TOWARD STRUCTURAL AND FUNCTIONAL PREDICTIONS

FROM BIOLOGICAL SEQUENCES

APPROVED BY SUPERVISORY COMMITTEE

Nick V. Grishin, Ph.D.: Advisor

Zbyszek Otwinowski, Ph.D.: Committee Chair

Philip J. Thomas, Ph.D.

Daniel Rosenbaum, Ph.D.

DEDICATION

To those who care

TOWARD STRUCTURAL AND FUNCTIONAL PREDICTIONS FROM BIOLOGICAL SEQUENCES

by

WENLIN LI

DISSERTATION

Presented to the Faculty of the Graduate School of Biomedical Sciences

The University of Texas Southwestern Medical Center at Dallas

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

The University of Texas Southwestern Medical Center at Dallas

Dallas, Texas

May, 2018

TOWARD STRUCTURAL AND FUNCTIONAL PREDICTIONS

FROM BIOLOGICAL SEQUENCES

Publication No. _____

Wenlin Li, Ph.D.

The University of Texas Southwestern Medical Center at Dallas, 2018

Supervising Professor: Nick V. Grishin, Ph.D.

Biological sequences, including DNA and protein sequences, are believed to encode sufficient information to determine the structure and function of biological molecules, which in turn decide the phenotypic traits of animals. Deciphering the biological sequences is an important and multiscale problem that connecting the information flow from genotypes to phenotypes. Current advances in next-generation sequence technology provided tons of sequencing data, demanding innovations in computational algorithm for better interpretation. I developed computational methodologies to understand the biological sequences in various

levels. In the primary sequence level, I analyzed the evolutionary information encoded in protein families and predicted the function (and active sites) of the proteins. To aid my sequence analysis, I developed a set of computational methodologies and deployed them as public web-servers. In the protein structure level, I studied the plasticity of the 3D structures, as well as demonstrated its effect on the uncertainty of computational scoring algorithms. In the organism level, I innovated the computational methodology to assemble and analyze complete genomes of butterflies and discovered convergence evolution in butterfly wing patterns. In conclusion, I advanced the knowledge of biological sequences in multi-layers by computational approaches.

ACKNOWLEDGEMENTS

I am so grateful to have Dr. Nick Grishin as my mentor. Nick showed very strong support, which is far more than what I can imagine, for both my research and my personal development. He encouraged and inspired me to become a better man. He was always so patient and tolerant when I made mistakes, and walked me through those dark days. I felt so lucky to be able to do researches based on my personal interests and participated in multiple projects. He is more like the sincerest friend who always be there to have my back. I also thank my committee members, Drs. Phillip Thomas, Denial Rosenbaum, and Zbyszek Otwinowski for kind supports. You guys are so good that I always received more than what I expected. I also thank Dr. Dominika Borek for her patience to explain the naïve problems I asked in great details.

Great thanks to everyone in the Grishin lab, particularly Lisa Kinch, Jimin Pei, Qian Cong, Jeremy Semeiks, Bong-Hyun Kim, Jinhui Shen, Jing Zhang and Ming Tang, for their friendship, collaboration, help with many questions, and being a wonderful team of colleagues. I especially acknowledged Dr. Qian Cong, who is the model of successful graduate student I am running after and the closest colleague helping me a lot, Drs. Lisa Kinch and Jimin Pei, who selflessly offer me help in both scientific topics and everyday life. Thanks to everybody in UT Southwestern Medical Center, and particularly, people located in ND10, for maintaining such an open, friendly, and inspiring working environment.

Finally, I am grateful to my parents, Zongchi Li and ShuiLi Li, for their constant support and complete trust on my choices in life. Special thanks to my wife, Dr. Fang Zhang, who listens to my words in my sleepless night and embraces me with warm arms.

TABLE OF CONTENTS

TABLE OF CONTENTS viii

PRIOR PUBLICATIONS 1

CHAPTER 1 GENERAL INTRODUCTION 3

CHAPTER 2 SEQ2REF: A WEB SERVER TO FACILITATE FUNCTIONAL INTERPRETATION 11

CHAPTER 3 PCLUST: PROTEIN NETWORK VISUALIZATION HIGHLIGHTING EXPERIMENTAL DATA 29

CHAPTER 4 THE ABC TRANSPORTERS IN CANDIDATUS LIBERIBACTER ASIATICUS 36

CHAPTER 5 CONSERVED EVOLUTIONARY UNITS IN THE HEME - COPPER OXIDASE SUPERFAMILY REVEALED BY NOVEL HOMOLOGOUS PROTEIN FAMILIES 82

CHAPTER 6 ESTIMATION OF UNCERTAINTIES IN THE GLOBAL DISTANCE TEST (GDT_TS) FOR CASP MODELS 124

CHAPTER 7 CHSEQ: A DATABASE OF CHAMELEON SEQUENCES 152

CHAPTER 8 ASSESSMENT OF CASP11 CONTACT - ASSISTED PREDICTIONS 188

CHAPTER 9 GENOMES OF 250 SKIPPER BUTTERFLIES REVEAL RAMPANT CONVERGENCE IN WING PATTERNS 234

PRIOR PUBLICATIONS

1. **Li W**, Cong Q, Pei J, Kinch LN, Grishin N V. The ABC transporters in Candidatus Liberibacter asiaticus. Proteins Struct Funct Bioinforma 2012;80(11):2614–2628.
2. Cong Q, Kinch LN, Pei J, Shi S, Grishin VN, **Li W**, Grishin N V. An automatic method for CASP9 free modeling structure prediction assessment. Bioinformatics 2011;27(24):3371–3378.
3. **Li W**, Kinch LN, Grishin N V. Pclust: protein network visualization highlighting experimental data. Bioinformatics 2013;29(20):2647–2648.
4. **Li W**, Kinch LN, Karplus PA, Grishin N V. ChSeq: a database of chameleon sequences. Protein Sci 2015;24(7):1075–1086.
5. Pei J*, **Li W***, Kinch LN, Grishin N V. Conserved evolutionary units in the heme-copper oxidase superfamily revealed by novel homologous protein families. Protein Sci 2014;23(9):1220–1234.
6. **Li W**, Schaeffer RD, Otwinowski Z, Grishin N V. Estimation of uncertainties in the global distance test (GDT_TS) for CASP Models. PLoS One 2016;11(5):e0154786.
7. Ji R, Cong Q, **Li W**, Grishin N V. M2SG: mapping human disease-related genetic variants to protein sequences and genomic loci. Bioinformatics 2013;29(22):2953–2954.
8. Kinch LN, **Li W**, Monastyrskyy B, Kryshtafovych A, Grishin N V. Evaluation of free modeling targets in CASP11 and ROLL. Proteins Struct Funct Bioinforma 2016;84(S1):51–66.

9. Cong Q, Shen J, **Li W**, Borek D, Otwinowski Z, Grishin N V. The first complete genomes of Metalmarks and the classification of butterfly families. Genomics 2017.

10. Kinch LN*, **Li W***, Monastyrskyy B, Kryshtafovych A, Grishin N V. Assessment of CASP11 Contact-Assisted Predictions. Proteins 2016.

11. Kinch LN, **Li W**, Schaeffer RD, Dunbrack RL, Monastyrskyy B, Kryshtafovych A, Grishin N V. CASP 11 Target Classification. Proteins 2016.

12. **Li W**, Cong Q, Kinch LN, Grishin N V. Seq2Ref: a web server to facilitate functional interpretation. BMC Bioinformatics 2013;14(1):30.

* Authors contributed equally

CHAPTER 1

GENERAL INTRODUCTION

Biological sequences, including DNA and protein sequences, are believed to encode all the information needed to determine the phenotype of an organism. Understanding what nature says in such biological words and how the words define the molecular function has drawn board attention in biological sciences. Recent advances in the next-generation sequencing technology generated avalanche of sequencing data and motivated biologists to use computational methods to decipher the sequence information. Achievements have been made in different levels, including predicting protein tertiary structures, predicting the functional site of a protein, and understanding the gene determinants for morphological traits, but yet far from satisfactory.

I decoded the information in the biological sequences, both protein and DNA, for structural and functional prediction using computational approaches. To better process the biological sequences, I innovated computational data mining methodologies, which were applied on protein sequences to bring medical insights from evolutionary perspectives. I also studied the conformation ambiguity in the protein structures and was invited to assess the performance of the 11st Critical Assessment of Structure Prediction experiment, the community-wide blind test to evaluate advances in structure prediction. Using the butterfly as the model system for evolution studies, I revolutionized the next-generation sequencing analysis algorithms and discovered the butterfly wing pattern divergence.

The size of the protein sequence database has been exponentially increasing due to advances in genome sequencing. However, experimentally characterized proteins only constitute a small portion of the database, such that the majority of sequences have been annotated by computational approaches. Current automatic annotation pipelines inevitably introduce errors, making the annotations unreliable. Instead of such error-prone automatic annotations, functional interpretation should rely on annotations of 'reference proteins' that have been experimentally characterized or manually curated. The Seq2Ref server uses BLAST to detect proteins homologous to a query sequence and identifies the reference proteins among them. Seq2Ref then reports publications with experimental characterizations of the identified reference proteins that might be relevant to the query. Furthermore, a plurality-based rating system is developed to evaluate the homologous relationships and rank the reference proteins by their relevance to the query. The reference proteins detected by our server will lend insight into proteins of unknown function and provide extensive information to develop in-depth understanding of uncharacterized proteins. Seq2Ref is available at: http://prodata.swmed.edu/seq2ref.

One approach to infer functions of new proteins from their homologs utilizes visualization of an all-against-all pairwise similarity network (A2ApsN) that exploits the speed of BLAST and avoids the complexity of multiple sequence alignment. However, identifying functions of the protein clusters in A2ApsN is never trivial, due to a lack of linking characterized proteins to their relevant information in current software packages. Given the database errors introduced by automatic annotation transfer, functional deduction should be made from proteins with experimental studies, i.e. 'reference

proteins'. Here, we present a web server, termed Pclust, which provides a user-friendly interface to visualize the A2ApsN, placing emphasis on such 'reference proteins' and providing access to their full information in source databases, e.g. articles in PubMed. The identification of 'reference proteins' and the ease of cross-database linkage will facilitate understanding the functions of protein clusters in the network, thus promoting interpretation of proteins of interest.

Candidatus Liberibacter asiaticus (*Ca.* L. asiaticus) is a Gram-negative bacterium and the pathogen of Citrus Greening disease (Huanglongbing, HLB). As a parasitic bacterium, *Ca*. L. asiaticus harbors ABC transporters that play important roles in exchanging chemical compounds between *Ca.* L. asiaticus and its host. Here, we analyzed all the ABC transporter-related proteins in *Ca*. L. asiaticus. We identified 14 ABC transporter systems and predicted their structures and substrate specificities. In-depth sequence and structure analysis including multiple sequence alignment, phylogenetic tree reconstruction, and structure comparison further support their function predictions. Our study shows that this bacterium could use these ABC transporters to import metabolites (amino acids and phosphates) and enzyme cofactors (choline, thiamine, iron, manganese, and zinc), resist to organic solvent, heavy metal, and lipid-like drugs, maintain the composition of the outer membrane (OM), and secrete virulence factors. Although the features of most ABC systems could be deduced from the abundant experimental data on their orthologs, we reported several novel observations within ABC system proteins. Moreover, we identified seven nontransport ABC systems that are likely involved in virulence gene expression regulation, transposon excision regulation, and

DNA repair. Our analysis reveals several candidates for further studies to understand and control the disease, including the type I virulence factor secretion system and its substrate that are likely related to *Ca*. L. asiaticus pathogenicity and the ABC transporter systems responsible for bacterial OM biosynthesis that are good drug targets.

The heme-copper oxidase (HCO) superfamily includes HCOs in aerobic respiratory chains and nitric oxide reductases (NORs) in the denitrification pathway. The HCO/NOR catalytic subunit has a core structure consisting of 12 transmembrane helices (TMHs) arranged in three-fold rotational pseudosymmetry, with six conserved histidines for heme and metal binding. Using sensitive sequence similarity searches, we detected a number of novel HCO/NOR homologs and named them HCO Homology (HCOH) proteins. Several HCOH families possess only four TMHs that exhibit the most pronounced similarity to the last four TMHs (TMHs 9–12) of HCOs/NORs. Encoded by independent genes, four-TMH HCOH proteins represent a single evolutionary unit (EU) that relates to each of the three homologous EUs of HCOs/NORs comprising TMHs 1–4, TMHs 5–8, and TMHs 9–12. Single-EU HCOH proteins could form homotrimers or heterotrimers to maintain the general structure and ligand-binding sites defined by the HCO/NOR catalytic subunit fold. The remaining HCOH families, including NnrS, have 12-TMHs and three EUs. Most three-EU HCOH proteins possess two conserved histidines and could bind a single heme. Limited experimental studies and genomic context analysis suggest that many HCOH proteins could function in the denitrification pathway and in detoxification of reactive molecules such as nitric oxide. HCO/NOR catalytic subunits exhibit remarkable structural similarity to the homotrimers of MAPEG

(membrane-associated proteins in eicosanoid and glutathione metabolism) proteins. Gene duplication, fusion, and fission likely play important roles in the evolution of HCOs/NORs and HCOH proteins.

The Critical Assessment of techniques for protein Structure Prediction (or CASP) is a community-wide blind test experiment to reveal the best accomplishments of structure modeling. Assessors have been using the Global Distance Test (GDT_TS) measure to quantify prediction performance since CASP3 in 1998. However, identifying significant score differences between close models is difficult because of the lack of uncertainty estimations for this measure. Here, we utilized the atomic fluctuations caused by structure flexibility to estimate the uncertainty of GDT_TS scores. Structures determined by nuclear magnetic resonance are deposited as ensembles of alternative conformers that reflect the structural flexibility, whereas standard X-ray refinement produces the static structure averaged over time and space for the dynamic ensembles. To recapitulate the structural heterogeneous ensemble in the crystal lattice, we performed time-averaged refinement for X-ray datasets to generate structural ensembles for our GDT_TS uncertainty analysis. Using those generated ensembles, our study demonstrates that the time-averaged refinements produced structure ensembles with better agreement with the experimental datasets than the averaged X-ray structures with B-factors. The uncertainty of the GDT_TS scores, quantified by their standard deviations (SDs), increases for scores lower than 50 and 70, with maximum SDs of 0.3 and 1.23 for X-ray and NMR structures, respectively. We also applied our procedure to the high accuracy version of GDT-based score and produced similar results with slightly higher SDs. To

facilitate score comparisons by the community, we developed a user-friendly web server that produces structure ensembles for NMR and X-ray structures and is accessible at http://prodata.swmed.edu/SEnCS. Our work helps to identify the significance of GDT_TS score differences, as well as to provide structure ensembles for estimating SDs of any scores.

Chameleon sequences (ChSeqs) refer to sequence strings of identical amino acids that can adopt different conformations in protein structures. Researchers have detected and studied ChSeqs to understand the interplay between local and global interactions in protein structure formation. The different secondary structures adopted by one ChSeq challenge sequence-based secondary structure predictors. With increasing numbers of available Protein Data Bank structures, we here identify a large set of ChSeqs ranging from 6 to 10 residues in length. The homologous ChSeqs discovered highlight the structural plasticity involved in biological function. When compared with previous studies, the set of unrelated ChSeqs found represents an about 20-fold increase in the number of detected sequences, as well as an increase in the longest ChSeq length from 8 to 10 residues. We applied secondary structure predictors on our ChSeqs and found that methods based on a sequence profile outperformed methods based on a single sequence. For the unrelated ChSeqs, the evolutionary information provided by the sequence profile typically allows successful prediction of the prevailing secondary structure adopted in each protein family. Our dataset will facilitate future studies of ChSeqs, as well as interpretations of the interplay between local and nonlocal interactions. A user-friendly web interface for this ChSeq database is available at prodata.swmed.edu/chseq.

As CASP11 assessors, we present an overview of contact-assisted predictions in the eleventh round of critical assessment of protein structure prediction (CASP11), which included four categories: predicted contacts (Tp), correct contacts (Tc), simulated sparse NMR contacts (Ts), and cross-linking contacts (Tx). Comparison of assisted to unassisted model quality highlighted a relatively poor overall performance in CASP11 using predicted Tp and crosslinked Tx contact information. However, average model quality significantly improved in the correct Tc and simulated NMR Ts categories for most targets, where maximum improvement of unassisted models reached an impressive 70 GDT_TS. Comparison of the performance in the correct Tc category to CASP10 suggested the improvement in CASP11 model quality originated from an increased number of provided contacts per target. Group rankings based on a combination of scores used in the CASP11 free modeling (FM) assessment for each category highlight four top-performing groups, with three from the Lee lab and one from the Baker lab. We used the overall performance of these groups in each category to develop hypotheses for their relative outperformance in the correct Tc and simulated NMR Ts categories, which stemmed from the fraction of correct contacts provided (correct Tc category) and a reduced fraction of correct contacts offset by an increased coverage of the correct contacts (simulated NMR Ts category).

For centuries, biologists relied on phenotypes to reason about evolution. For decades, a handful of gene markers gave us a glimpse at the genotype to add to phenotypic traits. Today, we can sequence entire genomes of hundreds of animals to gain the ultimate knowledge of their biology. Choosing a family of Skipper butterflies

(Hesperiidae) as an example, we show the power of genomics to learn about their phylogeny and evolution. Genomes of 250 Hesperiidae species from all major phylogenetic lineages focusing on the subfamily Eudaminae reveal rampant inconsistencies between their current classification and genome-based phylogeny. We use timed genomic tree to define tribes (5 new) and subtribes (7 new), overhaul genera (9 new) and subgenera (3 new), and study convergence in wing patterns that fooled researchers (and birds) for decades (or millennia). We find that many skippers with similar looks are distantly related and should belong to different genera. Conversely, we see that several skippers with distinct morphology are close relatives and should be grouped within one genus. These conclusions are strongly and invariably supported by different genomic regions, both nuclear and mitochondrial, coding genes and non-coding segments, and are consistent with some morphological traits. Similar to our study, genomic biology will soon revolutionize biodiversity research.

CHAPTER 2 SEQ2REF: A WEB SERVER TO FACILITATE FUNCTIONAL INTERPRETATION[1]

INTRODUCTION

Due to the avalanche of protein sequences made available by high-throughput genome sequencing, complete manual annotation is unfeasible, leaving a large fraction of protein functions to be predicted by automatic functional annotation pipelines [1]. However, without experimental characterization, the quality of annotation is often questionable, owing to errors in automatic annotation transfer and lack of updates from new findings. In spite of recent advances in highly integrative functional prediction methods [2], a recent investigation [3] into the annotation quality of well-characterized enzyme families revealed that the average percentage of misannotation for the haloacid dehalogenase (HAD) superfamily in the three largest public databases, i.e. non-redundant (nr) [4], TrEMBL [5] and KEGG [6], is over 60%. The possible causes of such annotation errors include multi-domain problems [7], experimental data misinterpretations, threshold relativity problems, and paralog-ortholog misclassifications [8-12]. Moreover, the simplified descriptions recorded in protein sequence and protein family databases are usually inadequate for understanding the precise function of a protein [1].

---

[1] This chapter was published as:
Li W, Cong Q, Kinch LN, Grishin N V. *Seq2Ref: a web server to facilitate functional interpretation*. BMC Bioinformatics 2013;14(1):30.

Such errors and omissions make database annotations insufficient for complete functional interpretation of a protein. A more accurate source of annotations is the 'reference proteins' closely related to the protein of interest. We define 'reference, proteins' as proteins that have been experimentally studied, manually curated, and reported in the literature. Information about reference proteins is essential for accurate functional interpretation and experimental design. The cross-links between proteins, genes, and associated literature available from National Center for Biotechnology Information (NCBI) provide a basis for reference protein identification. However, it is not trivial to identify a good set of reference proteins and supporting literature because such reference proteins constitute only a small portion of protein databases. Additionally, many proteins linked to large-scale studies (such as genome sequencing) do not provide sufficient functional information.

We have developed a web server named Seq2Ref to assist the identification of applicable reference proteins. Seq2Ref employs BLAST [13] to perform homology searches and exploits crosslinks created by NCBI between proteins and literature to detect reference proteins. Homologs from the Protein Data Bank (PDB) [14] and Swiss-Prot (SP) [15] databases are detected as well, as these databases contain experimental data on 3D protein structures and manually curated annotation on sequence records, respectively. Moreover, we developed a plurality-based rating system integrating reciprocal BLAST and Multiple Sequence Comparisons (MSC) to rank the reference proteins. By retrieving homologous reference proteins, Seq2Ref can contribute to

precisely inferring unknown protein function and developing detailed functional interpretation.

