12-2015

Application of Hidden Markov Model based methods for gaining insights into protein domain evolution and function

Amit Anil Upadhyay
University of Tennessee - Knoxville, aupadhy1@vols.utk.edu

To the Graduate Council:

I am submitting herewith a dissertation written by Amit Anil Upadhyay entitled "Application of Hidden Markov Model based methods for gaining insights into protein domain evolution and function." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Life Sciences.

Igor B. Jouline, Major Professor

We have read this dissertation and recommend its acceptance:

Elias Fernandez, Jerome Baudry, Tongye Shen, Mircea Podar

Accepted for the Council:
Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Application of Hidden Markov Model based methods for gaining insights into protein domain evolution and function

A Dissertation Presented for the

Doctoral of Philosophy

Degree

The University of Tennessee, Knoxville

Amit Anil Upadhyay

December 2015

DEDICATION

This dissertation is dedicated to my family and my best friend, Deepika.

ACKNOWLEDGEMENTS

I am extremely grateful to my advisor, Dr. Igor Jouline, for all the guidance and support. His passion and approach towards science has been an inspiration. I would also like to thank Dr. Bhanu Rekepalli for his mentoring and support and giving me an opportunity to expand my computational skills. I am grateful to my committee members, Dr. Elias Fernandez, Dr. Jerome Baudry, Dr. Tongye Shen and Dr. Mircea Podar for their time and helpful suggestions. I would like to extend special thanks to Dr. Leonid Sukharnikov who acted as a mentor and a friend when I joined Dr. Jouline's lab. I also extend my gratitude to other lab members – Dr. Luke Ulrich, Dr. Davi Ortega, Dr. Kirill Borziak, Dr. Aaron Fleetwood and Ogun Adebali for stimulating discussions and their help in various projects. Finally I would like to thank the Genome Science and Technology program for giving me the opportunity to achieve my career goals.

ABSTRACT

With the explosion in the amount of available sequence data, computational methods have become indispensable for studying proteins. Domains are the fundamental structural, functional and evolutionary units that make up proteins. Studying protein domains is an important part of understanding protein function and evolution. Hidden Markov Models (HMM) are one of the most successful methods that have been applied for protein sequence and structure analysis. In this study, HMM based methods were applied to study the evolution of sensory domains in microbial signal transduction systems as well as functional characterization and identification of cellulases in metagenomics datasets. Use of HMM domain models enabled identification of the ambiguity in sequence and structure based definitions of the Cache domain family. Cache domains are extracellular sensory domains that are present in microbial signal transduction proteins and eukaryotic voltage gated calcium channels. The ambiguity in domain definitions was resolved and more accurate HMM models were built that detected more than 50,000 new members. It was discovered that Cache domains constitute the largest family of extracellular sensory domains in prokaryotes. Cache domains were also found to be remotely homologous to PAS domains at the level of sequence, a relationship previously suggested purely based on structural comparisons. We used HMM-HMM comparisons to study the diversity of extracellular sensory domains in prokaryotic signal transductions systems. This approach allowed annotation of more than 46,000 sequences and reduced the percentage of unknown domains from 64% to 15%. New relationships were also discovered between domain families that were otherwise thought to be unrelated. Finally,

HMM models were used to retrieve Family 48 glycoside hydrolases (GH48) from sequence databases. Analysis of these sequences, enabled the identification of distinguishing features of cellulases. These features were used to identify GH48 cellulases from metagenomics datasets. In summary, HMM based methods enabled domain identification, remote homology detection and functional characterization of protein domains.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ATTACHMENTS

Appendix-1.1.xlsx - HHsearch results for Cache superfamily

Appendix-1.2.xlsx - Overlapping hits with Cache domain prediction using new models

Appendix-1.3.xlsx - Phyletic distribution of Cache, PAS and GAF superfamilies

Appendix-2.1.xlsx - Supplementary Information for Chapter II

Appendix-3.1.pdf – Supplementary Information for Chapter III

Appendix-3.2.xlsx – Supplementary Information for Chapter III

INTRODUCTION

Proteins are complex biomolecules that perform a myriad of functions such as catalysis, transport of nutrients, recognition and transmission of signals, structural elements and molecular machines [1, 2]. Thus, an accurate understanding of protein function is crucial for understanding life at the molecular level [3]. With the advent of next generation sequencing technologies, there is an explosion in the amount of available information [4]. Presently more than 13,000 genomes are available from National Center for Biotechnology Information (NCBI) [5] and the RefSeq database [6] contains 52 million protein sequences. The number of unannotated proteins are two orders of magnitude larger compared to the annotated proteins and this difference is only getting larger [7]. It is not feasible to experimentally characterize each protein thus making computational methods indispensable for studying proteins [3].

Protein Domains

Proteins are composed of a linear chain of amino acids that are connected by covalent peptide bonds. Each protein has a unique sequence of amino acids that ultimately determines its three-dimensional structure and function [8]. Domains are the fundamental units of proteins that have been defined based on structure or sequence [9]. Based on the structural aspect, domains are defined as the part of the polypeptide chain that can fold independently into a functional compact stable 3D structure [9-11]. The two most important databases that classify domains on the basis of structure include Structural Classification of Proteins (SCOP) [12] and CATH [13]. Domains are also defined as evolutionarily conserved independent units that can be present in different molecular

contexts [9]. The main repositories for sequence based domain families include Protein family database (Pfam) [14], Simple Modular Architecture Research Tool (SMART) [15] and Conserved Domains Database (CDD) [16]. There is usually a general agreement and overlap between structure based and sequence based definitions [17, 18] but the domain boundaries are rarely in agreement [19].

The length of domains vary from 40 to 700 amino acids with an average of 100 [20-22]. It has been estimated that two-thirds of all prokaryotic proteins and four-fifths of eukaryotic proteins have more than one domain [23]. Protein domains serve as building blocks and a relatively small number of domains can form different combinations giving rise to much larger number of unique proteins [24]. Thus, classifying proteins based on the domain composition is an efficient way to manage protein data [18]. Since multi-domain proteins may have domains from different families, an accurate prediction of function would require characterization of individual domains [9].

Homology

Detection of homologs is one of the most important aspects of protein sequence analysis with applications in protein function prediction, protein structure prediction and protein evolution [25]. Homologous proteins are those that have descended from a common ancestor and are expected to have similar amino acid sequences which in turn would confer similar structure and function [26, 27]. Homologs are further classified into different groups based on evolutionary events – (i) orthologs result from speciation events; (ii) paralogs result from divergence after gene duplication events and (iii) xenologs result from divergence after horizontal gene transfer events [28]. Orthologs usually perform the

same function while paralogs and xenologs have related but slightly different functions [28].

Sequence similarity can be used to find homologous proteins. However, it should be noted that sequence similarity can either be due to homology or convergent evolution [29]. A direct relationship between sequence, 3D structure and function has been proven only in case of globular proteins [29]. In case of non-globular segments such as coiled coils, transmembrane helices and disordered regions, the similarity may be a result of physico-chemical constraints resulting in amino acid composition bias or repetitive patterns [29]. Similarly, protein structure similarity may not always be due to homology [30]. The proteins that share similar structures but have little or no sequence similarity are known as analogs [28].

Methods for detecting homologs

In spite of comprehensive efforts such as the Protein Structure Initiative [31], there is an ever increasing gap between the number of experimentally determined structures and the number of gene sequences. The number of available structures is 200 times smaller than the number of sequences and it is estimated that achieving a coverage of 55% will take another 15 years [31]. Even though structure-based methods may provide a more complete and well-defined domain definition, it will be severely limited to a small number of proteins [29]. Thus sequence-based methods continue to be an important part of studying proteins.

The first generation of tools for finding homologs were based on sequence-sequence comparison where a query sequence was compared to a protein sequence database.

The earliest sequence alignment methods were based on dynamic programming that performed local alignment (Smith-Waterman) [32] or global alignment (Needleman-Wunsch) [33]. Since proteins may undergo several evolutionary events such as domain duplication, insertion or permutation, resulting in different parts of the protein being homologous to domains from different proteins [34], local alignment methods are more preferred. The dynamic programming approach could not keep up with the increasing size of protein databases and as a result heuristic methods were developed. The most popular tool for local alignment is BLAST (Basic Local Alignment Search Tool) [35]. The BLAST tool breaks down the query into a set of words and compares these words to a set of words generated from the protein sequence database. Matching words are used to initiate a gap free extension. The extensions meeting a specified score are then used to seed gapped extensions. Finally gapped extensions that meet a specified score are further used to calculate insertions and deletions. The BLAST algorithm returns an expect value that can be used to determine the significance of the match. The expect value estimates the number of matches that may occur randomly with a given score. Pairwise alignment methods such as BLAST assume that all positions are equally important [36]. Although sequence-sequence comparison is one of the easiest ways to detect related proteins, it cannot be used to detect remote relationships. Only half of the proteins known to share evolutionary relationships based on sequence, structure and function with sequence identity in the range of 20%-30% can be detected by pairwise comparisons [37]. In order to detect more distantly related sequences, multiple sequence alignment based methods such as profiles [38] and hidden Markov models (HMM) [39, 40] were

developed. These methods are more sensitive since they take into account evolutionary history of the proteins by identifying conserved and variable positions in the protein sequence [30]. The profile based methods can be used in three different ways – query profile against sequence database, query sequence against profile database and query profile against profile database [41].

Position Specific Iterated BLAST (PSI-BLAST) [42, 43] is one of the most popular methods for detecting remote homologs that uses Position Specific Scoring Matrix (PSSM) called profiles that are created dynamically each time the search is initiated. The PSSM incorporates the frequency of amino acids at each position. It starts with a BLAST search in the first iteration and for each subsequent iteration, a profile is built using similar sequences found in the previous iteration. It is a very sensitive method and can detect homologs that can only be retrieved by structural comparison [44]. However, due to the iterative nature, addition of non-homologous sequence may not be easily detected which may sometimes result in good scores for even unrelated proteins [30].

Hidden Markov Model (HMM) is a general statistical modeling technique [45]. It can be used for linear problems such as time series and sequences and has been widely used in protein classification, motif detection, multiple sequence alignments and protein structural modeling [36, 40]. Profile HMMs are used for modeling sequence conservation. Profile HMMs consist of a linear left-to-right structure with three different states – match, delete and insert [36, 40, 45]. Each match state has an emission distribution that corresponds to the probabilities of observing an amino acids in a given position. Every match state is also accompanied by a delete state and an insert state. Since HMMs can

handle insertions and deletions, they are considered to be superior to matrix-based profile methods. The HMMer package is the most widely used tool for using HMM for sequence analysis [46].

The profile based methods have also been extended to carry out profile-profile comparisons. These methods can identify 20-30% more homologs than PSI-BLAST [47-49]. FFAS [25], COACH [50] and COMPASS [48, 51] are examples of profile-profile comparison tools. HHsearch [52, 53] is one of the most popular tools for HMM-HMM comparison [54].

Scope of dissertation

This dissertation will describe three studies of protein domains where HMM-based methods played a vital role. Chapter I deals with overcoming the ambiguity between sequence based Cache domains and structure based PDC domains. The relationship between Cache and PAS domains at the level of sequence was also investigated. Using HMM along with other bioinformatics methods, new models were built, thousands of new members identified and remote relationship between PAS and Cache domains established. Chapter II focusses on understanding the diversity of extracellular sensory domains in all prokaryotic signal transduction systems. The most sensitive sequence comparison technique of HMM-HMM comparison was applied towards this problem. Almost 75 percent of all sensory domains were found to belong to either the Cache clan or the 4HB_MCP clan. In addition, the percentage of unannotated domains was reduced from 64% to 15% and several relationships between unrelated families were also discovered. In Chapter III, the goal was to understand the distinguishing characteristics

of cellulases so that they can be identified unambiguously from genomic datasets. The glycoside hydrolase family 48 (GH48) was chosen for this study. The features that distinguish cellulases from other enzymes in GH48 family were determined. These features were subsequently used to screen metagenomic datasets to identify GH48 cellulases. The conclusion will summarize these studies and also discuss future prospects.

References

1. Rost, B., et al., *Automatic prediction of protein function.* Cell Mol Life Sci, 2003. **60**(12): p. 2637-50.
2. Eisenberg, D., et al., *Protein function in the post-genomic era.* Nature, 2000. **405**(6788): p. 823-826.
3. Radivojac, P., et al., *A large-scale evaluation of computational protein function prediction.* Nat Methods, 2013. **10**(3): p. 221-7.
4. Ståhl, P.L. and J. Lundeberg, *Toward the single-hour high-quality genome.* Annual review of biochemistry, 2012. **81**: p. 359-378.
5. Wheeler, D.L., et al., *Database resources of the national center for biotechnology information.* Nucleic acids research, 2007. **35**(suppl 1): p. D5-D12.
6. Pruitt, K.D., et al., *NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy.* Nucleic Acids Research, 2012. **40**(D1): p. D130-D135.
7. Clark, W.T. and P. Radivojac, *Analysis of protein function and its prediction from amino acid sequence.* Proteins, 2011. **79**(7): p. 2086-96.
8. Alberts, B., et al., *Molecular Biology of the Cell (Garland Science, New York, 2002).* There is no corresponding record for this reference, 1997.
9. Ponting, C.P. and R.R. Russell, *The natural history of protein domains.* Annual review of biophysics and biomolecular structure, 2002. **31**(1): p. 45-71.
10. Wetlaufer, D.B., *Nucleation, rapid folding, and globular intrachain regions in proteins.* Proc Natl Acad Sci U S A, 1973. **70**(3): p. 697-701.
11. Richardson, J.S., *The anatomy and taxonomy of protein structure*. Vol. 34. 1981: Academic Press.
12. Murzin, A.G., et al., *SCOP: a structural classification of proteins database for the investigation of sequences and structures.* Journal of molecular biology, 1995. **247**(4): p. 536-540.
13. Orengo, C.A., et al., *CATH–a hierarchic classification of protein domain structures.* Structure, 1997. **5**(8): p. 1093-1109.
14. Finn, R.D., et al., *Pfam: the protein families database.* Nucleic acids research, 2013: p. gkt1223.
15. Schultz, J., et al., *SMART, a simple modular architecture research tool: identification of signaling domains.* Proceedings of the National Academy of Sciences, 1998. **95**(11): p. 5857-5864.
16. Marchler-Bauer, A., et al., *CDD: a Conserved Domain Database for the functional annotation of proteins.* Nucleic acids research, 2011. **39**(suppl 1): p. D225-D229.
17. Elofsson, A. and E. Sonnhammer, *A comparison of sequence and structure protein domain families as a basis for structural genomics.* Bioinformatics, 1999. **15**(6): p. 480-500.
18. Zhang, Y., et al., *Comparative mapping of sequence-based and structure-based protein domains.* BMC Bioinformatics, 2005. **6**: p. 77.
19. Ezkurdia, I. and M.L. Tress, *Protein structural domains: definition and prediction.* Curr Protoc Protein Sci, 2011. **Chapter 2**: p. Unit2 14.

20. Jones, S., et al., *Domain assignment for protein structures using a consensus approach: characterization and analysis.* Protein Science, 1998. **7**(2): p. 233-242.
21. Wheelan, S.J., A. Marchler-Bauer, and S.H. Bryant, *Domain size distributions can predict domain boundaries.* Bioinformatics, 2000. **16**(7): p. 613-618.
22. Rekapalli, B., et al., *Dynamics of domain coverage of the protein sequence universe.* BMC genomics, 2012. **13**(1): p. 634.
23. Chothia, C., et al., *Evolution of the protein repertoire.* Science, 2003. **300**(5626): p. 1701-3.
24. Teichmann, S.A., et al., *Small-molecule metabolism: an enzyme mosaic.* TRENDS in Biotechnology, 2001. **19**(12): p. 482-486.
25. Jaroszewski, L., et al., *FFAS03: a server for profile--profile sequence alignments.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W284-8.
26. Devos, D. and A. Valencia, *Practical limits of function prediction.* Proteins, 2000. **41**(1): p. 98-107.
27. Todd, A.E., C.A. Orengo, and J.M. Thornton, *Evolution of function in protein superfamilies, from a structural perspective.* J Mol Biol, 2001. **307**(4): p. 1113-43.
28. Wan, X.F. and D. Xu, *Computational methods for remote homolog identification.* Curr Protein Pept Sci, 2005. **6**(6): p. 527-46.
29. Wong, W.-C., S. Maurer-Stroh, and F. Eisenhaber, *More than 1,001 problems with protein domain databases: transmembrane regions, signal peptides and the issue of sequence homology.* PLoS Comput Biol, 2010. **6**(7): p. e1000867.
30. Pearson, W.R. and M.L. Sierk, *The limits of protein sequence comparison?* Curr Opin Struct Biol, 2005. **15**(3): p. 254-60.
31. Montelione, G.T., *The Protein Structure Initiative: achievements and visions for the future.* F1000 Biol Rep, 2012. **4**: p. 7.
32. Smith, T.F. and M.S. Waterman, *Identification of common molecular subsequences.* J Mol Biol, 1981. **147**(1): p. 195-7.
33. Needleman, S.B. and C.D. Wunsch, *A general method applicable to the search for similarities in the amino acid sequence of two proteins.* J Mol Biol, 1970. **48**(3): p. 443-53.
34. Russell, R.B. and C.P. Ponting, *Protein fold irregularities that hinder sequence analysis.* Current opinion in structural biology, 1998. **8**(3): p. 364-371.
35. Altschul, S.F., et al., *Basic local alignment search tool.* Journal of molecular biology, 1990. **215**(3): p. 403-410.
36. Eddy, S.R., *Hidden Markov models.* Curr Opin Struct Biol, 1996. **6**(3): p. 361-5.
37. Brenner, S.E., C. Chothia, and T.J. Hubbard, *Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships.* Proceedings of the National Academy of Sciences, 1998. **95**(11): p. 6073-6078.
38. Gribskov, M., A.D. McLachlan, and D. Eisenberg, *Profile analysis: detection of distantly related proteins.* Proc Natl Acad Sci U S A, 1987. **84**(13): p. 4355-8.
39. Eddy, S.R., *Profile hidden Markov models.* Bioinformatics, 1998. **14**(9): p. 755-763.
40. Yoon, B.-J., *Hidden Markov models and their applications in biological sequence analysis.* Current genomics, 2009. **10**(6): p. 402.

41. Fariselli, P., et al., *The WWWH of remote homolog detection: the state of the art.* Brief Bioinform, 2007. **8**(2): p. 78-87.
42. Schäffer, A.A., et al., *Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements.* Nucleic acids research, 2001. **29**(14): p. 2994-3005.
43. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
44. Aravind, L. and E.V. Koonin, *Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches.* Journal of molecular biology, 1999. **287**(5): p. 1023-1040.
45. Krogh, A., et al., *Hidden Markov models in computational biology. Applications to protein modeling.* J Mol Biol, 1994. **235**(5): p. 1501-31.
46. Eddy, S.R., *Accelerated profile HMM searches.* PLoS Comput Biol, 2011. **7**(10): p. e1002195.
47. Yona, G. and M. Levitt, *Within the twilight zone: a sensitive profile-profile comparison tool based on information theory.* Journal of molecular biology, 2002. **315**(5): p. 1257-1275.
48. Sadreyev, R. and N. Grishin, *COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance.* Journal of molecular biology, 2003. **326**(1): p. 317-336.
49. Rychlewski, L., et al., *Comparison of sequence profiles. Strategies for structural predictions using sequence information.* Protein Science, 2000. **9**(2): p. 232-241.
50. Edgar, R.C. and K. Sjölander, *COACH: profile–profile alignment of protein families using hidden Markov models.* Bioinformatics, 2004. **20**(8): p. 1309-1318.
51. Sadreyev, R.I., D. Baker, and N.V. Grishin, *Profile–profile comparisons by COMPASS predict intricate homologies between protein families.* Protein Science, 2003. **12**(10): p. 2262-2272.
52. Soding, J., *Protein homology detection by HMM-HMM comparison.* Bioinformatics, 2005. **21**(7): p. 951-60.
53. Soding, J., A. Biegert, and A.N. Lupas, *The HHpred interactive server for protein homology detection and structure prediction.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W244-8.
54. Remmert, M., et al., *HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment.* Nature methods, 2012. **9**(2): p. 173-175.

CHAPTER I: CACHE DOMAINS ARE DOMINANT EXTRACELLULAR SENSORS FOR SIGNAL TRANSDUCTION IN PROKARYOTES

This chapter was taken from a manuscript in preparation:

Amit A. Upadhyay, Aaron D. Fleetwood, Ogun Adebali, Robert D. Finn and Igor B. Zhulin, **Cache domains are dominant extracellular sensors for signal transduction in prokaryotes.**

I.B.Z and A.A.U designed research, A.A.U performed research, A.A.U, A.D.F, O.A, R.D.F and I.B.Z analyzed data and A.A.U., A.D.F. and I.B.Z. wrote the manuscript. O.A. generated the figure for Phyletic distribution.

Abstract

Cellular receptors usually contain a designated sensory domain that recognizes the signal. Per/Arnt/Sim (PAS) domains are ubiquitous sensors in thousands of species ranging from bacteria to humans. Although PAS domains were described as intracellular sensors, recent structural studies revealed PAS-like domains in extracytoplasmic regions in several transmembrane receptors. Here we show that structurally defined extracellular PAS-like domains belong to the Cache superfamily, which is homologous to, but distinct from the PAS superfamily. Our newly built computational models enabled identification of Cache domains in tens of thousands of signal transduction proteins including those from important pathogens and model organisms. Furthermore, we show that Cache domains comprise the dominant mode of extracellular sensing in prokaryotes.

Introduction

Signal transduction is a universal feature of all living cells. It is initiated by specialized receptors that detect various extracellular and/or intracellular signals, such as nutrients, and transmit information to regulators of different cellular functions [1, 2]. Receptors are usually comprised of several domains and one or more of them are designated sensors that physically interact with the signal. There is a great diversity in the sensory domain repertoire, but a few of these domains appear to be dominant. The most abundant sensory module that is found in tens of thousands of signal transduction proteins throughout the Tree of Life is the Per/Arnt/Sim (PAS) domain [3, 4]. PAS domains are related to another large group of dedicated sensors – cGMP phosphodiesterase/adenylyl cyclase/FhlA (GAF) domains [5, 6]: both superfamilies belong to the profilin-like fold [6, 7] and are found in similar types of signal transduction proteins in eukaryotes and prokaryotes. PAS and GAF are amongst the largest superfamilies of small molecule-binding domains in general, and the largest among those solely dedicated to signal transduction [8]. Originally, PAS domains were discovered as exclusively intracellular sensors [9, 10]; however more recent studies have identified several extracytoplasmic PAS domains. Members of this group include quorum- [11], dicarboxylate- [12, 13] and osmo-sensing [14] receptor kinases, and chemotaxis receptors [15, 16] from bacteria as well as the *Arabidopsis* cytokinin receptor [17] among others. As commonly accepted in structure-based approaches, these domains were termed PAS (or PAS-like) based on expert's visual inspection of three-dimensional structures. Surprisingly, none of these structurally defined domains matched any sequence-derived PAS domain models.

Furthermore, novel structural elements previously unseen in PAS domains have been noticed in some of these structures and a new name, PDC (acronym of three founding members, PhoQ, DcuS and CitA), has been suggested for these extracellular domains [18]. On the other hand, several independent observations suggest a link between extracellular PAS-like structures and yet another sensory domain superfamily, Cache [19]. Cache was originally described as a ligand-binding domain common to bacterial chemoreceptors [20] and animal voltage-dependent calcium channel subunits [21] that are targets for anti-neuropathic drugs [22]. First, the authors of the original Cache publication suggested a circular permutation of the Cache domain in extracellular regions of DcuS and CitA [19], proteins that later became the founding members of the proposed PDC domain [18]. Second, in their structural classification of PAS domains, Henry and Crosson [4] noted that a few sequences corresponding to structures included in their analysis were annotated as Cache in domain databases. Third, Zhang and Hendrickson reported that a conserved domain search detected the presence of a single Cache domain in their two related structures of the double PDC domain, namely 3LIA and 3LIB (PDB accession numbers), but not in the other three closely related structures of this domain, 3LIC, 3LID and 3LIF [23]. Nevertheless, these potential relationships with Cache have never been explored further and extracellular PAS-like domains are being referred to as PAS [4], PAS-like [14], PDC [18], PDC-like [24], and PDC/PAS [25] (Table 1.S1). Furthermore, there is no agreement between sequence- and structure-based classifications of these domains and associated structures provided by leading databases (Fig. 1.S1, Tables 1.S2 and 1.S3). The real problem beyond classification issues and

semantics is that other than a handful of examples with solved 3D structure, receptors containing these domains cannot be identified in current genomic datasets. This, in turn, is a barrier for practical applications, such as a proposed use of bacterial receptors as drug targets [26].

Here we show that extracellular PAS (PDC)-like domains belong not to PAS, but to the Cache superfamily. By building new Cache domain models utilizing structural information, we implicated more than 50,000 signaling proteins from all three domains of life as new members of this superfamily thus more than doubling the space of its current computational coverage. We also provide evidence that while being a distinct superfamily, Cache is homologous to the PAS superfamily and propose that the Cache domain emerged in bacteria from a simpler intracellular PAS ancestor as a benefit of extracellular sensing. Finally, we show that Cache domains are the dominant mode of extracellular sensing in prokaryotes.

Results

"Extracellular PAS" is Cache. To illustrate the level of ambiguity in classification of extracellular PAS/PDC-like domains (Table 1.S2) we compared it to that of diverse intracellular PAS domains from bacteria, archaea and eucarya (Table 1.S3). The results show a nearly perfect classification coverage and agreement between sequence- and structure-based definitions for the latter and a state of disarray for the former (Fig. 1.S1). We subjected protein sequences of all twenty-one single and double extracellular PAS-like domains[4] with known 3D structure to similarity searches against the Pfam database (v.27.0) using sequence-to-profile search tool, HMMscan [27] and a more sensitive,

profile-to-profile search tool HHpred [28]. None of the sequences had any PAS domain models as the best hit in any type of search. For fourteen of them (including both single and double domains), best hits were to domain models from the Cache superfamily, whereas for the remaining seven structures, best hits are not assigned to any domain superfamily (Table 1.S4).

Mapping regions matched to Cache domains onto corresponding structures revealed the nature of ambiguity between sequence- and structure-based domain definitions. Single domain structures showed better agreement with sequence-based domain models (Fig. 1.S2), although some of them still had substantial discrepancies. For example, the full-length Cache_2 model does not include the last three β-strands of the PAS-like domain (Fig. 1.1A). Dual domain structures showed major disagreements with sequence-based domain models. The Cache_1 model captures the last three strands from the membrane distal PAS-like domain, the first two strands of the membrane proximal domain, and the connecting elements between the two domains (Fig. 1.1B). Some of the most conserved structural elements, such as the long N-terminal helix captured in the Cache_2 model and connecting elements between two globular domains captured in the Cache_1 model, are never seen in proteins that belong to the PAS domain superfamily, which led to a suggestion that these domains are different from PAS [23]. We also confirmed that the long N-terminal helix in some of the double domain structures (Fig. 1.1B) matches a Pfam model MCP_N (Table 1.S4).

