=3 ,xlim=(-1), force=10, fontface="italic", colour="white",
segment.colour = "black")+
#geom_label_repel(data = subset(HaloDom2, Residue=="Thermofilum pendens"), show.legend =
FALSE, inherit.aes = TRUE, ylim=4 ,xlim=(-3.5), force=10, fontface="italic", colour="white",
segment.colour = "black")
```

R script A2. “PCA_analysis_EIGEN.r”

```
#PCA analysis for amino acid profiles of all data
HaloDom = read.csv("C:\\Users\\Alex\\Desktop\\Perl works\\Phylogenomics\\Amino acid
compositions\\Clustering of all Halophiles\\PCA\\GC content
test\\PCA_wGC_HAvsHalanerobiales.csv", header = TRUE)
data(HaloDom)
head(HaloDom)
summary(HaloDom)
View(HaloDom)
Haloactive <- HaloDom[1:35, 3:23]
head(Haloactive[, 1:6])
library(factoextra)
res.pca <- prcomp(Haloactive, scale = TRUE)
fviz_eig(res.pca)

fviz_pca_var(res.pca,
        col.var = "contrib", # Color by contributions to the PC
        gradient.cols = c("green", "orange", "purple", "red"),
        repel = TRUE,     # Avoid text overlapping
        arrowsize = 1,
        labelsize = 6
)
```

R script A3. “protein_length_historgam.r”

```
lengths = read.csv("C:\\Users\\Alex\\Desktop\\Perl works\\Phylogenomics\\Amino acid
compositions\\Protein location,function,size and AA residues\\protein
lengths\\all_halo_lengths.csv", header = TRUE)
x<-lengths$Length
max(x)
y=seq(0,14000,by=1000)

# histogram with added parameters
hist(x,
    main="Halophilic protein lengths",
    xlab = "Length (AA)",
    ylab = "Frequency",
    font.lab = 2,
    xaxt="n",
    yaxt="n",
    col="red",
    breaks = 1000,
    xlim=c(0,1500),
    ylim=c(0,14000),
    freq=TRUE
)
  axis(1,at=seq(0,1500,by=100),font=2)
  axis(2,at=y,font=2)



#newdata <- x[(x > 5000)]
#View(newdata)
```

R script A4. “PCA_GC_analysis.r”