RESULTS AND DISCUSSION

Server interface

The input and output interfaces are shown in Figure 1. An email address and the query protein are the minimal requirements to initiate a job. Options for BLAST search parameters and selection of server modes (fast/slow) are available in the PARAMETERS panel. We recommend manually selecting the organism of the input sequence for reciprocal BLAST if the input sequence is not in the nr database. The total run time is usually 5 to 15 min for fast mode and 1 hour or more for slow mode. When the job completes, an email notification will be sent to the address provided by the user.

The results page (shown in Figure 1) lists the reference proteins and relevant information in a ranked order. Reference proteins from three sources are shown, respectively, as: (1) a summary table containing protein definition, rating score and BLAST statistics (expectation value, sequence identity and coverage); (2) and a detailed description panel with the rating records, BLAST statistics and scores, and relevant database information. Reference proteins are ranked first by the rating score and second by the expectation value; the publications associated with each protein are sorted by the publication date. As functional studies of remote homologs may not be applicable to the

query protein, by default we do not display reference proteins with rating scores lower than 3 in the detailed description panel.

Benchmark

To assess the performance of the Seq2Ref server, especially the ability of our plurality-based rating system to sort out the most relevant references, we applied our algorithm to the enolase superfamily, which has been thoroughly characterized and recorded in the Structure-Function Linkage Database (SFLD) [16-18]. The enolase superfamily contains seven subgroups, which are further divided into 20 families. Proteins within one family share the same substrate specificity and can be considered orthologs; proteins within one subgroup share the same general base(s) in the active site and have the similar catalytic mechanism [19]. For each family, we selected one representative sequence, usually the one with an available 3D structure, as the input for benchmark.

At each rating score cutoff, coverage and average accuracy were used as parameters to evaluate the performance of Seq2Ref (Table 1). The coverage is defined as the percentage of tested sequences that detect reference proteins above the score cutoff. The accuracy is defined as the average of the true positive rates among tested sequences above the score cutoff. Two criteria were used to define true positives: (1) in a stringent (family) context, a true positive hit must be from the same family as the query; and (2) in a broader (subgroup) context, hits from the same subgroup but from different families are

also considered true. As shown in Table 1, the accuracy is always 100% with score cutoffs no less than 4; when the cutoff drops to 3, Seq2Ref reaches 100% coverage but starts to include those hits from the same subgroup but different families. Although the biased dataset from only one family might cause overfitting of the statistics, the benchmark suggests that accurate functional interpretation at family level should be achieved by utilizing the reference proteins with a score no less than 4. The information for marginal hits with scores between 3 and 4 is valuable to understand the broad function of the protein subgroup. However, one should not directly transfer the specific functions of marginal hits, such as substrate specificity, to the query.

Case study and examples

Due to its ability to retrieve reference proteins and their relevant information in a ranked order, the Seq2Ref server is useful for finding PubMed references relevant to proteins of unknown function, as well as obtaining a deeper understanding of proteins than that revealed by short annotations, as illustrated by the following examples.

Organizing new information for proteins of unknown function

Hypothetical proteins of unknown function constitute a remarkably large portion of the database [20]. Novel studies on uncharacterized proteins and their orthologs provide new insights about their functions, but sequence databases often do not

incorporate this information in a timely manner. By retrieving literature, Seq2Ref helps to obtain the most recent information about proteins.

The *Macaca mulatta* protein, corresponding to gi|355567738, is annotated as a hypothetical protein EGK_07670 in the NCBI Protein database (Seq2Ref results: http://prodata.swmed.edu/wenlin/server/user_data/seq2ref/S2Rnv4cun/result.html). This hypothetical protein contains three conserved domains of unknown function (two DUF3730 and one DUF3028). Our server detects one close homolog, a hypothetical protein (gi|23345097) in human, which has been experimentally studied. The highly confident statistics in BLAST (e-value around 0; 98% identity, 100% coverage) and similar protein domain composition support an orthologous relationship between these two proteins. The human protein was recently (in 2012) reported to be a tumor suppressor in gliomas. It was named 'focadhesin', due to its cellular localization at the focal adhesion of the cell membrane [21]. As a likely ortholog, the *M. mulatta* protein might also be a tumor suppressor and localized at the focal adhesion. Thus, by finding a homolog with the latest experimental publication not yet incorporated in sequence databases, Seq2Ref can serve as a basis for reliable functional prediction of unknown proteins.

Providing detailed information about a protein's function

Although conserved domains in proteins usually suggest their functions, overly broad descriptions of domain functions are less informative than more specific

descriptions. By presenting reference proteins and associated literature, Seq2Ref can offer more definitive and reliable information about protein functions.

One example is the hlyA gene product in *Cronobacter turicensis z3032* (gi|260595828, Seq2Ref results: http://prodata.swmed.edu/wenlin/server/user_data/seq2ref/S2RGsaCMG/result.html). A search of the Conserved Domain Database (CDD) merely suggests that this protein contains a 'haemolytic domain', with the most similar hit (lowest expectation value) annotated as a 'hypothetical protein' and one possible informative hit as 'conserved hypothetical protein YidD'. The 'conserved hypothetical protein YidD' domain (TIGR00278) shows neither functional studies nor a detailed functional description. The publication [22] associated with a Pfam domain record (pfam01809) in the CDD search result suggests that the name 'haemolytic domain' originated because one protein (ytjA from *Bacillus subtilis*) containing this domain can cause cells to lyse in culture. Unfortunately, this study failed to suggest a specific molecular function. Seq2Ref provided more information by detecting (e-value 8.0e-49; 90% identity; reciprocal best hit) the experimentally studied protein YidD from *E. coli* (gi|67476547), which is identical to the NCBI nr database representative protein (gi|16767126) from *Salmonella enterica*. This orthology is reinforced by the common conserved genomic context [23] and the CLANS [24] protein similarity network, in which *E. coli* YidD and *C. turicensis* hlyA cluster tightly together among their homologs. The reference [23] associated with *E. coli* YidD detected by our server suggests that YidD assists YidC, the protein insertase, in insertion of inner membrane proteins. As a confident ortholog of *E.*

coli YidD, the hlyA gene from *C. turicensis* very likely shares the same function. Thus, the homologous reference protein, detected by Seq2Ref, contributes to understanding the protein function more specifically.

Limitations

As shown in the examples, the Seq2Ref server detects reference proteins, which can facilitate deeper understanding of the protein function. However, we should keep in mind the limitations. The main concern regards the quality of cross-links between the NCBI Protein and PubMed databases. Missing or wrong links defined by NCBI would result in the loss of or the inappropriate assignment of relevant literature. Another concern is that although the top ranked reference proteins are very likely functionally similar to the query proteins, one should still be careful in directly transferring the information from the hit to the query, as verification of orthology requires additional diligent analysis. To come to the best conclusions about a protein's function, one should critically inspect the relevance of the publications and the homology of the reference proteins to the query.

CONCLUSIONS

Seq2Ref is a homology-based tool to identify reference proteins from PubMed, PDB and SP databases. We have developed a plurality-based rating system that evaluates

homologous relationships to indicate the degree of confidence one should have in transferring annotations from a well-studied reference protein to a similar new protein. Thus, by retrieving both experimental studies and high-quality functional annotations of reference proteins, our server provides a solid basis for correct function interpretation of novel proteins.

METHODS

Detection of homologs and identification of reference proteins

Seq2Ref performs the BLAST search against the NCBI nr database to detect homologs of the query protein. Based on BLAST search results, reference proteins are identified as: (1) the hits linked to PubMed literature by NCBI (those publications associated with more than 100 protein records are excluded); (2) the hits from PDB; (3) and the hits from SP. Reference proteins from PDB and SP databases are obtained by parsing the protein descriptions recorded in nr. Retrieval from PubMed requires fast but thorough searching of cross-links between NCBI databases. To implement this search, Seq2Ref has two modes: a “fast mode” based on searching a pre-processed local database (updated every 6 months) that consists of the reference proteins in nr, and a “slow mode” in which the most updated reference proteins are retrieved in real-time via NCBI Entrez [25].

Analysis of homologous relationship

We assign orthology firstly by the approximate method of reciprocal best hits [26]. In this method, it is necessary to know the the source-organism of the query protein. To automatically detect the species, Seq2Ref identifies the taxon of the first BLAST hit with at least 97% identity and 90% coverage. Alternatively, the user can manually specify the organism of the input sequence. To avoid possible false negatives caused by variants of the same gene in reciprocal BLAST, such as alleles containing a single nucleotide polymorphism, we pre-cluster the proteins from each genome using CD-HIT [27] (identity cutoff: 97%; coverage cutoff: 90%).

Reference proteins are further analyzed by the method of multiple sequence comparison (MSC) shown in Figure 2. Specifically, we retrieve the sequences most closely related to the query, and then compare the reference proteins to those closely related sequences. Such multiple comparisons allow us to obtain more robust statistics in evaluating homology compared to simple pairwise comparison.

Rank reference proteins by relevance to the query

We developed a plurality-based rating system with scores ranging from 1 to 6, with 6 indicating the most relevant hits (shown as Table 2). To note, our rating system aims to provide intuitive indicators for the level of similarity, but not act as a statistical predictor of functionality. Four features are considered: reciprocal BLAST, MSC,

pairwise comparison between the query and the hit, and whether the hit protein is a "reference protein", e.g. if there are PubMed citations linked to the protein in the current version of NCBI databases. The maximal rating score for each aspect is 2, 1.5, 1.5 and 1, respectively. A higher total rating score indicates the query protein is closer to the hit and is more likely to function similarly. Proteins with scores lower than 3 would be considered more distant homologs whose functions may have diverged, because they are neither reciprocal BLAST best hits nor with confident statistics in MSC and pairwise comparison.

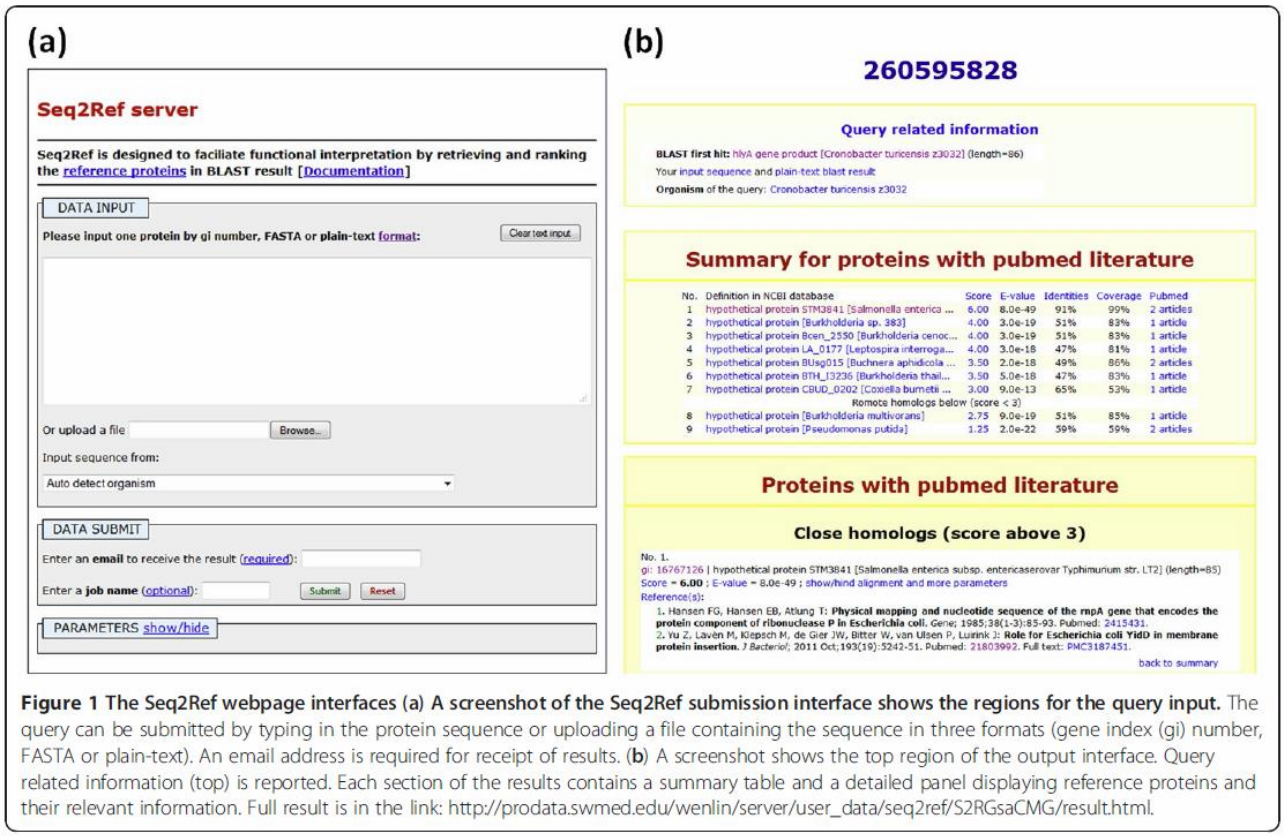


Figure 1 The Seq2Ref webpage interfaces (a) A screenshot of the Seq2Ref submission interface shows the regions for the query input. The query can be submitted by typing in the protein sequence or uploading a file containing the sequence in three formats (gene index (gi) number, FASTA or plain-text). An email address is required for receipt of results. (**b**) A screenshot shows the top region of the output interface. Query related information (top) is reported. Each section of the results contains a summary table and a detailed panel displaying reference proteins and their relevant information. Full result is in the link: http://prodata.swmed.edu/wenlin/server/user_data/seq2ref/S2RGsaCMG/result.html.

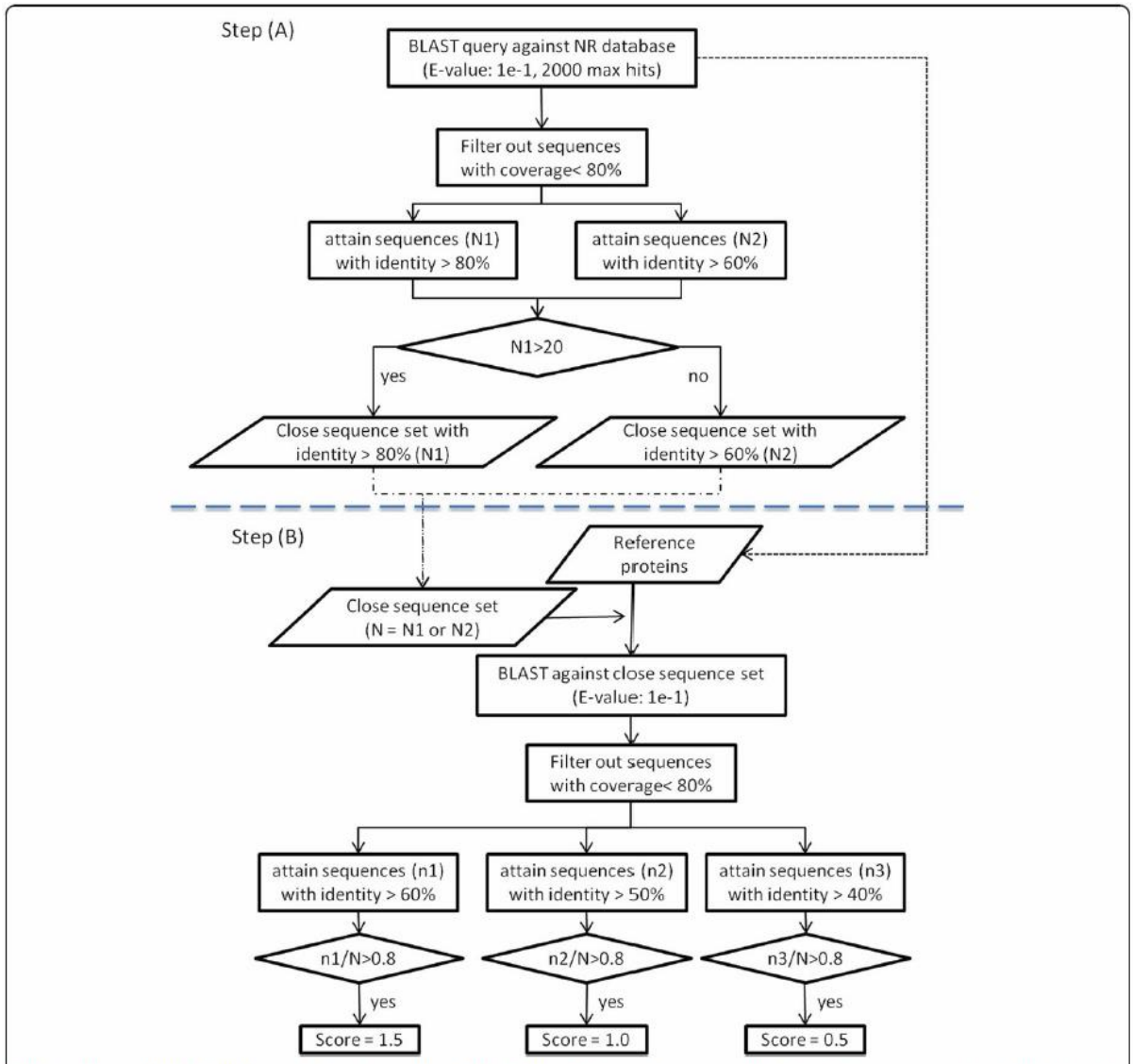


Figure 2 Flow chart of multiple sequence comparison (MSC) method. In step (**A**), MSC uses BLAST with filters as indicated to define a group of closely related sequences to the query (close sequence set, N); in Step (**B**), the Reference protein set consists of protein sequences identified without filters by BLAST in step A that are 1) from the Swiss-Prot database, 2) from the Protein Data Bank (PDB), and 3) linked to PubMed articles in the NCBI. Reference protein sequences are then used as queries against a database consisting of the close sequence set. Hits are filtered as indicated. If the filtered hit ratio is larger than 0.8, then a score is assigned to the reference protein.

Table 1 The accuracy and coverage of the rating system

Score	Subgroup accuracy[1]	Family accuracy[1]	Coverage[2]
6	100%	100%	80%
>=5	100%	100%	85%
>=4	100%	100%	95%
>=3	100%	78.5%	100%

1: accuracy calculated by averaging the family/subgroup true positive rates.
2: coverage calculated by taking the ratio of testing sequences that detect reference proteins above the score cutoff.

Table 2 The rating system

Feature	Criterion	Points		
		True	False	NA
Reciprocal BLAST	Query-to-hit-genome[1] best hit	+1	0	+0.25[3]
	Hit-to-query-genome[2] best hit	+1	0	+0.25[3]
Multiple Sequence Comparison (MSC) Method[4]	Accept with identity cutoff 60%	+0.5	0	\
	Accept with identity cutoff 50%	+0.5	0	\
	Accept with identity cutoff 40%	+0.5	0	\
Pairwise comparison to the query[4]	Identity>60%	+0.5	0	\
	Identity>50%	+0.5	0	\
	Identity>40%	+0.5	0	\
others	Reference proteins	+1	0	\

1: use the input sequence to BLAST against the genome of the hit.
2: use the hit to BLAST against the genome of the input sequence.
3: the whole genome sequences of the protein are unavailable.
4: coverage in both the input sequence and the hits must larger than 80%.

REFERENCES

1. Valencia A: **Automatic annotation of protein function**. *Current opinion in structural biology* 2005, **15**(3):267-274.
2. Rentzsch R, Orengo CA: **Protein function prediction--the power of multiplicity**. *Trends in biotechnology* 2009, **27**(4):210-219.
3. Schnoes AM, Brown SD, Dodevski I, Babbitt PC: **Annotation error in public databases: misannotation of molecular function in enzyme superfamilies**. *PLoS computational biology* 2009, **5**(12):e1000605.
4. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW: **GenBank**. *Nucleic acids research* 2010, **38**(Database issue):D46-51.
5. **The Universal Protein Resource (UniProt) in 2010**. *Nucleic acids research* 2010, **38**(Database issue):D142-148.
6. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T *et al*: **KEGG for linking genomes to life and the environment**. *Nucleic acids research* 2008, **36**(Database issue):D480-484.
7. Kim BH, Cong Q, Grishin NV: **HangOut: generating clean PSI-BLAST profiles for domains with long insertions**. *Bioinformatics* 2010, **26**(12):1564-1565.

8. Galperin MY, Koonin EV: **Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption**. *In silico biology* 1998, **1**(1):55-67.

9. Sasson O, Kaplan N, Linial M: **Functional annotation prediction: all for one and one for all**. *Protein science : a publication of the Protein Society* 2006, **15**(6):1557-1562.

10. Bork P, Bairoch A: **Go hunting in sequence databases but watch out for the traps**. *Trends in genetics : TIG* 1996, **12**(10):425-427.