New Cache Domain Models. We used newly uncovered relationships between structure and sequence characteristics to construct new Cache domain models. Three key facts

about Cache domains were taken into account. First, structural studies revealed that both single and double Cache domains occupy the entire extracellular region between two transmembrane helices. Second, Cache domains have been identified exclusively in proteins that contain output signaling domains. Third, the vast majority of Cache domains are found in prokaryotes. Consequently, in order to identify potential Cache domains, we retrieved a non-redundant set of prokaryotic sequences that contained at least one output signaling domain and a predicted extracellular region flanked by two transmembrane helices (see Methods for details). The final set of predicted extracellular regions (non-redundant at 90% identity) was used in the hidden Markov model (HMM) construction. Models were built in three stages using sequence-to-sequence and HMM-to-HMM comparisons (see Methods for details). We constructed eight new Cache models to replace the current three models (Cache_1, Cache_2, and Cache_3) from Pfam 27.0 (Table 1.S5). The fourth current Pfam model from the Cache clan, YkuI_C, was found to adequately capture the domain structure and to perform well (Fig. 1.S2B). Two other members of the clan, DUF4153 and DUF4173 were found to be unrelated to Cache based on both sequence similarity and secondary structure prediction (Appendix 1.1, spreadsheet 3). Consequently, these models will be removed from the clan.

The new models revealed complex relationships between single and double Cache domains. HMM-HMM comparison showed that the membrane distal subdomain of dCache_1 was more similar to sCache_3, whereas the membrane proximal subdomain was more similar to sCache_2 (Fig. 1.S3). On the other hand, dCache_2 and dCache_3 domains appear to be a result of sCache_2 and sCache_3 duplication, respectively.

Finally, the Cache_3-Cache_2 domain likely originated as a fusion of sCache_3 and sCache_2 domains.

The new models demonstrated dramatically improved sensitivity by identifying more than 50,000 Cache domains in the NCBI non-redundant database that escaped detection by Pfam 27.0 models (Table 1.S6). A small number of newly identified Cache domains (~4%) overlapped with other non-Cache Pfam domains, such as MCP_N. TarH, VGCC_alpha2 and few others (Appendix 1.2, spreadsheet 1). As already discussed earlier, we consider MCP_N as a part of the Cache domain. Overlap with TarH is caused by inclusion of several Cache-domain containing sequences in the seed alignment for a model depicting an all alpha-helical TarH domain (Appendix 1.2, spreadsheet 2). VGCC_alpha2 is usually present C-terminal to the Cache domain in Calcium channel subunits and in fact is a C-terminal part of the Cache domain missing from a Pfam 27.0 seed alignment. Both Mcp_N and TarH models will be retired from Pfam. After correcting for these artifacts, the overlap of newly defined Cache domains with unrelated Pfam domains is less than 0.3%.

New models also showed a significantly improved average coverage (Table 1.S7). The average length of single and double Cache domains is 140 and 271 amino acid residues, respectively. Occasionally, single Cache domain models match to extracellular regions that are significantly larger than the average length of single Cache domains (Fig. 1.S4). Similarly, double Cache domain models occasionally match to extracellular regions with a size of a single Cache domain. This is likely due to the complex modular nature of these domains (Fig. 1.S3). We used sequences with known 3D structures as controls to visualize the increased specificity and coverage of the newly built Cache models (Table

1.S8). All new models will be available in the Pfam 29.0 release (January 2016) upon further refinement of seed alignments according to Pfam standard protocols.

New Members of the Cache Superfamily and Its Relationship to PAS and GAF. When carrying out sensitive profile-to-profile searches initiated with the sequences of extracellular “PAS-like” structures, we noticed statistically significant (although never the best) hits with profiles corresponding to several Pfam domains other than members of the current Cache clan. We explored this indication of potential remote homology further by consistently analyzing all statistically significant HHpred matches for all nineteen structures. The results show that statistically significant hits belong either to the PAS and GAF superfamilies or to small families that have not been assigned to any domain superfamily, for example LuxQ-periplasm, CHASE, Diacid_rec, etc (Appendix 1.1, spreadsheet 1). Nearly the same repertoire of small families and members of PAS and GAF superfamilies were statistically significant hits in HHpred searches initiated with newly constructed Cache models (Appendix 1.1, spreadsheet 2). Finally, we have performed a reverse search, where queries were models from small families as well as PAS and GAF superfamilies identified as statistically significant hits in the previous two types of searches (Appendix 1.1, spreadsheet 3). These searches have identified nine additional current Pfam families that lacked any superfamily assignments. We now assign these families to the Cache superfamily (Table 1.S5, Appendix 1.1, spreadsheet 4). Relationships between all members of the three superfamilies at profile and sequence levels are shown in Fig. 1.2. While being closely related to PAS and GAF, members of

the Cache superfamily are more related to each other, thus fully justifying a separate superfamily designation.

Cache are Ubiquitous Extracellular Sensors. By performing the HMMscan search against the Pfam 27.0 database using eighteen domain models from the newly defined Cache superfamily, we have identified 31,570 protein sequences containing these domains. Thus, the size of the Cache superfamily is comparable to that of PAS (88,093 sequences) and GAF (47,618 sequences) superfamilies. Phyletic distribution of Cache domains is also similar to that of PAS and GAF (Fig. 1.S5, Appendix 1.1, spreadsheet 2). We have used the TMHMM2 tool to identify transmembrane regions in all 31,570 sequences with detectable Cache domains and determined that members of all Cache families are predicted to be principally extracellular, except for two small families, Diacid_rec and YkuI_C that are principally intracellular (Table 1.S9). Altogether, 78% of all Cache domains were confidently predicted to be extracellular. For comparison, 74% of all PAS domains were confidently predicted to be intracellular. Analysis of the domain architecture of all Cache domain-containing protein sequences revealed known output domains of signal transduction systems, except for the SMP_2 family members (Table 1.S5). The SMP_2 domain is the closest relative of the DUF2222 domain (mutual best hits in HHpred searches) and both are found exclusively in proteobacteria. While DUF2222 is the sensory module of the BarA/GacS/VarA-type histidine kinases that are global regulators of pathogenicity in gamma-proteobacteria [29], SMP_2 appears to be a sensory module that was cut off from the rest of the protein. The likelihood of this scenario is further supported by the nearly identical phyletic distribution of both domains and the

fact that SMP_2 proteins are also implicated in virulence in gamma-proteobacteria [30]. Apart from this neofunctionalization, all other Cache domains appear to serve as extracellular sensory modules for all major modes and brands of signal transduction proteins in prokaryotes, including sensor histidine kinases, cyclic di-GMP cyclases and diesterases, chemotaxis transducers, adenylate and guanylate cyclases, etc. Furthermore, Cache domains are dominant extracellular sensory domains in prokaryotes (Fig. 1.3, Table 1.S10), significantly outnumbering the best studied such domain, a four-helix bundle [31, 32].

Newly Identified Cache Domains. Among tens of thousands of newly identified Cache domains, many are present in signal transduction proteins from important human pathogens and model systems (Fig. 1.4). For example, we have confidently detected the Cache domain in the extracellular region of the WalK sensor histidine kinase from low G+C Gram positive bacteria, which plays a critical role in regulating cell division and wall stress responses [33]. WalK is a novel target for antibacterial agents against multidrug-resistant bacteria, including methicillin-resistant *Staphylococcus aureus* [26, 34]. We newly identified the double Cache domain in the YedQ diguanylate cyclase, which regulates cellulose biosynthesis and biofilm formation in *Escherichia coli* and *Salmonella enterica* [35, 36]. This domain was also identified in the Rv2435c adenylate cyclase in *Mycobacterium tuberculosis*, which is a part of the cAMP network involved in virulence [37]. Our new dCache_1 model has identified the double Cache domain in the extracellular region of the osmosensing histidine kinase Sln1 from *Saccharomyces*

cerevisiae, which controls activity of the HOG1 pathway [38]. The region, which is now designated as the Cache domain, was shown to be essential for its sensory function [39].

Evolutionary Scenario for Cache Origins. Several lines of evidence suggest that Cache domain(s) evolved from simpler intracellular PAS/GAF-like ancestor(s). We have shown that Cache is homologous to PAS and GAF (Fig. 1.2A). PAS and GAF (that are homologous to each other) or their common ancestor originated in the last universal common ancestor [5, 8, 40]. Our results show that Cache likely originated in the bacterial lineage after its separation from the archaeal/eukaryotic lineage. Every incidence of Cache in archaea and eukaryotes can be attributed to horizontal gene transfer. For example, Cache domains in Metazoa are limited to a single type of protein – a voltage-dependent calcium channel alpha-2-delta subunit [21] (Appendix 1.3, spreadsheet 5), whereas vertically inherited PAS and GAF domains are found in diverse signal transduction proteins [3, 41]. In plants and fungi, Cache is limited to histidine kinases that are known to be horizontally transferred from bacteria [42, 43] (Appendix 1.3, spreadsheets 3 and 4). In *Naegleria*, a representative of Excavates, the Cache domain is found in a bacterial-type adenylate cyclase (Fig. 1.4). Finally, the Cache-to-PAS ratio in archaea and eukaryotes is nearly five times smaller than that in bacteria (Appendix 1.3, spreadsheet 2). Taken together, these observations suggest that PAS and GAF predate Cache, which is consistent with the previous suggestion that intracellular sensing predates extracellular sensing [44].

Discussion

Our findings show that experimentally solved three-dimensional structures of so-called "extracellular PAS domains" belong not to PAS, but to Cache superfamily. Our new sequence profile models for the Cache superfamily dramatically improve computational coverage and enable identification of Cache domains in tens of thousands of signal transduction proteins including those from human pathogens and model systems. Consequently, we demonstrated that Cache is the most abundant extracellular sensory domain in prokaryotes, which originated from a simpler intracellular PAS/GAF ancestor as a benefit of extracellular sensing. The key structural innovation in Cache domains, when compared to PAS and GAF, is the long N-terminal alpha helix (Fig. 1.1), which is a direct extension of the transmembrane helix. It appears that this simple innovation was sufficient to convert an intracellular sensor to an extracellular sensor. However, this also placed significant physical constraints on the ability of the sensor to transmit information. Intracellular PAS and GAF domains have multiple options for interacting with downstream signaling domains, including direct domain-to-domain binding. In a striking contrast, the only option for an extracellular Cache to transmit signals is via its C-terminal transmembrane helix, similarly to the sensory four-helix bundle exemplified by the *E. coli* aspartate chemoreceptor [45]. It is highly likely that these physical constraints dictated some re-wiring of the PAS/GAF-like core in Cache domains resulting in evolutionary conservation of amino acid positions that are not under such constraints in cytoplasmic PAS and GAF domains. Finally, our results demonstrate that solving ambiguous sequence- and structure-based domain definitions can dramatically improve

computational models and significantly accelerate computational coverage of the protein sequence space [46].

Materials and Methods

Data sources and Bioinformatics software. The central data source for all analyses was the local MySQL Pfam 27 [47] database based on Uniprot 2012_06 release. The database files for PfamScan were downloaded in December 2014. The Non-redundant database fasta file was retrieved from NCBI on April 2015. Uniref90 (April 2015) was used for running Psipred [48, 49]. The following software packages were used in this study: BLAST 2.2.28+[50, 51], HHsuite-2.0.16 [28, 52, 53], CD-HIT 4.5.7[54], Cytoscape 2.8.3 [55], BLAST2SimilarityGraph plugin for Cytoscape [56], Graph-0.96_01 (UnionFind) Perl library, MAFFT v7.154b [57], Jalview v2.7 [58], TMHMM 2.0c [59], Phobius v1.01[60], DAS-TMfilter (December 2012) [61], HMMER 3.0 (March 2010) [27] , PfamScan (October 2013) [47], MEGA 5.05 [62], Circos v0.64 [63] and Psipred v3.5. The multiple sequence alignments were built with MAFFT-LINSI using legacygappenalty option. Maximum likelihood trees were constructed to aid in the model building using MEGA with pairwise deletion and the JTT substitution. Domain predictions with PfamScan were carried out at sequence evalue and domain evalue thresholds of 1E-3.

Hidden Markov model construction. A flow chart showing the model building approach is shown in Fig. 1.S5. More than 1 million sequences containing at least one signal transduction output domain as defined in MiST2 database [64] were retrieved from a local copy of the Pfam database (Fig. 1.S5). Eukaryotic sequences were discarded, because domain boundaries for Cache domains in eukaryotes are unclear. Predicted

extracytoplasmic regions that were longer than 50 amino acids were scanned for Pfam domains and redundancy (at 90% identity) was removed resulting in 36,320 sequences. In the next step, a similarity network was built using the BLAST2similarityGraph Cytoscape plugin. Clusters of similar sequences were obtained using the E-value threshold of 1e-10 and query coverage of 95% using Cytoscape and the Perl Graph library. 38 clusters comprising of at least ten members and containing at least one Cache domain (7577 sequences in total) were further chosen for building models. Representative sequences were obtained using a custom script (Fig. 1.S6) for each cluster and the sequences in each cluster were aligned using MAFFT-LINSi with the legacygappenalty option [65]. In case of the largest cluster, which was primarily comprised of sequences with the Cache_1 domain, the alignment was improved by dividing the cluster into smaller groups based on a maximum-likelihood tree generated using MEGA [66]. Individual groups were realigned using MAFFT-LINSi.

HMM models for each cluster were built using hhmake and all-against-all HMM-HMM comparison was carried out using HHsearch [53]. Based on the probability scores and coverage, the clusters were then merged using mafft-profile. Representatives of each cluster were chosen to construct HMMs using the hmmbuild utility in the HMMER3 package [27]. The sensitivity of the models was improved by incorporating remote homologs that were identified by a more sensitive HMM-HMM comparison using HHblits [52].

New Members of the Cache Superfamily and Its Relationship to PAS and GAF. The sequences of extracellular PAS-like domains with available PDB structures were used as

queries for HHpred search using default parameters against Pfam 27 database. Only those hits were considered which had a probability score greater than 95 for at least one of the PDB queries.

The alignments used for creating the new Cache models were also used as queries for HHPred search against with Pfam 27 database with 0 iterations of hhblits. All hits with a probability score greater than 70 were considered to be potentially homologous. To further explore the relationship between the families, we retrieved models for these hits along with new Cache models and the PAS and GAF clan. All-against-all HMM-HMM comparison was carried out using hhsearch. A similarity network was created with the domain families as nodes and hits representing reciprocal hhsearch hits with (i) evalue less than 1E-3 (ii) evalue less than 1E-1 and (iii) probability score >= 90. Families were assigned to the Cache clan when the evalue from HHPred was less than 1E-3 (LuxQ-periplasm, CHASE4, Diacid_rec and DUF2222) or when Cache was the closest superfamily (CHASE, Stimulus_sens_1 and 2CSK_N). SMP_2 and PhoQ_Sensor were included in Cache clan as they are mutual best hits with DUF2222 and 2CSK_N respectively.

We also performed sequence-sequence comparisons using all-against-all BLAST. The sequences for PAS clan, GAF clan and Cache clan comprising of new families were retrieved. For Cache clan, sequences that have overlapping domain prediction with other sensory Pfam domains were disregarded. 100% redundant sequences were removed using CD-HIT. The similarities between different domains were demonstrated using Circos tool.

Phyletic Distribution of Cache, PAS and GAF families

In order to show the phyletic distribution, only those organisms having more than 1000 proteins in Pfam 27.0 database were selected to exclude organisms with relatively incomplete genomes. The Sunburst was created by clustering the main level taxonomic ranks retrieved from NCBI Taxonomy database with the lowest rank used that of species. The domains were considered to be present if any strain of a given organism was found to contain a given domain. The Sunburst was generated using a custom script. PAS and GAF clans includes all the families as defined in Pfam 27.0. However, the Cache domains indicated comprise of those identified by the new models, YkuI_C as well as the other families (2CSK_N, CHASE, CHASE4, Diacid_rec, DUF2222, LuxQ-periplasm, PhoQ_Sensor, SMP_2 and Stimulus_sens_1) identified to be a part of the Cache clan in this study.

Cache Dendrogram

The secondary structure prediction by Psipred was mapped on to the alignment for each model. Only the PAS-like regions comprising of five beta strands were extracted. HMM profiles were built for each alignment using hhmake tool in the HHsuite. All-against-all HMM-HMM comparison was performed using hhsearch. A distance matrix was generated using probability scores from hhsearch. The dendrogram showing similarity between single Cache domains and the membrane-distal and membrane proximal domains of double Cache was generated using the DendroUPGMA web server.

References

1. Stock, A.M., V.L. Robinson, and P.N. Goudreau, *Two-component signal transduction.* Annu Rev Biochem, 2000. **69**: p. 183-215.
2. Chantranupong, L., R.L. Wolfson, and D.M. Sabatini, *Nutrient-Sensing Mechanisms across Evolution.* Cell, 2015. **161**(1): p. 67-83.
3. Taylor, B.L. and I.B. Zhulin, *PAS domains: internal sensors of oxygen, redox potential, and light.* Microbiol Mol Biol Rev, 1999. **63**(2): p. 479-506.
4. Henry, J.T. and S. Crosson, *Ligand-binding PAS domains in a genomic, cellular, and structural context.* Annu Rev Microbiol, 2011. **65**: p. 261-86.
5. Aravind, L. and C.P. Ponting, *The GAF domain: an evolutionary link between diverse phototransducing proteins.* Trends Biochem Sci, 1997. **22**(12): p. 458-9.
6. Ho, Y.S., L.M. Burden, and J.H. Hurley, *Structure of the GAF domain, a ubiquitous signaling motif and a new class of cyclic GMP receptor.* EMBO J, 2000. **19**(20): p. 5288-99.
7. Pellequer, J.L., et al., *Photoactive yellow protein: a structural prototype for the three-dimensional fold of the PAS domain superfamily.* Proc Natl Acad Sci U S A, 1998. **95**(11): p. 5884-90.
8. Anantharaman, V., E.V. Koonin, and L. Aravind, *Regulatory potential, phyletic distribution and evolution of ancient, intracellular small-molecule-binding domains.* J Mol Biol, 2001. **307**(5): p. 1271-92.
9. Ponting, C.P. and L. Aravind, *PAS: a multifunctional domain family comes to light.* Curr Biol, 1997. **7**(11): p. R674-7.
10. Zhulin, I.B., B.L. Taylor, and R. Dixon, *PAS domain S-boxes in Archaea, Bacteria and sensors for oxygen and redox.* Trends Biochem Sci, 1997. **22**(9): p. 331-3.
11. Neiditch, M.B., et al., *Ligand-induced asymmetry in histidine sensor kinase complex regulates quorum sensing.* Cell, 2006. **126**(6): p. 1095-108.
12. Reinelt, S., et al., *The structure of the periplasmic ligand-binding domain of the sensor kinase CitA reveals the first extracellular PAS domain.* J Biol Chem, 2003. **278**(40): p. 39189-96.
13. Cheung, J. and W.A. Hendrickson, *Crystal structures of C4-dicarboxylate ligand complexes with sensor domains of histidine kinases DcuS and DctB.* J Biol Chem, 2008. **283**(44): p. 30256-65.
14. Wu, R., et al., *Insight into the sporulation phosphorelay: crystal structure of the sensor domain of Bacillus subtilis histidine kinase, KinD.* Protein Sci, 2013. **22**(5): p. 564-76.
15. Pokkuluri, P.R., et al., *Structures and solution properties of two novel periplasmic sensor domains with c-type heme from chemotaxis proteins of Geobacter sulfurreducens: implications for signal transduction.* J Mol Biol, 2008. **377**(5): p. 1498-517.
16. Goers Sweeney, E., et al., *Structure and proposed mechanism for the pH-sensing Helicobacter pylori chemoreceptor TlpB.* Structure, 2012. **20**(7): p. 1177-88.
17. Hothorn, M., T. Dabi, and J. Chory, *Structural basis for cytokinin recognition by Arabidopsis thaliana histidine kinase 4.* Nat Chem Biol, 2011. **7**(11): p. 766-8.

18. Cheung, J., et al., *Crystal structure of a functional dimer of the PhoQ sensor domain.* J Biol Chem, 2008. **283**(20): p. 13762-70.
19. Anantharaman, V. and L. Aravind, *Cache - a signaling domain common to animal Ca(2+)-channel subunits and a class of prokaryotic chemotaxis receptors.* Trends Biochem Sci, 2000. **25**(11): p. 535-7.
20. Hazelbauer, G.L., J.J. Falke, and J.S. Parkinson, *Bacterial chemoreceptors: high-performance signaling in networked arrays.* Trends Biochem Sci, 2008. **33**(1): p. 9-19.
21. Dolphin, A.C., *Calcium channel auxiliary alpha2delta and beta subunits: trafficking and one step beyond.* Nat Rev Neurosci, 2012. **13**(8): p. 542-55.
22. Stahl, S.M., et al., *The diverse therapeutic actions of pregabalin: is a single mechanism responsible for several pharmacological activities?* Trends Pharmacol Sci, 2013. **34**(6): p. 332-9.
23. Zhang, Z. and W.A. Hendrickson, *Structural characterization of the predominant family of histidine kinase sensor domains.* J Mol Biol, 2010. **400**(3): p. 335-53.
24. Pineda-Molina, E., et al., *Evidence for chemoreceptors with bimodular ligand-binding regions harboring two signal-binding sites.* Proc Natl Acad Sci U S A, 2012. **109**(46): p. 18926-31.
25. Shah, N., et al., *Reductive evolution and the loss of PDC/PAS domains from the genus Staphylococcus.* BMC Genomics, 2013. **14**: p. 524.
26. Gotoh, Y., et al., *Two-component signal transduction as potential drug targets in pathogenic bacteria.* Curr Opin Microbiol, 2010. **13**(2): p. 232-9.
27. Eddy, S.R., *Accelerated Profile HMM Searches.* PLoS Comput Biol, 2011. **7**(10): p. e1002195.
28. Soding, J., A. Biegert, and A.N. Lupas, *The HHpred interactive server for protein homology detection and structure prediction.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W244-8.
29. Lapouge, K., et al., *Gac/Rsm signal transduction pathway of gamma-proteobacteria: from RNA recognition to regulation of social behaviour.* Mol Microbiol, 2008. **67**(2): p. 241-53.
30. Cox, A.J., et al., *Cloning and characterisation of the Pasteurella multocida ahpA gene responsible for a haemolytic phenotype in Escherichia coli.* Vet Microbiol, 2000. **72**(1-2): p. 135-52.
31. Milburn, M.V., et al., *Three-dimensional structures of the ligand-binding domain of the bacterial aspartate receptor with and without a ligand.* Science, 1991. **254**(5036): p. 1342-7.
32. Ulrich, L.E. and I.B. Zhulin, *Four-helix bundle: a ubiquitous sensory module in prokaryotic signal transduction.* Bioinformatics, 2005. **21 Suppl 3**: p. iii45-8.
33. Dubrac, S., et al., *A matter of life and death: cell wall homeostasis and the WalKR (YycGF) essential signal transduction pathway.* Mol Microbiol, 2008. **70**(6): p. 1307-22.
34. Watanabe, T., et al., *Isolation and characterization of signermycin B, an antibiotic that targets the dimerization domain of histidine kinase WalK.* Antimicrob Agents Chemother, 2012. **56**(7): p. 3657-63.

35. Da Re, S. and J.M. Ghigo, *A CsgD-independent pathway for cellulose production and biofilm formation in Escherichia coli.* J Bacteriol, 2006. **188**(8): p. 3073-87.
36. Garcia, B., et al., *Role of the GGDEF protein family in Salmonella cellulose biosynthesis and biofilm formation.* Mol Microbiol, 2004. **54**(1): p. 264-77.
37. Bai, G., G.S. Knapp, and K.A. McDonough, *Cyclic AMP signalling in mycobacteria: redirecting the conversation with a common currency.* Cell Microbiol, 2011. **13**(3): p. 349-58.
38. Posas, F., et al., *Yeast HOG1 MAP kinase cascade is regulated by a multistep phosphorelay mechanism in the SLN1-YPD1-SSK1 "two-component" osmosensor.* Cell, 1996. **86**(6): p. 865-75.
39. Reiser, V., D.C. Raitt, and H. Saito, *Yeast osmosensor Sln1 and plant cytokinin receptor Cre1 respond to changes in turgor pressure.* J Cell Biol, 2003. **161**(6): p. 1035-40.
40. Montgomery, B.L. and J.C. Lagarias, *Phytochrome ancestry: sensors of bilins and light.* Trends Plant Sci, 2002. **7**(8): p. 357-66.
41. Martinez, S.E., J.A. Beavo, and W.G. Hol, *GAF domains: two-billion-year-old molecular switches that bind cyclic nucleotides.* Mol Interv, 2002. **2**(5): p. 317-23.
42. Koretke, K.K., et al., *Evolution of two-component signal transduction.* Mol Biol Evol, 2000. **17**(12): p. 1956-70.
43. Wuichet, K., B.J. Cantwell, and I.B. Zhulin, *Evolution and phyletic distribution of two-component signal transduction systems.* Curr Opin Microbiol, 2010. **13**(2): p. 219-25.
44. Ulrich, L.E., E.V. Koonin, and I.B. Zhulin, *One-component systems dominate signal transduction in prokaryotes.* Trends Microbiol, 2005. **13**(2): p. 52-6.
45. Chervitz, S.A. and J.J. Falke, *Molecular mechanism of transmembrane signaling by the aspartate receptor: a model.* Proc Natl Acad Sci U S A, 1996. **93**(6): p. 2545-50.
46. Rekapalli, B., et al., *Dynamics of domain coverage of the protein sequence universe.* BMC Genomics, 2012. **13**: p. 634.
47. Finn, R.D., et al., *Pfam: the protein families database.* Nucleic acids research, 2013: p. gkt1223.
48. Buchan, D.W., et al., *Scalable web services for the PSIPRED Protein Analysis Workbench.* Nucleic acids research, 2013. **41**(W1): p. W349-W357.
49. Jones, D.T., *Protein secondary structure prediction based on position-specific scoring matrices.* Journal of molecular biology, 1999. **292**(2): p. 195-202.
50. Camacho, C., et al., *BLAST+: architecture and applications.* BMC bioinformatics, 2009. **10**(1): p. 421.
51. Altschul, S.F., et al., *Basic local alignment search tool.* Journal of molecular biology, 1990. **215**(3): p. 403-410.
52. Remmert, M., et al., *HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment.* Nat Methods, 2012. **9**(2): p. 173-5.
53. Soding, J., *Protein homology detection by HMM-HMM comparison.* Bioinformatics, 2005. **21**(7): p. 951-60.