```
#PCA analysis for amino acid profiles of data with GC content added
PCA_GC = read.csv("C:\\Users\\Alex\\Desktop\\Perl works\\Phylogenomics\\Amino acid
compositions\\Clustering of all Halophiles\\PCA\\GC content
test\\PCA_wGC_HAvsHalanerobiales_test.csv", header = TRUE)
data(PCA_GC)
head(PCA_GC)
summary(PCA_GC)
View(PCA_GC)
PCA <- prcomp(PCA_GC[3:22])
summary(PCA)
plot(PCA)
plot(PCA, type='l')
biplot(PCA, scale=0)
str(PCA)
PCA$x
PCA_GC2 <- cbind(PCA_GC, PCA$x[,1:2])
head(PCA_GC2)
library(ggplot2)
require("ggrepel")
ggplot(PCA_GC2, aes(PC1, PC2, label=Residue, fill=category))+
guides(fill=guide_legend(title="Taxonomic Groups", title.theme = element_text(colour = "black",
size = 12, angle = 0, face = "bold"), title.hjust = -3))+
scale_fill_manual(values=c("darkgreen", "red"))+
#stat_ellipse(data = subset(PCA_GC2, category=="Halophilic_Archaea"), geom="polygon",
col="black", alpha="0.5")+
#stat_ellipse(data = subset(HaloDom2, category=="Halophilic_Archaea"), geom="polygon",
col="black", alpha="0.5")+
geom_point(shape=21,color="black",size=4)
#geom_label_repel(data = subset(HaloDom2, category=="Bacteroidetes"), show.legend = FALSE,
inherit.aes = TRUE, colour="black", ylim = y_limits, xlim=x_limits, force=10)+
#geom_label_repel(data = subset(HaloDom2, Residue=="B_Salinivenus_lutea"), show.legend =
FALSE, inherit.aes = TRUE, colour="black", ylim =(-5.7), xlim = (1) , force=10)+
#geom_label_repel(data = subset(HaloDom2, Residue=="B_Salinivenus_iranica"), show.legend =
FALSE, inherit.aes = TRUE, colour="black", ylim =(-5.7), xlim = (-3) , force=10)+
#geom_label_repel(data = subset(HaloDom2, Residue=="B_Haliangium_ochraceum"), show.legend
= FALSE, inherit.aes = TRUE, colour="black", ylim =(-5.7), xlim=10, force=10)+
#geom_label_repel(data = subset(HaloDom2, Residue=="B_Halorhodospira_halophila"),
show.legend = FALSE, inherit.aes = TRUE, colour="black", ylim =(-5.7), xlim=(-8.9), force=10)
#geom_label_repel(data = subset(HaloDom2, Residue=="HA_Halorubrum_ezzemoulense"),
show.legend = FALSE, inherit.aes = TRUE, colour="black",xlim = x_limits5, ylim = y_limits5,
force=10)+
#geom_label_repel(data = subset(PCA_GC2, Residue=="Methanosarcina acetivorans"),
show.legend = FALSE, inherit.aes = TRUE, colour="black", ylim =(1.2), xlim=(-2.8), force=10)+
#geom_label_repel(data = subset(PCA_GC2, Residue=="Methanosalsum zhilinae"), show.legend =
FALSE, inherit.aes = TRUE, colour="black", ylim=(0.5) ,xlim=(-2.8), force=10, fontface="italic")+
#geom_label_repel(data = subset(PCA_GC2, Residue=="Methanohalophilus halophilus"),
show.legend = FALSE, inherit.aes = TRUE, colour="black", ylim=(0.25) ,xlim=(-1), force=10,
fontface="italic")+
#geom_label_repel(data = subset(PCA_GC2, Residue=="Methanohalobium evestigatum"),
show.legend = FALSE, inherit.aes = TRUE, colour="black", ylim=(-0.95) ,xlim=(-2), force=10,
fontface="italic")+
#geom_label_repel(data = subset(PCA_GC2, Residue=="Methanohalophilus mahii"), show.legend =
FALSE, inherit.aes = TRUE, colour="black", ylim=(0) ,xlim=(-1), force=10, fontface="italic")
PCA_GC2$PC1
```

R script A5. “LDA_with_unknown.r”

```
#LDA
data(iris)
View(iris)
fix(iris)
attach(iris)
library(MASS)
out1=lda(Species~., iris)
scores=predict(out1, iris)$x
plot(scores, col=rainbow(3)[iris$Species], asp=1)

#LDA : Classifying new observations
unknown=read.csv("C:\\Users\\Alex\\Documents\\R\\iris_un.csv")
fix(unknown)
predict(out1, unknown)$class
#add LDA prediction to the plot
out1p=predict(out1, unknown)
scores_unknown=out1p$x
points(scores_unknown, pch=19)
```

R script A6. “halopredictor_local_LDA.r”

```
#LDA
halo=read.csv("C:/Halopredictor_Local_V1/LDA_data.csv")
#View(halo)
#fix(halo)
#attach(halo)
library(MASS)
out1=lda(type~., halo)
#out1
scores=predict(out1, halo)$x
predictions <- predict(out1, halo)
#predictions
#mean(scores)

#LDA : Classifying new observations
unknown=read.csv("C:/Halopredictor_Local_V1/halo_un.csv")
#fix(unknown)
pred = predict(out1, unknown)$class
pred_for_plot = predict(out1, unknown)
query_seq=pred_for_plot$x
y<-query_seq[1,1]
vector <- c(pred,y)
write.table(rbind(vector), file = "C:/Halopredictor_Local_V1/r_out.csv", row.names =FALSE,
col.names = FALSE,sep = ",")
```

R script A7. ”stats.r”

```
library(scales)
data <- read.csv("C:/Halopredictor_Local_V1/results/halo_results.csv")


total <- nrow(data) #get the number of rows for data.frame "data".