11. Doerks T, Bairoch A, Bork P: **Protein annotation: detective work for function prediction**. *Trends in genetics : TIG* 1998, **14**(6):248-250.

12. Smith TF, Zhang X: **The challenges of genome sequence annotation or "the devil is in the details"**. *Nature biotechnology* 1997, **15**(12):1222-1223.

13. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic acids research* 1997, **25**(17):3389-3402.

14. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank**. *Nucleic acids research* 2000, **28**(1):235-242.

15. Bairoch A, Boeckmann B, Ferro S, Gasteiger E: **Swiss-Prot: juggling between evolution and stability**. *Briefings in bioinformatics* 2004, **5**(1):39-55.

16. Brown SD, Gerlt JA, Seffernick JL, Babbitt PC: **A gold standard set of mechanistically diverse enzyme superfamilies**. *Genome biology* 2006, **7**(1):R8.

17. Pegg SC, Brown S, Ojha S, Huang CC, Ferrin TE, Babbitt PC: **Representing structure-function relationships in mechanistically diverse enzyme superfamilies**. *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing* 2005:358-369.

18. Pegg SC, Brown SD, Ojha S, Seffernick J, Meng EC, Morris JH, Chang PJ, Huang CC, Ferrin TE, Babbitt PC: **Leveraging enzyme structure-function relationships for functional inference and experimental design: the structure-function linkage database**. *Biochemistry* 2006, **45**(8):2545-2555.

19. Babbitt PC, Hasson MS, Wedekind JE, Palmer DR, Barrett WC, Reed GH, Rayment I, Ringe D, Kenyon GL, Gerlt JA: **The enolase superfamily: a general strategy for enzyme-catalyzed abstraction of the alpha-protons of carboxylic acids**. *Biochemistry* 1996, **35**(51):16489-16501.

20. Jaroszewski L, Li Z, Krishna SS, Bakolitsa C, Wooley J, Deacon AM, Wilson IA, Godzik A: **Exploration of uncharted regions of the protein universe**. *PLoS biology* 2009, **7**(9):e1000205.

21. Brockschmidt A, Trost D, Peterziel H, Zimmermann K, Ehrler M, Grassmann H, Pfenning PN, Waha A, Wohlleber D, Brockschmidt FF *et al*: **KIAA1797/FOCAD encodes a novel focal adhesion protein with tumour suppressor function in gliomas**. *Brain : a journal of neurology* 2012, **135**(Pt 4):1027-1041.

22. Liu J, Fang C, Jiang Y, Yan R: **Characterization of a hemolysin gene ytjA from Bacillus subtilis**. *Current microbiology* 2009, **58**(6):642-647.

23. Yu Z, Laven M, Klepsch M, de Gier JW, Bitter W, van Ulsen P, Luirink J: **Role for Escherichia coli YidD in membrane protein insertion**. *Journal of bacteriology* 2011, **193**(19):5242-5251.

24. Frickey T, Lupas A: **CLANS: a Java application for visualizing protein families based on pairwise similarity**. *Bioinformatics* 2004, **20**(18):3702-3704.

25. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Federhen S *et al*: **Database resources of the National Center for Biotechnology Information**. *Nucleic acids research* 2012, **40**(Database issue):D13-25.

26. Moreno-Hagelsieb G, Latimer K: **Choosing BLAST options for better detection of orthologs as reciprocal best hits**. *Bioinformatics* 2008, **24**(3):319-324.

27. Li W, Godzik A: **Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences**. *Bioinformatics* 2006, **22**(13):1658-1659.

CHAPTER 3 PCLUST: PROTEIN NETWORK VISUALIZATION HIGHLIGHTING EXPERIMENTAL DATA[2]

INTRODUCTION

A common practice to formulate hypotheses for a protein of unknown function includes searching for annotations among homologous proteins. Although sequence similarity does not necessarily correlate with functional similarity (Clark and Radivojac, 2011), all-against-all pairwise similarity network (A2ApsN) works best to illustrate functional relationships among large numbers of proteins (Atkinson *et al.*, 2009), and meanwhile avoids computational complexity and problems of aligning non-homologous sequences (Frickey and Lupas, 2004). Software packages, such as CLANS (Frickey and Lupas, 2004), Pythoscape (Barber and Babbitt, 2012) and Cytoscape (Shannon *et al.*, 2003), provide powerful repositories to manage the A2ApsN. However, they require either programming basics or expertise in program setups to generate the network. Numerous efforts aim to visualize the protein–protein interaction (PPI) network (Agapito *et al.*, 2013). But these packages build the network by PPI data (not by sequence similarity) and assign functions by analysing the network structure, such as dissecting functional modules (Sharan *et al.*, 2007). Given the high misannotation rate in current databases (Schnoes *et al.*, 2009), the simple protein descriptions that current packages

[2] This chapter was published as:
Li W, Kinch LN, Grishin N V. *Pclust: protein network visualization highlighting experimental data*. Bioinformatics 2013;29(20):2647–2648.

offer are somewhat suspect, which hinders the understanding of the protein clusters. Thus, to avoid working with a network of uncertainty, one has to tediously verify the functions of nodes in the network before getting into interesting biology.

Here, we developed a web server named Pclust for visualization of the A2ApsN, which emphasizes those 'reference proteins' with experimental studies. Pclust works with the Seq2Ref server (Li *et al.*, 2013) to identify the 'reference proteins' and highlight them in the network. The web interface bypasses the pain of software installation and the requirement of programming expertise. The highlighted 'reference proteins' and easy access to their functional studies simplify the process of relating functions to protein clusters, thus facilitating hypothesis driven research of proteins of interest.

RESULTS

Pclust has four modes; a user can specify the input as (i) a single sequence; (ii) multiple sequences; (iii) a Seq2Ref result link; and (iv) customized network data. The four above modes are designed to provide A2ApsNs (i) for any single sequence; (ii) with an advanced interface similar to CLANS; (iii) for previously generated Seq2Ref jobs; and (iv) with a customized input that grants users the flexibility to design. An email is required to keep track of the job submission. Once the network is ready, an email containing the result link will be sent to the provided address.

Figure 1 (current E-value cutoff: 2e-38) shows the interface for an A2ApsN, as well as an example where the input protein (brown node), annotated as 'mandelate

racemase' (gi|17987990), should be a 'fuconate dehydratase', as previously described (Schnoes *et al.*, 2009). Merely reading the brief protein descriptions within the cluster, such as 'RTS beta protein', 'mandelate racemase' and 'enolase superfamily member' (as in the CLANS interface), results in confusion about the function of the cluster. Pclust alleviates this confusion by highlighting the reference proteins and referring to their annotation sources linked by our server. For example, the cluster circled in Figure 1 containing the questionable 'mandelate racemase' (brown) includes two solved crystal structures of known fuconate dehydratases with provided links to their experimental data. Thus, with the convenience of locating reference proteins in A2ApsN and accessing their database links, more accurate hypotheses about the function of protein queries can be generated, potentiating biological discovery.

METHODS AND IMPLEMENTATION

Preparation of the protein network

According to the type of user inputs, the protein sets shown in the A2ApsN are taken from (i) Seq2Ref BLAST results; (ii) user input sequences; or (iii) user customized data (e.g. http://prodata.swmed.edu/pclust/help/format.html#custom). If a single sequence is given, proteins will be taken from its BLAST result against NR. To speed up A2ApsN generation, CD-HIT (Fu *et al.*, 2012) (optional, default identity cutoff: 95%) reduces the redundancy of the protein set that is used for all-against-all BLAST clustering.

Reference protein detection

Proteins either detected by BLAST or input by the user are submitted to the Seq2Ref server to detect 'reference proteins'. As the user input format is flexible, we submit protein sequences to the Protein Identifier Cross-Reference (PICR, Wein *et al.*, 2012) service to detect their IDs in PDB, Swiss-Prot and RefSeq databases.

Protein network generation

A2ApsN is calculated with force-directed graph drawing algorithms implemented by Vivagraph (https://github.com/anvaka/VivaGraphJS) and rendered using WebGL library (supported by most browsers). Reference proteins are colored according to the data sources. Keyword search of the annotations, adjustment for the link number and on-the-fly reference panels describing functional studies are implemented by asynchronous request to our server through AJAX (Asynchronous JavaScript and XML).

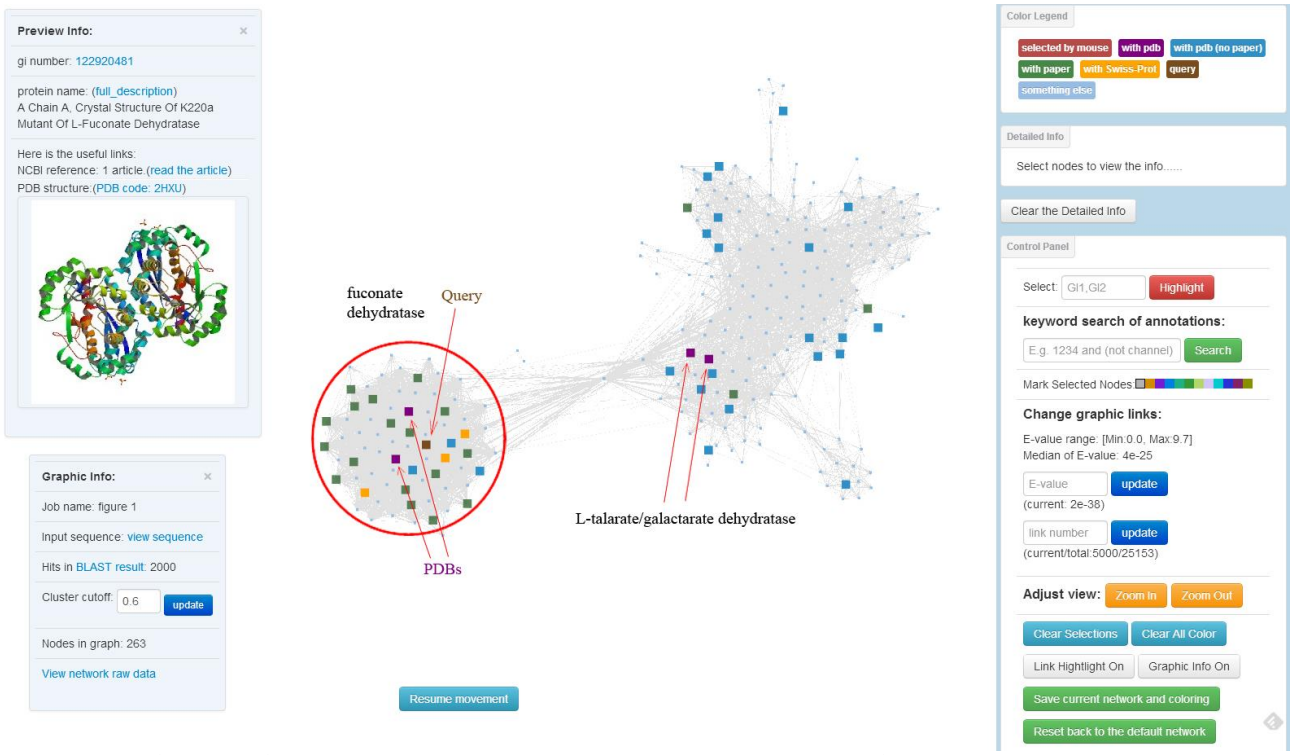


Fig. 1. Snapshot of an A2ApsN from Pclust (web link: http://prodata.swmed.edu/wenlin/server/paper_data/pclust/fig1). Protein nodes are colored, hierarchically, brown (input by the user, if applicable), red (selected by mouse), purple (with PDB structures and their articles), green (with PubMed articles), yellow (with Swiss-Prot functional comments) and blue (with PDB structures of no article, such as those from structural genomics) and light blue (without any reference, smaller size). A "preview" panel and a "detailed Info" panel appear for the protein on which your mouse hovers and your mouse clicks, respectively. Batch selection of proteins is available by inputting the gi numbers (if applicable) separated by commas, and keyword search is available for annotations from the NCBI protein database. By default, Pclust will include the first quarter or 5000 (whichever is smaller) network links (ordered by E-value) and report the corresponding E-value cutoff; a panel to adjust network links by varying the E-value cutoff or the link number is also available. To know more about the control panel, please refer to: http://youtu.be/XLkfEg2jGOc.

REFERENCES

Agapito,G. *et al.* (2013) Visualization of protein interaction networks: problems and solutions. *BMC bioinformatics*, **14 Suppl 1**, S1.

Atkinson,H.J. *et al.* (2009) Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. *PloS one*, **4**, e4345.

Barber,A.E. and Babbitt,P.C. (2012) Pythoscape: a framework for generation of large protein similarity networks. *Bioinformatics (Oxford, England)*, **28**, 2845–6.

Clark,W.T. and Radivojac,P. (2011) Analysis of protein function and its prediction from amino acid sequence. *Proteins*, **79**, 2086–96.

Frickey,T. and Lupas,A. (2004) CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics (Oxford, England)*, **20**, 3702–4.

Fu,L. *et al.* (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics (Oxford, England)*, **28**, 3150–2.

Li,W. *et al.* (2013) Seq2Ref: a web server to facilitate functional interpretation. *BMC bioinformatics*, **14**, 30.

Schnoes,A.M. *et al.* (2009) Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS computational biology*, **5**, e1000605.

Shannon,P. *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, **13**, 2498–504.

Sharan,R. *et al.* (2007) Network-based prediction of protein function. *Molecular systems biology*, **3**, 88.

Wein,S.P. *et al.* (2012) Improvements in the Protein Identifier Cross-Reference service. *Nucleic acids research*, **40**, W276–80.

CHAPTER 4 THE ABC TRANSPORTERS IN CANDIDATUS LIBERIBACTER ASIATICUS[3]

INTRODUCTION

Citrus Greening, also known as Huanglongbing (HLB), is one of the most destructive diseases of citrus. It was first reported in the early 20th century[1,2] and has developed into a major threat for the citrus industry in China, Brazil, and the Eastern United States.[3,4] The symptoms of HLB mainly include yellow shoots, chlorosis leaves, premature defoliation, and aborted fruits, followed by the eventual death of the entire plant.[5,6] The causal agents of HLB are believed to be three closely related bacteria in the *Candidatus* Liberibacter (*Ca.* L.) genus, that is, *Ca.* L. asiaticus, *Ca.* L. americanus, and *Ca.* L. africanus.[5] Among them, *Ca.* L. asiaticus is the most widespread and thus attracts the most attention from researchers.[7]

Ca. L. asiaticus is a Gram-negative alphaproteobacterium. Phylogenetic studies, using 16S rRNA and other genes,[8,9] placed this bacterium in the family of *Rhizobiaceae*. The long branch of *Ca.* L. asiaticus in the phylogenetic tree reveals rapid evolution of this pathogen.[8] The bacterium is transmitted among citrus plants by the piercing-sucking insects, citrus psyllids (*Diaphorina citri* Kuwayama and *Trioza erytreae*). In the plant, *Ca.* L. asiaticus resides mainly in the phloem tissue.[5,6] Efforts have been made to

[3] This chapter was published as:
Li W, Cong Q, Pei J, Kinch LN, Grishin N V. *The ABC transporters in Candidatus Liberibacter asiaticus*. Proteins Struct Funct Bioinforma 2012;80(11):2614–2628.

understand the mechanism of the disease.[10–12]However, difficulty in maintaining the bacterium in culture makes it challenging to carry out any experiments directly on *Ca.* L. asiaticus. Recently, the complete genome sequence[8] of the bacterium was obtained, which opened the possibility of getting insight into the pathogen and the disease by careful analysis of the genome with computational methods.

Here, we focus on the ATP-binding cassette (ABC) systems of the bacterium. ABC systems function in several central cellular processes such as nutrient uptake, drug export, and gene regulation.[13] Based on current phylogenetic analysis, the ABC systems can be divided into three classes: exporters, nontransporting ABC proteins, and a third class that is mostly composed of importers.[14,15] The essential ABC system component is an ABC-type ATPase (also named ABC protein or Nucleotide Binding Domain, NBD). The ABC-type ATPase contains a series of highly conserved sequence motifs, including Walker A and Walker B, which are common for all P-loop NTPases,[16] and Walker C, the signature of the ABC-type ATPase.[17] Walker A and Walker B are crucial for binding and hydrolyzing ATP. As ABC-type ATPases mostly function as homodimers, Walker C is responsible for binding ATP on the side opposite to Walker A and Walker B and is essential for cross-talks between the two monomers.

Most ABC systems include transmembrane proteins and function as transporters. One ABC transporter consists of at least four domains, that is, two transmembrane domains (TMDs) and two NBDs. Pseudo-centrosymmetric dimers formed by two homologous TMDs with similar structures are prevalent. The only known asymmetrically dimeric ABC transporters, the ECF transporters,[18,19] have two possibly structurally

different TMDs, a T-component TMD and an S-component TMD. The TMDs of most ABC-type transporters fall into four clans in the Protein families database (Pfam)[20]: "ABC transporter membrane domain" clan (CL0241), "ABC-2-transporter-like" clan (CL0181), "membrane and transport protein" clan (CL0142), and "BPD transporter like" clan (CL0404). Several known structures suggest that the TMD should dock into the NBD by a "coupling helix",[21] which coordinates a conformational change caused by ATP-hydrolysis. A number of ABC systems also include periplasmic components that are responsible for transporting the substrates across the periplasmic space. Periplasmic-binding proteins (PBPs) are used by many importers to recognize substrates and initialize the transporting cycle by interacting with the TMD.[22] Similarly, some exporters, especially those from Gram-negative bacteria, use a series of auxiliary proteins [periplasmic proteins and/or outer membrane (OM) proteins] to route the cargo across the periplasmic space. Here, we refer to the PBPs, auxiliary proteins, TMD-containing proteins, and NBD-containing proteins in the following text as "ABC system proteins."

In addition to primary active transport, ABC transporter activity is thought to be related to virulence in some Gram-negative bacteria.[23–26] In plants infected by *Ca.*L. asiaticus, the ABC transporters may contribute to host metabolic imbalances and thus the Citrus Greening disease symptoms.[8] Given the important roles of ABC transporters and their possible involvement in pathogenicity, analysis of these ABC transporters will help us to understand the metabolism of the bacterium and the mechanism of the disease. In this article, we report a detailed study of all ABC transporters in the *Ca.* L. asiaticus. We collected all potential ABC system proteins in the proteome and identified 14 ABC

transporter systems and 7 nontransporting ABC proteins. Combining different computational methods, we predicted the structure and substrate specificity of each ABC transporter.

RESULTS AND DISCUSSION

Detection and annotation of ABC-transporters in *Ca.* L. asiaticus

A total of 55 ABC system proteins were detected in the whole genome. We identified 14 complete ABC transport systems consisting of 42 ABC-system proteins and 7 ABC-type ATPases that are likely involved in cellular processes other than transport (Table I). The remaining six potential ABC transporter components do not have confident NBD partners in the proteome, and we thus name them "orphan" ABC components.

Evidence for annotations

To ensure functional annotations deduced by homologous proteins, we identified close homologs with experimentally verified function for each proposed ABC transporter NBD. Their close relationships are reinforced by reciprocal best hits detected by BLAST. Most NBDs of *Ca.* L. asiaticus share more than 40% sequence identity (shown in Table I) with their close homologs (proposed orthologs) with experimentally verified function, meeting the suggested threshold for precise function annotation transfer.[48] Each *Ca.* L.

asiaticus NBD clustered into a group (e-value cutoff: $1e^{-40}$) with its proposed ortholog as revealed by clustering on the basis of all-to-all BLAST sequence comparison, with an exception of the type I secretion system (discussed below). Four NBDs, that is, CLIBASIA_02415 (Nrt/Ssu/Tau-like system NBD), CLIBASIA_0135 (type I secretion system NBD), CLIBASIA_05125 (Uup nontransport system), and CLIBASIA_05400 (RecN nontransport system) are more variable and show marginal or low-sequence identity (<40%) to their proposed orthologs. We will discuss their predictions in the functional detail section.

Novel predictions of ABC system proteins

The original annotations of these ABC-system proteins from NCBI, SEED, COG, and KEGG were able to place them as ABC-transporter components. However, clear predictions on their substrate specificity and polarity of the transporters were absent in many cases. Although for 86% of all proteins, the most specific annotation carefully chosen from all these databases could successfully indicate the same substrate or function predictions as ours, our manual study provided or modified the annotation for seven ABC system proteins from four ABC systems including choline/acetylcholine importer (Cho system), possible oxoacid ion importer (Nrt/Ssu/Tau-like system), lipoprotein exporter (Lol system), and Uup nontransport ABC protein. The systems revised with new annotations are described later.