54. Li, W. and A. Godzik, *Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.* Bioinformatics, 2006. **22**(13): p. 1658-1659.
55. Smoot, M.E., et al., *Cytoscape 2.8: new features for data integration and network visualization.* Bioinformatics, 2011. **27**(3): p. 431-432.
56. Wittkop, T., et al., *Comprehensive cluster analysis with Transitivity Clustering.* Nature protocols, 2011. **6**(3): p. 285-295.
57. Katoh, K. and D.M. Standley, *MAFFT multiple sequence alignment software version 7: improvements in performance and usability.* Molecular biology and evolution, 2013. **30**(4): p. 772-780.
58. Waterhouse, A.M., et al., *Jalview Version 2—a multiple sequence alignment editor and analysis workbench.* Bioinformatics, 2009. **25**(9): p. 1189-1191.
59. Krogh, A., et al., *Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes.* Journal of molecular biology, 2001. **305**(3): p. 567-580.
60. Käll, L., A. Krogh, and E.L. Sonnhammer, *A combined transmembrane topology and signal peptide prediction method.* Journal of molecular biology, 2004. **338**(5): p. 1027-1036.
61. Cserzo, M., et al., *TM or not TM: transmembrane protein prediction with low false positive rate using DAS-TMfilter.* Bioinformatics, 2004. **20**(1): p. 136-137.
62. Tamura, K., et al., *MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods.* Molecular biology and evolution, 2011. **28**(10): p. 2731-2739.
63. Krzywinski, M., et al., *Circos: an information aesthetic for comparative genomics.* Genome research, 2009. **19**(9): p. 1639-1645.
64. Ulrich, L.E. and I.B. Zhulin, *The MiST2 database: a comprehensive genomics resource on microbial signal transduction.* Nucleic Acids Res, 2010. **38**(Database issue): p. D401-7.
65. Katoh, K., et al., *MAFFT version 5: improvement in accuracy of multiple sequence alignment.* Nucleic Acids Res, 2005. **33**(2): p. 511-8.
66. Tamura, K., et al., *MEGA6: Molecular Evolutionary Genetics Analysis version 6.0.* Mol Biol Evol, 2013. **30**(12): p. 2725-9.
67. Zhulin, I.B., B.L. Taylor, and R. Dixon, *PAS domain S-boxes in Archaea, Bacteria and sensors for oxygen and redox.* Trends in Biochemical Sciences, 1997. **22**(9): p. 331.
68. Anantharaman, V. and L. Aravind, *Cache-a signaling domain common to animal Ca (2+)-channel subunits and a class of prokaryotic chemotaxis receptors.* Trends in Biochemical Sciences, 2000. **25**(11): p. 535.
69. Pappalardo, L., *The NMR Structure of the Sensory Domain of the Membranous Two-component Fumarate Sensor (Histidine Protein Kinase) DcuS of Escherichia coli.* Journal of Biological Chemistry, 2003. **278**(40): p. 39185-39188.
70. Zhou, Y.F., et al., *C4-dicarboxylates sensing mechanism revealed by the crystal structures of DctB sensor domain.* J Mol Biol, 2008. **383**(1): p. 49-61.

71. Finn, R.D., et al., *Pfam: the protein families database.* Nucleic Acids Research, 2014. **42**(D1): p. D222-D230.
72. Pappalardo, L., et al., *The NMR structure of the sensory domain of the membranous two-component fumarate sensor (histidine protein kinase) DcuS of Escherichia coli.* J Biol Chem, 2003. **278**(40): p. 39185-8.
73. Cho, U.S., et al., *Metal bridges between the PhoQ sensor domain and the membrane regulate transmembrane signaling.* J Mol Biol, 2006. **356**(5): p. 1193-206.
74. Minasov, G., et al., *Crystal structures of YkuI and its complex with second messenger cyclic Di-GMP suggest catalytic mechanism of phosphodiester bond cleavage by EAL domains.* J Biol Chem, 2009. **284**(19): p. 13174-84.
75. Emami, K., et al., *Regulation of the xylan-degrading apparatus of Cellvibrio japonicus by a novel two-component system.* J Biol Chem, 2009. **284**(2): p. 1086-96.
76. Chang, C., et al., *Extracytoplasmic PAS-like domains are common in signal transduction proteins.* J Bacteriol, 2010. **192**(4): p. 1156-9.
77. Mougel, C. and I.B. Zhulin, *CHASE: an extracellular sensing domain common to transmembrane receptors from prokaryotes, lower eukaryotes and plants.* Trends Biochem Sci, 2001. **26**(10): p. 582-4.
78. Getzoff, E.D., K.N. Gutwin, and U.K. Genick, *Anticipatory active-site motions and chromophore distortion prime photoreceptor PYP for light activation.* Nat Struct Biol, 2003. **10**(8): p. 663-8.
79. Scheuermann, T.H., et al., *Artificial ligand binding within the HIF2alpha PAS-B domain of the HIF2 transcription factor.* Proc Natl Acad Sci U S A, 2009. **106**(2): p. 450-5.
80. Park, H., et al., *Insights into signal transduction involving PAS domain oxygen-sensing heme proteins from the X-ray crystal structure of Escherichia coli Dos heme domain (Ec DosH).* Biochemistry, 2004. **43**(10): p. 2738-46.
81. Miyatake, H., et al., *Sensory mechanism of oxygen sensor FixL from Rhizobium meliloti: crystallographic, mutagenesis and resonance Raman spectroscopic studies.* J Mol Biol, 2000. **301**(2): p. 415-31.
82. Rajagopal, S. and K. Moffat, *Crystal structure of a photoactive yellow protein from a sensor histidine kinase: conformational variability and signal transduction.* Proc Natl Acad Sci U S A, 2003. **100**(4): p. 1649-54.
83. Fedorov, R., et al., *Crystal structures and molecular mechanism of a light-induced signaling switch: The Phot-LOV1 domain from Chlamydomonas reinhardtii.* Biophys J, 2003. **84**(4): p. 2474-82.
84. Razeto, A., et al., *Structure of the NCoA-1/SRC-1 PAS-B domain bound to the LXXLL motif of the STAT6 transactivation domain.* J Mol Biol, 2004. **336**(2): p. 319-29.
85. Jaiswal, R.K., G. Manjeera, and B. Gopal, *Role of a PAS sensor domain in the Mycobacterium tuberculosis transcription regulator Rv1364c.* Biochem Biophys Res Commun, 2010. **398**(3): p. 342-9.

86. Amezcua, C.A., et al., *Structure and interactions of PAS kinase N-terminal PAS domain: model for intramolecular kinase regulation.* Structure, 2002. **10**(10): p. 1349-61.
87. Hennig, S., et al., *Structural and functional analyses of PAS domain interactions of the clock proteins Drosophila PERIOD and mouse PERIOD2.* PLoS Biol, 2009. **7**(4): p. e94.
88. Morais Cabral, J.H., et al., *Crystal structure and functional analysis of the HERG potassium channel N terminus: a eukaryotic PAS domain.* Cell, 1998. **95**(5): p. 649-55.
89. Essen, L.O., J. Mailliet, and J. Hughes, *The structure of a complete phytochrome sensory module in the Pr ground state.* Proc Natl Acad Sci U S A, 2008. **105**(38): p. 14709-14.
90. Crosson, S. and K. Moffat, *Photoexcited structure of a plant photoreceptor domain reveals a light-driven molecular switch.* Plant Cell, 2002. **14**(5): p. 1067-75.
91. Yang, X., J. Kuk, and K. Moffat, *Crystal structure of Pseudomonas aeruginosa bacteriophytochrome: photoconversion and signal transduction.* Proc Natl Acad Sci U S A, 2008. **105**(38): p. 14715-20.
92. Vaidya, A.T., et al., *Structure of a light-activated LOV protein dimer that regulates transcription.* Sci Signal, 2011. **4**(184): p. ra50.
93. Burgie, E.S., et al., *Crystal structure of the photosensing module from a red/far-red light-absorbing plant phytochrome.* Proc Natl Acad Sci U S A, 2014. **111**(28): p. 10179-84.
94. Yamada, S., et al., *Structure of PAS-linked histidine kinase and the response regulator complex.* Structure, 2009. **17**(10): p. 1333-44.
95. Card, P.B., P.J. Erbel, and K.H. Gardner, *Structural basis of ARNT PAS-B dimerization: use of a common beta-sheet interface for hetero- and homodimerization.* J Mol Biol, 2005. **353**(3): p. 664-77.
96. Preu, J., et al., *The sensor region of the ubiquitous cytosolic sensor kinase, PdtaS, contains PAS and GAF domain sensing modules.* J Struct Biol, 2012. **177**(2): p. 498-505.
97. Conrad, K.S., A.M. Bilwes, and B.R. Crane, *Light-induced subunit dissociation by a light-oxygen-voltage domain photoreceptor from Rhodobacter sphaeroides.* Biochemistry, 2013. **52**(2): p. 378-91.
98. Ukaegbu, U.E. and A.C. Rosenzweig, *Structure of the redox sensor domain of Methylococcus capsulatus (Bath) MmoS.* Biochemistry, 2009. **48**(10): p. 2207-15.

Appendix

.

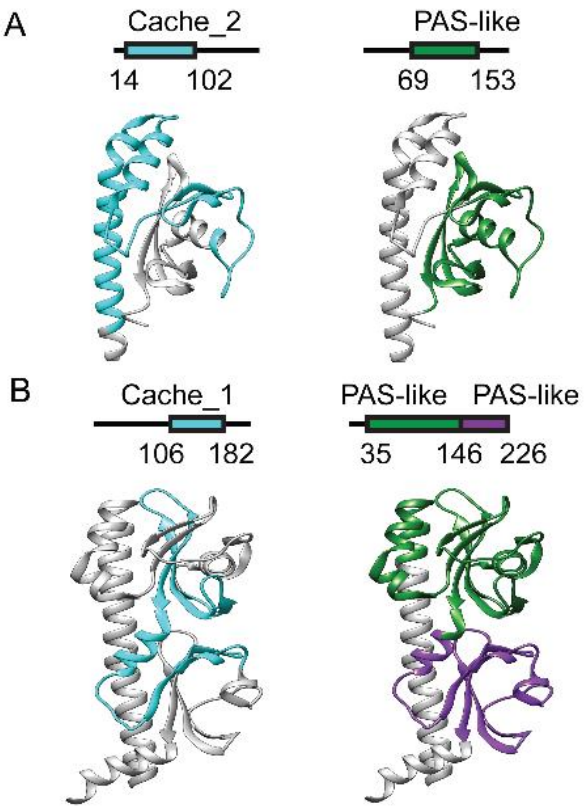

Fig. 1.1. Comparison of sequence- and structure-based definitions for extracellular PAS-like domains.

(A) Vibrio parahaemolyticus chemoreceptor (PDB: 2QHK); (B) Vibrio cholerae chemoreceptor (PDB: 3C8C). Domains are visualized on sequences with corresponding amino acid positions (top) and structures (bottom). Cache (cyan) domains are defined by Pfam; PAS domains (green and magenta) were defined by visual inspection of corresponding structures.

Fig. 1.2. Relationship between Cache (red), PAS (blue) and GAF (green) superfamilies.

(A) HMM-to-HMM comparisons. The nodes represent domain families. Links represent reciprocal hits in hhsearch. Hits with an evalue <1E-3 are shown as thick lines, those with evalue <1E-1 are shown as thin lines and dotted lines are used to represent hits with > 90 probability score. Filled circles represent hits from HHPred search using new models. Empty circles are other members of the clans that were later included (B) Sequence-to-sequence comparisons. The outer circle represents domain families. Links between individual sequences represent reciprocal BLAST hits with an evalue threshold of 1E-8.

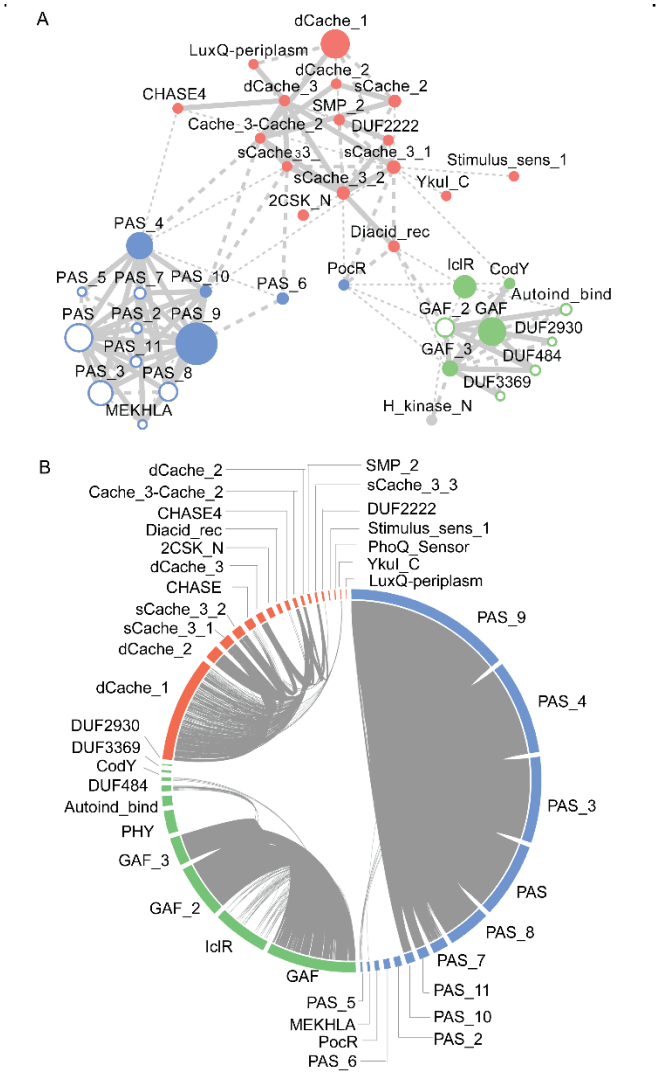

Fig. 1.2 continued

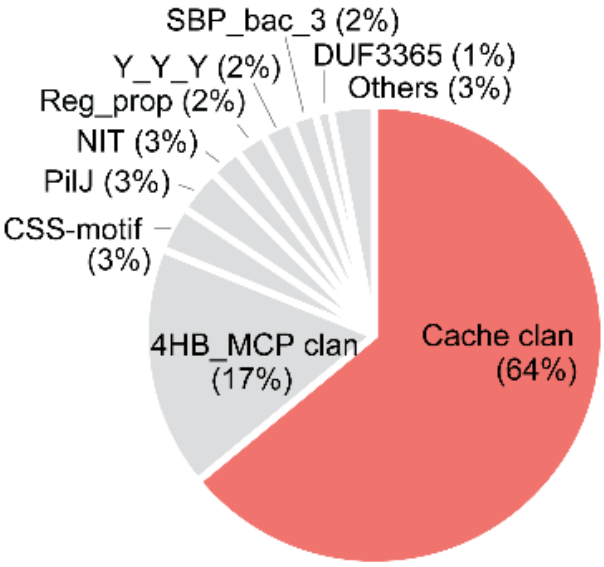

Fig. 1.3. Relative abundance of known extracellular sensory domains in prokaryotes.

Domain counts were obtained by running Pfamscan against a dataset of non-redundant prokaryotic extracellular sequences. Relative abundance is shown considering only known domains.

Fig. 1.4. Examples of newly identified and better defined Cache domains in diverse signal transduction proteins from bacteria, archaea and eukaryotes.

Domain architectures for representative sequences from model organisms are shown along with their UniProt accession numbers. Newly defined Cache domains are shown in red. Cache boundaries defined by the previous Pfam models are shown in pink (Cache) and green (MCP_N). HAMP domains are shown as grey circles, PAS domains as cyan circles, and HisKA domains as white circles. Other Pfam domains are abbreviated as follows: MCP, MCPsignal; GGDEF, GGDEF; GC, guanylate cyclase; HK, the histidine kinase HATPase_c domian; RR, response regulator; VWA, a combination of VWA_N and VWA domains; VGCC, VGCC_alpha2.

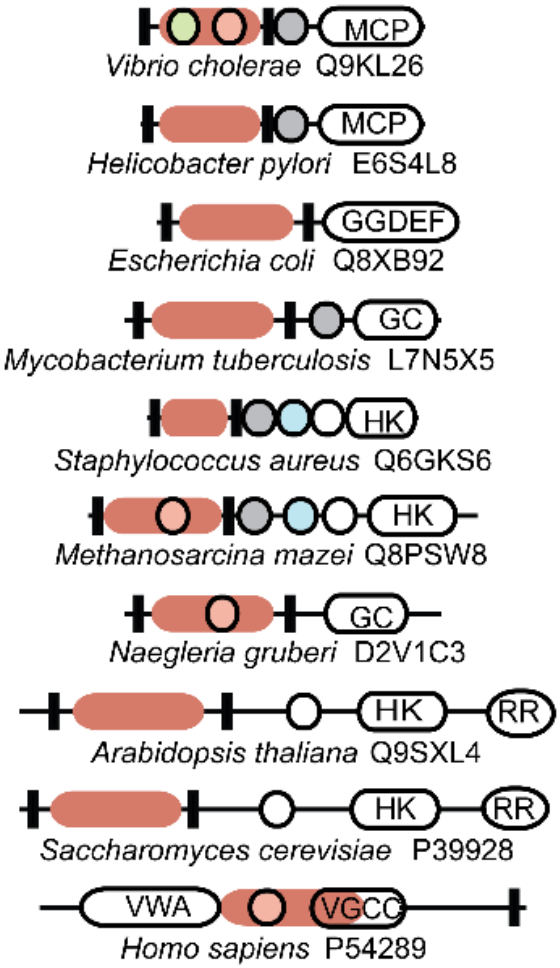

Fig. 1.4 continued

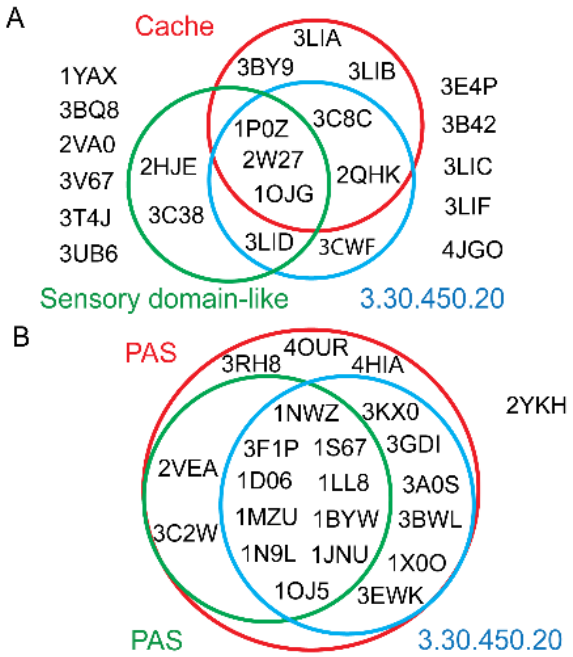

Fig. 1.S1. Superfamily assignment of PAS domains in sequence and structure classification databases.

(A) Extracellular PAS-like domains; (B) intracellular PAS domains. Assignments of PDB structures by Pfam (red), SCOP (green) and CATH (blue) are shown as Venn diagrams to scale.

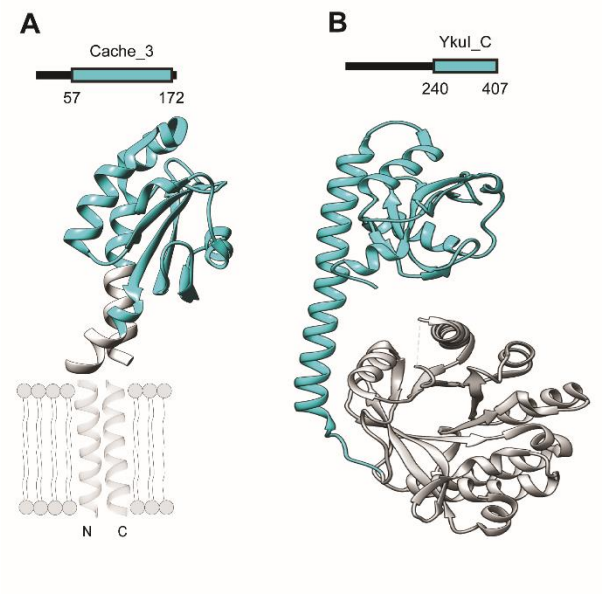

Fig. 1.S2. Comparison of sequence- and structure-based definitions for extracellular and intracellular single PAS-like domains.

(A). Periplasmic domain of CitA from *Klebsiella pneumoniae* (PDB-1P0Z). Cache_3 domain predicted by Pfam is shown in cyan, (B) YkuI comprising of EAL and YkuI_C domains from *Bacillus subtilis* (PDB-2W27). The EAL domain is shown in gray and YkuI_C domain predicted by Pfam is shown in cyan.

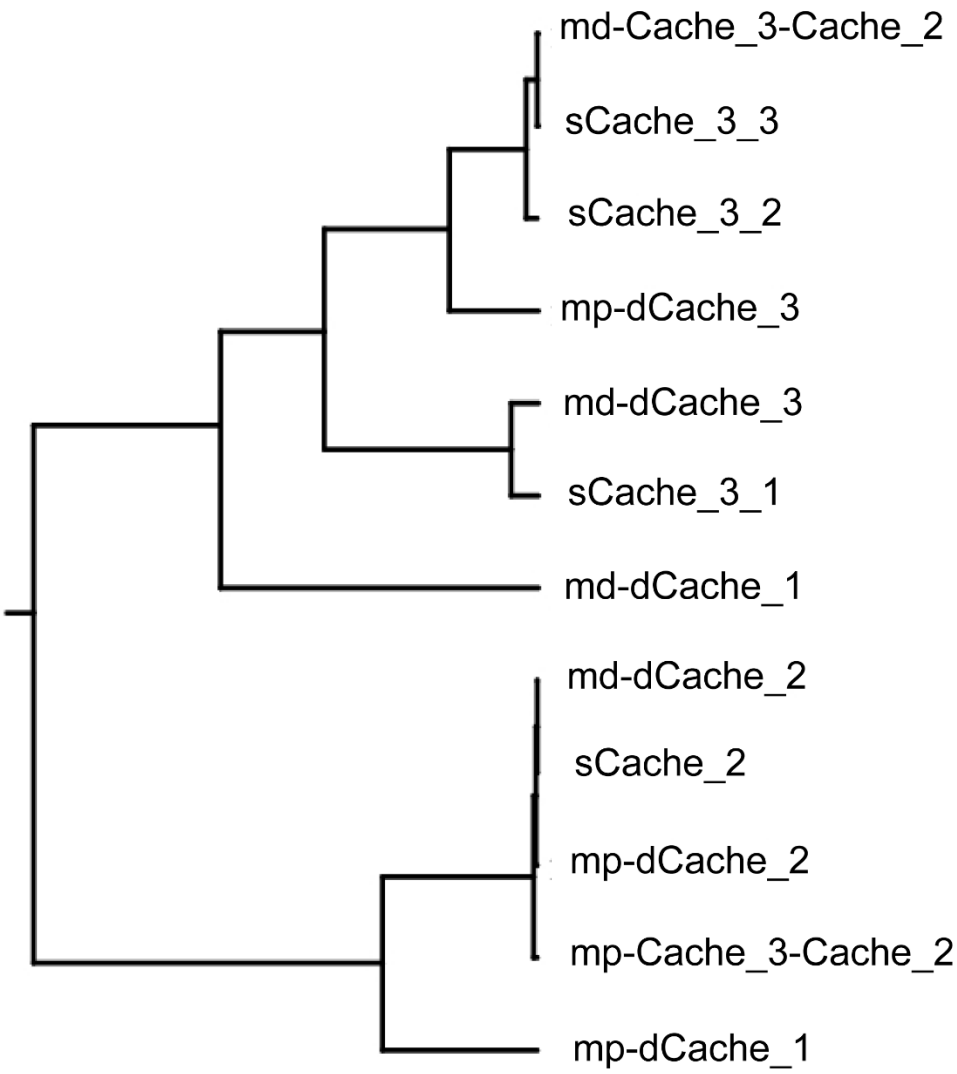

Fig. 1.S3. Relationship between single Cache domains and the membrane distal and membrane proximal domains of double Cache.

The PAS-like regions were extracted for each model based on secondary structure prediction and all-against-all HHsearch comparison was carried out. The dendrogram was generated by using the probability scores as similarity measure.

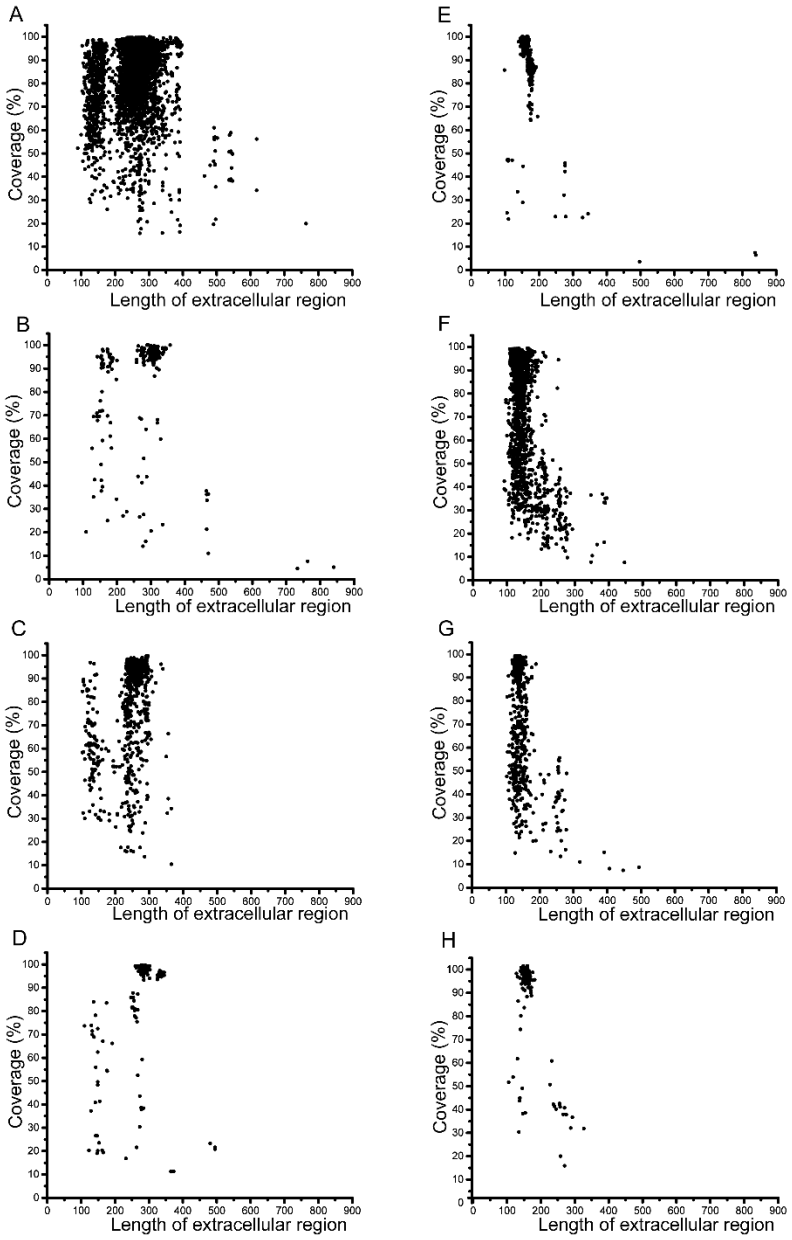

Fig. 1.S4. Coverage of extracellular region by new Cache models.

(A) dCache_1, (B) dCache_2, (C) dCache_3, (D) Cache_3-Cache_2, (E) sCache_2, (F) sCache_3_1, (G) sCache_3_2, (H) sCache_3_3.