LDA_groups <- split(data, data$Group) #split the data frame according to halophilic prediction

LDA_halo <- as.data.frame(LDA_groups[1]$`1`$LDA) #keep only halophilic rows

perc_halo <- (nrow(LDA_halo))/total #calculate percentage of halophilic predictions

perc_halo_formated <- percent_format(big.mark = ",", suffix = " %")(perc_halo) #use function for formating the percentage

LDA_nonhalo <- as.data.frame(LDA_groups[2]$`2`$LDA) #keep only non-halophilic rows

perc_nonhalo <- (nrow(LDA_nonhalo))/total #calculate percentage of halophilic predictions

halo_stats <- data.frame(total,perc_halo_formated) #create data frame with values

names(halo_stats) <- c('Total sequences', 'Percentage of halophilic sequences') #change the name of columns

write.csv(halo_stats, 'C:/Halopredictor_Local_V1/results/statistics.csv' , row.names = FALSE)
```

R script A8. ”halopredictor_local_visualize.r”

```
# Multiple plot function
#
# ggplot objects can be passed in ..., or to plotlist (as a list of ggplot objects)
# - cols:   Number of columns in layout
# - layout: A matrix specifying the layout. If present, 'cols' is ignored.
#
# If the layout is something like matrix(c(1,2,3,3), nrow=2, byrow=TRUE),
# then plot 1 will go in the upper left, 2 will go in the upper right, and
# 3 will go all the way across the bottom.
#
sink("C:/Halopredictor_Local_V1/logfile.txt")
multiplot <- function(..., plotlist=NULL, file, cols=1, layout=NULL) {
  require(grid)

  # Make a list from the ... arguments and plotlist
  plots <- c(list(...), plotlist)

  numPlots = length(plots)

  # If layout is NULL, then use 'cols' to determine layout
  if (is.null(layout)) {
    # Make the panel
    # ncol: Number of columns of plots
    # nrow: Number of rows needed, calculated from # of cols
    layout <- matrix(seq(1, cols * ceiling(numPlots/cols)),
                    ncol = cols, nrow = ceiling(numPlots/cols))
  }

  if (numPlots==1) {
    print(plots[[1]])
  } else {
    # Set up the page
    grid.newpage()
    pushViewport(viewport(layout = grid.layout(nrow(layout), ncol(layout))))

    # Make each plot, in the correct location
    for (i in 1:numPlots) {
      # Get the i,j matrix positions of the regions that contain this subplot
      matchidx <- as.data.frame(which(layout == i, arr.ind = TRUE))

      print(plots[[i]], vp = viewport(layout.pos.row = matchidx$row,
                                      layout.pos.col = matchidx$col))
    }
  }
}
#library ("Hmisc")
#LDA : Classifying results from halo_results.csv
results=read.csv("C:/Halopredictor_Local_V1/results/halo_results.csv")
LDAs<-results$LDA

#Data frames editing for ggplot
#checking if both groups are present
new_type <- results$Group
#supress any error messages
try(Proteins <- factor(new_type, labels = c("Halophilic", "Non-halophilic")), silent = TRUE)
try(df <- data.frame(LDAs,Proteins, stringsAsFactors = FALSE), silent = TRUE)
#setting limits for LDA density graph
min<-min(results$LDA)
min_final<-min+(min*0.5)
max<-max(results$LDA)
max_final<-max+(max*0.5)
```

```
#trying to plot separate densities - did it !
library(ggplot2)
myplot<-ggplot(df,aes(x=LDAs,fill=Proteins))+geom_density(alpha=0.5)+
  scale_fill_manual( values = c("red","green"))+
  labs(title="Density plot of linear discriminants of input proteins")+xlab('Linear discriminants')+
  ylab('Density')+xlim(min_final,max_final)
#now plot to a file
ggsave("C:\\Halopredictor_Local_V1\\results\\density_plot.png", width = 28, height = 8, units =
"cm")


#Violin plot

p <- ggplot(results, aes(x=Prediction, y=LDA, fill=Prediction)) +
  geom_violin(trim=FALSE)+stat_summary(fun.data="mean_sdl", mult=1,geom="pointrange",
color="black")+
  scale_fill_manual(values=c("red", "green"))
#now plot to a file
ggsave("C:\\Halopredictor_Local_V1\\results\\violin_plot.png", width = 17, height = 12, units =
"cm")


pdf("C:/Halopredictor_Local_V1/results/graphs.pdf")
multiplot(myplot,p)
dev.off()
```

R script A9