The TMD of lipoprotein exporter (Lol system) was absent in current NCBI and KEGG databases, possibly due to the fact that this TMD consists of two open reading frames that were considered as pseudogenes by the NCBI gene prediction pipeline. The gene prediction pipeline of the SEED, in contrast, detected these two protein fragments but failed to predict the function of the second half. Because of the presence of the intact Lol system ATPase and other essential components in *Ca.* L. asiaticus, it is unlikely that the TMD of Lol system has lost its function. Instead, in the absence of any potential sequencing error, the two protein halves may interact with each other after translation, or some type of translational frame shift mechanism may allow the successful expression of the full protein.

Sequence and structural analysis of the NBDs in *Ca.* L. asiaticus

The NBDs are the most conserved domains among various ABC system proteins. The MSA [Fig 1(a)] of NBDs from *Ca.* L. asiaticus, including those from the nontransporting ABC proteins, reveals characteristic conservation patterns. To note, three ATPases, that is, MutS, RecF, and RecN, are not included in the MSA due to their diverse sequences. The conserved motifs match known motifs in ABC-type ATPases,[17] including A-loop, Walker A, Q-loop, Walker C, Walker B, H-loop, and D-loop from the N-terminus to C-terminus, suggesting that the NBDs in *Ca.* L. asiaticus are functional ABC-type ATPases. All these NBDs are evolutionarily related, and their predicted structures all belong to the same family (ABC transporter ATPase domain like)

in structural classification of proteins. ABC transporters function as dimers, as shown in Figure 1(b). In the structure, all the sequence motifs are clustered on the interface of the two NBDs [Fig 1(b)]. To bind one ATP molecule, motifs from both sides are involved [Fig 1(c)], allowing the co-ordinate movements of two NBDs upon the binding and hydrolyzing of ATP.[49] Noticeably, the Walker C motif of PrtD is deteriorated. Whether the substitution disables the function or develops a new functional theme remains to be explored experimentally.

To confirm the close relationships between the *Ca.* L. asiaticus NBDs and their experimentally studied orthologs, we constructed a phylogenetic tree of those NBDs, together with a set of previously analyzed NBDs in Ref.[14] (Fig 2). Similar to the previous phylogenetic studies,[13,14] the constructed tree topology revealed three major groups colored red, green, and blue, respectively. To note, some nontransport systems (i.e., MutS, RecF, and RecN) are not included due to their diverse sequences. The first major group (red) contains ABC-type exporters mainly for multiple drugs, lipids, peptides, and proteins and corresponds to class I ABC systems in the previous classification. The second major group (green) contains NBDs from both importers (majority) and exporters and corresponds to class III ABC systems in the previous classification. It is possible that these mixed exporters in the second group originated from ancient ABC-type importers and adopted the function of working in efflux systems later in evolution. The third major group (blue) contains mainly nontransport ABC proteins, corresponding to class II ABC systems in the previous classification. Three ATPases (Uniprot ID: CCMA_ECOLI, WHIT_DROME, PDR5_ECOLI) form a small clade (colored orange). The TMDs of the

three export systems happen to be in the same clan "ABC-2-transporter-like clan" (CL0181) in the Pfam database while the TMDs of the other exporters (red) are from the clan "ABC transporter membrane domain" (CL0241).

The exhilarating message the phylogenetic tree conveys is that, except for PrtD, all other *Ca.* L. asiaticus proteins are placed closely to the proposed experimentally studied orthologs. Although marginal bootstrap values exist due to the diverse sequences between different ABC-type ATPase families, branches with confident bootstrap values suggest a positive correlation between similarity in substrate preference and similarity in sequence. Six groups with similar substrate preference formed individual clades with good bootstrap probabilities (as indicated by the black dots). However, a few transporters of similar substrates appear to be phylogenetically far from each other. These dispersed branches of similar functions may reveal a real complexity in functional divergence or merely be incorrect tree topology due to nonconfident statistics and insufficient data in the process of evolutionary tree reconstruction. Some sequences placed in long branches (purple frame in Fig 2) are more diverse among the other ATPases. They failed to group with other ATPases possibly due to the insufficient number of representative sequences and long branch attraction problems associated with tree construction.

Classification of the ABC transporter TMDs in *Ca.* L. asiaticus

In contrast to the conserved NBDs, the TMDs are more divergent in both sequence and structure. For 15 of the 19 TMDs, we were able to generate homology-

based structure models and classify them into three groups. Within each group, the TMDs adopt the same fold, and the representative structure templates for these three groups are shown in Figure 3(a). In a recent review, the authors classified solved TMD crystal structures into three different folds,[49] that is, type I importer, type II importer, and exporter. Nevertheless, the newly established S-component structure of ECF transporter[18,19] exhibited a new fold and thus extended the TMD classification. The three groups of TMDs in *Ca.* L. asiaticus are consistent with these three structure folds in the review[49] and correspond to three nonhomologous Pfam families (Table I). For each group, the MSAs of the TMDs together with their representative homologs were generated. Although the sequences appear to be rather diverse, hydrophobic and hydrophilic patterns are preserved. Small residues mediating interhelix interactions and other characteristic residues, such as proline involved in helix kinks, are highly conserved as well.

The coupling helices from different folds exhibit varied sequence features and structures, thus serving as the signature of each fold (Fig 3). TMDs in the first group are from the “binding-protein-dependent transport system inner membrane (IM) component” Pfam family (PF00528), and they adopt the type I importer fold. This group of TMDs possesses five core TMHs, and the essential coupling helix responsible for the interaction between the TMD and NBD is located between the third and fourth TMH [Fig 3(a), left panel]. The coupling helix of the type I importer adopts a semiperpendicular interaction with the following helix [Fig 3(b), left panel]. The short helical pair is connected to their connecting TMHs by kinks. All the TMD sequences in the second group belong to the “ABC 3 transport” family (PF00950) and assume the type II importer fold.

Representative type II structures include 10 TMHs [Fig 3(a), center panel]. The type II coupling helix differs from that found in the type I importer fold. It follows a short helix that extends from the sixth TMH by a kink and is connected to the seventh TMH by a short loop [Fig 3(b), center panel]. Compared to the representative structure template consisting of 10 TMHs, the *Ca.* L. asiaticus sequences lack one peripheral TMH at the very N-terminus [colored gray in Fig 3(a) center panel], suggested by HHsearch alignments. As this TMH does not participate in the structure core, its absence should not affect the general fold. All the TMDs in the third group are from "TMD of ABC transporters" family (PF00664) and adopt the exporter fold. They consist of six TMHs that extend into the cytoplasm [Fig 3(a), right panel]. These TMDs form swapped dimers by exchanging the fourth and fifth helices. Upon binding ATP, the TMDs switch from an inward-facing conformation to an outward-facing conformation and release the substrate to the periplasmic space.[50] Unlike the importers, the exporter fold has two coupling helices [Fig 3(b), right panel]: one is located between the second and the third TMHs and interact with both NBDs in the closed conformation, while the other is located between the swapped fourth and fifth TMHs and inserted into the groove of the NBD on the opposite side.[51]

The other four *Ca.* L. asiaticus TMDs fall into "Permease" (Permease, CLIBASIA_00085), "Predicted Permease YjgP/YjgQ family" (YjgP/Q, CLIBASIA_01390 and CLIBASIA_01395), and "FtsX-like Permease family" (FtsX, peg.788&peg.789) in Pfam. FtsX and YjgP/Q belong to the same Pfam clan "BPD transporter-like" (BPD-like). This suggested homologous relationship is further

supported by the pairwise HHsearch probability over 90%. Moreover, the third Pfam family, Permease, is likely related to FtsX and YjgP/Q families, as suggested by HHsearch (probability over 90%). The suggested sequence relationships are limited to the coupling helix and its surrounding TMHs. The HHsearch alignment between YjgP/Q and Permease extended to the N-terminal TMH (marked by a plus symbol in Fig 4), while the extension in the alignment between FtsX and Permease is one TMH at the C-terminus of surrounding TMHs (marked by asterisk in Fig 4). All three families include a similar predicted minimal transmembrane topology displayed in FtsX [Fig 4(c)]. The presumed core topology includes four helices, with the coupling helix located between the second and third TMH. With respect to this core FtsX TMH topology, YjgP/Q includes an inserted extracellular domain following the third TMH and two additional C-terminal TMHs, and Permease includes an N-terminal cytoplasmic domain and an additional C-terminal TMH. Thus, the three families, FtsX, YjgP/Q, and Permease, share a similar core TMH topology in addition to the type I importer coupling helix motif, suggesting that they adopt similar structures. Intriguingly, the ABC systems consisting of these TMDs in *Ca.* L. asiaticus are all noncanonical transporters. They are involved in shuttling substrates between the IM and the OM, by either releasing (Lol and Lpt) or inserting (Lin) molecules that are lipids or with lipid moieties from/to the outer leaflet of the IM. Such unique and similar transported substrates may serve as evidence to reinforce their relationships.

The Pfam-defined BPD-like clan includes solved structures within the family “Binding-protein-dependent transport system IM component” (PF00528, BPD_transp_1,

abbreviated as BPD below). Because proteins in the same Pfam clan indicate that they are evolutionarily related, it raises the question whether the structures of the three families look similar to the known structure in BPD family. The predicted TMD topology of FtsX, YjgP/Q, and Permease differs from the BPD family structure topology (Fig 4). To maintain both the position of the coupling helix and a similar TMH topology, the N-terminal TMH of BPD must be deleted. However, this N-terminal TMH plays an integral role in the BPD fold, maintaining interactions with all other TMHs and positioning the coupling helix (Fig 4). Given the central role of this TMH, its deletion would not likely be tolerated and relating BPD to FtsX would therefore require a less parsimonious pathway of losing the peripheral BPD C-terminal TMH (red), followed by a circular permutation to replace the N-terminal BPD TMH with the C-terminal helix of FtsX. Given this complex requirement for maintaining topology, we could not confidently infer the relationship between the BPD structure and FtsX, bringing into question the Pfam clan assignment.

Another special common feature of the TMDs from Lpt, Lol, and Lin systems is the presence of fused soluble domains (Fig 4). The TMDs of the Lpt system and the Lol system in *Ca.* L. asiaticus are fused with periplasmic domains that likely participate in delivering substrates from the IM to the periplasmic space.[52] The fused periplasmic domain in the IM proteins (CLIBASIA_01390, CLIBASIA_01395) of the *Ca.* L. asiaticus Lpt system is predicted by HHsearch to be structurally similar to other auxiliary proteins of the *Ca.* L. asiaticus Lpt system, including LptA (CLIBASIA_03160), LptC (CLIBASIA_03165), and one domain in the OM auxiliary protein LptD

(CLIBASIA_01400). It is likely that the Lpt system has evolved by duplications to allow efficient conveying of substrates.[53] The TMD of the Lin system, on the contrary, is fused with a cytoplasmic "Anti-Sigma factor antagonist" (STAS) domain.[54] In the SulP family transporters,[55] STAS is suggested to sequester acyl-carrier protein, an essential protein for fatty acid biosynthesis, and thus links transport with fatty acid metabolism. Similarly, the STAS domain in the IM proteins might be able to recruit other proteins and contribute to the regulation of the Lin system or the cross-talks between transporting and other processes.

Function details of the predicted ABC transporters in *Ca.* L. asiaticus

ABC-type importers

In the *Ca.* L. asiaticus proteome, we detected eight ABC-type importers that should be responsible for uptaking essential nutrients from the environment. The substrates of these ABC type importers include amino acids, B family vitamins, ions, and lipids (the first eight systems of Table I). Because it is suggested that *Ca.* L. asiaticus might deplete the host's nutrient supply, which results in disease symptoms,[8] these ABC-type importers might help contribute to the death of the plant.

Canonical importer systems

The substrate specificities of six nutrient importers can be confidently inferred from their prominent sequence similarity to close homologs with experimentally verified functions. They are general L-amino acid transporter (Aap), phosphate transporter (Pst),

thiamine transporter (Thi), choline transporter (Cho), zinc transporter (Znu), and manganese and iron transporter (Sit). One NBD, one PBP, and one or two TMDs are present in each system. Among them, Cho and Znu contain one TMD that should act as a homodimer, and the other four systems contain two homologous TMDs (fused TMDs for Thi system and separated TMDs for others). Because the metabolic pathways for some amino acids are missing in *Ca.* L. asiaticus, the presence of an amino acid transporter (Aap) suggests that this bacterium requires external amino acid supply for its survival. However, we could not deduce the amino acid preference of this transporter, because the orthologous Aap system[56] in *Rhizobium leguminosarum* is reported to have broad substrate specificity. Thus, understanding the substrate preference of the Aap system experimentally might illuminate the minimal amino acid requirement of *Ca.* L. asiaticus.

Among the six importers, the system with revised annotations is the choline importer (Cho system), which was annotated as a glycine-betaine transporter in existing databases. Although choline and glycine-betaine are chemically similar compounds (glycine-betaine is the oxidized form of choline), the experiments on the orthologous system[57] in a closely related bacterium *Sinorhizobium melilotis*howed a high specificity for choline, rather than the annotated glycine-betaine substrate. Given the close relationship of the *Ca.* L. asiaticus CLIBASIA_01125 (NBD) component to that in *S. meliloti* as well as the similar operon arrangement (ChoX-ChoW-ChoV), we annotate these components as a Cho system. The highly similar TMD (ChoW, identity: 62%, e-value: $1e^{-96}$) and PBP (ChoX, identity: 49%, e-value: $5e^{-87}$) between *Ca.* L. asiaticus and *S. meliloti* further warrant our prediction.

A possible novel ABC system

The seventh proposed *Ca.* L. asiaticus transporter (Nrt/Ssu/Tau-like system), consisting of one special ATPase containing a characteristic C-terminal domain and another protein containing two TMDs, cannot be classified in terms of a specific substrate. The NBD of the ATPase (CLIBASIA_02415) shows a close relationship to three experimentally studied oxoacid ion transporters, that is, *Synechococcus elongatus* nitrate transporter (Nrt),[58–60] *Bacillus subtilis* alkanesulfonate transporter (Ssu),[61] and *Escherichia coli* taurine transporter (Tau).[62] However, this Nrt/Ssu/Tau-like system in *Ca.* L. asiaticus shows significantly different component arrangements and protein domain contents from any of the three closely related systems (Fig 5). No PBP has been detected for this system in *Ca.* L. asiaticus, and the membrane component (CLIBASIA_02420) contains two tandem TMDs instead of one TMD in the other three systems. Moreover, the *Ca.* L. asiaticus NBD (CLIBASIA_02415) has a unique fused C-terminal domain that is conserved among a small group of homologs. This additional domain differs from the ABC-type NBD and is classified as "ABC nitrate/sulfonate/bicarbonate family transporter, ATPase subunit" (PF09821), which is not a member of the P-loop NTPase clan (CL0023) in the Pfam database. HHsearch suggests that it adopts a "winged helix" DNA-binding fold with over 95% probability. On the basis of its relatively close relationship to Nrt, Ssu, and Tau, we annotated this system as Nrt/Ssu/Tau-like ABC transporter without a specific functional annotation. The phylogenetic positions of the NBDs (Fig 2) imply that the Nrt/Ssu/Tau-like system might

have diverged from the three systems at an early time point. Possibly, the ATPases with this characteristic C-terminal domain may be components of novel ABC transporters that have not been experimentally studied. Given the dramatic differences in its operon organization and the domain structures of the TMDs and the ATPase from the three related systems, whether the Nrt/Ssu/Tau-like system is still a transporter remains questionable and requires further experimental exploration.

A noncanonical importer system possibly involved in lipid trafficking

Another special ABC-type transporter in *Ca.* L. asiaticus is the Lin system composed of one NBD, one TMD, and two PBPs. Its closely related protein[63] in *Sphingobium japonicum* is reported to be involved in the utilization of gamma-hexachlorocyclohexane, presumably by controlling membrane hydrophobicity. However, the detailed mechanism of the Lin transporter in *S. japonicum* remains unclear. Our phylogenetic analysis reveals that the Lin system is closely related to the MKL family of lipid importers.[15] Experimentally studied systems in the MKL family include phospholipid importer (Mla)[64] in *Escherichia coli*, cholesterol importer (Mce4)[65,66] in *Mycobacterium tuberculosis*, lipid importer (TGD)[67] in the chloroplasts of *Arabidopsis*, and a transporter involved in toluene tolerance (Ttg2)[68]*in Pseudomonas putida*. Unlike the other importers that transport substrates across the cell membrane, MKL family importers only insert their substrates, mostly lipid-like compounds, into the IM (Mla and TGD) or cell membrane (Mce4). Inferred from the relationship to MKL family members, the predicted Lin system in *Ca.* L. asiaticus is likely to insert certain

lipid components, possibly cargoed from OM like the Mla system, into the IM, thus contributing to the maintenance of membrane hydrophobicity and resistance to organic solvent. Because one canonical transporter should translocate the substrate across the cell membrane, we categorize it as a noncanonical importer.

ABC-type exporters

Six ABC-type exporters were detected in the *Ca.* L. asiaticus proteome. Compared to ABC-type importers, exporters generally have a wider spectrum for substrates. The six exporters in *Ca.* L. asiaticus are predicted to contribute mainly to the biogenesis of the OM, multiple drug resistance, and toxin protein secretion.

Noncanonical exporters involved in OM biogenesis

The outer membrane (OM) is an essential component of Gram-negative bacteria. To complete the biogenesis of the OM, two types of compounds, that is, lipopolysaccharide (LPS) and lipoprotein, have to get anchored in the OM. The transporting process of these two compounds has two similar steps: first, the precursors anchor into the outer leaflet of the IM and mature to be LPS or lipoprotein in the IM; second, the compounds detach from the outer leaflet of the IM and anchor to the inner leaflet of the OM. For LPS biogenesis, a third step is taken to flip the LPS from the inner leaflet of the OM to the outer leaflet of the OM. Several ABC-type transporters are involved in LPS and lipoprotein translocation.[69]

LPS biogenesis requires two ABC-type exporters, although more than two ABC-type transporters are involved.[69,70] For the first step, lipid A, one of the LPS precursors, is flipped by MsbA.[70] In *Ca.* L. asiaticus, two copies of MsbA with fused TMD and NBD have been detected. One or both of them could be responsible for this step. The second step involves the Lpt system, consisting of an ABC-type transporter in the IM and a set of auxiliary proteins in the periplasmic space and the OM. The ABC transporter in the *Ca.* L. asiaticus Lpt system includes one NBD, two homologous TMDs, and all other necessary components. For lipoprotein biosynthesis, only one ABC-type transporter (LolD) is involved in the second step of shuttling the substrate between the membranes, while the first step is carried out by Sec translocase.[69] Compared to the *E. coli* Lol system, the *Ca.* L. asiaticus Lol system, which harbors one NBD, one TMD, and one auxiliary protein, lacks the LolE (TMD) and LolB (auxiliary) genes. This difference is consistent with the previous observation that alphaproteobacteria generally lack LolB and only gammaproteobacteria harbor LolE.[69] Because LolC (TMD) and LolA (auxiliary) are homologous to LolE and LolB, respectively, it is possible that LolC forms a homodimeric TMD instead of a heterodimer with LolE, and LolA could compensate for the function of LolB.

Unlike other canonical ABC exporters in the proteome, Lpt and Lol systems help substrates to detach from the outer leaflet of the IM, rather than transporting the substrates directly from the cytoplasm to the periplasm.[69] Thus, Nagao et al.[71]named these processes “projections” to distinguish those noncanonical exporters. It has been reported that three OM-biogenesis-related systems, namely, Lol, MsbA, and Lpt, are

required for the viability of *E. coli*.[72–75] Thus, they could be promising drug targets for Citrus Greening disease treatment.

Exporters related to drug resistance

Three *Ca.* L. asiaticus ABC systems with fused NBD and TMD are suggested to be associated with drug resistance. Among them, the NBDs of MsbA1 (CLIBASIA_04080) and MsbA2 (CLIBASIA_00390) show high pairwise identity (43%) to each other, and both show high identity to *E. coli* MsbA (Table I). Although their substrate preferences might differ, the experimental studies on this family of proteins do not provide enough information to distinguish their substrates. Knowing that the *E. coli* MsbA, the proposed ortholog for *Ca.* L. asiaticus MsbA1 and MsbA2, is capable of generating multidrug resistance,[76] we proposed that those two copies of MsbA could also be involved in exporting multiple drugs.

Another *Ca.* L. asiaticus system, whose NBD is fused with its TMD and shows 45% identity to that of *Ca.* L. asiaticus MsbA1, is the AtmA exporter (CLIBASIA_02315). Its close homolog (AtmA)[77] in *Cupriavidus metalliduran*functions as a transporter that is related to cobalt and nickel resistance. Therefore, it is likely that the *Ca.* L. asiaticus AtmA also functions in heavy metal resistance, possibly by exporting heavy metals.