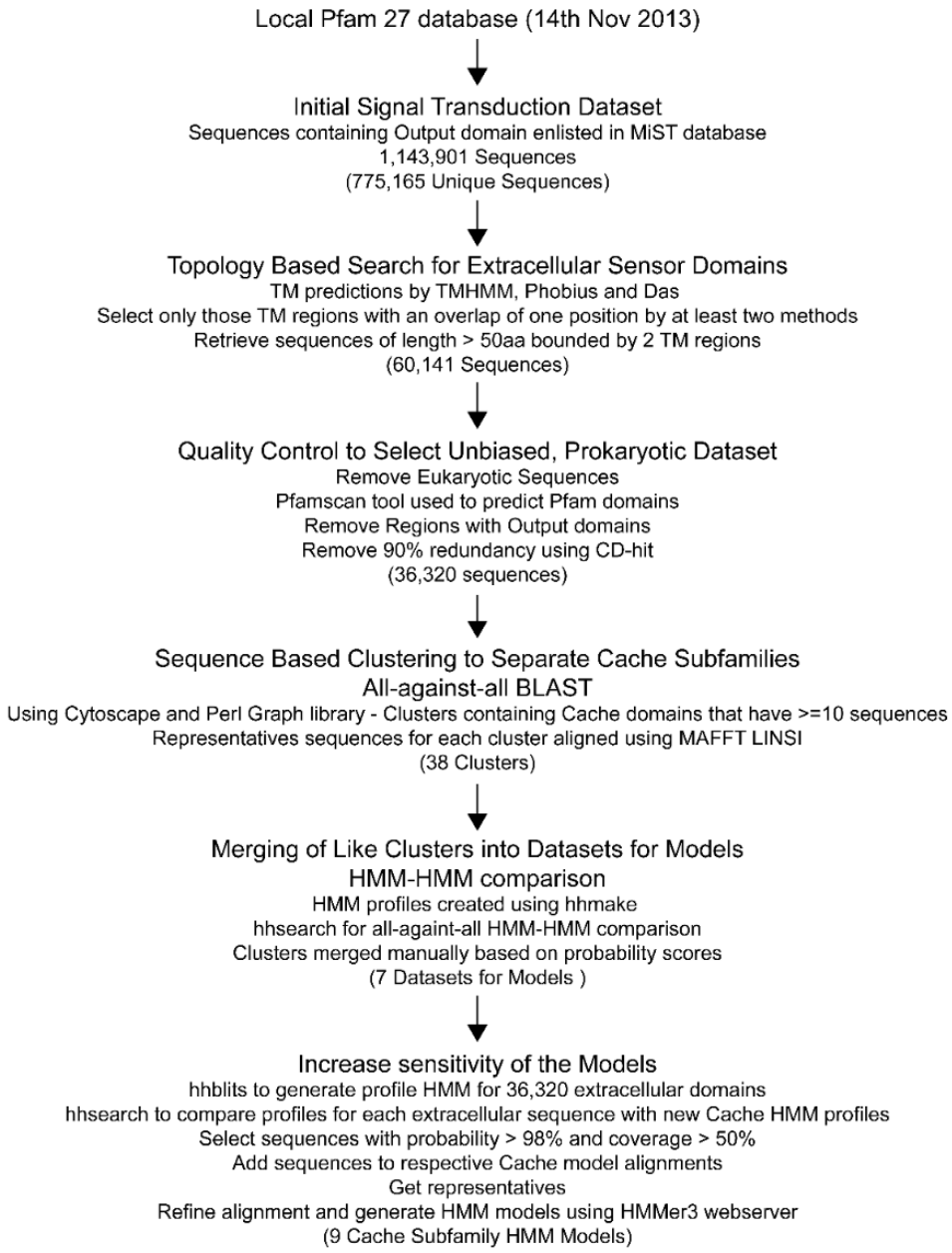

Fig. 1.S5. Flow chart of the HMM construction process.

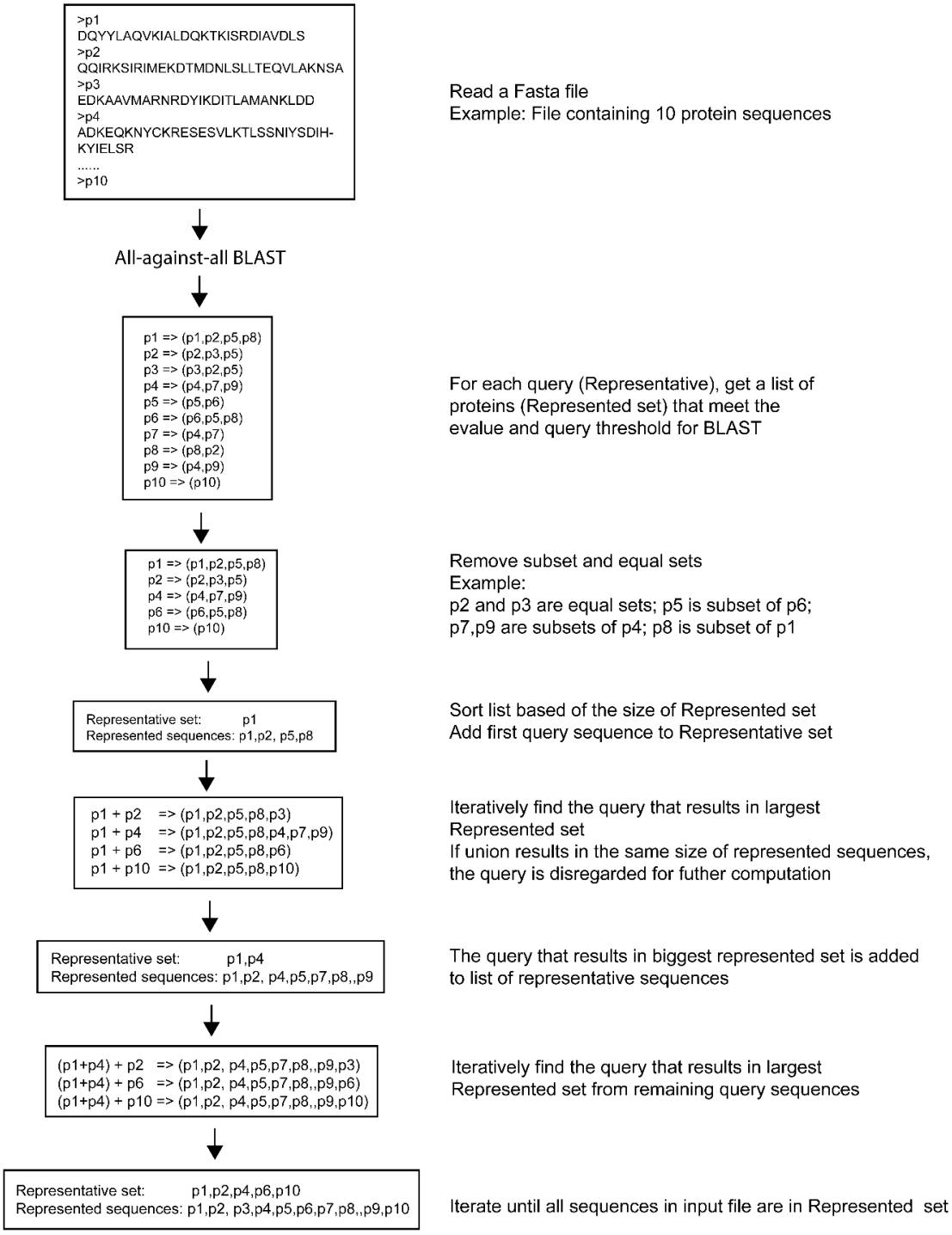

Fig. 1.S6. Algorithm for selecting representatives from a given set of sequences based on all-against-all BLAST results

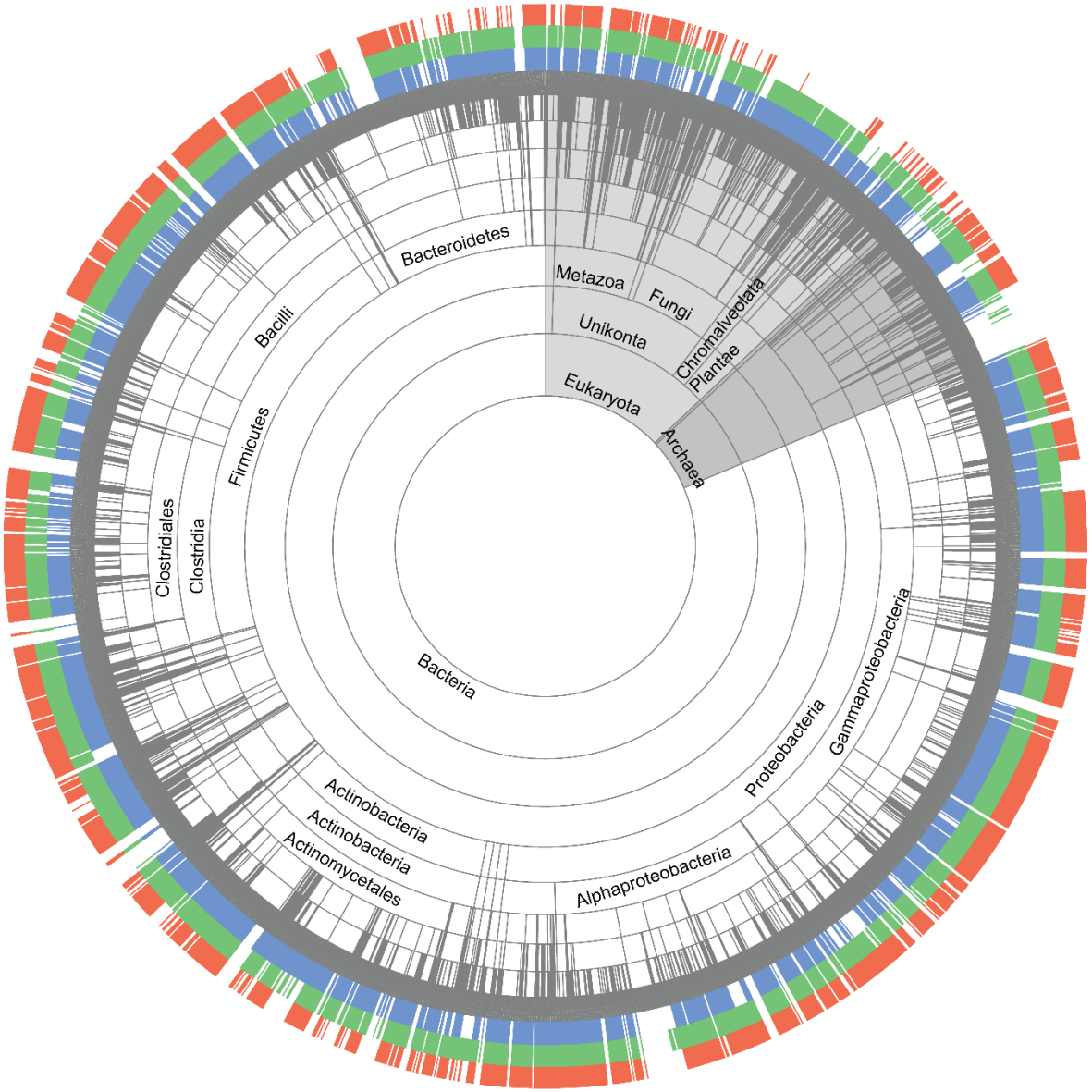

Fig. 1.S7. Phyletic distribution of PAS (blue), GAF (green) and Cache (red) domains.

Flags at the outer three layers represent the domain presence in a corresponding genome. The tree was built using taxonomic ranks retrieved from NCBI.

Table 1.S1. Timeline of PAS, Cache and PDC domains

Year	Highlights	References
1997	PAS domain defined	[9, 67]
2000	Cache domain defined based on sequence similarity CitA circular permutation of Cache Cache N-terminal similar to PAS core	[68]
2003	Structure of CitA solved Suggested to be PAS based on structural similarity to PYP No PAS detected by BLAST, 3D-PSSM, LOOPP, manual searching of S1/S2 boxes First structure for extracytoplasmic PAS	[12]
2003	Structure of DcuS (belongs to CitA family) solved Suggested to possess a novel domain	[69]
2008	Structure of PhoQ solved New family PDC (PhoQ-DcuS-CitA) PDC family subset of PAS – difference in N-terminal helix and the region between 2nd and 3rd strands PDC and PAS belong to separate superfamilies	[18]
2008	PDC family comprised of single PDC (DcuS) and double PDC (*Vibrio cholerae* DctB)	[13]
2008	*Sinorhizobium meliloti* DctB comprised of tandem PAS domains and one N-terminal helical region	[70]
2010	Structures of dPDC proteins solved PhoQ, CitA, DcuS, AbsF, PhoR – sPDC LuxQ, DctB, KinD - dPDC	[23]
2013	CitA and DcuS part of Cache_3, a new family in the Cache clan	[71]

Table 1.S2. Family (domain) and superfamily assignments for extracellular PAS-like domains.

PDB	Organism	Gene	Domain name	Refs	SCOP superfamily	CATH superfamily	Pfam clan
1P0Z	*Klebsiella pneumoniae*	CitA	PAS Cache PDC	[12] [19] [18]	Sensory domain-like	3.30.450.20	Cache
3BY8	*Escherichia coli*	DcuS	PAS-like Cache PDC	[72] [19] [18]	Sensory domain-like	3.30.450.20	Cache
3C8C	*Vibrio cholerae*	VCA0923	PAS PDC	[4] [23]	N/A	3.30.450.20	Cache
1YAX	*Salmonella enterica*	PhoQ	PAS	[73]	N/A	N/A	N/A
3BQ8	*Escherichia coli*	PhoQ	PDC PAS	[18] [4]	N/A	N/A	N/A
2HJE	*Vibrio harveyi*	LuxQ	PAS PDC	[11] [23]	Sensory domain-like	N/A	N/A
3C38	*Vibrio cholerae*	LuxQ	PAS PDC	[4] [23]	Sensory domain-like	N/A	N/A
3E4P	*Sinorhizobium meliloti*	DctB	PAS PDC	[70] [23]	N/A	N/A	N/A
3BY9	*Vibrio cholerae*	DctB	PDC	[13]	N/A	N/A	Cache
3B42	*Geobacter sulfurreducens*	GSU0935	PAS	[15]	N/A	N/A	N/A
2W27	*Bacillus subtilis*	YkuI	PAS-like PAS	[74] [4]	Sensory domain-like	3.30.450.20	Cache
2VA0	*Cellvibrio japonicus*	AbsF	PAS PDC	[75] [23]	N/A	N/A	N/A
3LIA	*Methanosarcina mazei*	MM2955	PDC PAS	[23] [4]	N/A	N/A	Cache
3LIB	*Methanosarcina mazei*	MM2965	PDC PAS	[23] [4]	N/A	N/A	Cache
3LIC	*Shewanella oneidensis*	SO0859	PDC PAS	[23] [4]	N/A	N/A	N/A
3LID	*Vibrio parahaemolyticus*	VP0354	PDC PAS	[23] [4]	Sensory domain-like	3.30.450.20	N/A
3LIF	*Rhodopseudomonas palustris*	RPA3616	PDC PAS	[23] [4]	N/A	N/A	N/A
3CWF	*Bacillus subtilis*	PhoR	PAS-like PDC PAS	[76] [23] [4]	N/A	3.30.450.20	N/A
3T4J	*Arabidopsis thaliana*	AHK4	PAS-like CHASE	[17] [77]	N/A	N/A	N/A
4JGO	*Bacillus subtilis*	KinD	PAS-like PDC PAS	[14] [23] [4]	N/A	N/A	N/A
2QHK	*Vibrio parahaemolyticus*	VP0183	PAS	[4]	N/A	3.30.450.20	Cache

N/A – not assigned; Pfam clan assignments are based on default E-values, Pfam 27.0 release.

Table 1.S3. Family (domain) and superfamily assignments for intracellular PAS domains.

PDB	Organism	Gene	Domain name	Refs	SCOP superfamily	CATH superfamily	Pfam clan
1NWZ	*Ectothiorhodospira halophila*	PYP	PAS	[78]	PYP-like sensor (PAS domain)	3.30.450.20	PAS
3F1P	*Homo sapiens*	Hif2a	PAS	[79]	PYP-like sensor (PAS domain)	3.30.450.20	PAS
1S67	*Escherichia coli*	DosP	PAS	[80]	PYP-like sensor (PAS domain)	3.30.450.20	PAS
1D06	*Rhizobium meliloti*	FixL	PAS	[81]	PYP-like sensor (PAS domain)	3.30.450.20	PAS
1MZU	*Rhodospirillum centenum*	Pph	PAS	[82]	PYP-like sensor (PAS domain)	3.30.450.20	PAS
1N9L	*Chlamydomonas reinhardtii*	Phot	LOV	[83]	PYP-like sensor (PAS domain)	3.30.450.20	PAS
1OJ5	*Mus musculus*	NCOA1	PAS	[84]	PYP-like sensor (PAS domain)	3.30.450.20	PAS
3KX0	*Mycobacterium tuberculosis*	Rv1364c	PAS	[85]	N/A	3.30.450.20	PAS
1LL8	*Homo sapiens*	KIAA0135	PAS	[86]	PYP-like sensor (PAS domain)	3.30.450.20	PAS
3GDI	*Mus musculus*	Per2	PAS	[87]	N/A	3.30.450.20	PAS
1BYW	*Homo sapiens*	HERG	PAS	[88]	PYP-like sensor (PAS domain)	3.30.450.20	PAS
2VEA	*Synechocystis sp. PCC6803*	Cph1	PAS	[89]	PYP-like sensor (PAS domain)	N/A	PAS
1JNU	*Adiantum capilus-veneris*	PHY3	LOV	[90]	PYP-like sensor (PAS domain)	3.30.450.20	PAS
3C2W	*Pseudomonas aeruginosa*	BhpP	PAS	[91]	PYP-like sensor (PAS domain)	N/A	PAS
3RH8	*Neurospora crassa*	vvd	LOV	[92]	N/A	N/A	PAS
4OUR	*Arabidopsis thaliana*	PHYB	PAS	[93]	N/A	N/A	PAS
3A0S	*Thermotoga maritima*	TM_1359	PAS	[94]	N/A	3.30.450.20	PAS
3BWL	*Haloarcula marismortui*	HtlD	N/A		N/A	3.30.450.20	PAS
1X0O	*Homo sapiens*	BHLHE2	PAS	[95]	N/A	3.30.450.20	PAS
2YKH	*Mycobacterium tuberculosis*	Rv3220c	PAS	[96]	N/A	N/A	N/A
4HIA	*Rhodobacter sphaeroides*		LOV	[97]	N/A	N/A	PAS
3EWK	*Methylococcus capsulatus*	MmoS	PAS	[98]	PYP-like sensor (PAS domain)	3.30.450.20	PAS

N/A – not assigned; LOV, a subfamily of the PAS domain

Table 1.S4. Best Pfam database matches for extracellular PAS-like domains in sequence-profile and profile-profile searches.

PDB	Organism	HMM scan			HHpred		
		Best hit	**E-value**	**Clan**	**Best hit**	**Probability**	**Clan**
1P0Z	*Klebsiella pneumoniae*	Cache_3	4.9e-34	Cache	Cache_3	99.7	Cache
3BY8	*Escherichia coli*	Cache_3	6.2e-38	Cache	Cache_3	99.9	Cache
3C8C	*Vibrio cholerae*	Mcp_N Cache_1	3.9e-32 3.6e-21	Cache	Mcp_N Cache_1	98 99.8	Cache
1YAX	*Salmonella enterica*	PhoQ_ sensor	1.6e-69	N/A	PhoQ_ sensor	100	N/A
3BQ8	*Escherichia coli*	PhoQ_ sensor	2.9e-69	N/A	PhoQ_ sensor	100	N/A
2HJE	*Vibrio harveyi*	LuxQ-periplasm	6.5e-109	N/A	LuxQ-periplasm	100	N/A
3C38	*Vibrio cholerae*	LuxQ-periplasm	6.6e-106	N/A	LuxQ-periplasm	100	N/A
3E4P	*Sinorhizobium meliloti*	Cache_3	0.83	Cache	Cache_1	99.4	Cache
3BY9	*Vibrio cholerae*	YkuI_C	4.1e-5	Cache	Cache_1	99.5	Cache
3B42	*Geobacter sulfurreducens*	DUF3365	0.984	N/A	DUF3365	99.3	N/A
2W27	*Bacillus subtilis*	YkuI_C	2.2e-73	Cache	YkuI_C	100	Cache
2VA0	*Cellvibrio japonicus*	no hit		N/A	DUF2286	74.2	N/A
3LIA	*Methanosarcina mazei*	Cache_1	3.6e-13	Cache	Cache_1	99.7	Cache
3LIB	*Methanosarcina mazei*	Cache_1	2.5e-13	Cache	Cache_1	99.7	Cache
3LIC	*Shewanella oneidensis*	Cache_1	1.3e-3	Cache	Cache_1	99.7	Cache
3LID	*Vibrio parahaemolyticus*	no hit		N/A	Cache_1	99.6	Cache
3LIF	*Rhodopseudomonas palustris*	Cache_1	3.1e-3	Cache	Cache_1	99.7	Cache
3CWF	*Bacillus subtilis*	no hit		N/A	Cache_3	98.6	Cache
3T4J	*Arabidopsis thaliana*	CHASE	1.4e-21	N/A	CHASE	100	N/A
4JGO	*Bacillus subtilis*	Cache_1	5.4e-3	Cache	Cache_1	99.6	Cache
2QHK	*Vibrio parahaemolyticus*	Cache_2	5.6e-27	Cache	Cache_2	99.5	Cache

Table 1.S5. Newly defined Cache superfamily. Number of sequences in UniProt 2012_06 release are shown.

Family	Total	HK	MCP	GCD	AC / GC	SP	STK	IC	TF	PDB
Double domains										
dCache_1	15569	4958	4908	2880	265	204	25	467	300	3C8C, 2P7J, 2ZBB, 3BY9, 3E4P, 3LIA, 3LIB, 3LIC, 3LID, 3LIF, 4JGO
dCache_2	299	71	92	89	-	21	-	-	1	-
dCache_3	883	327	236	248	4	8	1	-	-	-
Cache_3-Cache_2	407	17	330	10	-	-	-	-	-	-
CHASE	1214	607	-	549	9	3	5	-	-	3T4J
LuxQ-periplasm	115	112	-	1	-	-	-	-	-	2HJE, 3C38
Single domains										
sCache_2	2243	356	1534	29	-	2	-	-	-	2QHK
sCache_3_1	2854	2799	3	15	-	1	2	-	3	3CWF
sCache_3_2	2499	2189	64	40	-	60	2	-	3	1P0Z, 2J80, 2V9A, 3BY8
sCache_3_3	276	14	201	15	-	-	-	-	14	-
YkuI_C	277	-	-	178	-	-	-	-	-	-
CHASE4	529	79	7	387	3	-	-	-	-	-
Stimulus_sens_1	203	202	-	-	-	-	-	-	-	-
DUF2222	713	705	-	1	-	-	-	-	-	-
SMP_2	788	-	-	-	-	-	-	-	-	-
Diacid_rec	1274	-	30	3	-	-	-	-	1192	-
2CSK_N	966	952	-	-	-	-	-	-	-	-
PhoQ_sensor	556	551	-	-	-	-	-	-	-	3BQ8, 1YAX

Abbreviations: MCP, methyl-accepting chemotaxis proteins (MCPsignal); HK, histidine kinases (HATPase_c, HATPase_c_2, HATPase_c_3, HATPase_c_5, HisKA, HisKA_2, HisKA_3, HWE_HK); GCD, c-di-GMP-cyclases and

diesterases (GGDEF, EAL, HD); SP, serine phosphatases (SpoIIE, PP2C, PP2C_2); AC, adenylate- and guanylate cyclases (guanylate_cyc): STK, serine/threonine kinases (Pkinase); TF, transcription factors (HTH clan, LytTR); IC, ion channels (VWA_N, VGCC_alpha2).

Table 1.S6. Number of Cache domains predicted by Pfam 27 Cache models and new models against Pfam 27 database and NCBI non-redundant (NR) database (April 2015 release)

HMM models	Model	Pfam 27	NR
New models	dCache_1	15569	60390
	dCache_2	299	995
	dCache_3	883	2706
	Cache_3-Cache_2	407	1733
	sCache_2	2243	8043
	sCache_3_1	2854	6493
	sCache_3_2	2499	7979
	sCache_3_3	276	1038
	Total	**25030**	**89377**
Pfam 27 models	Cache_1	5381	18940
	Cache_2	2250	7705
	Cache_3	2608	8265
	Total	**10239**	**34910**

Table 1.S7. Query coverage of extracellular region by Cache hits with new models.

Query Coverage of Extracellular region (%)	Percent of Cache hits
100	65.27
90	13.8
80	7.02
70	4.86
60	3.06
50	2.71
40	1.81
30	1.04
20	0.38
10	0.06

Table 1.S8. Computational coverage of Cache domains in proteins with known 3D structure

PDB	Length	Pfam models				New models			
		Domain	Start	End	Coverage	Domain	Start	End	Coverage
2QHK	174	Cache_2	14	102	51.15	sCache_2	14	160	84.48
1P0Z	131	Cache_3	15	127	86.26	sCache_3_2	3	128	96.18
2J80	135	Cache_3	21	133	83.7	sCache_3_2	6	133	94.81
3BY8	142	Cache_3	18	133	81.69	sCache_3_2	5	142	97.18
2V9A	133	Cache_3	19	131	84.96	sCache_3_2	4	131	96.24
3CWF	122	-	-	-	0	sCache_3_1	8	113	86.89
3BY9	260	YkuI_C	71	114	16.92	dCache_1	5	215	81.15
3C8C	240	MCP_N	2	70	28.75	dCache_1	5	237	97.08
		Cache_1	106	182	32.08				
3LI8	291	Cache_1	140	207	23.37	dCache_1	9	279	93.13
3LIB	290	Cache_1	138	207	24.14	dCache_1	9	278	93.1
3LIC	274	-	-	-	0	dCache_1	10	269	94.89
3LID	295	-	-	-	0	dCache_1	12	272	88.47
3LIF	254	-	-	-	0	dCache_1	9	249	94.88
4JGO	217	-	-	-	0	dCache_1	14	182	77.88
2P7J	287	-	-	-	0	dCache_1	12	272	90.94
3E4P	305	-	-	-	0	dCache_1	53	240	61.64

Table 1.S9. Cellular localization prediction for members of the Cache superfamily.

Family	Total	Between 2 TM or TM and HAMP (Extracellular)	No TM (Intracellular)
dCache_1	15568	12252	1004
dCache_2	299	249	14
dCache_3	882	787	11
Cache_3-Cache_2	407	351	11
sCache_2	2243	1783	174
sCache_3_1	2854	2658	9
sCache_3_2	2499	2253	17
sCache_3_3	276	249	3
Diacid_rec	1274	0	1268
CHASE	1214	861	58
2CSK_N	966	834	19
SMP_2	788	439	31
DUF2222	713	695	1
PhoQ_Sensor	556	542	0
CHASE4	529	411	35
Stimulus_sens_1	203	201	1
YkuI_C	184	0	180
LuxQ-periplasm	115	84	2
Cache clan	**31570**	**24648 (78.07 %)**	**2838 (8.99%)**
PAS clan	**88093**	**573 (0.65%)**	**65496 (74.34%)**

Table 1.S10. Abundance of the two largest clans in prokaryotic extracellular sensory domains. Domain models were searched against non-redundant prokaryotic extracellular sequences.