```
library(micropan)
library(R.utils)    # for de-compressing files
library(ggdendro)   # plotting dendrogram tree
library(ggplot2)
library(tidyr)
library(readr)
library(purrr)
library(tibble)
library(stringr)
library(forcats)
library(ape)
library(cluster)
library(dendextend)
library(circlize)


setwd('C:\\Users\\Alex\\Desktop\\test_my_data')



#The genome table
suppressMessages(read_delim("rawdata/halobacteria_complete_genomes.txt", delim = ",")) %>%
  select(Name = `#Organism Name`, Strain, Level, GenBank_FTP = `GenBank FTP`, Taxonomy =
'my_taxonomy') %>%
  mutate(GID.tag = str_c("GID", 1:n())) %>%
  mutate(GenBank_ID = str_remove(GenBank_FTP, "^.+/")) %>%
  slice(1:80) -> gnm.tbl


gnm.tbl$Taxonomy <- as.factor(gnm.tbl$Taxonomy)
```

```
df <- data.frame(gnm.tbl)


#Pan-matrix creation
panmat.blast<-read.csv('C:\\Users\\Alex\\Desktop\\test_my_data\\nr_panmat.csv', header = TRUE, row.names = 1)

#Relation between genomes
#weighted dendrogram for core-cloud genes
pm <- panmat.blast                                        # make a copy

#sort the pan-matrix without GID-tags, only numbers as rownames
ordered_pm <- pm[order(as.numeric(as.character(rownames(pm)))), ]
#add GID prefix from csv file
prefix<-read.csv('C:\\Users\\Alex\\Desktop\\test_my_data\\GID_codes.csv')
prefix <- as.matrix(prefix)
row.names(ordered_pm) <- prefix


rownames(ordered_pm) <- df$Name[match(rownames(ordered_pm), df$GID.tag)] # new rownames
#replace the original pan-matrix with the ordered one
pm<-ordered_pm
weights <- geneWeights(pm, type = "shell")


#my ultimate circular dendrogram

dist_circ<-distManhattan(pm, weights = weights)
```

```
#test<-
c(2,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1
,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,2)
#test<-as.numeric(df$Taxonomy)

colors<-as.numeric(df[,5])
#colors <- c(2, 3, 1, 1, 3, 1, 3, 1, 3, 2, 3, 2, 3, 1, 1, 1, 2, 2, 1, 3, 1, 1, 2, 1, 2, 1, 2, 2, 3, 2, 1, 1, 1, 3, 2, 1,
3, 1, 2, 1, 1, 3, 2, 2, 1, 2, 2, 1, 2, 1, 2, 3, 1, 3, 2, 1, 2, 3, 3, 3, 3, 2, 1, 1, 2, 2, 1, 2, 1, 1, 1, 3, 1, 1, 1, 1)

#changing the colors to my preferrence
library(plyr) #PROBLEMA -> must be called here, not on the top.
colors<-mapvalues(colors, 1, "slateblue1")
colors<-mapvalues(colors, 2, "turquoise4")
colors<-mapvalues(colors, 3, "violetred2")



hc <- hclust(dist_circ, method = "average")  #methods used complete in default and average
dend <- as.dendrogram(hc)%>%
  set("branches_k_color", colors, order_value = TRUE) %>%
  set("branches_lwd", 1.9) %>%
  set("labels_colors", colors, order_value = TRUE) %>%
  set("labels_cex", 0.8)
  #set("nodes_pch", 19) %>%
  #set("nodes_col", c("black","red","orange"))

fig1<-circlize_dendrogram(dend,dend_track_height = 0.3,labels = TRUE,labels_track_height = 0.6)
```

R script A10

```
library(ggplot2)

mydata<-read.csv('C:\\Users\\Alex\\Documents\\R\\Pan-genome
graphs\\full_pangenome_stats.csv')

ggplot(mydata, aes(factor(Annotation), Percentage, fill = Category)) +
  geom_bar(stat="identity", position = "dodge")+
  theme(axis.text.x = element_text(angle = 90, hjust = 1,face = "plain",size=13))+
  labs(title = "Functional annotation of Halobacteria pan-genome", x = "Functional group", y =
"Percentage(%)", fill= "Gene category")
  #scale_fill_manual(values = c("yellow", "red"))
```