A type I secretion system in Ca. L. asiaticus

A special ABC-type exporter in *Ca.* L. asiaticus is the type I secretion system (PrtD). It is one of only two protein secretion systems (the other is the Sec protein secretion system) present in *Ca.* L. asiaticus.[8] Type I secretion systems can export proteins of varied sizes and are responsible for secreting RTX (repeat in toxin) proteins.[78,79] Although *Ca.* L. asiaticus PrtD (CLIBASIA_01350) has been annotated as a type I secretion system ATPase in the current database, its low-sequence identity to the experimentally studied ortholog (27%) and the substitution of Walker C motif (Fig 1) question its function. One explanation for the low similarity might be that the *Ca.* L. asiaticus PrtD NBD exhibits an elevated evolutionary rate compared to the other type I secretion systems, as suggested by the distant relationships to its homologous type I secretion systems in CLANS clusters. More importantly, the presence of a highly characteristic type I secretion system substrate, RTX protease serralysin[79,80] (CLIBASIA_01345, NCBI gi: 254780384) next to the *Ca.* L. asiaticus PrtD NBD, suggests that this type I secretion system should be capable of exporting substrates, at least RTX proteases. Because the transporting cycle requires energy provided by ATP hydrolysis, the Walker C substitution of this possibly functional type I secretion system would be an intriguing issue to investigate in terms of the structure, function, and co-operativity between two NBDs. In ABCG5-ABCG8 heterodimeric sterol transporter, an intact Walker A and Walker B from ABCG5 functions together with an intact Walker C from ABCG8 that is essential for transport activity. The second nucleotide-binding site is deteriorated, and substitution of Walker C in ABCG5 does not affect the sterol secretion.[81] Thus, it is possible that the *Ca.* L. asiaticus PrtD with the deteriorated Walker

C developed a partnership with other NBDs and only contributes its Walker A and Walker B to form the ATP-binding site.

Nontransport ABC proteins

Seven nontransport ABC-type ATPases are detected in *Ca.* L. asiaticus. Although not involved in transporting, they are related to important cellular processes such as Fe-S assembly (SufC),[82] virulence gene regulation (ChvD),[83] transposon excision regulation (Uup), and DNA repair regulation (UvrA, MutS, RecF, and RecN).[15] The four ATPases involved in DNA repair show diverse sequence features compared to the ABC-type ATPases. For three of them (UvrA, RecF, and RecN), a long insertion between the Q-loop and Walker C motif makes it difficult to detect their relationships to ABC-type ATPases, while the Q-loop and Walker C motif are lacking in the MutS sequence. We also detected a short *Ca.* L. asiaticus protein (CLIBASIA_02635, 110 residues, not listed in Table I) similar to the N-terminal region of Rad50 (1312 residues in *Saccharomyces cerevisiae*), a structure maintenance of chromosome family protein. This protein may be a relic of evolution.

The novel *Ca.* L. asiaticus nontransport ATPase is now annotated as Uup. The closest experimentally studied ortholog is the *E. coli* Uup (gi: 16128916), a soluble protein involved in transposon excision regulation.[84–87] Although the overall identity of *Ca.* L. asiaticus Uup to *E. coli* Uup is marginal (35%), possibly due to the different insertion length between the Q-loop and Walker C motif in the first NBD, its homologous

relationship is supported by the high sequence similarity of the second NBD (about 49%). Another piece of evidence is a coiled-coil domain in the C-terminus of the *Ca.* L. asiaticus Uup predicted by COILS,[88] which is consistent with the presence of a similar coiled-coil domain at the C-terminus of *E. coli* Uup. This coiled-coil domain in *E. coli* Uup is essential for overall structure stability and participates in binding DNA.[89] Therefore, similar to *E. coli* Uup, the *Ca.* L. asiaticus Uup might also have the ability to regulate the DNA excision. A recent gene deletion study also proposes the Uup protein to be involved in bacterial quorum-sensing[90] mediated by direct contact between the cells, which makes this predicted Uup an interesting target for experimental study as the quorum-sensing phenomenon is thought to play an important role in bacterial virulence.[91]

Incomplete systems

Five ABC-system proteins in *Ca.* L. asiaticus do not have confident NBD partners. Considering the small genome size of this bacterium, these proteins might be the evolutionary relics of genome reduction. Alternatively, these “orphan” ABC system proteins, either TMD or PBP, may be able to adopt functions other than ABC transport. A recent study[92] in *Arabidopsis* ABCG family transporters reported that ABCG11 could form a homodimer with itself or a heterodimer with ABCG12. Considering the promiscuous dimerization of ABC transporter in *Arabidopsis*, it is also possible that the

ATPases from complete systems could exhibit multiple functions and might be capable of forming a functional ABC-transporter with these “orphan” components.

CONCLUSIONS

Combining various computational methods, we identified a complete set of ABC transporters and several other nontransport ABC systems in the *Ca.* L. asiaticus proteome, confirmed annotations for most of the ABC system proteins, predicted the polarity and structure of each ABC transporter, and generated new annotations for seven proteins from four ABC systems. Although the features of most ABC systems could be deduced from the abundant experimental data on their orthologs, we reported several novel observations, including a Nrt/Ssu/Tau-like transporter that has never been studied, a deterioration of the Walker C motif in the type I secretion system, the duplication events in the periplasmic components of the Lpt system, and the remote homology relationships between the FtsX, YjgP/Q, and Permease Pfam families. In addition, our analysis reveals several proteins likely important for controlling the Citrus Greening disease, including the type I secretion system and its substrate, and the essential ABC transporter systems involved in bacterial OM biosynthesis. Further studies targeting these proteins might lead to better understanding and treatment of HLB.

MATERIALS AND METHODS

Identification of ABC system proteins

Ca. L. asiaticus protein sequences were downloaded from the NCBI Genbank database (ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria/Candidatus_Liberibacter_asiaticus_psy62_uid29835), and additional proteins predicted by the SEED[27](http://pseed.theseed.org/seedviewer.cgi), but missed by NCBI, were added. The relevant information and annotations of these proteins from NCBI (http://www.ncbi.nlm.nih.gov/nuccore/CP001677), Cluster of Orthologous Groups (COGs),[28] and the SEED were taken as references. The protein annotations by NCBI, COG and the SEED were examined manually to obtain a primary list of ABC system proteins, followed by two additional approaches to ensure a complete list. First, starting from the proteins in the primary list and the sequence profiles of known ABC system protein families in Pfam, we used PSI-BLAST[29,30] and HHsearch[31] to identify homologous proteins from the *Ca.* L. asiaticus proteome. Second, assisted by our comprehensive database of the *Ca.* L. asiaticus proteome (http://prodata.swmed.edu/liberibacter_asiaticus/), we manually curated all the proteins to ensure that all ABC system proteins were included in our list.

Substrate specificity and structure prediction of assembled ABC systems

For each predicted *Ca.* L. asiaticus ABC system protein, we applied PSI-BLAST, RPS-BLAST,[32] and HHsearch to detect homologous proteins, protein families, and conserved domains,[33] paying special attention to close homologs with experimentally verified functions. Sequence comparison between *Ca.* L. asiaticus and these homologs assisted by the sequence clustering tool CLANS[34] served as the primary evidence for our function assignments. Based on these assignments, the genomic context of each protein retrieved from the SEED, and the functional association networks between proteins detected by STRING,[35] we assembled these protein components into ABC transport complexes. The closest homologous proteins with 3D crystal structures judged by confidence, coverage, and match of conserved residues were manually selected as structure templates. MODELLER[36] was then applied to these templates to obtain a structure model for each protein. For potential TMDs and PBPs, the transmembrane helices and signal peptides were detected by TMHMM,[37] TOPPRED,[38] HMMTOP,[39] MEMSAT,[40]MEMSAT_SVM,[41] Phobius,[42] and SignalP[43] to reveal their topologies and confirm their subcellular localization.

Multiple sequence alignment and phylogenetic tree of NBDs in *Ca.* L. asiaticus

We generated a multiple sequence alignment (MSA) for all the NBDs in the *Ca.* L. asiaticus proteome together with five representative protein structure templates by PROfile Multiple Alignment with predicted Local Structures and 3D constraints (PROMALS3D),[44] followed by manual adjustments. The sequences of well-characterized

ABC-type ATPases from other organisms were then added to these predicted NBD sequences to generate a common MSA. For phylogenetic reconstruction, positions with gap fraction over 10% were discarded. Phylogenetic estimation using Maximum Likelihood (PhyML) program[45] was used to build an evolutionary tree with LG model[46] for the substitution model, four discrete rate categories for the rate heterogeneity among sites, Nearest Neighbor Interchange for the tree improvement, and SH-like approximate Likelihood-Ratio Test[47] for estimating the branch support.

Sequence and structure analysis of the TMDs

We performed structure comparison on the TMDs to classify them manually on the basis of their topology and architecture. We further used PROMALS3D to construct a MSA of TMD sequences in each group, together with homologous structure templates and representative sequences that share more than 30% sequence identity with *Ca.* L. asiaticus proteins. The MSAs were then improved by manual adjustment considering both sequence patterns and structure features.

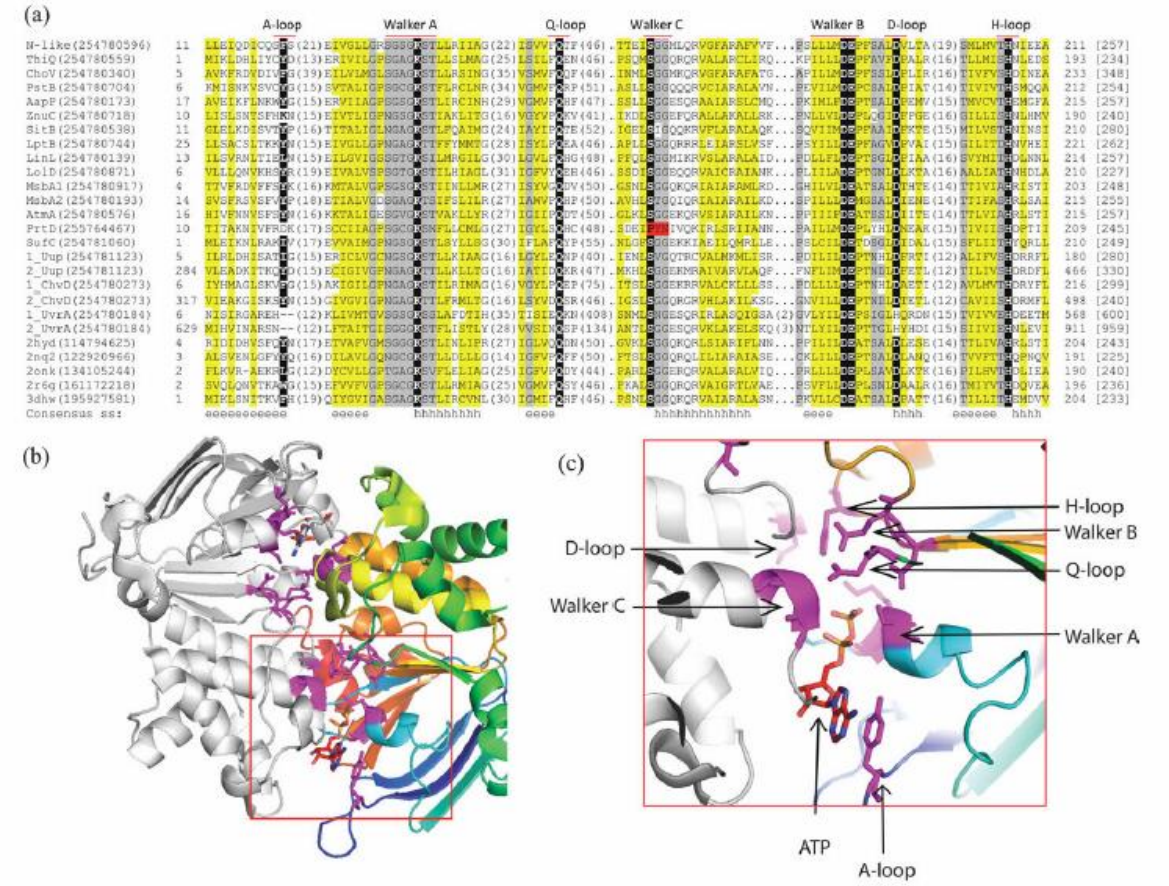


Figure 1

MSA and representative structure of NBDs in *Candidatus* Liberibacter asiaticus proteome. (**a**) Simplified version of the MSA of all NBDs of *Candidatus* Liberibacter asiaticus and representative homologous structures (only the segments containing sequence motifs of the NBD are shown). The names of motifs are labeled on the top of the MSA. Protein name abbreviations or PDB IDs, with gi number in the parentheses, are used as sequence identifiers at the beginning of each line. N-like is short for Nrt/Ssu/Tau-like system NBD. For ABC-type ATPases with two NBDs, we assign a number in front of the identifier to distinguish between the two domains. In the sequences, hydrophobic residues are highlighted in yellow, small residues positions are colored in gray, and the most essential residues for the function are represented as white letters on black backgrounds. Starting/ending residue numbers and sequence length are shown in italic font and in brackets, respectively. Numbers of residues between the segments are indicated in the parentheses. Dots are used to adjust the space for the MSA. Gaps are shown in dash lines. The PYN residue marked red indicates the substituted Walker C motif. The "consensus ss" line shows the consensus secondary structures predicted by PROMALS3D. For the secondary structure, "e" means beta sheet and "h" stands for alpha helix. (**b**) Structure of ABC transporter nucleotide-binding domain homodimer with ATP molecules. The structure is adapted from Sav1866 (PDB: 2hyd). The right NBD is colored in rainbow from N to C terminus while the left NBD is colored in gray. Residues essential for the function are shown as sticks in magenta. (**c**) Close-up of ATP-binding site enlarged from the red frame in (b). ATP and sequence motifs of the NBDs are pointed out.

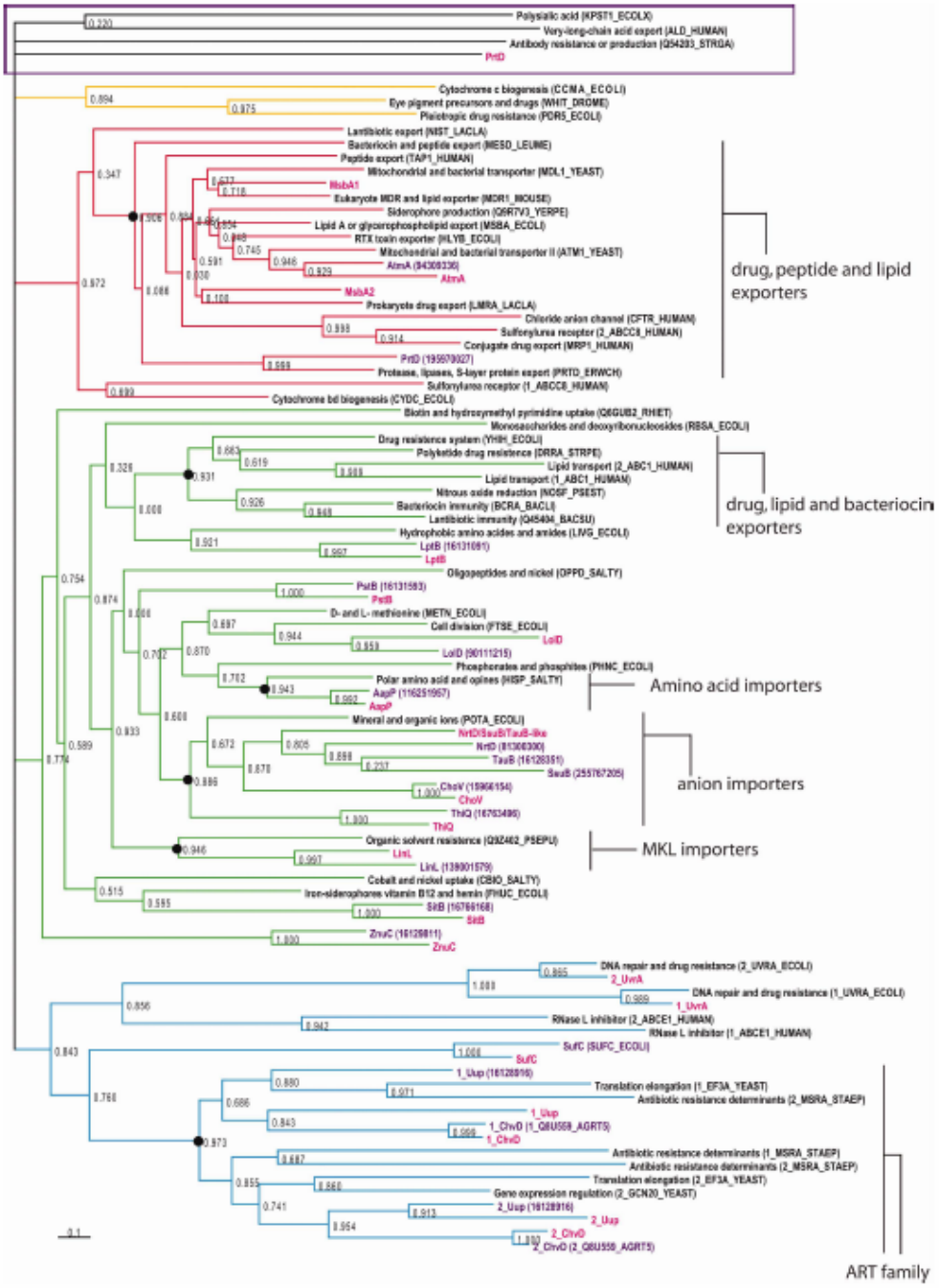


Figure 2

Phylogenetic tree of ABC transporter NBDs. *Ca.* L. asiaticus NBDs and their experimentally studied orthologous NBDs are labeled by the gene abbreviation and colored in pink and purple, respectively. Other NBDs are denoted by their functional definitions. In the parenthesis, gi or unprot id is indicated. For the ABC proteins with multiple NBDs, a domain number is assigned in front of the gene abbreviation as well as the uniprot ID. Bootstrap support values are shown for the internal nodes. The black dots mark the roots of several clades made of proteins with similar substrates. Three major groups are colored in green, red, and blue, respectively. The small group of exporters is colored in orange. The purple box highlights the sequences that are remote from others. ART is short for "antibiotic resistance and translation regulation," and MKL family represents the family of "retrograde transport of lipids, organic solvent resistance, and steroid uptake."[15] The multiple sequence alignment used to generate the tree is available in Supporting Information Figure S9. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

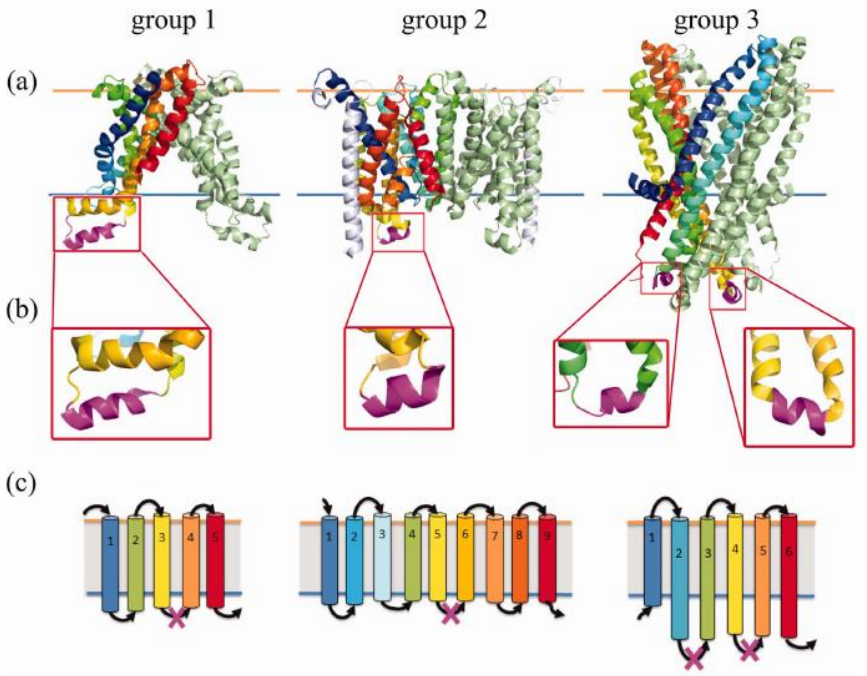


Figure 3

Three groups of ABC transporter TMD structures in *Ca.* L. asiaticus. Representative structure templates for three groups from left to right are 3dhw, 2qi9, and 3b60, respectively. (a) Structure overviews of transmembrane helices. Coupling helices are colored magenta. For each structure dimer, the left TMD is colored from N terminus to C terminus in rainbow; the right TMD is colored in pale green. In Group 2, the residues unaligned by any *Ca.* L. asiaticus sequences are colored in gray. Coupling helix regions are colored magenta and marked in red frames. (b) Enlarged views of the coupling helices. The coupling helices are colored magenta and reorientated for a better view. (c) Topologies of the representative structures colored in rainbow. The coupling helices are marked as magenta crosses. The topology diagram of group 2 does not include the gray N-terminal TMH in the structure. In (a) and (c), the inner membrane region is indicated between the orange and blue lines with the cytoplasm on the bottom. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