Clan	Family	Count
Cache	dCache_1	9570
	sCache_3_1	1481
	sCache_2	1347
	sCache_3_2	1221
	CHASE	690
	dCache_3	678
	2CSK_N	670
	CHASE4	365
	Cache_3-Cache_2	302
	DUF2222	294
	dCache_2	226
	sCache_3_3	208
	Sensor_TM1	153
	Stimulus_sens_1	149
	PhoQ_Sensor	133
	LuxQ-periplasm	63
	Total	**17550**
4HB_MCP	4HB_MCP_1	2484
	TarH	1164
	CHASE3	1034
	Total	**4682**

CHAPTER II: DIVERSITY OF EXTRACELLULAR SENSORY DOMAINS IN PROKARYOTIC SIGNAL TRANSDUCTION

This chapter was taken from a manuscript in preparation:

Amit A. Upadhyay and Igor B. Zhulin. **Diversity of extracellular sensory domains in prokaryotic signal transduction**

A.A.U and I.B.Z. designed research; A.A.U. performed research; A.A.U. and I.B.Z. analyzed data; and A.A.U. and I.B.Z. wrote the manuscript.

Abstract

Bacteria need to constantly sense changes in environment in order to regulate various cellular processes which is essential to their survival. Towards this end, they make use of several membrane bound sensors to relay information from the environment to the regulatory machinery inside the cell. Since these sensory domains bind to a vast array of ligands, they have highly divergent sequences. Over the years, structures of several extracellular domain have become available and many sequence-based domain families such as Cache, 4HB_MCP and CHASE have been defined. However, owing to their high sequence divergence, a large majority of the extracellular sensory domains are still unannotated. It is not known if the unannotated domains constitute novel domain or they are divergent forms of known domain. Here we show that using sensitive profile-profile comparisons, about 85% of sensory domains, can be classified into known Pfam domain families with relatively high confidence. We also found that the ubiquitous extracellular Cache domains are even more widely distributed than reported earlier. Almost three/fourths of all sensory domains belong to Cache or 4HB_MCP clan. Our results will help in improving existing HMM models which will enable easy identification of these

domains. The remaining 15% unannotated sequences are also potential targets for structural genomics studies.

Introduction

Sensing and responding to environmental stimuli is central to the survival of microbial cells. In order to respond to different environmental stimuli, bacteria employ several signal transduction systems. The most widely used are the two component systems comprising of histidine kinases and response regulators. Other output modules include methyl accepting chemotaxis proteins (MCP), adenylate and guanylate cyclases, c-di-GMP associated cyclases and diesterases and Ser/Thr kinases [1]. Many of these systems comprise of transmembrane proteins that have an extracellular/periplasmic sensory domain that is responsible for binding small molecules or protein-protein interactions. An N-terminal extracellular domain flanked by two transmembrane regions is the most common topology for both histidine kinases as well as chemoreceptors [2, 3].

There are limited number of output domains that are well conserved and can be easily identified by existing domain models. Sensory domains on the other hand evolve much more rapidly in order to bind to diverse ligands and show considerably higher sequence diversity [4, 5]. Structures for several extracellular sensory domains have become available over the years. Based on the overall fold, theses extracellular sensors can be grouped in to three different classes – mixed αβ, all-helical and periplasmic solute-binding protein fold [6]. A combination of two seven-blade β-propellers along with a C-terminal Ig-type fold, mostly limited to Bacteroidetes and Proteobacteria, has also been observed [7].

An additional class of all-β sensors has been predicted based on sequence analysis alone [6].

The mixed αβ class comprises of sensors with the PAS-like/PDC fold [6] as seen in PhoQ [8, 9], DcuS [10, 11] and CitA [12, 13]. The all-helical sensor domains are typified by the four antiparallel helix bundle as seen in Tar [14, 15] and NarX [16]. HK29 was the first available structure for periplasmic solute-binding protein fold [17]. An interesting feature observed in sensory domains is the presence of duplicated forms of each of the former domains – LuxQ [18], DctB [10, 19] and KinD [20] possess double PAS-like/PDC domains [21]; TorS [22] and McpS [23] comprise of double four-helix bundles; BvgS from *Bordetella pertussis* [24] has a double periplasmic-solute binding protein fold. The Pfam database [25] has several domain models that enable identification of these domains using sequence information alone. Some models for these structural classes include: Cache_1 [26], Cache_2, Cache_3, YkuI_C, CHASE, CHASE4 [27-29], LuxQ-periplasm and PhoQ_Sensor for the PAS-like/PDC domains; 4HB_MCP_1 [30], TarH, CHASE3, KinB_sensor, NIT [31] and HBM [32] for all-helical domains, PBP clan for periplasmic solute-binding proteins; Reg_prop for β-propeller and Y_Y_Y for the Ig fold.

In spite of the increase and improvements in the number of models for these extracellular domains, a large number of extracellular sequences remain unannotated. A study in 2010 reported that almost 89% of the ligand binding extracellular region was unannotated in bacterial MCP alone [2]. It has been suggested that PDC/PAS-like domains are the predominant sensory domains on the basis of unpublished results [5, 21]. However, to our knowledge, no comprehensive analysis for the relative abundance of extracellular

sensory domains is publically available. We undertook this study to investigate the extent of coverage of the sensory domains by existing Pfam models and also to determine if the unannotated regions are divergent sequences not covered by existing models or they constitute novel domains. We used sensitive profile-profile comparisons and found out that almost 85% of sensory domains belong to known domain folds. We also found that Cache domains are much more widespread and ubiquitously used in all prokaryotic signal transduction systems.

Results and Discussion

Distribution of membrane associated sensors in different output classes. We obtained a non-redundant set of prokaryotic sequences from Pfam containing at least one output signaling domain. We classified sequences into following major groups based on the Pfam output domains: transcription factors, histidine kinases, c-di-GMP-cyclases and diesterases, methyl-accepting chemotaxis proteins, adenylate- and guanylate cyclases, serine phosphatases and serine/threonine kinases and RNA-binding (see Methods). Transcription factors comprised 69% of all the signal transduction proteins followed by histidine kinases (15%), c-di-GMP metabolism associated proteins (8%) and chemoreceptors (3%) (Table 2.1). We used three methods for transmembrane (TM) prediction – TMHMM, DAS-TMfilter and Phobius. Using a consensus of at least two methods, we found that 81% of signal transduction proteins are intracellular (Table 2.2). Most transcription factors were intracellular (97%) (Table 2.3). However, small number of membrane associated proteins with transcription factor output domains were also identified. Most of these sequences also had the histidine kinase output domains. Majority

of histidine kinases and chemoreceptors are membrane associated. 38% of histidine kinases and 17% of chemoreceptors respectively were cytoplasmic with no predicted TM regions (Table 2.3). Earlier studies have also reported similar figures with 33% histidine kinases being cytoplasmic and 13-16% MCPs being cytoplasmic [1, 2]. Serine/threonine kinases have almost equal proportion of cytoplasmic and membrane associated proteins while remaining groups including c-di-GMP cyclases and diesterases, adenylate and guanylate cyclases and serine proteases are mostly cytoplasmic. A study in 2005 reported that about half of adenylate and diguanylate cyclases and c-di-GMP diesterases were membrane bound and majority of serine/threonine kinases and HD-GYP domains were soluble [1]. The discrepancies maybe attributed to the changes in Pfam domain as well as the starting dataset.

The number of TM regions in membrane associated proteins ranged between one and twenty which is in agreement with previous study [3]. Overall the most common topology was class I [4], comprising of an extracellular region flanked by two TM regions with an intracellular output domain, which has also been reported earlier for histidine kinases [3] as well as MCPs [2]. Class I topology was the most common in membrane bound histidine kinases (61%), MCPs (86%), c-di-GMP cyclases and diesterases (42%) and serine phosphatases (26%). In case of adenylate and guanylate cyclases, the most common topology was that with six TM regions (29%) followed by class I (21%). For transcription factors, serine/threonine kinases and other small groups, RseA_N and RHH_1 most common topology was with one TM region. Since topology I was the most common and

it is a more reliable way to extract the extracellular regions, we limited our analysis to class I proteins with only two TM regions.

Assigning domains to unannotated sequences. We extracted the region between two TM regions and obtained a non-redundant set of 95,139 sequences. Using HMMer tool [33] that uses sequence-profile comparison, we were able to assign known Pfam domains to 36% sequences with no domain predicted in 60,464 sequences (Table 2.4). We then utilized a more sensitive HMM-HMM based method – HHsearch [34, 35]. Fig. 2.1 shows the distribution of probability scores for the extracellular sequences using HHsearch. Using a more stringent threshold of 98 for the probability score, we can assign 59% of sequences to Pfam domain families while using the more relaxed threshold of 95, we can assign 71% of the extracellular sequences to known domain families. In either case, we are left with 38,831 and 27,678 unannotated sequences (Table 2.4).

The high sequence variability of sensory domains can severely limit the diversity of sequences used in the seed alignment that is used to generate models which may consequently limit our ability to reliably predict the presence of these domains. To overcome this limitation, instead of completely relying on the models, we opted to compare profiles generated for each sequence. We created profile for each sequence using hhblits [34] and then carried out all-against-all hhsearch for all sequences. The domain family of the best known hit from list of annotated domains was assigned to the unannotated sequence in each iteration. The list of annotated sequences was updated after each iteration until no domains could be further assigned (see Methods for details). Using this approach we were able to increase the percentage of annotated sequences

from 36 to 85. The number of unannotated sequences was reduced from 60,464 to 14,602 sequences. Thus majority of the unannotated sequences are divergent forms of already known domains and do not constitute novel folds.

Using the new annotation scheme we see some striking changes in the abundance of known sensory domains (Fig. 2.2, Appendix 2.1). Cache_1 domain (double PAS-like) which constituted 7% of sensor domains is found to actually comprise 21% of all sensor domains and is the most abundant sensory domain. It is followed by the four helix bundle, 4HB_MCP_1, which also shows a two-fold increase in the abundance of the family. It thus appears that in case of mixed αβ, the double domain fold has been evolutionarily favored while in case of four helix bundle the single domain fold is more preferred. Other families that showed a large increase in number of newly classified members include – PAS_12, Cache_3, 2CSK_N, SMP_2, CHASE3, DUF2222, HBM, NIT, PhoQ_Sensor and RisS_PPD. The availability of newly identified members of these families should help in improving domain models so that these domains can be more readily identified. In some cases such as TarH, the number of family members was reduced after using the annotation pipeline compared to the earlier Pfam annotations, as these members were reassigned to other families such as 4HB_MCP_1.

Relationship between families of extracellular sensory domains. We created a similarity network using all-against-all HHsearch results. Nodes represent domain families in each cluster. Nodes are connected by an edge if the hhsearch probability score is >=95 and the query coverage is >=90 reciprocally for any pair of sequences from two different domain families in a cluster. Using this threshold, most families were found to

cluster with known clan members (Fig. 2.3). In addition, we also observed some new relationships between sensor domain families. An interesting observation is that PAS_12, a newly defined family in the intracellular PAS clan, is actually part of the Cache clan. DUF3365, Sensor_TM1 and LapD_MoxY_N also appear to be related to the Cache clan. KinB_sensor domain which is unique to *Pseudomonas* genus [36] and has an all-helical structure was found to cluster with the 4HB clan. In addition we also found relationships with other all-helical domains such as CZB and NIT. There is one DUF3365 sequence which clusters with 4HB_MCP_1, one 4HB_MCP_1 which clusters with PilJ domains and three Abhydrolase_1 sequences that cluster with CHASE3. Since the number of sequences is very low for these, we cannot be certain about the relationships between these domains. In all other cases, families were found to cluster with known clan members. Thus based on these results and our previous study, we consider the Cache clan to comprise of the following domain families: Cache_1, Cache_2, Cache_3, CHASE, CHASE4, Sensor_TM1, Stimulus_sens_1, PAS_12, DUF3365, DUF2222, 2CSK_N, SMP_2, Diacid_rec, PhoQ_Sensor and LapD_Moxy_N, LuxQ-periplasm and YkuI_C. The 4HB_MCP clan comprises of 4HB_MCP_1, TarH, CHASE3, KinB_sensor, CZB, HBM and NIT.

Diversity of extracellular domains in different signal transduction systems.

The distribution of sensory domains in different signal transduction systems in shown in Fig. 2.4. Overall, Cache domains are the most widely distributed in prokaryotic signal transduction systems accounting for almost 55% of all sensory domains. The 4HB_MCP clan comprising of all-helical domains are the next most widely distributed (19%). In case

of Archaea, almost all sensory domains belong to the Cache clan (Appendix 2.1). 15% of sequences were not assigned to any domains. Based on the HHPred probability scores (Fig. 2.1), it is unlikely that all these constitute novel domains and are likely even more divergent sequences. Cache domains are the most abundant in histidine kinases, c-di-GMP cyclases and diesterases, adenylate and guanylate cyclases and serine phosphatases. Interestingly, 4HB_MCP domains are the most abundant in chemoreceptors. The most common domain used in case of membrane associated sensors with transcription factor related output domains is the double 7 blade β-propeller Reg_prop domain and the Ig like Y_Y_Y domain.

The double PAS-like Cache_1 domains are ubiquitous and are present in association with all output domains. Several members of the Cache domain show strong association with a single class of output domains – Cache_3, PAS_12, 2CSK_N, DUF2222, PhoQ_Sensor, Sensor_TM1 and Stimulus_sens_1 with histidine kinases, Cache_2 with MCPs; CHASE4 and LapD_MoxY_N with c-di-GMP cyclases and diesterases. The single four helix bundle domain 4HB_MCP_1 is most prevalent in MCP while CHASE3 is most prevalent in histidine kinases. TarH is almost always found in association with MCP and KinB_sensor with histidine kinases. The double four-helix bundle domains of HBM and NIT are also present in MCPs and histidine kinases but are much more abundant in MCPs. The periplasmic solute-binding protein fold is almost absent in MCPs with only two instances of Phosphonate-bd domain. Some smaller domain families are specialized for specific output domain classes – RisS_PPD and CpxA_peri for histidine kinases and the CSS-motif domain for c-di-GMP cyclases and diesterases.

Conclusion

We showed here that almost 75% of all prokaryotic extracellular sensory domains belong to either the Cache clan or the 4HB_MCP clan. Almost 77% of unannotated sequences were divergent forms of known domains that were not picked up by existing domain models. We were able to reduce the percentage of unannotated sequences from 64% to 15%. These newly annotated domain can be used to improve existing models. Also, the unannotated sequences can be used for establishing new domain families. These would be useful targets for structural genomics efforts.

Materials and Methods

Data sources and Bioinformatics software. A local copy of MySQL Pfam 28 database based on Uniprot 2014_07 release served as the central data source. The following software packages were used in this study: HHsuite-2.0.16 16 [34, 35, 37], CD-HIT 4.5.7[38], Cytoscape 2.8.3 [39], Graph-0.96_01 (UnionFind) Perl library, TMHMM 2.0c [40], Phobius v1.01 [41], DAS-TMfilter (December 2012) [42], PfamScan (July 2015) [25] and HMMER 3.1b2 [33].

Retrieving extracellular sequences in signal transduction proteins and domain prediction. We retrieved 4,420,149 prokaryotic sequences from Pfam 28 which had at least one output domain as listed in MiST2.0 database. 100% redundant sequences were removed using CD-Hit which resulted in 1,365,467 sequences. TM regions were predicted using TMHMM, DAS and Phobius. Only those TM regions were considered that had an overlap of one position by at least two methods. We selected only those proteins that were predicted to contain two TM regions and retrieved the region between the two

TM if it was greater than 50 amino acids. After removing redundant sequences, the dataset comprised of 95,273 sequences. PfamScan tool was used to predict domains using default threshold. Some sequences in the dataset were found to belong to output domains and were discarded. The hhblits tool from HHsuite was used to generate profiles using uniprot20_2015_06 database with two iterations for the remaining 95,139 sequences. The hhsearch tool was then used to compare these profiles to Pfam28 database to predict domains. Non-overlapping domains that had a probability score of >=98 were assigned to each sequence.

Pipeline for annotation of unannotated sequences. The profiles generated for each of the 95,139 sequences were used carry to out an all-against-all comparison using hhsearch. A list of annotated sequences was compiled using hhsearch as described above. For unannotated sequences, the best hit to an annotated sequence was determined from the all-against-all hhsearch results. For earlier iterations, if the probability score was >=98 and the query coverage >=90, then the unannotated sequence was assigned the domain of the best annotated hit. The newly annotated sequence were added to the list of annotated sequences and the process was repeated until no domains could be assigned. Subsequently, a probability score threshold of >=95 and query coverage >=90 was used until no domains could be assigned to the unannotated sequences. The details of each iteration are shown in Table 2.5.

Abundance of output domains and sensory domains. Sequences were classified into the following major groups based on the Pfam output domains: transcription factors (HTH clan, LytTR, ROS_MUCR, Arc, CtsR, ArsD, ComK,); histidine kinases (HATPase_c,

HATPase_c_2, HATPase_c_3, HATPase_c_5, HisKA, HisKA_2, HisKA_3, HWE_HK); c-di-GMP-cyclases and diesterases (GGDEF, EAL, HD); methyl-accepting chemotaxis proteins (MCPsignal); adenylate- and guanylate cyclases (Guanylate_cyc, CYTH); serine phosphatases (SpoIIE) and serine/threonine kinases (Pkinase, Pkinase_Tyr) and RNA-binding (ANTAR, CsrA). When determining the abundance of output domains, if multiple output domains from different groups were present in a sequence, they were counted separately for each class. In case of sensory domains, the percentage was calculated by determining the total number of domains instead of total number of sequences. The Reg_prop domain is usually present in multiple copies in a sequence and since it is a small motif, it was counted only once for each sequence to prevent overestimation.

References

1. Galperin, M.Y., *A census of membrane-bound and intracellular signal transduction proteins in bacteria: bacterial IQ, extroverts and introverts.* BMC Microbiol, 2005. **5**: p. 35.
2. Lacal, J., et al., *Sensing of environmental signals: classification of chemoreceptors according to the size of their ligand binding regions.* Environ Microbiol, 2010. **12**(11): p. 2873-84.
3. Mascher, T., J.D. Helmann, and G. Unden, *Stimulus perception in bacterial signal-transducing histidine kinases.* Microbiol Mol Biol Rev, 2006. **70**(4): p. 910-38.
4. Wuichet, K., R.P. Alexander, and I.B. Zhulin, *Comparative Genomic and Protein Sequence Analyses of a Complex System Controlling Bacterial Chemotaxis.* 2007. **422**: p. 3-31.
5. Szurmant, H., R.A. White, and J.A. Hoch, *Sensor complexes regulating two-component signal transduction.* Curr Opin Struct Biol, 2007. **17**(6): p. 706-15.
6. Cheung, J. and W.A. Hendrickson, *Sensor domains of two-component regulatory systems.* Curr Opin Microbiol, 2010. **13**(2): p. 116-23.
7. Lowe, E.C., et al., *A scissor blade-like closing mechanism implicated in transmembrane signaling in a Bacteroides hybrid two-component system.* Proc Natl Acad Sci U S A, 2012. **109**(19): p. 7298-303.
8. Cheung, J., et al., *Crystal structure of a functional dimer of the PhoQ sensor domain.* J Biol Chem, 2008. **283**(20): p. 13762-70.
9. Cho, U.S., et al., *Metal Bridges between the PhoQ Sensor Domain and the Membrane Regulate Transmembrane Signaling.* J Mol Biol, 2006. **356**(5): p. 1193-1206.
10. Cheung, J. and W.A. Hendrickson, *Crystal structures of C4-dicarboxylate ligand complexes with sensor domains of histidine kinases DcuS and DctB.* J Biol Chem, 2008. **283**(44): p. 30256-65.
11. Pappalardo, L., *The NMR Structure of the Sensory Domain of the Membranous Two-component Fumarate Sensor (Histidine Protein Kinase) DcuS of Escherichia coli.* Journal of Biological Chemistry, 2003. **278**(40): p. 39185-39188.
12. Sevvana, M., et al., *A ligand-induced switch in the periplasmic domain of sensor histidine kinase CitA.* J Mol Biol, 2008. **377**(2): p. 512-23.
13. Reinelt, S., et al., *The structure of the periplasmic ligand-binding domain of the sensor kinase CitA reveals the first extracellular PAS domain.* J Biol Chem, 2003. **278**(40): p. 39189-96.
14. Bowie, J.U., A.A. Pakula, and M.I. Simon, *The three-dimensional structure of the aspartate receptor from Escherichia coli.* Acta Crystallogr D Biol Crystallogr, 1995. **51**(Pt 2): p. 145-54.
15. Milburn, M.V., et al., *Three-dimensional structures of the ligand-binding domain of the bacterial aspartate receptor with and without a ligand.* Science, 1991. **254**(5036): p. 1342-7.
16. Cheung, J. and W.A. Hendrickson, *Structural analysis of ligand stimulation of the histidine kinase NarX.* Structure, 2009. **17**(2): p. 190-201.

17. Cheung, J., M. Le-Khac, and W.A. Hendrickson, *Crystal structure of a histidine kinase sensor domain with similarity to periplasmic binding proteins.* Proteins, 2009. **77**(1): p. 235-41.
18. Neiditch, M.B., et al., *Ligand-induced asymmetry in histidine sensor kinase complex regulates quorum sensing.* Cell, 2006. **126**(6): p. 1095-108.
19. Zhou, Y.F., et al., *C4-dicarboxylates sensing mechanism revealed by the crystal structures of DctB sensor domain.* J Mol Biol, 2008. **383**(1): p. 49-61.
20. Wu, R., et al., *Insight into the sporulation phosphorelay: Crystal structure of the sensor domain of Bacillus subtilis histidine kinase, KinD.* Protein Science, 2013.
21. Zhang, Z. and W.A. Hendrickson, *Structural characterization of the predominant family of histidine kinase sensor domains.* J Mol Biol, 2010. **400**(3): p. 335-53.
22. Moore, J.O. and W.A. Hendrickson, *Structural analysis of sensor domains from the TMAO-responsive histidine kinase receptor TorS.* Structure, 2009. **17**(9): p. 1195-204.
23. Pineda-Molina, E., et al., *Evidence for chemoreceptors with bimodular ligand-binding regions harboring two signal-binding sites.* Proceedings of the National Academy of Sciences, 2012. **109**(46): p. 18926-18931.
24. Herrou, J., et al., *Periplasmic domain of the sensor-kinase BvgS reveals a new paradigm for the Venus flytrap mechanism.* Proc Natl Acad Sci U S A, 2010. **107**(40): p. 17351-5.
25. Finn, R.D., et al., *Pfam: the protein families database.* Nucleic acids research, 2013: p. gkt1223.
26. Anantharaman, V. and L. Aravind, *Cache-a signaling domain common to animal Ca (2+)-channel subunits and a class of prokaryotic chemotaxis receptors.* Trends in Biochemical Sciences, 2000. **25**(11): p. 535.
27. Anantharaman, V. and L. Aravind, *The CHASE domain: a predicted ligand-binding module in plant cytokinin receptors and other eukaryotic and bacterial receptors.* Trends Biochem Sci, 2001. **26**(10): p. 579-82.
28. Mougel, C. and I.B. Zhulin, *CHASE: an extracellular sensing domain common to transmembrane receptors from prokaryotes, lower eukaryotes and plants.* Trends Biochem Sci, 2001. **26**(10): p. 582-4.
29. Zhulin, I.B., A.N. Nikolskaya, and M.Y. Galperin, *Common Extracellular Sensory Domains in Transmembrane Receptors for Diverse Signal Transduction Pathways in Bacteria and Archaea.* Journal of Bacteriology, 2003. **185**(1): p. 285-294.
30. Ulrich, L.E. and I.B. Zhulin, *Four-helix bundle: a ubiquitous sensory module in prokaryotic signal transduction.* Bioinformatics, 2005. **21 Suppl 3**: p. iii45-8.
31. Shu, C.J., L.E. Ulrich, and I.B. Zhulin, *The NIT domain: a predicted nitrate-responsive module in bacterial sensory receptors.* Trends Biochem Sci, 2003. **28**(3): p. 121-4.
32. Ortega, A. and T. Krell, *The HBM domain: introducing bimodularity to bacterial sensing.* Protein Sci, 2014. **23**(3): p. 332-6.
33. Eddy, S.R., *Accelerated Profile HMM Searches.* PLoS Comput Biol, 2011. **7**(10): p. e1002195.

34. Remmert, M., et al., *HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment.* Nat Methods, 2012. **9**(2): p. 173-5.
35. Soding, J., *Protein homology detection by HMM-HMM comparison.* Bioinformatics, 2005. **21**(7): p. 951-60.
36. Tan, K., et al., *Sensor domain of histidine kinase KinB of Pseudomonas: a helix-swapped dimer.* J Biol Chem, 2014. **289**(18): p. 12232-44.
37. Soding, J., A. Biegert, and A.N. Lupas, *The HHpred interactive server for protein homology detection and structure prediction.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W244-8.
38. Li, W. and A. Godzik, *Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.* Bioinformatics, 2006. **22**(13): p. 1658-1659.
39. Smoot, M.E., et al., *Cytoscape 2.8: new features for data integration and network visualization.* Bioinformatics, 2011. **27**(3): p. 431-432.
40. Krogh, A., et al., *Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes.* Journal of molecular biology, 2001. **305**(3): p. 567-580.
41. Käll, L., A. Krogh, and E.L. Sonnhammer, *A combined transmembrane topology and signal peptide prediction method.* Journal of molecular biology, 2004. **338**(5): p. 1027-1036.
42. Cserzo, M., et al., *TM or not TM: transmembrane protein prediction with low false positive rate using DAS-TMfilter.* Bioinformatics, 2004. **20**(1): p. 136-137.