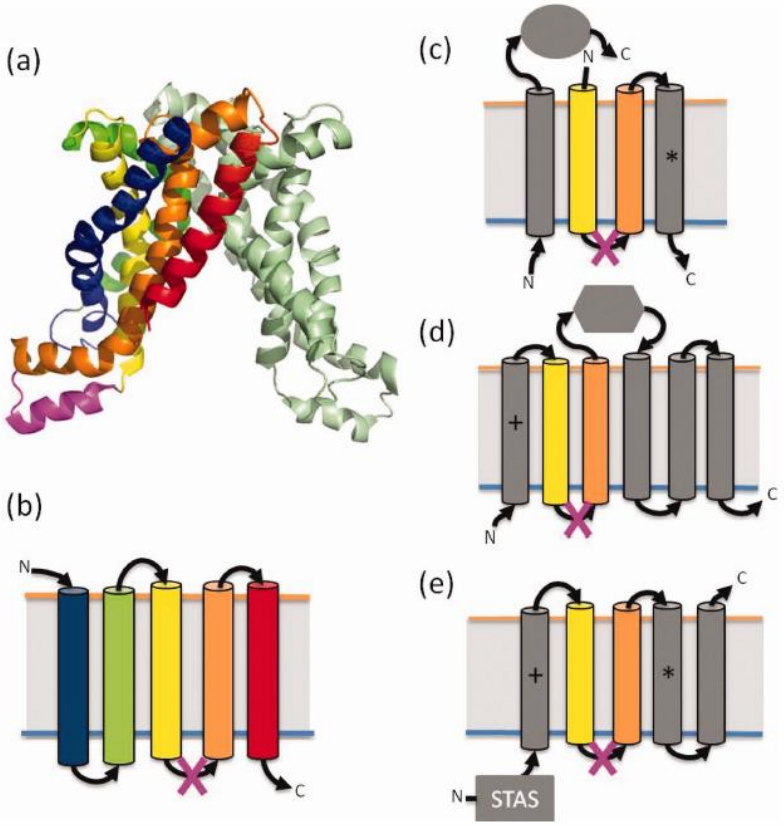


Figure 4

The predicted topology diagram of the TMDs that are related to Pfam clan BPD-like (CL0404). (**a**) The 3D structure representative (PDB: 3dhw) of TMDs in BPD family (PF00528); (**b**) the topology diagram of the TMDs in Pfam family BPD (PF00528) colored in rainbow; (**c**) the topology diagram of the TMD in *Ca.* L. asiaticus Lol system (FtsX family, PF02687); (**d**) the topology diagram of the TMDs in *Ca.* L. asiaticus Lpt system (YjgP/Q family, PF03739); (**e**) the topology diagram of the TMD in *Ca.* L. asiaticus Lin system (Permease family, PF02405). In (a), one domain is colored in pale green while the other is colored in rainbow. In (b–e), the cylinders represent the transmembrane helices. Letter N and C indicates the terminus of the proteins. Orange and blue lines define the membrane region with the cytoplasm on the bottom. In (c–e), nontransmembrane domains are represented by filled oval, hexagon, and rectangle, respectively. The coupling helix is colored magenta in the structure and marked by the magenta cross in the topology diagrams. The characteristic helices in HHsearch alignments are colored in yellow and orange. The extended TMHs in the HHsearch alignments between FtsX and Permease, YjgP/Q, and Permease are labeled by asterisks and plus symbols, respectively. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

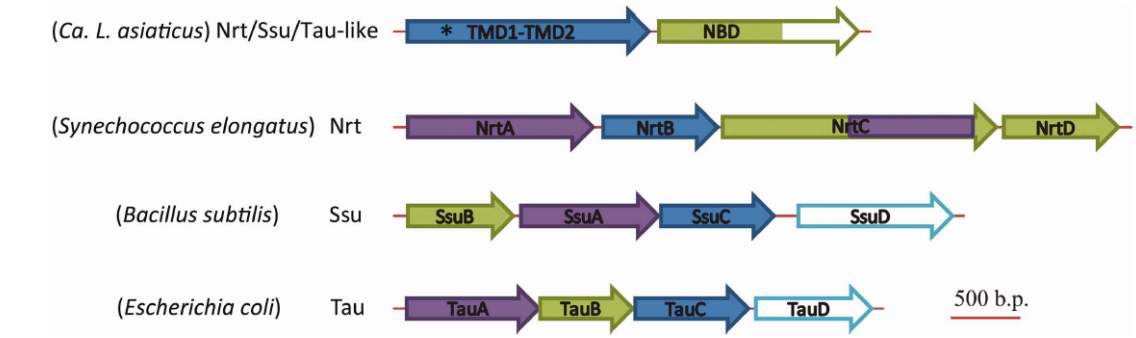


Figure 5

Operon structures of Nrt/Ssu/Tau-like system and its closely related systems. The abbreviation name of each system is shown on the left with the species in the parenthesis. Domain types are shown on the arrows for Nrt/Ssu/Tau-like system while the gene names are shown for the rest three systems. Each gene is shown as an arrow indicating the gene transcription direction. Genes marked with asterisk are proteins with two domains. The NBD, TMD, and PBP are colored green, blue, and purple, respectively. Only NBD, TMD, and PBP are filled. Other operon components not in transporters are denoted in cyan. The PBP in NrtC is a PBP homolog in cytoplasm.[60] No PBP is detected for Nrt/Ssu/Tau-like system. The scale is illustrated on the bottom right. The three experimental verified systems are from: *Nrt*: *Synechococcus elongatus* PCC 7942[59]; Ssu: *Bacillus subtilis* subsp. subtilis str. 168[61]; Tau: *Escherichia coli* str. K-12 substr. MG1655.[62] [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Table I
ABC Systems in *Ca.* L. asiaticus

		NBD					TMD		Other	
			Identified ortholog[b]							
Class[a]	Family[a]	Gene locus	Name	Species[c]	e-Value	Identity[d] (%)	Gene locus	Pfam family (clan)	Gene locus (type[e])	Function prediction or substrate specificity
III(i)	PAO	CLIBASIA_00280	AapP	*R. leguminosarum*	$1e^{-107}$	68	CLIBASIA_00275, CLIBASIA_00270	PF00528 (CL0404)	CLIBASIA_00265 (PP)	General L-amino acid
III(i)	MOI	CLIBASIA_02955	PstB	*E coli*	$7e^{-75}$	56	CLIBASIA_02960, CLIBASIA_02965	PF00528 (CL0404)	CLIBASIA_02970 (PP)	Phosphate
III(i)	MOI	CLIBASIA_02230	ThiQ	*S. enterica*	$8e^{-68}$	52	CLIBASIA_02235	PF00528 (CL0404)	CLIBASIA_02240 (PP)	Thiamine (vitamin B_1)
III(i)	OTCN	CLIBASIA_01125	ChoV	*S. meliloti*	$1e^{-114}$	60	CLIBASIA_01130	PF00528 (CL0404)	CLIBASIA_01135 (PP)	Choline (vitamin B_4)
III(i)	OTCN	CLIBASIA_02415	NrtD	*S. elongatus*	$5e^{-42}$	39	CLIBASIA_02420	PF00528 (CL0404)	N/A	Possible oxoacid ion
			SsuB	*B. subtilis*	$4e^{-41}$	38				
			TauB	*E coli*	$3e^{-39}$	42				
III(i)	MET	CLIBASIA_03025	ZnuC	*E coli*	$3e^{-59}$	45	CLIBASIA_03030	PF00950 (CL0142)	CLIBASIA_03020 (PP)	Zinc
III(i)	MET	CLIBASIA_02125	SitB	*S. typhimurium*	$6e^{-71}$	52	CLIBASIA_02130, CLIBASIA_02135	PF00950 (CL0142)	CLIBASIA_02120 (PP)	Manganese and iron
III(i)	MKL	CLIBASIA_00090	LinL	*S. japonicum*	$6e^{-62}$	49	CLIBASIA_00085	PF02405(N/A)	CLIBASIA_00095 (PP), CLIBASIA_00100 (PP)	Membrane lipid
III(e)	o228	CLIBASIA_03840	LolD	*E coli*	$9e^{-45}$	43	peg.788, peg.789[f]	PF02687 (CL0404)	CLIBASIA_03445 (PP)	Lipoprotein
III(e)	LPT	CLIBASIA_03155	LptB	*E coli*	$1e^{-80}$	51	CLIBASIA_01390, CLIBASIA_01395	PF03739 (CL0404)	CLIBASIA_03160 (PP), CLIBASIA_03165 (PP), CLIBASIA_01400 (OM)	Lipopolysaccharide
I	LIP	CLIBASIA_04080	MsbA	*E coli*	$5e^{-51}$	48	CLIBASIA_04080	PF00664 (CL0241)	N/A	Multidrug/lipid
I	LIP	CLIBASIA_00390	MsbA	*E coli*	$7e^{-52}$	45	CLIBASIA_00390	PF00664 (CL0241)	N/A	Multidrug/lipid
I	HMT	CLIBASIA_02315	AtmA	*C. metallidurans*	$5e^{-74}$	54	CLIBASIA_02315	PF00664 (CL0241)	N/A	Heavy metal
I	PRT	CLIBASIA_01350	PrtD	*S. meliloti*	$9e^{-49}$	26	CLIBASIA_01350	PF00664 (CL0241)	CLIBASIA_01355 (PP), CLIBASIA_04145 (OM)	Type I protein secretion
III(s)	ISB	CLIBASIA_04810	SufC	*E coli*	$5e^{-72}$	56	N/A	N/A	N/A	Fe-S assembly
II	ART	CLIBASIA_00790	ChvD	*A. tumefaciens*	0	77	N/A	N/A	N/A	Virulence gene regulation
II	ART	CLIBASIA_05125	Uup	*E coli*	$6e^{-90}$	35	N/A	N/A	N/A	Transposon excision regulation
II	UVR	CLIBASIA_00335	UvrA	*S. meliloti*	0	67	N/A	N/A	N/A	DNA repair
New		CLIBASIA_03185	MutS	*R. etli*	0	58	N/A	N/A	N/A	DNA repair
New		CLIBASIA_03285	RecF	*R. etli*	$3e^{-107}$	49	N/A	N/A	N/A	DNA repair
New		CLIBASIA_05400	RecN	*E coli*	$1e^{-71}$	31	N/A	N/A	N/A	DNA repair

[a]Classes and families are defined by existing NBD family classification.[15] In class III, (i), (e), and (s) mean importers, exporters, and soluble proteins, respectively.
[b]Closest BLAST hit with experimentally determined function.
[c]Abbreviations for species: *R. leguminosarum, Rhizobium leguminosarum; E. coli, Escherichia coli; S. enterica, Salmonella enterica; S. meliloti, Sinorhizobium meliloti; S. elongates, Synechococcus elongatus; B. subtilis, Bacillus subtilis; S. typhimurium, Salmonella typhimurium; S. japonicum, Sphingobium japonicum; C. metallidurans, Cupriavidus metallidurans; A. tumefaciens, Agrobacterium tumefaciens; R. etli, Rhizobium etli.*
[d]Sequence identity calculated according to *Ca.* L. asiaticus NBD for the transporters and *Ca.* L. asiaticus full-length protein for nontransport ATPases.
[e]Auxiliary/periplasmic components according to the cellular localization. PP, periplasmic protein (PBP or other periplasmic auxiliary proteins); OM, outer membrane protein.
[f]Proteins detected by the SEED but missed in NCBI database. They are encoded by neighboring genes in the genome and likely to function together.

REFERENCES

1. Husain MA ND. The citrus psylla (Diaphorina citri, Kuw.) [Psyllidae:Homoptera]. Memoirs of the Department of Agriculture in India 1927;10:1–27 1927.
2. KH L. Observations on yellow shoot disease. Acta Phytopathologica Sinica 1956;2:1–42 1956.
3. Teixeira DD, Danet JL, Eveillard S, Martins EC, Junior WCJ, Yamamoto PT, Lopes SA, Bassanezi RB, Ayres AJ, Saillard C, Bove JM. Citrus huanglongbing in Sao Paulo State, Brazil: PCR detection of the 'Candidatus' Liberibacter species associated with the disease. Molecular and Cellular Probes 2005;19(3):173-179.
4. Halbert SE. The discovery of huanglongbing in Florida. Proc of 2nd International Citrus Canker and Huanglongbing Research Workshop Florida Citrus Mutual 2005.
5. Bove JM, Ayres AJ. Etiology of three recent diseases of citrus in Sao Paulo State: sudden death, variegated chlorosis and huanglongbing. IUBMB Life 2007;59(4-5):346-354.
6. Gottwald TR. Current epidemiological understanding of citrus Huanglongbing. Annu Rev Phytopathol 2010;48:119-139.
7. Zhou L, Powell CA, Hoffman MT, Li W, Fan G, Liu B, Lin H, Duan Y. Diversity and plasticity of the intracellular plant pathogen and insect symbiont "Candidatus

Liberibacter asiaticus" as revealed by hypervariable prophage genes with intragenic tandem repeats. Appl Environ Microbiol 2011;77(18):6663-6673.

8. Duan Y, Zhou L, Hall DG, Li W, Doddapaneni H, Lin H, Liu L, Vahling CM, Gabriel DW, Williams KP, Dickerman A, Sun Y, Gottwald T. Complete genome sequence of citrus huanglongbing bacterium, 'Candidatus Liberibacter asiaticus' obtained through metagenomics. Mol Plant Microbe Interact 2009;22(8):1011-1020.

9. Doddapaneni H, Liao H, Lin H, Bai X, Zhao X, Civerolo EL. Comparative Phylogenomics and Multi-gene cluster analyses of the Citrus Huanglongbing (HLB)-associated bacterium Candidatus Liberibacter. Phytopathology 2008;98(6):S47-S47.

10. Kim JS, Sagaram US, Burns JK, Li JL, Wang N. Response of sweet orange (Citrus sinensis) to 'Candidatus Liberibacter asiaticus' infection: microscopy and microarray analyses. Phytopathology 2009;99(1):50-57.

11. Cevallos-Cevallos JM, Rouseff R, Reyes-De-Corcuera JI. Untargeted metabolite analysis of healthy and Huanglongbing-infected orange leaves by CE-DAD. Electrophoresis 2009;30(7):1240-1247.

12. Trivedi P, Duan Y, Wang N. Huanglongbing, a systemic disease, restructures the bacterial community associated with citrus roots. Appl Environ Microbiol 2010;76(11):3427-3436.

13. Davidson AL, Dassa E, Orelle C, Chen J. Structure, function, and evolution of bacterial ATP-binding cassette systems. Microbiol Mol Biol Rev 2008;72(2):317-364.

14. Dassa E, Bouige P. The ABC of ABCs: a phylogenetic and functional classification of ABC systems in living organisms. Research in Microbiology 2001;152(3-4):211-229.

15. Dassa E. Natural history of ABC systems: not only transporters. Essays Biochem 2011;50(1):19-42.

16. Walker JE, Saraste M, Runswick MJ, Gay NJ. Distantly related sequences in the alpha- and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. EMBO J 1982;1(8):945-951.

17. Ambudkar SV, Kim IW, Xia D, Sauna ZE. The A-loop, a novel conserved aromatic acid subdomain upstream of the Walker A motif in ABC transporters, is critical for ATP binding. FEBS Lett 2006;580(4):1049-1055.

18. Erkens GB, Berntsson RPA, Fulyani F, Majsnerowska M, Vujicic-Zagar A, ter Beek J, Poolman B, Slotboom DJ. The structural basis of modularity in ECF-type ABC transporters. Nature Structural & Molecular Biology 2011;18(7):755-U723.

19. Zhang P, Wang JW, Shi YG. Structure and mechanism of the S component of a bacterial ECF transporter. Nature 2010;468(7324):717-U148.

20. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer ELL, Eddy SR,

Bateman A. The Pfam protein families database. Nucleic Acids Res 2010;38:D211-D222.

21. Hollenstein K, Dawson RJP, Locher KP. Structure and mechanism of ABC transporter proteins. Curr Opin Struct Biol 2007;17(4):412-418.

22. Oldham ML, Chen J. Crystal Structure of the Maltose Transporter in a Pretranslocation Intermediate State. Science 2011;332(6034):1202-1205.

23. Daigle F, Fairbrother JM, Harel J. Identification of a mutation in the pst-phoU operon that reduces pathogenicity of an Escherichia coli strain causing septicemia in pigs. Infect Immun 1995;63(12):4924-4927.

24. Mantis NJ, Winans SC. The chromosomal response regulatory gene chvI of Agrobacterium tumefaciens complements an Escherichia coli phoB mutation and is required for virulence. J Bacteriol 1993;175(20):6626-6636.

25. von Kruger WMA, Humphreys S, Ketley JM. A role for the PhoBR regulatory system homologue in the Vibrio cholerae phosphate limitation response and intestinal colonization. Microbiology-(UK) 1999;145:2463-2475.

26. Garrido ME, Bosch M, Medina R, Llagostera M, de Rozas AMP, Badiola I, Barbe J. The high-affinity zinc-uptake system znuA CB is under control of the iron-uptake regulator (fur) gene in the animal pathogen Pasteurella multocida. FEMS Microbiol Lett 2003;221(1):31-37.

27. Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, de Crecy-Lagard V, Diaz N, Disz T, Edwards R, Fonstein M, Frank ED, Gerdes S, Glass EM, Goesmann A, Hanson A, Iwata-Reuyl D, Jensen R, Jamshidi N,

Krause L, Kubal M, Larsen N, Linke B, McHardy AC, Meyer F, Neuweger H, Olsen G, Olson R, Osterman A, Portnoy V, Pusch GD, Rodionov DA, Ruckert C, Steiner J, Stevens R, Thiele I, Vassieva O, Ye Y, Zagnitko O, Vonstein V. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res 2005;33(17):5691-5702.

28. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA. The COG database: an updated version includes eukaryotes. Bmc Bioinformatics 2003;4.

29. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215(3):403-410.

30. Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25(17):3389-3402.

31. Soding J. Protein homology detection by HMM-HMM comparison. Bioinformatics 2005;21(7):951-960.

32. Marchler-Bauer A, Panchenko AR, Shoemaker BA, Thiessen PA, Geer LY, Bryant SH. CDD: a database of conserved domain alignments with links to domain three-dimensional structure. Nucleic Acids Res 2002;30(1):281-283.

33. Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Lu F, Marchler GH, Mullokandov M,

Omelchenko MV, Robertson CL, Song JS, Thanki N, Yamashita RA, Zhang D, Zhang N, Zheng C, Bryant SH. CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res 2011;39(Database issue):D225-229.

34. Frickey T, Lupas A. CLANS: a Java application for visualizing protein families based on pairwise similarity. Bioinformatics 2004;20(18):3702-3704.

35. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P, Jensen LJ, von Mering C. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res 2011;39:D561-D568.

36. Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, Shen M-Y, Pieper U, Sali A. Comparative protein structure modeling using MODELLER. Current protocols in protein science / editorial board, John E Coligan [et al] 2007;Chapter 2:Unit 2.9.

37. Krogh A, Larsson B, von Heijne G, Sonnhammer ELL. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. Journal of Molecular Biology 2001;305(3):567-580.

38. von Heijne G. Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. J Mol Biol 1992;225(2):487-494.

39. Tusnady GE, Simon I. Principles governing amino acid composition of integral membrane proteins: Application to topology prediction. Journal of Molecular Biology 1998;283(2):489-506.

40. Jones DT. Improving the accuracy of transmembrane protein topology prediction using evolutionary information. Bioinformatics 2007;23(5):538-544.

41. Nugent T, Jones DT. Transmembrane protein topology prediction using support vector machines. Bmc Bioinformatics 2009;10.

42. Kall L, Krogh A, Sonnhammer ELL. A combined transmembrane topology and signal peptide prediction method. Journal of Molecular Biology 2004;338(5):1027-1036.

43. Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nature Methods 2011;8(10):785-786.