Appendix

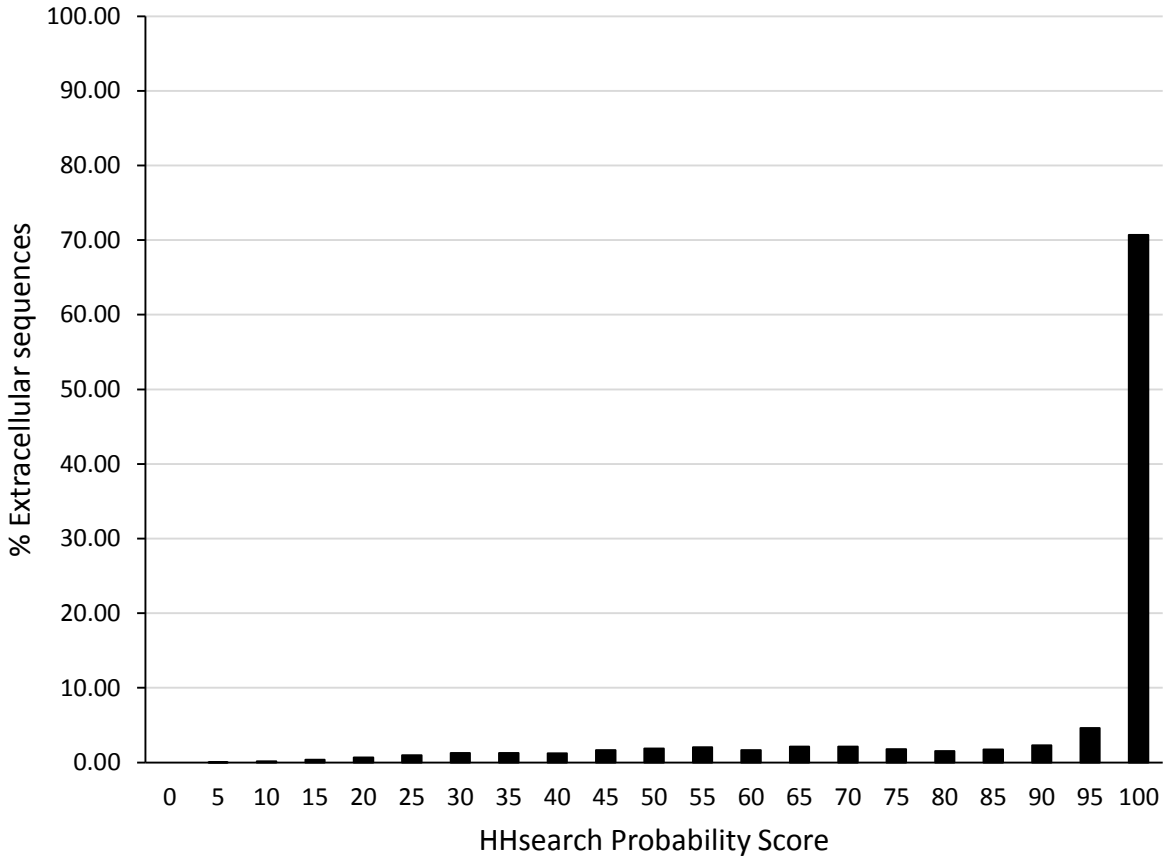

Fig. 2.1. Distribution of HHsearch probability scores (Pfam domains) for extracellular sequences

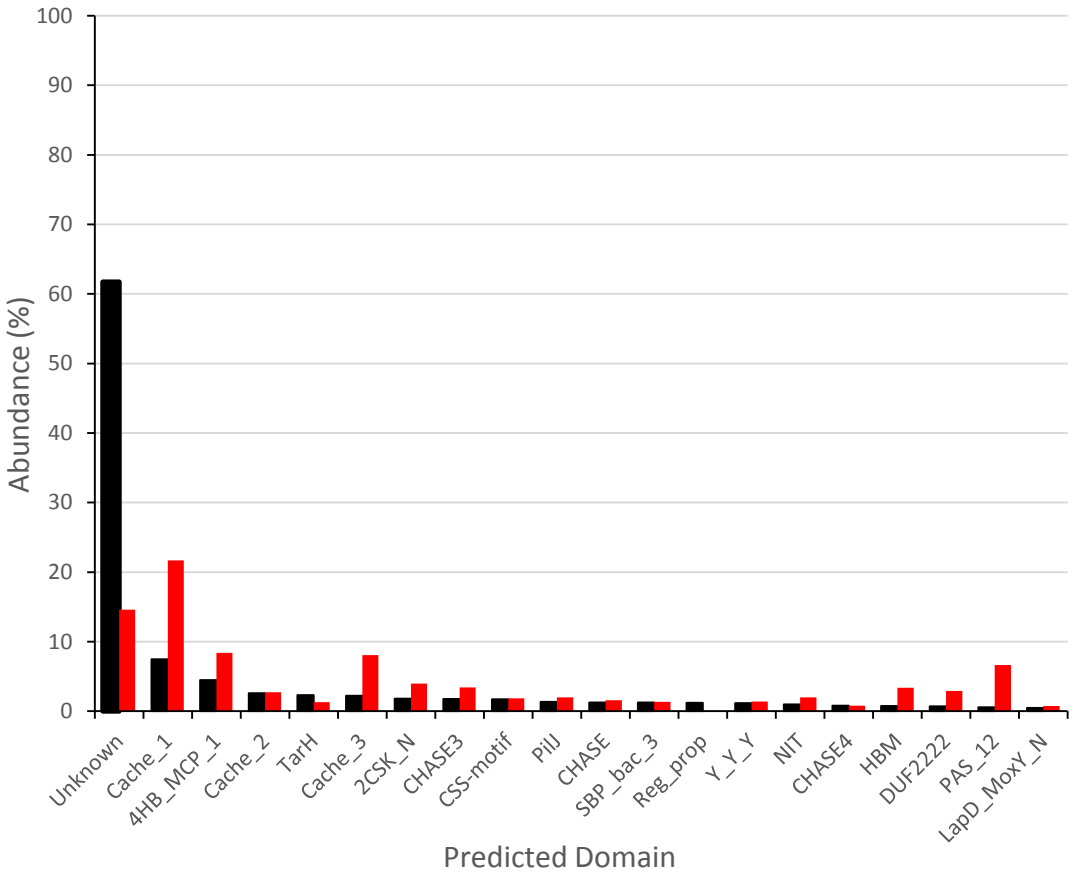

Fig. 2.2. Relative abundance of sensory domain families determined by HMMer (black) and HHsearch (red).

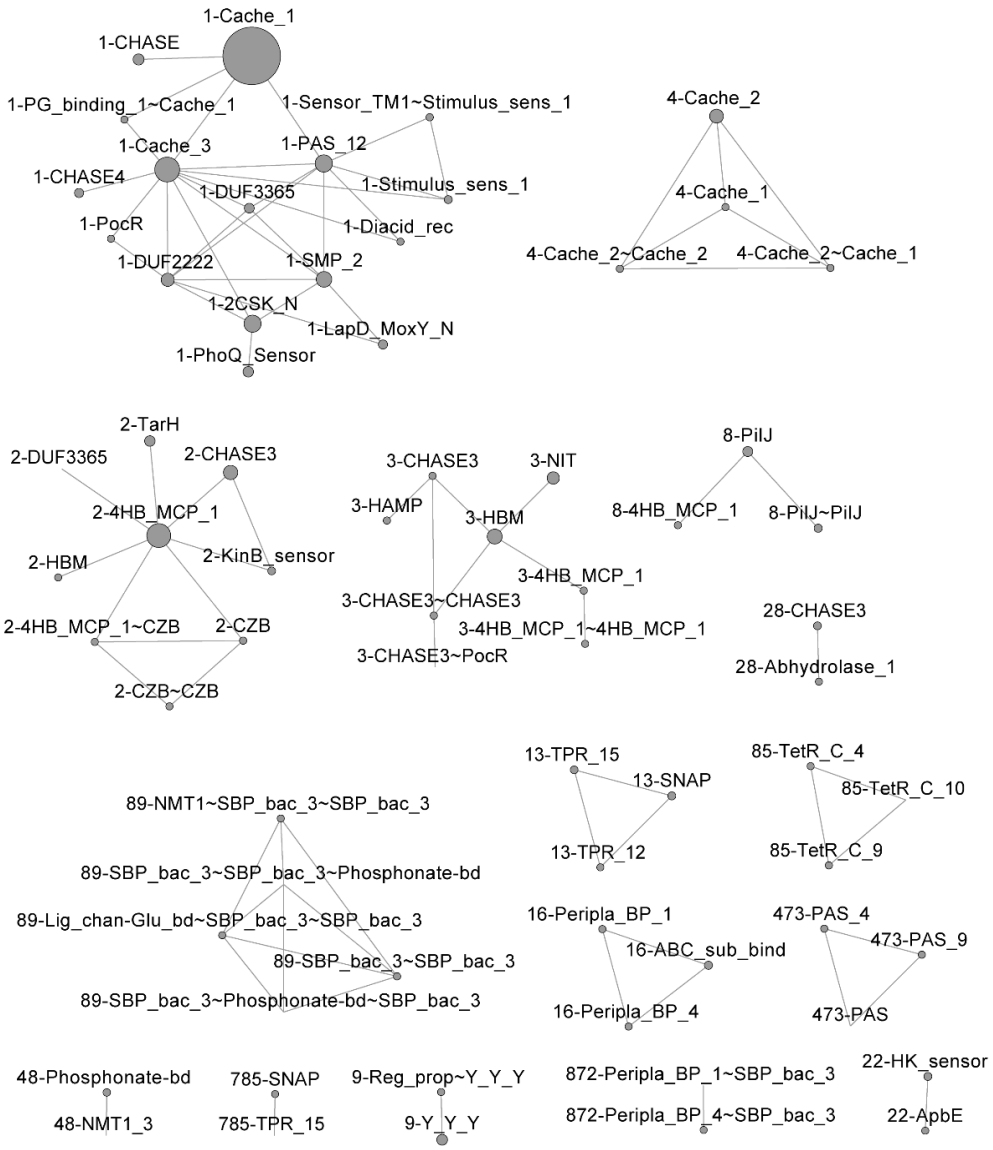

Fig. 2.3. Relationship between different extracellular sensory domain families.

A similarity network was generated with Perl Graph library using all-against-all hhsearch. Only those clusters that had domains from different families are shown. Each domain family is represented as a node. Edges represent reciprocal hhpred hits with a probability score >=95 and query coverage >=90 between a pair of sequences from two different families.

Fig. 2.4. Relative abundance of sensory domains in different output domain classes.

(A) All output domains, (B) Histidine kinases, (C) MCP, (D) c-di-GMP cyclases and diesterases, (E) Transcription factors, (F) Adenylate/Guanylate cyclases, (G) Serine phosphatases and (H) Ser/Thr kinases

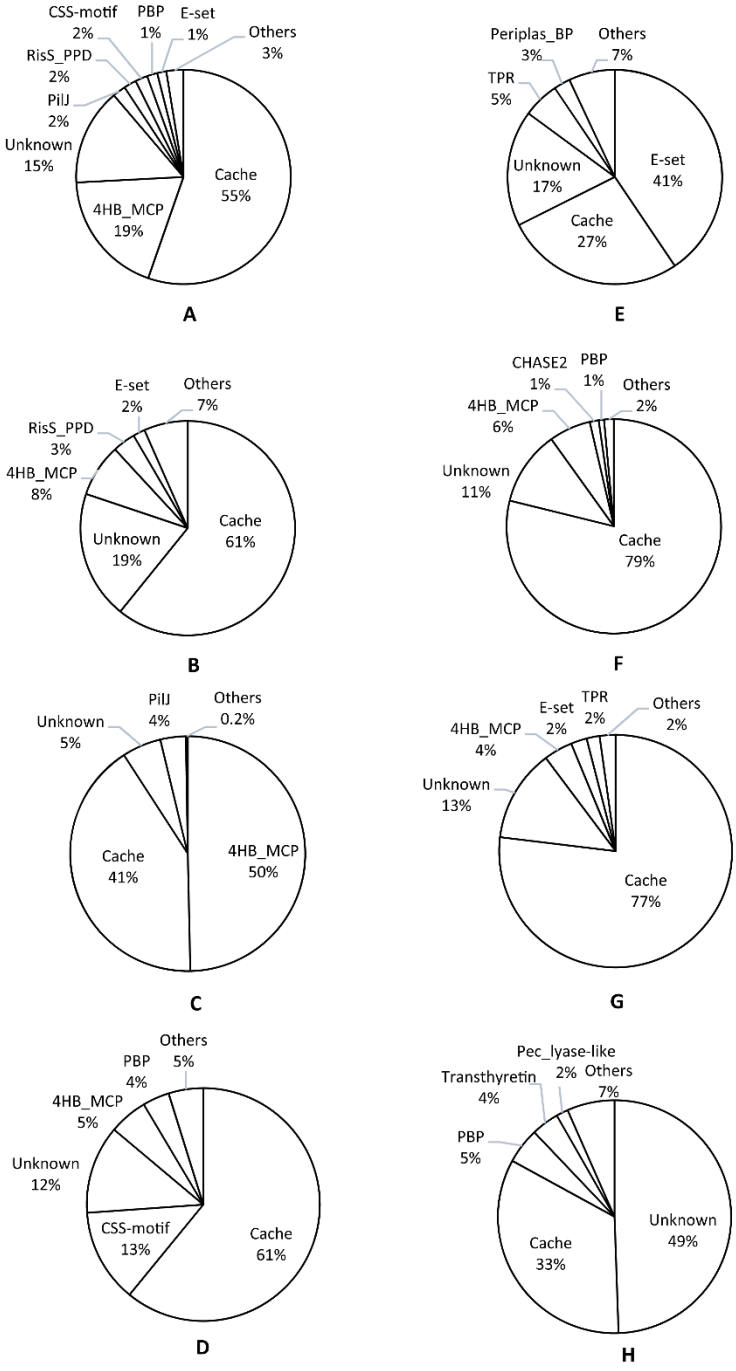

Fig. 2.4 continued

Table 2.1. Number of TM regions predicted in prokaryotic signal transduction proteins using TMHMM, Das and Phobius. Consensus column shows TM regions that overlapped with at least two methods

Number of TM	TMHMM	DAS	Phobius	Consensus (Counts)	Consensus (%)
0	1,114,844	1,115,855	1,097,426	1111439	81.40
1	57,876	41,549	95,207	50172	3.67
2	122,577	139,483	102,268	135477	9.92
3	11,611	9,714	10,691	8413	0.62
4	9,535	11,394	9,064	9877	0.72
5	13,360	11,083	9,745	12005	0.88
6	13,310	14,591	15,205	13704	1.00
7	10,789	8,853	13,473	11308	0.83
8	4,647	6,143	3,477	5286	0.39
9	2,213	2,100	3,272	2533	0.19
10	1,129	1,972	2,335	2111	0.15
11	988	844	648	516	0.04
12	1,627	925	601	689	0.05
13	279	431	1,089	1151	0.08
14	574	386	788	680	0.05
15	33	74	98	27	0.00
16	25	26	7	10	0.00
17	21	18	21	24	0.00
18	11	21	4	19	0.00
19	17	5	27	8	0.00
20	1	-	21	18	0.00

Table 2.2. Localization of prokaryotic signal transduction proteins for different output domain groups.

Output	Total		Cytoplasmic		Membrane-associated		2TM	
	Count	% of All	Count	% of Output	Count	% of Output	Count	% of Membrane associated
TF	949981	69.57	924971	97.4	25010	2.6	4354	17.4
HK	208177	15.25	78925	37.9	129252	62.1	78739	60.9
GCD	110042	8.06	64796	58.9	45246	41.1	18873	41.7
MCP	45233	3.31	7640	16.9	37593	83.1	32258	85.8
STK	24638	1.80	12431	50.5	12207	49.5	611	5.0
AC-GC	13194	0.97	8790	66.6	4404	33.4	943	21.4
SP	10489	0.77	6946	66.2	3543	33.8	904	25.5
RNA-binding	5474	0.40	5463	99.8	11	0.2	1	9.1
RHH_1	4621	0.34	4616	99.9	5	0.1	1	20.0
RseA_N	977	0.07	606	62.0	371	38.0	6	1.6

Abbreviations: TF, transcription factors; HK, histidine kinases; GCD, c-di-GMP-cyclases and diesterases; MCP, methyl-accepting chemotaxis proteins; STK, serine/threonine kinases; AC-GC, adenylate- and guanylate cyclases; SP, serine phosphatases

Table 2.3. Distribution of number of TM regions for membrane associated microbial signal transduction proteins.

Number of TM	HK	GCD	MCP	TF	STK	GC	SP	RseA_N	RNA-binding	RHH_1
1	12.24	13.71	9.39	60.86	85.66	9.33	8.35	98.38	90.91	80.00
2	60.92	41.71	85.81	17.41	5.01	21.41	25.52	1.62	9.09	20.00
3	3.95	2.62	0.83	3.32	2.58	6.61	11.77	0.00	0.00	0.00
4	4.62	2.80	0.23	5.11	2.71	18.39	4.71	0.00	0.00	0.00
5	5.84	6.93	1.45	1.38	1.44	6.95	2.12	0.00	0.00	0.00
6	4.89	9.40	1.94	2.99	1.29	28.95	6.77	0.00	0.00	0.00
7	3.27	11.00	0.10	5.68	0.52	5.04	10.61	0.00	0.00	0.00
8	1.51	5.79	0.06	0.77	0.47	2.45	9.79	0.00	0.00	0.00
9	0.57	2.63	0.03	0.24	0.11	0.32	14.20	0.00	0.00	0.00
10	0.48	2.71	0.03	0.22	0.06	0.14	5.22	0.00	0.00	0.00
11	0.08	0.56	0.11	0.41	0.01	0.05	0.31	0.00	0.00	0.00
12	0.21	0.04	0.00	1.54	0.07	0.02	0.06	0.00	0.00	0.00
13	0.87	0.04	0.00	0.02	0.02	0.00	0.31	0.00	0.00	0.00
14	0.51	0.00	0.00	0.00	0.05	0.00	0.23	0.00	0.00	0.00
15	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
16	0.00	0.00	0.00	0.00	0.01	0.02	0.03	0.00	0.00	0.00
17	0.00	0.00	0.00	0.01	0.00	0.32	0.00	0.00	0.00	0.00
18	0.01	0.02	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
19	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
20	0.00	0.04	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00

Abbreviations: TF, transcription factors; HK, histidine kinases; GCD, c-di-GMP-cyclases and diesterases; MCP, methyl-accepting chemotaxis proteins; STK, serine/threonine kinases; AC-GC, adenylate- and guanylate cyclases; SP, serine phosphatases

Table 2.4. Proportion of extracellular sequences assigned to existing Pfam domains using HMMer, HHsearch and the new annotation pipeline. The probability score cut off of 98 and 95 were used for HHsearch.

Pfam annotation	Pfam		HHsearch(P>=98)		HHsearch(P>=95)		New-pipeline	
	Count	%	Count	%	Count	%	Count	%
Assigned to Pfam domain	34675	36.45	56308	59.18	67461	70.91	81,077	85.22
Unannotated	60464	63.55	38831	40.82	27678	29.09	14062	14.78

Table 2.5. Iterative assignment of extracellular sequences to existing Pfam domain families using all-against-all HHsearch.

Iteration	HHSearch Probability threshold	Query coverage	Newly Assigned	Total Assigned	% Assigned	Unassigned	% Unassigned
0	98	-	56308	56308	59.2	38831	40.8
1	98	90	8805	65113	68.4	30026	31.6
2	98	90	2166	67279	70.7	27860	29.3
3	98	90	2387	69666	73.2	25473	26.8
4	98	90	1045	70711	74.3	24428	25.7
5	98	90	288	70999	74.6	24140	25.4
6	98	90	173	71172	74.8	23967	25.2
7	98	90	85	71257	74.9	23882	25.1
8	98	90	16	71273	74.9	23866	25.1
9	98	90	3	71276	74.9	23863	25.1
10	98	90	1	71277	74.9	23862	25.1
11	95	90	3103	74380	78.2	20759	21.8
12	95	90	4341	78721	82.7	16418	17.3
13	95	90	1632	80353	84.5	14786	15.5
14	95	90	388	80741	84.9	14398	15.1
15	95	90	263	81004	85.1	14135	14.9
16	95	90	70	81074	85.2	14065	14.8
17	95	90	3	81077	85.2	14062	14.8

CHAPTER III: SEQUENCE, STRUCTURE, AND EVOLUTION OF CELLULASES IN GLYCOSIDE HYDROLASE FAMILY 48

This research was originally published in The Journal of Biological Chemistry. Leonid O. Sukharnikov, Markus Alahuhta, Roman Brunecky, Amit Upadhyay, Michael E. Himmel, Vladimir V. Lunin, and Igor B. Zhulin. Sequence, Structure, and Evolution of Cellulases in Glycoside Hydrolase Family 48. *J. Biol. Chem.* 2012; 287: 41068-41077. © the American Society for Biochemistry and Molecular Biology."

A.U. performed analysis for metagenomics and contributed to writing for metagenomics sections. A.U. also assisted L.O.S. in some bioinformatics analysis and literature review

Abstract

Currently, the cost of cellulase enzymes remains a key economic impediment to commercialization of biofuels (1). Enzymes from glycoside hydrolase family 48 (GH48) are a critical component of numerous natural lignocellulose-degrading systems. Although computational mining of large genomic data sets is a promising new approach for identifying novel cellulolytic activities, current computational methods are unable to distinguish between cellulases and enzymes with different substrate specificities that belong to the same protein family. We show that by using a robust computational approach supported by experimental studies, cellulases and non-cellulases can be effectively identified within a given protein family. Phylogenetic analysis of GH48 showed non-monophyletic distribution, an indication of horizontal gene transfer. Enzymatic function of GH48 proteins coded by horizontally transferred genes was verified experimentally, which confirmed that these proteins are cellulases. Computational and structural studies of GH48 enzymes identified structural elements that define cellulases

and can be used to computationally distinguish them from non-cellulases. We propose that the structural element that can be used for ***in silico*** discrimination between cellulases and non-cellulases belonging to GH48 is an ω-loop located on the surface of the molecule and characterized by highly conserved rare amino acids. These markers were used to screen metagenomics data for “true” cellulases.

Introduction

The recent exponential growth of genomic data presents a unique opportunity to search for novel cellulolytic activities. However, the absence of a clear understanding of structural and functional features that are critical for decisive computational identification of cellulases prevents their identification in these data sets. True cellulases are defined as enzymes that show biochemical activity on cellulose substrates (*i.e.* crystalline or amorphous cellulose). Strikingly, all known cellulases have homologs that have similar protein folds and even amino acid sequences but do not show biochemical activity on cellulosic substrates (2), which makes computational-only identification of true cellulases error-prone. Glycoside hydrolase family 48 (GH48)[4] is one of the many families defined in the CAZy (Carbohydrate-Active EnZymes) database (3) that contains biochemically confirmed cellulases. Furthermore, GH48 cellulases are considered the key component of various cellulolytic systems (4–6). They are highly expressed in cellulolytic bacteria, such as *Clostridium cellulolyticum*, *Clostridium cellulovorans*, *Clostridium josui*, *Clostridium thermocellum*, and many others (4). In *C. thermocellum*, a bacterium that exhibits one of the highest rates of cellulose degradation among all known cellulolytic bacteria, GH48 cellulases are up-regulated during growth on crystalline cellulose (4).

Hence, these enzymes become the most abundant subunits in the *C. thermocellum* cellulosome, a complex of enzymes highly efficient in cellulose degradation (4, 7). Notably, complete knockout of both GH48 enzymes in *C. thermocellum* leads to a significant decrease in performance but does not completely abolish cellulolytic activity (4), whereas knockout of the GH48 gene in *Ruminococcus albus* (5) leads to nearly complete loss of cellulase activity.

Usually, only one (or rarely two or three) gene(s) encoding GH48 enzymes can be found in the genomes of cellulose-degrading bacteria (6), whereas genes for GH5 and GH9 cellulases are present in much higher numbers (8, 9). Interestingly, GH48 cellulases often act in synergy with GH9 cellulases, which increases their catalytic activity dramatically (10), a feature that may be utilized for industrial application of these enzymes (*e.g.* "designer cellulosomes") (11).

Experimental studies revealed that some GH48 cellulases have only cellulolytic activity and thus cannot hydrolyze other substrates (*i.e.* xylan and mannan) (12). A few GH48 cellulases have mixed substrate specificity (*e.g.* they are capable of degradation of xylan (13) or β-glucan (14) in addition to cellulose). There are two GH48 enzymes from the beetle *Gastrophysa atrocyanea* that are unable to hydrolyze cellulose-containing substrates (*e.g.* Avicel, carboxymethylcellulose, acid-swollen cellulose, etc.), whereas they showed distinct enzymatic activity toward chitin (15) (Appendix-3.1-Table S1).

Previous genomic studies have shown that GH48 enzymes are found in fungi as well as in bacteria, including Clostridia, Bacilli (both Firmicutes), and Actinobacteria. However,

the presence of the GH48 cellulase (16) in the evolutionarily distant deltaproteobacterium, *Myxobacter* sp. AL-1, was never explained.

Here we report evolutionary studies of GH48 enzymes, present a crystal structure of the GH48 enzyme encoded by a horizontally transferred gene, and characterize structural and functional differences between cellulases and chitinases in this group of enzymes. We also show that our computational approach can be used to search for true GH48 cellulases in metagenomic databases.

Experimental Procedures

Bioinformatics Software and Computer Programming Environment. The following software packages were used in this study: HMMER version 3.0 (17), MAFFT version 6.0 (18), MEGA version 5.0 (19), Jalview version 2.7.0 (20), and BLAST version 2.2.17 (21). All multiple-sequence alignments were built in MAFFT with its L-INS-i algorithm. All maximum likelihood phylogenetic trees were built in PhyML (22) with LG + Γ_4 + F parameters. Symmetrical best hits (SymBets) were assigned using the BLAST algorithm. All computational analyses were performed in a local computing environment, and custom scripts for data analysis were written in BioPerl. A remote version of the NCBI non-redundant database was used for direct queries using BioPerl scripts. A local version of the same database was used for querying with the *hmmsearch* algorithm of the HMMER package.

Data Sources and Literature Analysis. National Center for Biotechnology Information (NCBI) non-redundant database (nr) in FASTA format as of April 2011 was retrieved. A hidden Markov model of glycol_hydro_48 (PF02011) was retrieved from the Pfam database vPfam26 (23). Structures of Cel48S from *C. thermocellum* (24) and Cel48F from *C. cellulolyticum* (25) were retrieved from the Protein Data Bank (26).

Glycoside hydrolases of family 48 (classification according to the CAZy database (3)) with known activity were identified from the literature (Appendix-3.1-Table S1) and then mapped on the phylogenetic tree of GH48 enzymes in order to place the functional knowledge into the taxonomic context. Enzymes were considered to be of demonstrated cellulolytic function if their activity had been analyzed by *in vitro* biochemical studies.

Multiple Sequence Alignment and Construction of Phylogenetic Tree. 183 GH48 protein sequences were retrieved from the NCBI nr database using *hmmsearch* of the HMMER package (17) with Pfam gathering threshold and Pfam domain model glycol_hydro_48 (>600 amino acid residues). Then GH48 enzymatic domains corresponding to the Pfam model were excised from the protein sequences using BioPerl scripts and further analyzed. 69 domain sequences were found to be too short (<300 amino acid residues) and thus were discarded to improve the quality of the subsequent studies. 114 GH48 sequences were taken into further analysis.

A multiple sequence alignment (MSA) of 114 GH48 domains was constructed in MAFFT. The resulting alignment was used to build a maximum likelihood tree in PhyML. The conservation pattern in the MSA was analyzed in Jalview (20) with underlying tools, and the phylogenetic tree was analyzed using the MEGA5 package. Taxonomy assignments

for the proteins on the tree were taken from GenPept records from the NCBI protein database.

Identification of Orthologs, Paralogs, and Horizontally Transferred Genes. Because GH48 is typically present as one copy per genome, we assigned as orthologs all GH48 protein sequences that (i) form a monophyletic clade on the tree; (ii) were present as a single copy per genome; (iii) come from phyla with the same common ancestor after species-proteins tree topology reconciliation (Firmicutes, Actinobacteria, and Chloroflexi); and (iv) were characterized by symmetrical best matches (SymBets). Similar GH48 sequences that were present in two or more copies per genome were assigned as paralogs.

Horizontally transferred genes were defined in two ways: (i) by means of phylogenetic studies, where horizontally transferred genes were assigned based on phyletic distribution on the tree (27), and (ii) by means of a probabilistic approach (27), where the probability of occurrence of GH48 genes in prokaryotic genomes was calculated as the percentage of genomes containing GH48 genes divided by the total number of the available genomes, assuming that each genome contains only one GH48 gene (Table 3.1).