44. Pei J, Kim B-H, Grishin NV. PROMALS3D: a tool for multiple protein sequence and structure alignments. Nucleic Acids Res 2008;36(7):2295-2300.

45. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Systematic Biology 2010;59(3):307-321.

46. Le SQ, Gascuel O. An improved general amino acid replacement matrix. Mol Biol Evol 2008;25(7):1307-1320.

47. Anisimova M, Gascuel O. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. Syst Biol 2006;55(4):539-552.

48. Wilson CA, Kreychman J, Gerstein M. Assessing annotation transfer for genomics: Quantifying the relations between protein sequence, structure and

function through traditional and probabilistic scores. Journal of Molecular Biology 2000;297(1):233-249.

49. Rees DC, Johnson E, Lewinson O. ABC transporters: the power to change. Nat Rev Mol Cell Biol 2009;10(3):218-227.

50. Ward A, Reyes CL, Yu J, Roth CB, Chang G. Flexibility in the ABC transporter MsbA: Alternating access with a twist. Proc Natl Acad Sci U S A 2007;104(48):19005-19010.

51. Oldham ML, Davidson AL, Chen J. Structural insights into ABC transporter mechanism. Curr Opin Struct Biol 2008;18(6):726-733.

52. Okuda S, Tokuda H. Model of mouth-to-mouth transfer of bacterial lipoproteins through inner membrane LolC, periplasmic LolA, and outer membrane LolB. Proc Natl Acad Sci U S A 2009;106(14):5877-5882.

53. Suits MD, Sperandeo P, Deho G, Polissi A, Jia Z. Novel structure of the conserved gram-negative lipopolysaccharide transport protein A and mutagenesis analysis. J Mol Biol 2008;380(3):476-488.

54. Sharma AK, Rigby AC, Alper SL. STAS Domain Structure and Function. Cell Physiol Biochem 2011;28(3):407-422.

55. Babu M, Greenblatt JF, Emili A, Strynadka NC, Reithmeier RA, Moraes TF. Structure of a SLC26 anion transporter STAS domain in complex with acyl carrier protein: implications for E. coli YchM in fatty acid metabolism. Structure 2010;18(11):1450-1462.

56. Walshaw DL, Poole PS. The general L-amino acid permease of Rhizobium leguminosarum is an ABC uptake system that also influences efflux of solutes. Molecular Microbiology 1996;21(6):1239-1252.

57. Dupont L, Garcia I, Poggi MC, Alloing G, Mandon K, Le Rudulier D. The Sinorhizobium meliloti ABC transporter Cho is highly specific for choline and expressed in bacteroids from Medicago sativa nodules. J Bacteriol 2004;186(18):5988-5996.

58. Omata T. Structure, function and regulation of the nitrate transport system of the cyanobacterium Synechococcus sp. PCC7942. Plant Cell Physiol 1995;36(2):207-213.

59. Omata T, Gohta S, Takahashi Y, Harano Y, Maeda S. Involvement of a CbbR homolog in low CO2-induced activation of the bicarbonate transporter operon in cyanobacteria. J Bacteriol 2001;183(6):1891-1898.

60. Koropatkin NM, Pakrasi HB, Smith TJ. Atomic structure of a nitrate-binding protein crucial for photosynthetic productivity. Proc Natl Acad Sci U S A 2006;103(26):9820-9825.

61. van der Ploeg JR, Cummings NJ, Leisinger T, Connerton IF. Bacillus subtilis genes for the utilization of sulfur from aliphatic sulfonates. Microbiology 1998;144 (Pt 9):2555-2561.

62. van der Ploeg JR, Weiss MA, Saller E, Nashimoto H, Saito N, Kertesz MA, Leisinger T. Identification of sulfate starvation-regulated genes in Escherichia

coli: a gene cluster involved in the utilization of taurine as a sulfur source. J Bacteriol 1996;178(18):5438-5446.

63. Endo R, Ohtsubo Y, Tsuda M, Nagata Y. Identification and characterization of genes encoding a putative ABC-type transporter essential for utilization of gamma-hexachlorocyclohexane in Sphingobium japonicum UT26. J Bacteriol 2007;189(10):3712-3720.

64. Malinverni JC, Silhavy TJ. An ABC transport system that maintains lipid asymmetry in the gram-negative outer membrane. Proc Natl Acad Sci U S A 2009;106(19):8009-8014.

65. Mohn WW, van der Geize R, Stewart GR, Okamoto S, Liu J, Dijkhuizen L, Eltis LD. The actinobacterial mce4 locus encodes a steroid transporter. J Biol Chem 2008;283(51):35368-35374.

66. Pandey AK, Sassetti CM. Mycobacterial persistence requires the utilization of host cholesterol. Proc Natl Acad Sci U S A 2008;105(11):4376-4380.

67. Benning C. Mechanisms of lipid transport involved in organelle biogenesis in plant cells. Annu Rev Cell Dev Biol 2009;25:71-91.

68. Kim K, Lee S, Lee K, Lim D. Isolation and characterization of toluene-sensitive mutants from the toluene-resistant bacterium Pseudomonas putida GM73. J Bacteriol 1998;180(14):3692-3696.

69. Narita S. ABC Transporters Involved in the Biogenesis of the Outer Membrane in Gram-Negative Bacteria. Biosci Biotechnol Biochem 2011;75(6):1044-1054.

70. Raetz CRH, Whitfield C. Lipopolysaccharide endotoxins. Annu Rev Biochem 2002;71:635-700.

71. Nagao K, Kimura Y, Mastuo M, Ueda K. Lipid outward translocation by ABC proteins. FEBS Lett 2010;584(13):2717-2723.

72. Narita S, Tanaka K, Matsuyama S, Tokuda H. Disruption of lolCDE, encoding an ATP-binding cassette transporter, is lethal for Escherichia coli and prevents release of lipoproteins from the inner membrane. J Bacteriol 2002;184(5):1417-1422.

73. Karow M, Georgopoulos C. The essential Escherichia coli msbA gene, a multicopy suppressor of null mutations in the htrB gene, is related to the universally conserved family of ATP-dependent translocators. Mol Microbiol 1993;7(1):69-79.

74. Sperandeo P, Pozzi C, Deho G, Polissi A. Non-essential KDO biosynthesis and new essential cell envelope biogenesis genes in the Escherichia coli yrbG-yhbG locus. Research in Microbiology 2006;157(6):547-558.

75. Ruiz N, Gronenberg LS, Kahne D, Silhavy TJ. Identification of two inner-membrane proteins required for the transport of lipopolysaccharide to the outer membrane of Escherichia coli. Proc Natl Acad Sci U S A 2008;105(14):5537-5542.

76. Reuter G, Janvilisri T, Venter H, Shahi S, Balakrishnan L, van Veen HW. The ATP binding cassette multidrug transporter LmrA and lipid transporter MsbA have overlapping substrate specificities. J Biol Chem 2003;278(37):35193-35198.

77. Mikolay A, Nies DH. The ABC-transporter AtmA is involved in nickel and cobalt resistance of Cupriavidus metallidurans strain CH34. Antonie Van Leeuwenhoek 2009;96(2):183-191.

78. Delepelaire P. Type I secretion in gram-negative bacteria. Biochim Biophys Acta-Mol Cell Res 2004;1694(1-3):149-161.

79. Linhartova I, Bumba L, Masin J, Basler M, Osicka R, Kamanova J, Prochazkova K, Adkins I, Hejnova-Holubova J, Sadilkova L, Morova J, Sebo P. RTX proteins: a highly diverse family secreted by a common mechanism. FEMS Microbiol Rev 2010;34(6):1076-1112.

80. Zhang Y, Bak DD, Heid H, Geider K. Molecular characterization of a protease secreted by Erwinia amylovora. J Mol Biol 1999;289(5):1239-1251.

81. Zhang DW, Graf GA, Gerard RD, Cohen JC, Hobbs HH. Functional asymmetry of nucleotide-binding domains in ABCG5 and ABCG8. J Biol Chem 2006;281(7):4507-4516.

82. Saini A, Mapolelo DT, Chahal HK, Johnson MK, Outten FW. SufD and SufC ATPase Activity Are Required for Iron Acquisition during in Vivo Fe-S Cluster Formation on SufB. Biochemistry 2010;49(43):9402-9412.

83. Liu ZY, Jacobs M, Schaff DA, McCullen CA, Binns AN. ChvD, a chromosomally encoded ATP-binding cassette transporter-homologous protein involved in regulation of virulence gene expression in Agrobacterium tumefaciens. J Bacteriol 2001;183(11):3310-3317.

84. Reddy M, Gowrishankar J. Characterization of the uup locus and its role in transposon excisions and tandem repeat deletions in Escherichia coli. J Bacteriol 2000;182(7):1978-1986.

85. Hopkins JD, Clements M, Syvanen M. New class of mutations in Escherichia coli (uup) that affect precise excision of insertion elements and bacteriophage Mu growth. J Bacteriol 1983;153(1):384-389.

86. Murat D, Bance P, Callebaut I, Dassa E. ATP hydrolysis is essential for the function of the Uup ATP-binding cassette ATPase in precise excision of transposons. J Biol Chem 2006;281(10):6850-6859.

87. Reddy M, Gowrishankar J. Identification and characterization of ssb and uup mutants with increased frequency of precise excision of transposon Tn10 derivatives: nucleotide sequence of uup in Escherichia coli. J Bacteriol 1997;179(9):2892-2899.

88. Lupas A, Van Dyke M, Stock J. Predicting coiled coils from protein sequences. Science 1991;252(5010):1162-1164.

89. Burgos Zepeda MY, Alessandri K, Murat D, El Amri C, Dassa E. C-terminal domain of the Uup ATP-binding cassette ATPase is an essential folding domain that binds to DNA. Biochim Biophys Acta 2010;1804(4):755-761.

90. Murat D, Goncalves L, Dassa E. Deletion of the Escherichia coli uup gene encoding a protein of the ATP binding cassette superfamily affects bacterial competitiveness. Res Microbiol 2008;159(9-10):671-677.

91. Antunes LC, Ferreira RB, Buckner MM, Finlay BB. Quorum sensing in bacterial virulence. Microbiology 2010;156(Pt 8):2271-2282.

92. McFarlane HE, Shin JJ, Bird DA, Samuels AL. Arabidopsis ABCG transporters, which are required for export of diverse cuticular lipids, dimerize in different combinations. Plant Cell 2010;22(9):3066-3075.

CHAPTER 5 CONSERVED EVOLUTIONARY UNITS IN THE HEME-COPPER OXIDASE SUPERFAMILY REVEALED BY NOVEL HOMOLOGOUS PROTEIN FAMILIES[4]

INTRODUCTION

Aerobic respiration has evolved to use oxygen, which produces about 16 times more adenosine triphosphates (ATPs) than anaerobic respiration.[1] This advantage likely facilitated some critical evolutionary steps, such as the origin of eukaryotes and the increase of body size.[2, 3] Heme-copper oxidases (HCOs) are membrane-bound enzyme complexes functioning in the terminal step of aerobic respiratory chains.[4-6] They catalyze the reduction of dioxygen to water using electrons transferred from cytochrome c or a quinol derivative. The released energy is coupled to the translocation of protons across the membrane to generate an electrochemical gradient that can be used for ATP synthesis. All HCOs possess a catalytic subunit, an integral membrane protein with 12 core transmembrane helices (TMHs). Six conserved histidines in the TMHs of the catalytic subunit coordinate three co-factors: a high-spin heme and a copper ion in the binuclear catalytic site, and an additional low-spin heme functioning in the electron transfer pathway.[7, 8] Two of the six histidines function as axial ligands to coordinate the

[4] This chapter was published as:
Pei J*, **Li W***, Kinch LN, Grishin N V. *Conserved evolutionary units in the heme-copper oxidase superfamily revealed by novel homologous protein families*. Protein Sci 2014;23(9):1220–1234.
* Authors contributed equally

low-spin heme, while the rest participate in the catalytic site, with three histidines positioning the copper on one side of the high-spin heme and one histidine serving as the axial ligand on the other side. The 12-TMH core structure of the HCO catalytic subunit displays three-fold rotational pseudosymmetry and distributes the two heme groups into two of the three proposed pseudosymmetric units,[9], [10] each of which consists of four TMHs.

HCOs from various organisms have been discovered. They differ in heme types, electron donors (such as cytochrome c and ubiquinol), proton transfer pathways, and subunit composition. Three major types of HCOs (A, B, and C) have been defined based on sequence and structural analyses.[5], [11] A-type HCOs include cytochrome c oxidases in mitochondria, and cytochrome c oxidases and quinol oxidases in many bacteria and some archaea.[12] B-type HCOs are mainly found in the archaeal phylum of Crenarchaeota and appear to use one proton pathway compared to two proton pathways in A-type HCOs. A- and B-type HCOs share a conserved tyrosine residue residing in the sixth TMH. This tyrosine is covalently linked to a copper-binding histidine and was proposed to donate a fourth electron to the binuclear center in the catalytic process.[13] C-type HCOs, mainly from Proteobacteria, use a catalytic tyrosine residue located in a structurally different position (the seventh TMH) than that of A- and B-type HCOs.[14], [15]

Sequence and structural analyses revealed that the catalytic subunit of nitric oxide reductases (NORs) is homologous to that of the HCOs.[6], [16] NOR catalytic subunit also has 12 core TMHs sharing the same topology of HCO catalytic subunit and binds two hemes and a non-heme iron (instead of copper in HCOs) in a similar fashion by using six

conserved histidines. The HCO superfamily thus includes both HCOs and NORs. NORs catalyze the reduction of nitric oxide (NO) to nitrous oxide (N_2O) with the help of the heme groups and the non-heme iron in the denitrification pathway of the nitrogen cycle.[6, 17, 18] As NO is a toxic reactive agent, NORs in some pathogenic bacteria also play important roles in detoxifying exogenous NO generated by hosts.[19] Besides substrate preference, NORs differ from HCOs in that they do not translocate protons across the membrane and do not have a catalytic tyrosine due to fewer electrons required in one catalytic cycle. Two major subgroups of NORs have been described: the cytochrome c-dependent cNOR and the quinol-dependent qNOR.[20] NORs appear to be more closely related to C-type HCOs than A- and B-type HCOs in terms of sequence similarity and subunit composition.[21]

Although NORs and the three types of HCOs each form well-separated clades in the phylogeny reconstructed for the HCO superfamily, the position of the root remains controversial. Different evolutionary scenarios have been proposed for the origin and evolutionary order of HCOs and NORs. Several studies[18, 22-25] suggested that NORs may be more ancient than HCOs, consistent with the assumption that aerobic respiration evolved from denitrification after the emergence of atmospheric oxygen. Other researchers have proposed that the widely distributed A-type HCOs were present before the split of bacteria and archaea and are ancestors to B- and C-type HCOs and NORs.[12] These hypotheses, still in debate, explain how the 12 core TMHs developed into various types of oxidases. However, the origin of this pseudosymmetric helical architecture, an ancient event, is rarely discussed.

In this study, we used sensitive sequence similarity search methods such as transitive PSI-BLAST[26] searches and HHpred[27] to detect proteins homologous to the catalytic subunits of the HCO superfamily members. We called the newly found homologs HCO homology (HCOH) proteins. Interestingly, we discovered HCOH proteins with only four TMHs. These four-TMH proteins exhibit the highest similarity to the last four TMHs of HCOs (TMHs 9–12). They are considered to correspond to one evolutionary unit (EU) and are called single-EU HCOH proteins. Single-EU HCOH proteins may form homotrimers or heterotrimers to maintain the general structure and the ligand-binding sites defined by the fold of HCO/NOR catalytic subunits. HCO/NOR catalytic subunits are proposed to contain three homologous EUs made of TMHs 1–4, TMHs 5–8, and TMHs 9–12. We also discovered several groups of 12-TMH HCOH proteins that, like HCOs/NORs, contain three EUs. The majority of these three-EU HCOH proteins possess two conserved histidines that are predicted to bind a single heme. Most of the newly found remote homologs of HCOs/NORs are hypothetical proteins without experimental characterization. Only two of the seven major groups of HCOH proteins have been defined in current domain databases (DUF2871 and NnrS). Limited experimental studies and genomic context analysis suggest that they could function in the denitrification pathway and in the detoxification of reactive agents such as NO. Remarkably, the structural core of the three-EU assembly of HCOs/NORs resembles that of a diverse family of trimeric membrane-associated proteins in eicosanoid and glutathione metabolism (MAPEG).[28, 29]We propose the potential evolutionary scenarios

linking existing families, as well as the early evolutionary events of HCOs/NORs in aerobic respiration.

RESULTS AND DISCUSSION

Transitive PSI-BLAST[26] searches (see Materials and Methods) and HHpred[27] were used to detect proteins homologous to the catalytic subunits of the HCO superfamily members. A number of remote homologs of HCOs/NORs with different patterns of conserved histidines were discovered. We called these newly found superfamily members HCOH proteins and divided them into groups based on the number of TMHs in the homologous regions, patterns of conserved histidines, and the CLANS[30] sequence clustering results.

Four-TMH proteins homologous to HCOs/NORs help define EUs

The catalytic subunits of HCOs/NORs exhibit an approximate three-fold structural symmetry and are considered as a result of duplications of four-TMH units.[9, 10] In previous structure studies, three pseudosymmetric structural units (SUs) have been defined as TMHs 11/12/1/2, TMHs 3/4/5/6, and TMHs 7/8/9/10 [Fig. 1(A)] (TMHs 1–12 correspond to previously defined TMHs I–XII).[9, 10] We detected a set of four-TMH proteins homologous to the catalytic subunits of HCOs/NORs. Most of these four-TMH proteins possess the HxH motif at the beginning of the second TMH. These proteins exhibit the highest sequence similarity to the last four TMHs of HCOs/NORs

(TMHs 9/10/11/12). The HxH motif of these four-TMH proteins aligns to the HxH motif in the tenth TMH of the catalytic subunits of HCOs/NORs. We consider that these four TMHs correspond to an evolutionarily conserved unit and define them as one EU. Each of the four-TMH HCOH proteins possesses one EU and is thus called a single-EU HCOH protein. On the other hand, HCOs/NORs contain three EUs: TMHs 1/2/3/4 (EU1), TMHs 5/6/7/8 (EU2), and TMHs 9/10/11/12 (EU3) [Fig. 1(B,C)].

The second TMH in each of the three EUs in HCOs/NORs harbors conserved histidine(s) for heme or metal-binding, with characteristic three-residue motifs of Hxx, xxH, and HxH in EU1, EU2, and EU3, respectively (x: a variable residue) (Fig. 2). These motifs are homologous and occupy structurally equivalent positions in the superposition of EU1, EU2, and EU3. The third TMH of HCO/NOR EU2 additionally harbors a conserved HH motif (Fig. 2). The histidine in the Hxx motif of EU1 and the second histidine in the HxH motif of EU3 coordinate the low-spin heme in the ligand-binding pocket between EU1 and EU3 [Fig. 1(B,C)]. The histidines in the xxH motif and the HH motif of EU2 as well as the first histidine in the HxH motif of EU3 contribute to the binding of high-spin heme and copper/non-heme iron in the pocket between EU2 and EU3 [Fig. 1(B,C)].

Single-EU proteins with four TMHs may form homotrimers or heterotrimers

The single-EU proteins can be roughly divided into two major groups: HCOH-s1 (HCO homology proteins with a single EU, group 1) and HCOH-s2, according to CLANS sequence clustering results (Fig. 3) and sequence conservation.

HCOH-s1 (Fig. 3, red up triangles) consists of a main cluster (marked by A in Fig. 3) and several nearby small clusters (marked by letters B, C, D, E, F, G, and H in Fig. 3). HCOH-s1 proteins of the main cluster (cluster A) are mostly from Proteobacteria and Firmicutes. To maintain the structural compactness in the general fold of HCOs/NORs, these single-domain proteins most likely assemble as trimers. They could form homotrimers, since genomes of the main cluster of HCOH-s1 proteins contain only one single-EU protein. As this cluster of HCOH-s1 contains the HxH motif, three symmetric heme-binding sites at the interfaces of EUs can be inferred for the homotrimers [Fig. 1(D)], similar to the fashion of coordination of the low-spin heme by HCOs/NORs.

The B cluster (Fig. 3) of HCOH-s1 consists of closely related proteins encoded by gene pairs that are chromosomal neighbors [see three gene structure examples in Fig. 4(A–C)]. Interestingly, two proteins from the same species have the xxH and Hxx motifs, respectively (e.g., gi|183219809 and gi|183219810 in Fig. 2). These neighboring gene products could from heterotrimers, with the histidines contributing to the heme-binding site similar to the way the low-spin heme is coordinated in HCOs/NORs. Figure 1(E) depicts one possible way of forming a heterotrimer consisting of two proteins with the xxH motif and one protein with the Hxx motif. A single heme-binding site can be inferred for such a heterotrimer.