Metagenomic Data Analysis. We analyzed a publicly available data set of protein sequences from microbial communities in 295 metagenome samples retrieved from JGI/M (28) as of October, 2011 and the cow rumen data set from Ref. 29. Sequences encoding glycoside hydrolase family 48 proteins (glycol_hydro_48 (PF02011) Pfam

domain model) were collected from metagenome data sets with *hmmsearch*. Duplicate sequences were removed.

Cloning, Expression, and Purification of Hahella chejuensis GH48. A codon-optimized pMAL expression plasmid obtained from DNA2.0 (Menlo Park, CA) containing the *H. chejuensis* catalytic domain was transformed into *Escherichia coli* (BL21) (Agilent, Santa Clara, CA) and overexpressed at 37 °C in the presence of 0.3 mM IPTG. The recombinant fusion protein contained a C-terminal maltose-binding domain and was purified using an amylose high flow resin (New England Biolabs, Ipswich, MA). The eluted fusion protein was then cleaved using a Genenase protease site incorporated into the sequence (New England Biolabs). The *H. chejuensis* GH48 module was further purified by anion chromatography on a source 15Q column (GE Healthcare), using buffers A (20 mM Tris, pH 6.8) and B (20 mM Tris, pH 6.8, 2 M NaCl). Minor impurities were removed by size exclusion chromatography using HiLoad Superdex 75 (26/60) (GE Healthcare) in 20 mM acetate buffer, pH 5.0, containing 100 mM NaCl and 1 mM sodium azide. The purified protein solution was concentrated with a Vivaspin 5K concentrator (Vivaproducts, Littleton, MA), and its concentration was measured using the BCA assay (Pierce).

Model Substrate and Pretreated Biomass. Avicel (PH101), and phosphoric acid-swollen cellulose generated from Avicel, were used to evaluate the cellulolytic efficiency of *H. chejuensis* GH48. To provide a basis for the maximum theoretical sugar yield achievable from each substrate during enzymatic hydrolysis, portions of each of the pretreated solid samples were dried and subjected to the standard two-stage sulfuric acid hydrolysis method for determining structural carbohydrates in lignocelluloses, as

described by Sluiter *et al.* (30). In this method, the carbohydrate content of each pretreated sample is calculated from the carbohydrates released. In both cases, it is ~95% glucan.

Enzymatic Digestion Assays. GH48 activity was determined at 45 °C, at an enzyme loading of 15 mg/g glucan Avicel or 80 mg/g glucan phosphoric acid-swollen cellulose in 20 mM acetate buffer, pH 5.5, containing 10 mM $CaCl_2$ and 100 mM NaCl. The assay slurry was mixed by inversion. Digestions were run continuously for 7 days, and sugar release was monitored by removing aliquots. Samples taken at various time points and the enzymes were inactivated by boiling for 15 min. Samples were then filtered through 0.45-μm Acrodisc syringe filters and analyzed for glucose and cellobiose by HPLC. Samples were injected at 20 μl and run on an Agilent 1100 HPLC system equipped with a Bio-Rad Aminex HPX-87H 300 × 7.8-mm column heated to 55 °C. A constant flow of 0.6 ml/min was used with 0.1 M H_2SO_4 in water as the mobile phase to give separation of the analytes. Glucose and cellobiose were quantified against independent standard curves. All experiments were performed in triplicate, and the resulting extents of conversion are shown as percentage of glucan converted.

CD Methods. CD measurements were carried out using a Jasco J-715 spectropolarimeter with a jacketed quartz cell with a 1.0-mm path length. The cell temperature was controlled to within ±0.1 °C by circulating 90% ethylene glycol using a Neslab R-111m water bath (Neslab Instruments, Portsmouth, NH) through the CD cell jacket. The results were expressed as mean residue ellipticity, $[e]_{mrw}$. The spectra obtained were averages of five scans. The spectra were smoothed using an internal

algorithm in the Jasco software package, J-715 for Windows. Protein samples were studied in 20 mM sodium acetate buffer, pH 5.0, with 100 mM NaCl at a protein concentration of 0.25 mg/ml for the near-UV CD. Thermal denaturation of different constructs was monitored by CD in the near-UV (190–260 nm) region. For the analysis of thermal stability, the temperature was increased from 25 to 60 °C with a step size of 0.2 °C and monitored at a wavelength of 222 nm.

Crystallization. *H. chejuensis* GH48 (YP_433697) crystals were obtained with sitting drop vapor diffusion using a 96-well plate with Crystal Screen HT from Hampton Research (Aliso Viejo, CA). 50 µl of well solution was added to the reservoir, and drops were made with 0.2 µl of well solution and 0.2 µl of protein solution using a Phoenix crystallization robot (Art Robbins Instruments, Sunnyvale, CA). The crystals were grown at 20 °C with 0.05 M potassium phosphate monobasic and 20% (w/v) polyethylene glycol 8000 as the well solution. The protein solution contained 15 mg/ml protein, 20 mM acetic acid, pH 5, 100 mM NaCl, and 10 mM $CaCl_2$.

Data Collection and Processing. The *H. chejuensis* crystal was flash-frozen in a nitrogen gas stream at 100 K before home source data collection using a Bruker X8 MicroStar x-ray generator with Helios mirrors and a Bruker Platinum 135 CCD detector. Data were indexed and processed with the Bruker Suite of programs version 2011.2–0 (Bruker AXS, Madison, WI).

Structure Solution and Refinement. Intensities were converted into structure factors, and 5% of the reflections were flagged for R_{free} calculations using the programs F2MTZ, Truncate, CAD, and Unique from the CCP4 package of programs (31). The GH48

structure was solved using molecular replacement with the program Molrep (32) with Protein Data Bank entry 1G9G as a model. ARP/wARP (33) version 7.0 and Coot (34) version 0.6.2 were used for multiple cycles of automatic and manual model building. Further refinement and manual correction was performed using REFMAC5 (35) version 5.6.0117 and Coot. The MOLPROBITY method (36) was used to analyze the Ramachandran plot, and root mean square deviations of bond lengths and angles were calculated from ideal values of Engh and Huber stereochemical parameters (37). Wilson *B*-factor was calculated using CTRUNCATE (31) version 1.0.11. Average *B*-factors were calculated using the program ICM version 3.7–2a (Molsoft LLC, La Jolla, CA). The resulting structures have been deposited in the Protein Data Bank with code 4FUS. The data collection and refinement statistics are shown in Appendix 3.1-Table S2.

Results

Phyletic Distribution of GH48 Sequences and Horizontal Gene Transfer. GH48 enzymes that were retrieved from databases belong to only four prokaryotic phyla (Actinobacteria, Firmicutes, Chloroflexi, and Proteobacteria) and only two eukaryotic phyla (Fungi and Arthropoda), indicating a rather unusual evolutionary history. Taking into account that Firmicutes, Actinobacteria, and Chloroflexi (i) probably shared a common ancestor (38), (ii) showed GH48 enrichment compared with other phyla (Table 3.1), and (iii) contained a significant number of biochemically confirmed GH48 cellulases while lacking any confirmed non-cellulases, we hypothesize that the GH48 cellulase originated in the last common ancestor of Firmicutes, Actinobacteria, and Chloroflexi. Therefore, we

have first analyzed sequences only from these three phyla that satisfied two additional criteria: (i) they were present as the only GH48 gene in a genome, and (ii) they showed many-to-many symmetrical best hits (SymBets) relationships (39). As a result, 65 sequences, which included 12 biochemically confirmed cellulases, were taken into further analysis and aligned. The maximum likelihood tree constructed from this alignment was monophyletic (*i.e.* sequences from the same phylum were found in a single clade). In the next step, we determined the conserved residues in the alignment and found that all functionally important sites (including folding and substrate binding) were invariably conserved (Appendix-3.1-Table S3).

Because paralogs typically have a similar but not identical function, we asked whether paralogous GH48 sequences may represent enzymes with different substrate specificity. If so, they should show differences in some of the highly conserved sites, especially those implicated in substrate binding. Surprisingly, we found that paralogous GH48 sequences in genomes of Firmicutes and Actinobacteria were nearly identical (90–98% identity) and retained all conserved residues that were identified in the set of orthologous sequences. It appears that the functional innovation in paralogs resides not in the catalytic domain but in the repertoire of their auxiliary domains (Fig. 3.1).

The evidence of horizontal gene transfer emerges when a protein sequence from a particular organism shows high similarity to a homolog from a distant taxon (27). In the case of GH48, all sequences from Fungi were found in the middle of the Firmicutes clade, whereas all sequences from Insecta were found in the middle of the Actinobacterial clade

(Fig. 3.2). This non-monophyletic distribution clearly suggests horizontal gene transfer into eukaryotes from the two prokaryotic phyla.

Thus, a total of 23 horizontally transferred genes were identified through phylogenomic analysis, where an implicitly defined (see above) set of orthologs showed the presence of non-monophyletic clades with representatives of Proteobacteria, Fungi, and Insecta (Fig. 3.2). Additionally, in prokaryotes, they were also identified by a probabilistic approach (27), where relative increases in abundance of GH48 genes in the genomes of Actinobacteria, Firmicutes, and Chloroflexi were compared with that of Proteobacteria, as described under “Experimental Procedures” (Table 3.1). Notably, Actinobacteria, Firmicutes, and Chloroflexi genomes had much higher probability of occurrence of GH48 genes compared with Proteobacteria, Fungi, and insects, which along with their distribution on the phylogenetic tree presents additional evidence for horizontal gene transfer into the latter. In summary, here we define all GH48 orthologs and paralogs from Actinobacteria, Firmicutes, and Chloroflexi as true cellulases based on phylogenomic analysis, which correlates with their experimentally confirmed enzymatic activities (Fig. 3.2 and Appendix 3.1-Table S1).

Cellulose Digestion by the Horizontally Transferred GH48. A comprehensive list of all biochemically studied GH48 cellulases is presented in Appendix 3.1-Table S1. This list shows that previously studied cellulases are mostly present in Firmicutes and Actinobacteria, with a single representative of Proteobacteria (*Myxobacter* sp. Al-1). We determined the activity of the GH48 enzyme from a proteobacterium *H. chejuensis*, which was a subject of horizontal gene transfer, on both crystalline and amorphous substrates

(Appendix-3.1-Fig. S1). These results showed that *H. chejuensis* is a cellulase because it shows activity on the phosphoric acid-swollen cellulose substrate. The poor performance on the more crystalline substrate is probably due to the lack of the carbohydrate-binding module domains in our construct, which is critical for optimal performance on a crystalline substrate, such as Avicel.

Crystal Structure of *H. chejuensis* GH48. The structure of HcheGH48 was refined to a resolution of 1.75 Å with *R* and $R_{free}$of 0.154 and 0.205, respectively. There is one molecule in the asymmetric unit in complex with a cellobiose molecule bound at the product position. It has an $(\alpha/\alpha)_6$barrel fold with one calcium and two sodium atoms and multiple ethylene glycol, glycerol, acetate, and phosphate molecules. Due to the long crystallization time (more than 1 year), two residue modifications were observed: a 2-oxohistidine at position 352 and polyethylene glycol modification of Tyr-439. There were two outliers in the Ramachandran plot, Glu-72 and Ala-73, both of them well defined by the density and close to the allowed region.

Structural Comparison with Other Known GH48s. Pairwise secondary structure matching of structures with at least 70% secondary structure similarity by PDBefold (40) found 22 unique structural matches for HcheGH48 from the Protein Data Bank. All similar structures were CelF, CelS, or CelA GH48 variants with secondary structure similarity between 79 and 88%. Closer inspection of the structure shows that the overall fold (Fig. 3.3) and the catalytic tunnel are almost identical to *C. cellulolyticum* CelF, *C. thermocellum* CelS, and *Caldicellulosiraptor bescii* CelA. In HcheGH48, Glu-83 is the catalytic residue. The residues lining the tunnel, catalytic Glu-83, and the positions of the

sugar rings of the cellobiose molecule are mostly conserved when compared with the *C. cellulolyticum* CelF, *C. thermocellum* CelS, and *C. bescii* CelA GH48 structures (Protein Data Bank codes 1FCE (41), 1L2A (24), and 4EL8). The identical residues lining the pocket are Trp-344, Gln-247, Asp-241, Ser-245, Thr-239, Ser-136, Phe-346, Lys-303, Tyr-331, Thr-251, Gln-207, Phe-206, Asn-204, Trp-180, Trp-330, Tyr-357, Trp-450, Trp-453, Trp-447, His-64, Arg-648, Trp-650, Asp-529, Glu-83, and Glu-83. The biggest differences are Trp-450 and Ala-616. Trp-450 is a methionine in CelF GH48 and phenylalanine in *C. thermocellum* CelS and *C. bescii* CelA GH48s. Ala-616 is a histidine in CelS GH48 and alanine in the other structures.

Closer inspection of the ω-loop shows that it is defined by two anchor residues, Trp-508 and Asn-516 (Fig. 3.3). Comparison with *C. cellulolyticum* CelF, *C. thermocellum* CelS, and *C. bescii* CelA GH48 structures shows that these residues are conserved and have identical conformation in all four structures. The ω-loop of HcheGH48 differs from the others by having a proline at position 523, causing a local conformational change, where the other structures have a tyrosine, which further anchors the loop. This, however, does not change the overall conformation or position of the loop but does suggest that variability in the loop is possible without affecting activity.

Conserved Amino Acid Positions in the GH48 Family in the Context of Structure. We used sequence numbering of Cel48F from *C. cellulolyticum* H10 to designate amino acids in all multiple-sequence alignment studies because it is the most extensively studied GH48 structure currently available (25, 41, 42). Literature and MSA analysis showed that

all GH48 enzymes have 100% conserved catalytic acid and base positions (Glu-55 and Asp-230 in Cel48F); thus, these residues are not discussed.

There are three major types of amino acids that participate in substrate recognition and correct folding of the GH48 enzymes: hydrophobic stacking interactions, hydrogen bonding, and calcium coordination residues (Appendix-3.1-Table S3) (24, 25, 41, 42). All of these residues are highly conserved in orthologs from Actinobacteria, Firmicutes, and Chloroflexi as well as in Proteobacteria, which indicates that genes horizontally transferred to Proteobacteria code for cellulases, a statement confirmed biochemically (this work and see Ref. 16). Our results also revealed that the GH48 enzyme from *H. chejuensis* does not possess any additional elements that would differentiate it from other cellulases.

Consequently, we hypothesize that fungal GH48s are also cellulases due to their high sequence similarities with cellulolytic orthologs and the fact that almost all residues important for catalysis (Appendix-3.1-Table S3) are highly conserved in fungi (Appendix-3.1-Table S4) with only one exception, the Ca^{2+} coordination residues, which were considered to play a role in the thermal stability of GH48 enzymes (24) but not in substrate specificity. In contrast, GH48 enzymes from all insects are represented by non-cellulases because of the large number of amino acid substitutions in positions that are conserved among cellulases, one ω-loop deletion, and the lack of cellulolytic activity confirmed biochemically (15).

Mutations in critical positions were not found in all sequences from insects (Appendix-3.1-Table S4). Thus, MSA and structural analyses suggested that the major difference

between cellulases and non-cellulases (*i.e.* chitinases) from insects is the additional ω-loop located between Pro-469 and Ala-482 (as in Cel48F) in all cellulases. This ω-loop includes two residues highly conserved in all cellulolytic orthologs (99–100% conservation): Trp-472 and Asn-481. Residue Leu-484 (as in Cel48F), located adjacent to the loop and strictly conserved in cellulolytic orthologs, is also mutated in all insects. This ω-loop is located on the surface of the GH48 molecule and connects two β-strands that form one side of the catalytic tunnel near the exit of the product (Fig. 3.3). Thus, here we report structural differences that occurred after an event of horizontal gene transfer from Actinobacteria to Insecta that caused mutation of cellulases to chitinases.

Screening Metagenomic Data Sets for GH48 Cellulases. Sequences of 211 GH48 proteins were retrieved from the combined metagenome data set (>79 million sequences) with *hmmsearch* of HMMER (17) and glycol_hydro_48 Pfam domain model with the Pfam gathering threshold. Then 36 duplicates were removed, and the remaining 175 sequences were used in protein BLAST queries. BLAST results showed that these sequences belong to the same major phyla as sequences belonging to well defined genomes that were retrieved from the NCBI nr database (Actinobacteria, Firmicutes, Chloroflexi, Proteobacteria, and insects (Arthropoda)) except for Fungi (Fig. 3.4). These results indicate that fungal species are either absent from the metagenomes used in this study or significantly underrepresented. In summary, nine sequences from metagenomics samples belonged to insects and were classified as non-cellulases, and the other 166 sequences were classified as cellulases, based on the phylogenomic and structural evidence presented above.

To confirm the validity of this classification, 166 metagenomic GH48 sequences classified as cellulases were aligned by *hmmalign* of HMMER (17) with default Pfam parameters. MSA analysis (Appendix-3.2) showed that 93% of sequences have all of the residues important for protein folding and catalysis with very few conservative substitutions that were also found in some of the cellulolytic orthologs from complete genomes. A few non-conservative substitutions that were found in a small set of the sequences (~7% of all) could indicate potential differences in function or could simply be sequencing/assembly errors, a rather common problem in metagenomics (43, 44) Therefore, experimental evidence must be obtained to clarify this point.

Because metagenomic samples show a large variation in the total number of genes sequenced (*e.g.* a wastewater treatment plant metagenome has 30,169 genes, whereas a biofuel metagenome has 2,706,009 genes), the percentage of GH48 domains in each metagenome was calculated (Fig. 3.5). These metagenomes were also grouped together according to their habitats, and the percentage abundance of GH48 in each habitat was also calculated (Fig. 3.5).

Discussion

Using a phylogenomic approach, we have determined that the GH48-type enzymes might have originated in a common ancestor of three closely related phyla: Firmicutes, Actinobacteria, and Chloroflexi (38). We have determined a number of gene duplication events in representatives of these phyla and several cases of horizontal gene transfer. For example, fungi received these genes horizontally from a representative of Firmicutes, whereas insects received these genes from a representative of Actinobacteria. Similarly,

representatives of Proteobacteria also received their GH48 genes horizontally. By comparing orthologous sequences from Firmicutes, Actinobacteria, and Chloroflexi, we identified a number of amino acid positions that are uniquely conserved in this group of organisms. Satisfactorily, the only activity that was previously found in this group is that of a cellulase. Thus, we suggest that conserved positions in the catalytic domains from Firmicutes, Actinobacteria, and Chloroflexi can be used as a genomic signature for a GH48 cellulase.

We then wondered if this genomic signature for a cellulase remains intact in paralogs and horizontally transferred genes, because these types of genes often assume a slightly different function. For example, just one or a few mutations in a catalytic domain may lead to different substrate specificity. Notably, screening and study of paralogous sequences of GH48 proteins showed no significant differences in their catalytic domains but rather noticeable differences in their auxiliary domains (*i.e.* cellulose-binding domain, fibronectin type III-like domain, etc.). On the contrary, genes that were horizontally transferred from Actinobacteria to insects (Metazoa) acquired a new activity to hydrolyze chitin but lost the ability to degrade cellulose.

Following this initial evolutionary analysis, we extended our findings to structural analysis of GH48 enzymes. We found that all orthologs and paralogs have a 10–14 residue ω-loop (Pro-469 to Ala-482 as in Cel48F) that has no counterpart in enzymes from insects. Moreover, this ω-loop is constituted by highly conserved amino acids (Trp-472 and Asn-481 as in Cel48F) and located on the surface of the molecule. Thus, in accord with the

classical definition of ω-loops (45), it may play the following roles in this enzyme structure: folding, stability, or contribution to the dynamics of the enzyme during catalysis.

High conservation of the ω-loop residues in cellulases suggests its importance for the computational identification of cellulases, and the complete absence of the loop in all non-cellulases indicates that GH48 chitinases lost this structural element. We hypothesize that the absence of the loop in chitinases allows more conformational degrees of freedom in the active site tunnel upon binding of the substrate, which permits a bulkier chitin to “slide” freely. In contrast, cellulases may have more rigid structures “reinforced” by the ω-loop. Regardless of the exact role of the ω-loop, which can be determined only experimentally, we have suggested that it is important for cellulolytic activity, which has allowed us to design a strategy to identify new cellulases in metagenomic data.

Thus, phylogenomic and structural analyses of GH48 suggest that proteins from Actinobacteria, Firmicutes, Chloroflexi, and Proteobacteria are indeed cellulases. Biochemical activities of GH48 proteins from two Pyromyces species have never been studied; thus, it is unknown whether they are cellulases. However, because these proteins are not only homologous to known cellulases but also contain all conserved amino acids identified in our analysis, it is very likely that they also possess cellulolytic activities. On the other hand, GH48s from insects, where only chitinolytic activities were detected experimentally, are non-cellulases. Consequently, the existing Pfam model for GH48 can be used to retrieve true cellulases; however, there is one exception. GH48 proteins from insects should be annotated as non-cellulases. This approach allowed us to identify 166 true cellulases in the combined metagenomic data set of hundreds of environmental

samples. The largest number of cellulases came from the metagenomes of “engineered” microbial communities, such as enriched samples or bioreactors (*e.g.* the “mixed alcohol bioreactor” and the “cellulolytic enrichment from sediment of Great Boiling Springs”). Most of the environmental cellulases come from communities that typically include saprophytes (46), such as soil, wastewater, ant fungal gardens, and the rhizosphere (Fig. 3.5), which is in agreement with previously published research (47, 48). Interestingly, very few GH48 cellulases were identified in cow rumen microbial communities, which also correlates with previous extensive biochemical analysis of this classical cellulolytic community (29). Moreover, all of the GH48s from cow rumen, found in this study, belong to *Ruminococcus flavefaciens*, a highly specialized cellulose degrader. We hypothesize that because, collectively, major ruminal cellulolytic specialists are found to represent as little as 0.3% of the total bacterial population (49), and *R. flavefaciens* is typically one of the three most abundant cellulolytic bacteria in cow rumen (50), its GH48 gene was more selective for sequencing (51) when compared with the genes of other “rare” members of the community.

Conclusions

High-throughput computational screening for cellulases from genomic and metagenomic data sets is a challenge due to the absence of a clear understanding of structural and functional features that distinguish them from closely related enzymes with other substrate specificities (2). Here, we present a combined sequence-structure approach leading to the identification of clear markers that can be used to distinguish between cellulases and non-cellulases within the GH48 family. This approach was applied to

identify “true” GH48 cellulases in large metagenomic data sets, illustrating its feasibility in the search for novel cellulolytic capabilities.

Finally, we propose that this approach can be generalized to define genomic signatures for identifying cellulases in other CAZy families (2), such as GH5, GH9, GH12, GH45, and GH61 that are known to contain biochemically confirmed cellulases.

References

1. Aden A., Foust T. (2009) Technoeconomic analysis of the dilute sulfuric acid and enzymatic hydrolysis process for the conversion of corn stover to ethanol. Cellulose 16, 535–545
2. Sukharnikov L. O., Cantwell B. J., Podar M., Zhulin I. B. (2011) Cellulases. Ambiguous nonhomologous enzymes in a genomic perspective. Trends Biotechnol. 29, 473–479
3. Cantarel B. L., Coutinho P. M., Rancurel C., Bernard T., Lombard V., Henrissat B. (2009) The Carbohydrate-Active EnZymes database (CAZy). An expert resource for Glycogenomics. Nucleic Acids Res. 37, D233–D238
4. Olson D. G., Tripathi S. A., Giannone R. J., Lo J., Caiazza N. C., Hogsett D. A., Hettich R. L., Guss A. M., Dubrovsky G., Lynd L. R. (2010) Deletion of the Cel48S cellulase from Clostridium thermocellum. Proc. Natl. Acad. Sci. U.S.A. 107, 17727–17732
5. Devillard E., Goodheart D. B., Karnati S. K., Bayer E. A., Lamed R., Miron J., Nelson K. E., Morrison M. (2004) Ruminococcus albus 8 mutants defective in cellulose degradation are deficient in two processive endocellulases, Cel48A and Cel9B, both of which possess a novel modular architecture. J. Bacteriol. 186, 136–145
6. Izquierdo J. A., Sizova M. V., Lynd L. R. (2010) Diversity of bacteria and glycosyl hydrolase family 48 genes in cellulolytic consortia enriched from thermophilic biocompost. Appl. Environ. Microbiol. 76, 3545–3553
7. Gold N. D., Martin V. J. (2007) Global view of the Clostridium thermocellum cellulosome revealed by quantitative proteomic analysis. J. Bacteriol. 189, 6787–6795
8. Wisniewski-Dyé F., Borziak K., Khalsa-Moyers G., Alexandre G., Sukharnikov L. O., Wuichet K., Hurst G. B., McDonald W. H., Robertson J. S., Barbe V., Calteau A., Rouy Z., Mangenot S., Prigent-Combaret C., Normand P., Boyer M., Siguier P., Dessaux Y., Elmerich C., Condemine G., Krishnen G., Kennedy I., Paterson A. H., González V., Mavingui P., Zhulin I. B. (2011) Azospirillum genomes reveal transition of bacteria from aquatic to terrestrial environments. PLoS Genet. 7, e1002430
9. Dam P., Kataeva I., Yang S. J., Zhou F., Yin Y., Chou W., Poole F. L. 2nd., Westpheling J., Hettich R., Giannone R., Lewis D. L., Kelly R., Gilbert H. J., Henrissat B., Xu Y., Adams M. W. (2011) Insights into plant biomass conversion from the genome of the anaerobic thermophilic bacterium Caldicellulosiruptor bescii DSM 6725. Nucleic Acids Res. 39, 3240–3254
10. Irwin D. C., Zhang S., Wilson D. B. (2000) Cloning, expression and characterization of a family 48 exocellulase, Cel48A, from Thermobifida fusca. Eur. J. Biochem. 267, 4988–4997

11. Vazana Y., Moraïs S., Barak Y., Lamed R., Bayer E. A. (2010) Interplay between Clostridium thermocellum family 48 and family 9 cellulases in cellulosomal versus noncellulosomal states. Appl. Environ. Microbiol. 76, 3236–3243
12. Shen H., Gilkes N. R., Kilburn D. G., Miller R. C. Jr.., Warren R. A. (1995) Cellobiohydrolase B, a second exo-cellobiohydrolase from the cellulolytic bacterium Cellulomonas fimi. Biochem. J. 311, 67–74
13. Liu C. C., Doi R. H. (1998) Properties of exgS, a gene for a major subunit of the Clostridium cellulovorans cellulosome. Gene 211, 39–47
14. Berger E., Zhang D., Zverlov V. V., Schwarz W. H. (2007) Two noncellulosomal cellulases of Clostridium thermocellum, Cel9I and Cel48Y, hydrolyse crystalline cellulose synergistically. FEMS Microbiol. Lett. 268, 194–201
15. Fujita K., Shimomura K., Yamamoto K., Yamashita T., Suzuki K. (2006) A chitinase structurally related to the glycoside hydrolase family 48 is indispensable for the hormonally induced diapause termination in a beetle. Biochem. Biophys. Res. Commun. 345, 502–507
16. Ramírez-Ramírez N., Romero-García E. R., Calderón V. C., Avitia C. I., Téllez-Valencia A., Pedraza-Reyes M. (2008) Expression, characterization and synergistic interactions of Myxobacter sp. AL-1 Cel9 and Cel48 glycosyl hydrolases. Int. J. Mol. Sci. 9, 247–257
17. Finn R. D., Clements J., Eddy S. R. (2011) HMMER web server. Interactive sequence similarity searching. Nucleic Acids Res. 39, W29–W37
18. Katoh K., Toh H. (2010) Parallelization of the MAFFT multiple sequence alignment program. Bioinformatics 26, 1899–1900
19. Tamura K., Peterson D., Peterson N., Stecher G., Nei M., Kumar S. (2011) MEGA5. Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. 28, 2731–2739
20. Waterhouse A. M., Procter J. B., Martin D. M.., Clamp M., Barton G. J. (2009) Jalview version 2. A multiple sequence alignment editor and analysis workbench. Bioinformatics 25, 1189–1191
21. Altschul S. F., Madden T. L., Schäffer A. A., Zhang J., Zhang Z., Miller W., Lipman D. J. (1997) Gapped BLAST and PSI-BLAST. A new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402
22. Guindon S., Gascuel O. (2003) PhyML. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52, 696–704
23. Finn R. D., Mistry J., Tate J., Coggill P., Heger A., Pollington J. E., Gavin O. L., Gunasekaran P., Ceric G., Forslund K., Holm L., Sonnhammer E. L., Eddy S. R., Bateman A. (2010) The Pfam protein families database. Nucleic Acids Res. 38, D211–D222
24. Guimarães B. G., Souchon H., Lytle B. L., David Wu J. H., Alzari P. M. (2002) The crystal structure and catalytic mechanism of cellobiohydrolase CelS, the major enzymatic component of the Clostridium thermocellum cellulosome. J. Mol. Biol. 320, 587–596

25. Parsiegla G., Reverbel-Leroy C., Tardif C., Belaich J. P., Driguez H., Haser R. (2000) Crystal structures of the cellulase Cel48F in complex with inhibitors and substrates give insights into its processive action. Biochemistry 39, 11238–11246
26. Berman H. M., Westbrook J., Feng Z., Gilliland G., Bhat T. N., Weissig H., Shindyalov I. N., Bourne P. E. (2000) The Protein Data Bank. Nucleic Acids Res. 28, 235–242
27. Koonin E. V., Makarova K. S., Aravind L. (2001) Horizontal gene transfer in prokaryotes. Quantification and classification. Annu. Rev. Microbiol. 55, 709–742
28. Markowitz V. M., Chen I. M., Palaniappan K., Chu K., Szeto E., Grechkin Y., Ratner A., Jacob B., Huang J., Williams P., Huntemann M., Anderson I., Mavromatis K., Ivanova N. N., Kyrpides N. C. (2012) IMG. The integrated microbial genomes database and comparative analysis system. Nucleic Acids Res. 40, D115–D122
29. Hess M., Sczyrba A., Egan R., Kim T. W., Chokhawala H., Schroth G., Luo S., Clark D. S., Chen F., Zhang T., Mackie R. I., Pennacchio L. A., Tringe S. G., Visel A., Woyke T., Wang Z., Rubin E. M. (2011) Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. Science 331, 463–467
30. Sluiter A., Hames B., Ruiz R., Scarlata C., Sluiter J., Templeton D., Crocker D. (2006) Determination of structural carbohydrates and lignin in biomass. Laboratory Analytical Procedure (LAP). Technical Report NREL/TP-510–42618, National Renewable Energy Laboratory, Golden, CO
31. Winn M. D., Ballard C. C., Cowtan K. D., Dodson E. J., Emsley P., Evans P. R., Keegan R. M., Krissinel E. B., Leslie A. G., McCoy A., McNicholas S. J., Murshudov G. N., Pannu N. S., Potterton E. A., Powell H. R., Read R. J., Vagin A., Wilson K. S. (2011) Overview of the CCP4 suite and current developments. Acta Crystallogr. D Biol. Crystallogr. 67, 235–242
32. Vagin A., Teplyakov A. (2010) Molecular replacement with MOLREP. Acta Crystallogr. D Biol. Crystallogr. 66, 22–25
33. Langer G., Cohen S. X., Lamzin V. S., Perrakis A. (2008) Automated macromolecular model building for x-ray crystallography using ARP/wARP version 7. Nat. Protoc. 3, 1171–1179
34. Emsley P., Lohkamp B., Scott W. G., Cowtan K. (2010) Features and development of Coot. Acta Crystallogr. D Biol. Crystallogr. 66, 486–501
35. Murshudov G. N., Skubák P., Lebedev A. A., Pannu N. S., Steiner R. A., Nicholls R. A., Winn M. D., Long F., Vagin A. A. (2011) REFMAC5 for the refinement of macromolecular crystal structures. Acta Crystallogr. D Biol. Crystallogr. 67, 355–367
36. Chen V. B., Arendall W. B. 3rd., Headd J. J., Keedy D. A., Immormino R. M., Kapral G. J., Murray L. W., Richardson J. S., Richardson D. C. (2010) MolProbity. All-atom structure validation for macromolecular crystallography. Acta Crystallogr. D Biol. Crystallogr. 66, 12–21
37. Engh R. A., Huber R. (1991) Accurate bond and angle parameters for x-ray protein-structure refinement. Acta Crystallogr. A 47, 392–400

38. Gutiérrez-Preciado A., Henkin T. M., Grundy F. J., Yanofsky C., Merino E. (2009) Biochemical features and functional implications of the RNA-based T-box regulatory mechanism. Microbiol. Mol. Biol. Rev. 73, 36–61
39. Koonin E. V. (2005) Orthologs, paralogs and evolutionary genomics. Annu. Rev. Genet. 39, 309–338
40. Krissinel E., Henrick K. (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr. D Biol. Crystallogr. 60, 2256–2268
41. Parsiegla G., Juy M., Reverbel-Leroy C., Tardif C., Belaïch J. P., Driguez H., Haser R. (1998) The crystal structure of the processive endocellulase CelF of Clostridium cellulolyticum in complex with a thiooligosaccharide inhibitor at 2.0 Å resolution. EMBO J. 17, 5551–5562
42. Parsiegla G., Reverbel C., Tardif C., Driguez H., Haser R. (2008) Structures of mutants of cellulase Cel48F of Clostridium cellulolyticum in complex with long hemithiocellooligosaccharides give rise to a new view of the substrate pathway during processive action. J. Mol. Biol. 375, 499–510
43. Pignatelli M., Moya A. (2011) Evaluating the fidelity of de novo short read metagenomic assembly using simulated data. PLoS One 6, e19984
44. Rho M., Tang H., Ye Y. (2010) FragGeneScan. Predicting genes in short and error-prone reads. Nucleic Acids Res. 38, e191
45. Fetrow J. S. (1995) -Loops. Nonregular secondary structures significant in protein function and stability. FASEB J. 9, 708–717 Abstract
46. Mba Medie F., Davies G. J., Drancourt M., Henrissat B. (2012) Genome analyses highlight the different biological roles of cellulases. Nat. Rev. Microbiol. 10, 227–234
47. Suen G., Scott J. J., Aylward F. O., Adams S. M., Tringe S. G., Pinto-Tomás A. A., Foster C. E., Pauly M., Weimer P. J., Barry K. W., Goodwin L. A., Bouffard P., Li L., Osterberger J., Harkins T. T., Slater S. C., Donohue T. J., Currie C. R. (2010) An insect herbivore microbiome with high plant biomass-degrading capacity. PLoS Genet. 6, e1001129
48. Sessitsch A., Hardoim P., Döring J., Weilharter A., Krause A., Woyke T., Mitter B., Hauberg-Lotte L., Friedrich F., Rahalkar M., Hurek T., Sarkar A., Bodrossy L., van Overbeek L., Brar D., van Elsas J. D., Reinhold-Hurek B. (2012) Functional characteristics of an endophyte community colonizing rice roots as revealed by metagenomic analysis. Mol. Plant Microbe Interact. 25, 28–36
49. Brulc J. M., Yeoman C. J., Wilson M. K., Berg Miller M. E., Jeraldo P., Jindou S., Goldenfeld N., Flint H. J., Lamed R., Borovok I., Vodovnik M., Nelson K. E., Bayer E. A., White B. A. (2011) Cellulosomics, a gene-centric approach to investigating the intraspecific diversity and adaptation of Ruminococcus flavefaciens within the rumen. PLoS One 6, e25329
50. Huws S. A., Lee M. R., Muetzel S. M., Scott M. B., Wallace R. J., Scollan N. D. (2010) Forage type and fish oil cause shifts in rumen bacterial diversity. FEMS Microbiol Ecol. 73, 396–407

51. Cowan D., Meyer Q., Stafford W., Muyanga S., Cameron R., Wittwer P. (2005) Metagenomic gene discovery. Past, present and future. Trends Biotechnol. 23, 321–329

Appendix

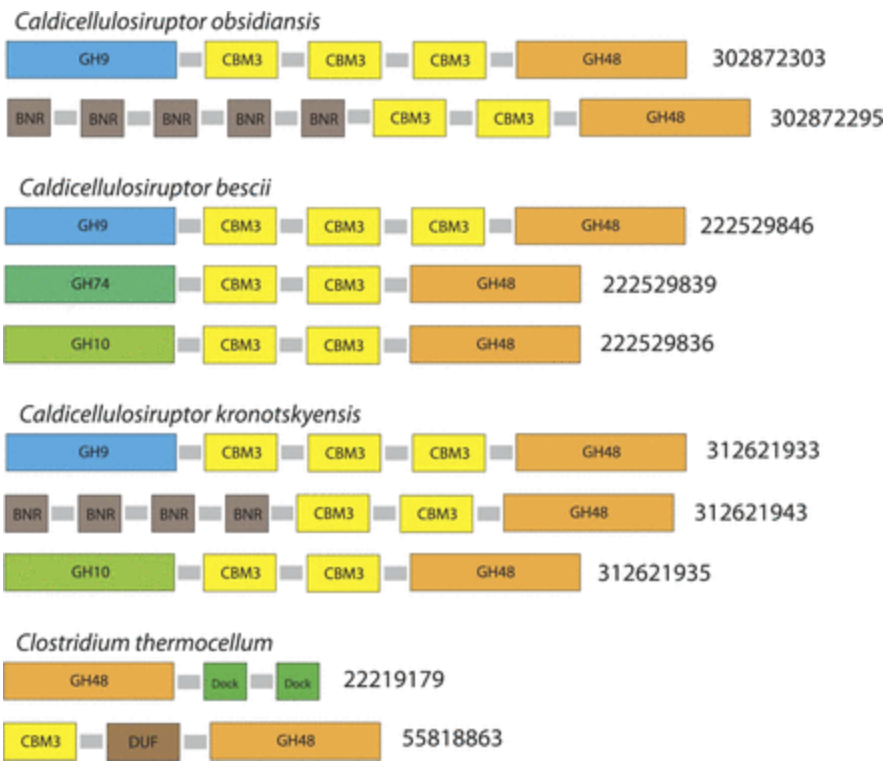

Fig. 3.1. Modular domain architecture of GH48 paralogs.

Representative examples of GH48 paralogs that contain different auxiliary domains are shown. The GenBank™ identifiers (GI numbers) are listed *beside* each protein. Protein domains are as follows: carbohydrate-binding module (*CBM*), dockerin (*Dock*), domain of unknown function (*DUF*), BNR repeat (*BNR*), and glycoside hydrolase (*GH*).

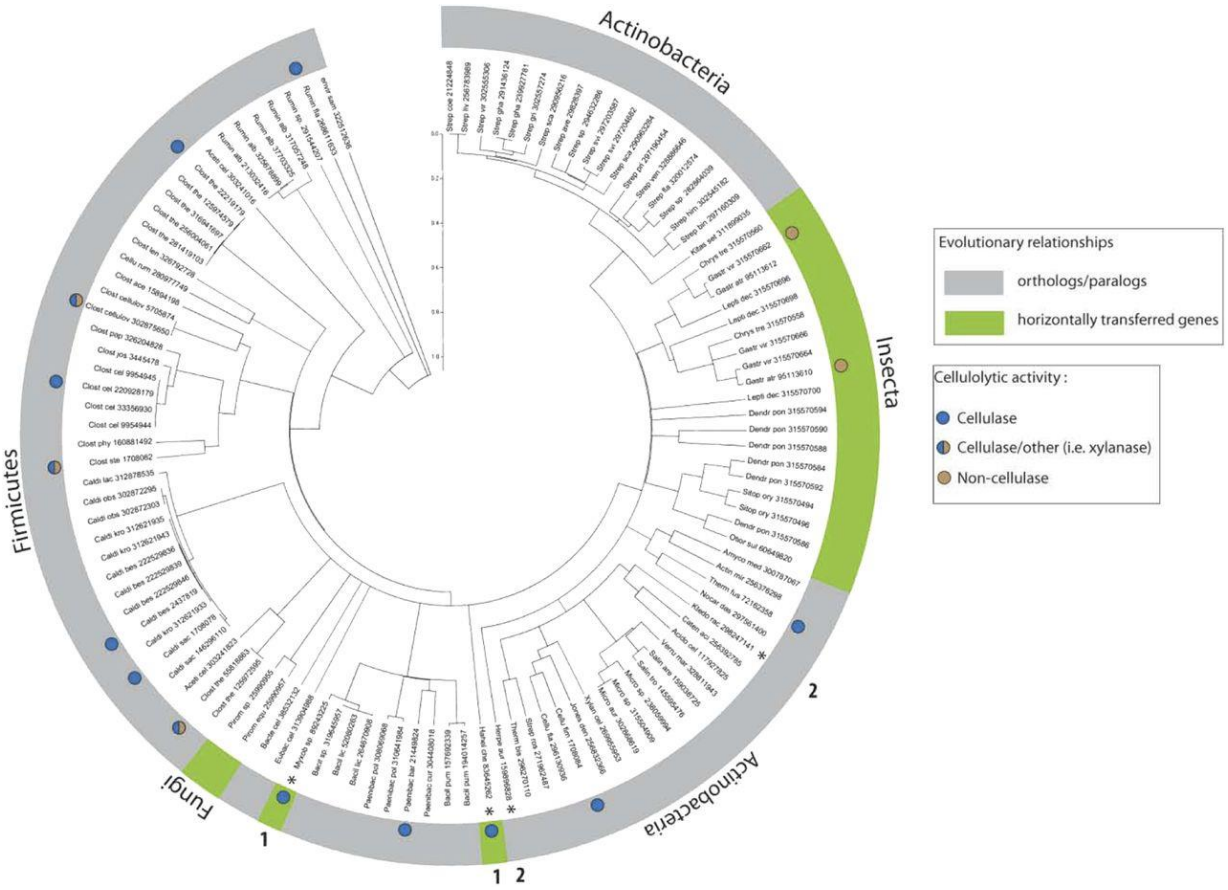

Fig. 3.2. Horizontal gene transfer of GH48 enzymes.

A maximum-likelihood phylogenetic tree constructed from multiple sequence alignment of GH48 sequences is shown. Known enzymatic activities, taxonomic information, and inferred evolutionary relationships are shown on the *outside circle*. Sequences from underrepresented phyla are marked with an *asterisk*: Proteobacteria (*1*) and Chloroflexi (*2*).

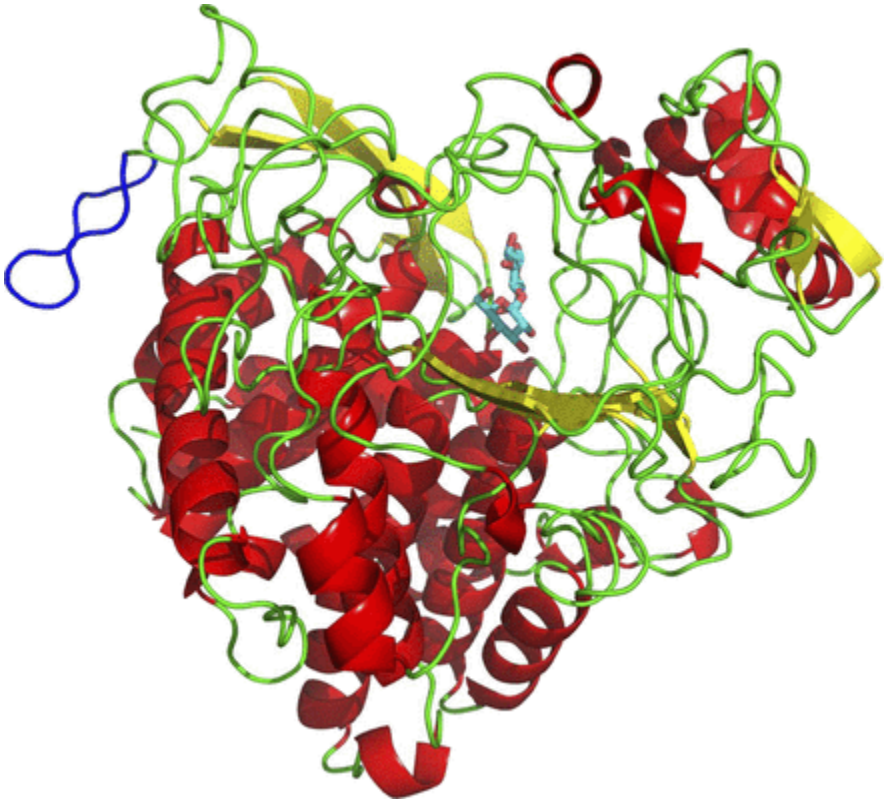

Fig. 3.3. Structure of GH48 from *H. chejunsis*.

The additional ω-loop identified in all cellulases is labeled in *blue*. The α-helices are shown in *red*, β-strands in *yellow*, and loops in *green*. The cellobiose molecule is shown with carbon atoms in *cyan* and oxygen atoms in *red*.

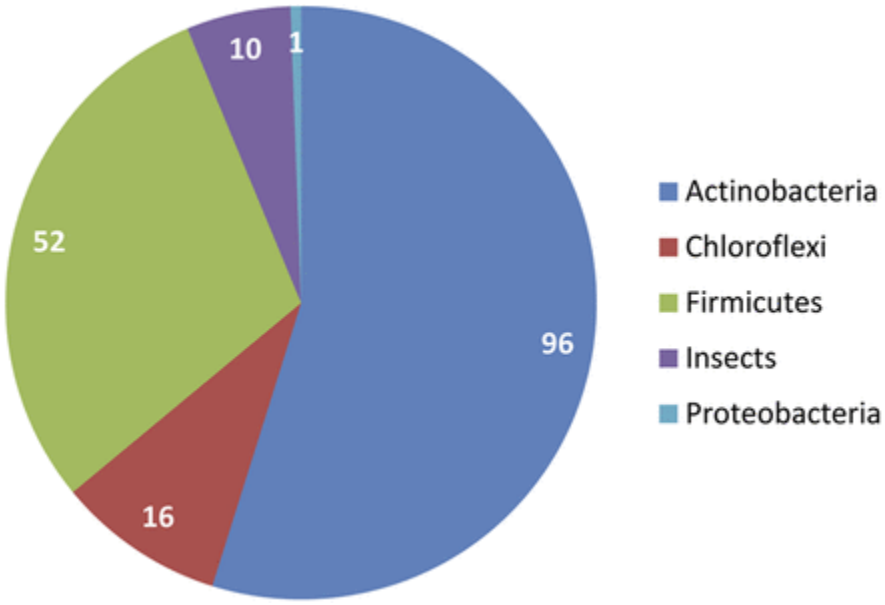

Fig. 3.4. Phyletic distribution of GH48 sequences retrieved from a combined metagenomic data set.

Nearly 95% of sequences belong to three closely related prokaryotic phyla: Actinobacteria, Firmicutes, and Chloroflexi.

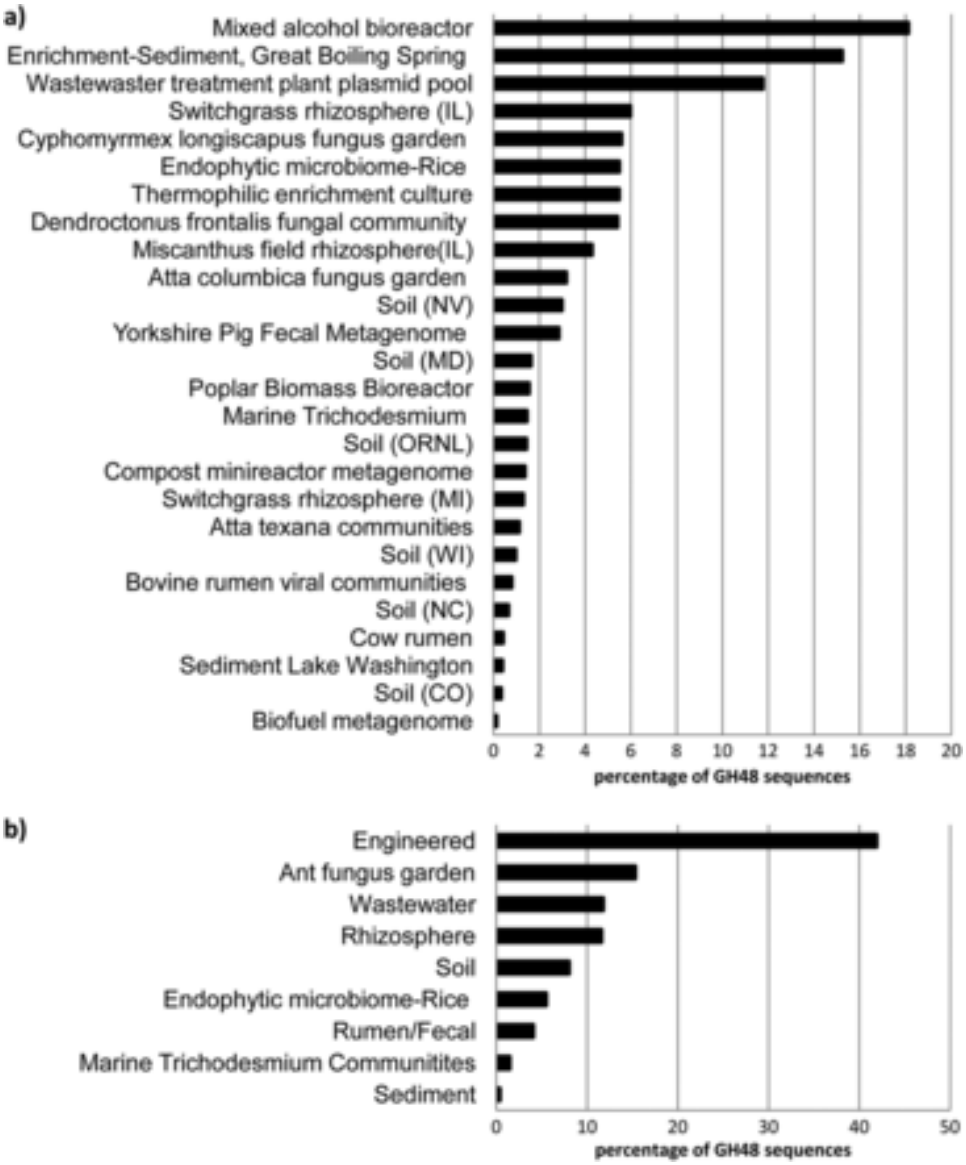

Fig. 3.5. Abundance of GH48 cellulases in metagenomes.

A, percentage of GH48 sequences in each metagenome (abundance) was calculated by dividing the number of GH48 hits by the total number of genes in each metagenome. *B*, the abundance of GH48 sequences in different habitats. The normalized percentage of GH48 genes was calculated as the percentage of GH48 sequences in a given metagenome divided by the sum of the percentage of GH48 for all metagenomes.

Table 3.1. Enrichment of GH48 genes in the prokaryotic genomes

Taxon	Total no. of genomes[a]/No. of genomes containing GH48 genes	Percentage of genomes containing GH48 genes
		%
Actinobacteria	218/38	17
Firmicutes	418/80	19
Chloroflexi	17/2	12
Proteobacteria	732/2	0.3

[a] Complete and draft genomes with a size of >1 Mb.

CONCLUSION

In this dissertation, protein domains were studied using HMM-based methods at three different scales. In case of cellulases, in chapter III, the efforts were directed towards gaining a molecular level understanding of domains of a small family of cellulases. For Cache domains in chapter I, the questions were addressed for a much larger superfamily of domains. Finally, in chapter II, the scope of the problem was at the level of all prokaryotic signal transduction systems.

In Chapter I, use of HMM models showed the ambiguity in sequence based Cache domains and the structure based PDC/PAS-like domains. This information aided in building more accurate and sensitive domain models which enabled identification of large number of new members which ultimately led to the discovery that Cache domains constitute the largest extracellular sensory domain family in prokaryotic signal transduction systems. HMM-HMM comparisons also revealed remote homology between Cache domains and structurally related PAS domains.

In Chapter II, use of the highly sensitive HMM-HMM comparison approach enabled annotation of a large fraction of previously unannotated sequences. It also revealed that three-fourths of all extracellular sensory domains belong to the Cache clan or the 4HB_MCP clan suggesting these folds are specialized for their roles as extracellular sensors.

In Chapter III, HMM model enabled retrieval of GH48 sequences from protein sequence database. Sequence and structure analysis revealed that an ω-loop was missing in chitinases but present in all cellulases. Using HMM based alignment it was possible to

align even short GH48 sequences obtained from metagenomic datasets. The ability to align these sequences to full length GH48 sequences allowed comparison of cellulase-specific features and prediction of cellulolytic activity of the sequences retrieved from metagenomic datasets.

Thus, overall, HMM-based methods proved crucial for analyzing domain families that showed very high sequence divergence as well as functional prediction of very short metagenomic sequences. The work here further emphasizes the utility of HMMs for studying protein evolution and function.

Future Directions

Although the new Cache models were able to identify a large number of new members, the models can still be improved by incorporating domains that are not fully covered. We have a much better understanding of Cache domains and their relationship to PAS and PDC domains. Future efforts can now focus on understanding how the families in the Cache clan differ from each other. Many of the Cache domains have been experimentally characterized with known ligands. Determining the molecular basis of ligand recognition in this highly diverse group of domains will be extremely useful for functional prediction in other members of this dominant family of prokaryotic extracellular sensory domains.

We were able to vastly reduce the unknown space in extracellular sensory domains in prokaryotic signal transduction systems. The domains that were not annotated in this study can be potential targets for structural genomics initiatives. These may also be used to define new domain families and create models for their identification. An interesting feature of sensory domains is the duplication of a domain fold such as double Cache and

double four-helical bundles which raises the questions - How does the signal transduction differ between single and double domains?.

The approach used in Chapter III can be considered as a framework for analyzing protein domains which may have wide applicability to other domain families including other families of cellulases that have experimentally characterized members.

VITA

Amit Upadhyay was born in Mumbai, India. He obtained his Bachelor's and Master's in Microbiology from University of Mumbai, India in 2006 and 2008 respectively. He then moved to USA and obtained a Master's in Bioinformatics from Northeastern University, Boston in 2010. He was accepted in the Genome Science and Technology Program at the University of Tennessee, Knoxville in 2010 and joined the laboratory of Dr. Igor Jouline. After obtaining his Ph.D., he would like to pursue a career in the area of precision medicine.