The remaining six small clusters (C□H) of the HCOH-s1 each have limited species distribution. HCOH-s1 proteins of clusters C and D contain the HxH motif and are from the Thermales order of the Deinococcus-Thermus phylum. In several species of the *Thermus*genus, a cluster C member and a cluster D member are products of genes not far away from each other, separated by gene clusters containing denitrification enzymes [see one example of Fig. 4(D)]. These proteins may form homotrimers or heterotrimers for genomes containing both C and D cluster members. On the other hand, two species (*Meiothermus silvanus* DSM 9946 and *Oceanithermus profundus* DSM 14977) have cluster C members and do not have cluster D members [see the example of Fig. 4(E)], suggesting that cluster C members can form homotrimers. HCOH-s1 proteins of clusters E and F, mostly from the *Thioalkalivibrio* genus, are also largely encoded by neighboring gene pairs [one example shown in Fig. 4(F)]. Although cluster E proteins have the HxH motif, cluster F proteins are characterized by the xxH motif. Products of these neighboring gene pairs may form heterotrimers. HCOH-s1 proteins of clusters G and H all possess the HxH motif and are encoded by pairs of genes in chromosomal vicinity [see two examples in Fig. 4(G,H)].

Members of the HCOH-s1 group have not been classified in publicly available domain databases, while the HCOH-s2 group corresponds to Pfam family DUF2871 and consists of proteins with unknown function. HCOH-s2 members have the conserved HxH motif in their second TMH. They likely form homotrimers with three symmetric sites that can coordinate three hemes [Fig. 1(D)]. Compared to HCOH-s1 proteins, HCOH-s2 proteins additionally possess a conserved “RE” motif at the end of the first TMH and a

conserved histidine in the fourth TMH (Fig. 2). HCOH-s2 proteins are mostly from bacterial phyla Firmicutes and Actinobacteria. Interestingly, several single-cell eukaryotes also possess HCOH-s2 proteins, including some species in the order of Trypanosomatida such as those of the *Leishmania* genus, *Angomonas deanei*, and *Strigomonas culicis*. Manual inspection of weak PSI-BLAST hits also revealed a divergent HCOH-s2 protein from *Naegleria gruberi* (gi|290974763, Fig. 2), a free-living single-cell eukaryotic species of the Heterolobosea class. HCOH-s2 may be present in the ancestor of eukaryotes, and its patchy phylogenetic distribution suggests that it may be lost independently in most eukaryotic lineages. *Leishmania* species are parasites for leishmaniasis, a disease that causes skin sores and visceral failure.[31, 32] *Leishmania* species only include the last three enzymes in the heme biosynthetic pathway.[33] Although *Leishmania* may be able to synthesize heme from heme precursors, it is thought to transport heme with an unknown mechanism[34, 35] and is uniquely dependent on the acquisition of exogenous heme for survival.[36, 37] Considering the membrane localization and potential heme-binding capability of HCOH-s2 proteins, *Leishmania* HCOH-s2 proteins might be involved in the maintenance of heme homeostasis in these parasites. One hypothesis about their function is that *Leishmania* HCOH-s2 proteins sequester hemes in the membrane to reduce heme toxicity[38] to the cell and increase heme accessibility to other membrane proteins.

The majority of HCOH-s1 and HCOH-s2 proteins consist of a single domain corresponding to one EU of four TMHs. As exceptions, one HCOH-s1 protein has an N-terminal divergent cupin_2 domain as suggested by HHpred (gi|91786010, Fig. 5), and

all F-cluster HCOH-s1 proteins contain an N-terminal thioredoxin domain (e.g., gi|220936290, Fig. 5). A small number of HCOH-s2 members have several additional TMHs (four or seven) (Fig. 5) in their N-termini that do not show detectable sequence similarity to the EUs of HCOs/NORs and HCOH proteins (based on PSI-BLAST and HHpred results).

HCOH groups with three EUs

We found a number of HCOH proteins with 12 TMHs (about 1500 proteins in the nre90 database). These HCOH proteins, like HCOs/NORs, consist of three EUs. However, they usually have fewer conserved histidines than HCOs/NORs. They form several clusters in the CLANS protein clustering diagram (Fig. 3). We divided them into seven groups: HCOH-t1 (HCO homology proteins with three EUs, group 1), HCOH-t2, HCOH-t3, HCOH-t4, HCOH-t5, HCOH-t6, and HCOH-t7.

HCOH-t1, HCOH-t2, and HCOH-t3 proteins possess two conserved histidine residues in motifs Hxx of EU1 and xxH of EU3 and do not have conserved histidines in EU2 [Figs. 1(F) and 2]. These proteins also possess a conserved arginine in the third TMH of EU3 (Fig. 2). The two conserved histidines correspond to the two residues in HCOs/NORs that bind the low-spin heme. HCOH-t1 proteins, mainly from Proteobacteria, correspond to the previously classified NnrS family (PF05940) in the Pfam database. The *nnrS* gene was identified as the neighboring gene of *nnrR* in *Rhodobacter sphaeroides* 2.4.1 and *R. sphaeroides* 2.4.3.[39] The *nnrR* gene

encodes a transcriptional regulator that responses to nitric oxide (NO) to activate the expression of the NOR gene *norB*. The expression of the *nnrS* gene is also dependent on the *nnrR* gene.[39] The STRING functional association server[40] suggests that the *nnrR* and *nnrS* genes co-occur with the *norB*-containing *nor*operon in various bacterial genomes, and they are often chromosomal neighbors, such as in the genome of *R. sphaeroides* 2.4.1 [Fig. 4(I)]. The purified NnrS protein appears to contain heme and copper.[41] Disruption of the *nnrS* gene affected taxis towards nitrate and nitrite, suggesting a role of NnrS in the denitrification process.[41]

Recent studies showed that NnrS contributes to NO resistance in the bacterial pathogen*Vibrio cholerae*.[42] NO is a host-generated reactive nitrogen species toxic to many bacterial pathogens such as *V. cholerae*. The *nnrS* gene in *V. cholerae* is up-regulated by the NorR transcriptional regulator in response to NO.[42] NorR also activates the expression of the *hmpA* gene that encodes a protein with nitric oxide dioxygenase (NOD) activity that turns NO to less toxic nitrogen oxides.[42] Gene disruption experiments suggest that *nnrS* and *hmpA*are important for *V. cholerae* colonization of intestines under the NO conditions, suggesting their roles in NO detoxification.[42] Unlike HmpA, *V. cholerae* NnrS does not remove NO, but it may protect cellular iron pool from NO damage.[43] The STRING server revealed that genes encoding NorR, HmpA, and NnrS are chromosomal neighbors in some bacteria, such as the opportunistic pathogen *Pseudomonas aeruginosa* [Fig. 4(J), *fhpR* and *fhp* encoding orthologous genes of *norR* and *hmpA* in *V. cholerae*, respectively), further supporting their functional association. The *nnrS* gene was also identified in a transposon

mutagenesis screen to be important in host colonization of *Neisseria meningitidis*, a bacterial pathogen that causes meningitis.[44]

The majority of HCOH-t2 proteins are from Actinobacteria and Proteobacteria. They also include some archaeal members mainly from the Crenarchaeota phylum. Most of the HCOH-t2 proteins are annotated as hypothetical proteins. A few HCOH-t2 proteins harbor additional domains. For example, the protein gi|408501395 has an additional TMH, a cupredoxin domain, and a copper-containing nitrite reductase domain C-terminal to the HCOH domain (Fig. 5). Such a domain composition suggests that HCOH-t2 may function in the denitrification process. Another protein, gi|392374446, has a different oxidoreductase domain (VKOR, vitamin K epoxide reductase) located N-terminally to the HCOH domain (Fig. 3). Genes encoding HCOH-t2 proteins are frequently found as neighbors of reductases in the denitrification process. For example, the HCOH-t2 gene is the neighbor to a NOR gene (*norB*) in *Burkholderia pseudomallei* K96243 [Fig. 4(K)] and to a copper-containing nitrite reductase gene in *Corynebacterium diphtheriae* NCTC 13129 [Fig. 4(N)]. In the hyperthermophilic crenarchaeon *Pyrobaculum aerophilum* str. IM2, the HCOH-t2 gene (*PAE3602*) is adjacent to a NOR gene (*norB*) and a nitrite reductase subunit (cytochrome D1) [Fig. 4(L)]. These three genes were all up-regulated after induction with nitrate.[45] The association with these reductase genes also suggests that HCOH-t2 may function in the denitrification process.

The HCOH-t3 group consists of 14 bacterial sequences and one archaeal sequence forming a loosely connected cluster in the CLANS diagram (Fig. 3). A few HCOH-t3 proteins also possess DUF1858 (Pfam: PF08984) and ScdA_N (Pfam: PF04405) domains

(Fig. 5). HHpred results suggest that DUF1858, a domain of unknown function, is distantly related to the ScdA_N domain. ScdA_N domain is named after the N-terminal domain of *Staphylococcus aureus* protein ScdA, which also contains the hemerythrin domain (Pfam: PF01814) that binds non-heme diirons.[46] The bacterial hemerythrin domain-containing RIC (repair of iron centers) family proteins, including ScdA from *S. aureus*, DnrN from *Neisseria gonorrhoeae* and YtfE from *Escherichia coli*, confer resistance to reactive nitrogen and oxygen molecules such as NO and H_2O_2 by repairing their damages to iron-sulfur centers.[47] The presence of the ScdA_N domain in many RIC proteins suggests that ScdA_N could aid in oxidative or nitrosative stress response. Such a domain in a few HCOH-t3 proteins indicates that they may also be involved in resistance to reactive nitrogen molecules such as NO, like some HCOH-t1 (NnrS) members. In *Solibacter usitatus* Ellin6076, the HCOH-t3 gene (*Acid_2923*) is in the neighborhood of genes of the RIC family and genes containing ScdA_N and DUF1858 domains [Fig. 4(M)], further supporting their functional associations.

HCOH-t4 proteins exhibit the histidine patterns of HxH, xxx, xxH in EU1, EU2, and EU3, respectively (Fig. 2). Similar to HCOH-t1, HCOH-t2, and HCOH-t3, such a pattern allows coordination of a heme group at the interface of EU1 and EU3. HCOH-t4 proteins are mainly from the Bacteroidetes phylum. A few of them contain a cytochrome c domain at the C-terminus (Fig. 5), suggesting that they may be involved in cytochrome c-dependent electron transfer. The HCOH-t4 gene is often located near the *nos*operon [one example shown in Fig. 4(O)] that encodes nitrous oxide (N_2O) reductase, which catalyzes the final step of the denitrification pathway: conversion of N_2O to dinitrogen. A

gene encoding the RIC family protein ScdA is also frequently found to be close to the HCOH-t4 gene [e.g., Fig. 4(O)], suggesting that HCOH-t4 could be involved in denitrification and detoxification of reactive molecules.

The majority of HCOH-t5 and HCOH-t6 proteins possess the xxH and Hxx motifs in the second TMH of EU1 and the second TMH of EU2, respectively, and lack conserved histidines in EU3 (Fig. 2). Such a pattern allows coordination of one heme group at the interface between EU1 and EU2 [Fig. 1(G)]. HCOH-t5 members are present in both bacteria and archaea. The bacterial members of HCOH-t5 are mainly from Proteobacteria and Firmicutes, while the archaeal members are all from the Halobacteria class of the Euryarchaeota phylum. The majority of HCOH-t5 proteins are annotated as hypothetical proteins and do not contain additional domains. As exceptions, a few HCOH-t5 proteins possess domains of unknown function such as DUF2249 (Pfam: PF10006) and DGC (Pfam: PF08859, a domain with four conserved cysteines that likely coordinate zinc). The STRING server revealed strong association of HCOH-t5 proteins with proteins containing DUF2249 domain, proteins with DUF59 domain, and RIC proteins based on evidence of gene neighborhood, gene fusion (in the case of DUF2249 domain), and gene co-occurrence. For example, the HCOH-t5 gene from *Rhizobium etli* CFN 42 (*RHE_PF00521*) is predicted to be functionally associated with two genes containing DUF2249 domains (*RHE_PF00520* and *RHE_PF00522*) and a gene with DUF59 domain (*RHE_PF00523*) [Fig. 4(P)]. These genes are also neighbors to *nnrS* (a HCOH-s1 gene), *nnrR*, and gene clusters encoding denitrification enzymes such as NOR and nitrite reductase [Fig. 4(O)].

HCOH-t6 proteins are mainly from Actinobacteria. Some HCOH-t6 proteins are annotated as “multicopper oxidase” or “nitrite reductase,” as they also possess a cupredoxin domain and a copper-containing nitrite reductase domain. Such a domain composition is similar to a few HCOH-t2 proteins described above (Fig. 5). Interestingly, HCOH-t6 and HCOH-t2 genes are often chromosomal neighbors, such as *DIP1877* (a HCOH-t6 gene) and *DIP1878* (a HCOH-t2 gene) of *Corynebacterium diphtheriae*NCTC 13129 [Fig. 4(N)]. In one case, HCOH-t2 and HCOH-t6 are fused together in one open reading frame (gi|493596372 from *Actinomyces urogenitalis*, Fig. 5). These observations suggest that HCOH-t2 and HCOH-t6 may have related functions.

HCOH-t7 proteins exhibit yet another pattern of histidine motifs with xxH in EU2 and Hxx in EU3, while lacking conserved histidines in EU1 (Fig. 2). Such a pattern would allow coordination of one heme group at the interface between EU2 and EU3 [Fig. 1(H)]. HCOH-t7 proteins form two small clusters in the CLANS diagram (Fig. 3). Each cluster has restricted phylogenetic distribution. They are from the bacterial phylum Aquificae and from the archaeal class Halobacteria of the Euryarchaeota phylum, respectively. These proteins are annotated as hypothetical proteins and do not have additional domains.

In summary, three-EU HCOH genes are often neighbors to genes involved in the denitrification process and/or detoxification of small reactive molecules such as NO (Fig. 4). Many of these neighboring genes encode various denitrification enzymes such as nitrate reductase, nitrite reductase, NOR, and nitrous oxide reductase (Fig. 4). The detoxification genes include those that encode NOD and the RIC family proteins with

hemerythrin domains. In addition, genes encoding a few domains of unknown functions, such as ScdA_N, DUF2249, DUF1858, and DUF59, are also frequently found in the neighborhood of three-EU HCOH genes. Three-EU HCOH proteins from different groups are sometimes gene neighbors (Fig. 4), suggesting that they are involved in the same biological process. The gene neighborhood associations are consistent with the domain contents of limited multi-domain three-EU HCOH proteins, as some of these proteins contain nitrite reductase domain, DUF2249 domain and DUF1858 domain (Fig. 5).

For some single-EU HCOH proteins such as HCOH-s2 (DUF2871) proteins and cluster A HCOH-s1 proteins, we did not find strong functional associations to denitrification/detoxification genes according to the results of the STRING server. On the other hand, genes encoding HCOH-s1 proteins of clusters B, C, D, E, and F are frequently neighbors to denitrification genes and potential detoxification genes such as those with the hemerythrin domain and the globin domain48 (Fig. 4(A–F)]. Like many three-EU HCOH genes, these single-EU HCOH genes often have neighbors with domains of unknown function such as ScdA_N, DUF2249 and DUF59 [Fig. 4(A–F)]. Other genes frequently in the vicinity of HCOH-s1 genes include those encoding proteins with cytochrome c domain (cyC), ferredoxin domain (fx), thioredoxin domain (Trx), and cupin_2 domain (cp2) [Fig. 4(A–F)]. Some three-EU HCOH genes are also found to be neighbors of HCOH-s1 genes [e.g., Fig. 4(B,C,E)].

Limited experimental studies on NnrS proteins, gene context analysis and domain content analysis indicate the involvement of many HCOH proteins in denitrification and

detoxification. The molecular mechanisms of their functions remain unknown and await further experimental investigations. As putative heme-binding proteins, HCOH proteins could contribute to denitrification and detoxification in several ways, such as maintenance of cellular iron and heme homeostasis, being part of the electron transfer pathways in various denitrification enzymes, binding/sequestering/exporting reactive nitrogen species, and possessing enzymatic activities that convert reactive nitrogen species to less toxic molecules.

Structure similarity between the catalytic subunits of HCOs/NORs and the MAPEG family proteins

After establishing the EUs for the HCO superfamily, we sought to explore its relationship to known structures. We queried the structures of HCO/NOR EUs against the SCOP[49] database using the HorA server,[50] which not only reports structural similarity, but also evaluates if the similarity represents homology or analogy. HorA identified a MAPEG structure as the top hit following the structures from the HCO superfamily, with relatively good scores for both structure comparison (Dali Z-score 7.6) and sequence comparison (HHpred probability 0.52), resulting in an overall score (5.042) that is consistent with scores derived for distant homologs.[51] MAPEG proteins are a group of membrane-bound enzymes with diverse functions, including glutathione transferase activity that provides protection from oxidative stress in the membrane.[28] The MAPEG fold, consisting of four TMHs, forms a homotrimer that binds three substrates in a symmetric manner. The four TMHs of a MAPEG subunit adopt the same topology as the four TMHs in each EU of HCOs/NORs, showing right-handed connections between them.

In addition, an unexpected structural similarity lies in the same arrangement of the three subunits of the MAPEG homotrimer compared to the three EUs in HCOs/NORs, as noticed before.[29] A superposition of a trimeric MAPEG structure (pdb id: 4al0) onto an HCO structure (pdb id: 3mk7) (Fig. 6) reveals a striking similarity (Dali Z-score: 20.6, RMSD: 3.3Å) covering all 12 TMHs. Each of the three EUs of HCO structure corresponds to one monomer of the MAPEG homotrimer (Fig. 6). Interestingly, the hemes from the HCO structure (red sticks, Fig. 6) overlap with the MAPEG substrate (glutathione, magenta sticks, Fig. 6), which is located at the interfaces of the monomers. Combined with the trimeric state required for MAPEG enzyme function,[52, 53] the structural similarity and similar active site position suggest that MAPEG proteins and HCOs/NORs are evolutionarily related.[29]

Evolutionary scenarios of HCOs/NORs and the HCOH proteins

Gene duplication, divergence, fusion, and fission are the main driving forces in the evolution of proteins with novel functions.[54-56] It is estimated that a large fraction of proteins form oligomers with functional importance.[57] Evolution of oligomeric protein complexes has drawn interest in both theoretic and experimental studies.[58, 59] Gene duplication is considered as the cause to generate heteromers (hetero-oligomers) from homomers (homo-oligomers), for example, in the evolution of proteasomes[60] and chaperonins.[61] Subunits in a heteromer have different sequences and structures that break the symmetry of homomers and allow more versatile functions.[62] It is estimated

that a large fraction of membrane proteins form oligomers or are internally pseudosymmetric,[63], [64] like the catalytic subunits of HCOs/NORs.

The structural similarity between the catalytic subunits of HCOs/NORs and the MAPEG trimers suggests that HCOs/NORs could have evolved from a four-TMH ancestral protein that forms MAPEG-like homotrimers (Fig. 7). It is likely that a fortuitous sequence divergence event of this ancestor [Fig. 7(A)] gave rise to the HxH motif in the second TMH that enabled binding of heme groups. The discovery of single-EU HCOH proteins with four TMHs and the HxH motif supports this hypothesis. The homotrimers of four-TMH single-EU proteins with the HxH motif would have three symmetric heme-binding sites located at the interfaces of the monomers, similar to those of the MAPEG proteins. Single-EU HCOH genes could have undergone gene duplications [Fig. 7(B)] followed by sequence divergence to generate multiple copies of single-EU HCOH genes encoding proteins that form heterotrimers. If the HxH motif is kept in all duplicated genes, a heterotrimer with three heme-binding sites could be formed. In case some histidines are mutated [Fig. 7(C)], like the HCOH-s1 cluster B proteins, heterotrimers that bind less than three hemes could evolve. The fusion of three single-EU genes would result in an open reading frame with three EUs [Fig. 7(D,F)]. Three-EU proteins could bind three hemes with six histidines if the HxH motif is kept in all three EUs [Fig. 7(D)]. Interestingly, we identified one archaeal protein in the HCOH-t5 group (gi|288930450 in Fig. 2, marked by a star in Fig. 3) that has three HxH motifs and could bind three hemes. Deterioration of some histidines, either in the stage of single-EU ancestors [Fig. 7(C)] or three-EU ancestors [Fig. 7(E)], could lead to three-EU HCOH

proteins with fewer than six conserved histidines and thus less than three heme-binding sites.

Most of extant three-EU HCOH proteins possess two conserved histidines and are inferred to bind a single heme. They mostly follow one of three histidine patterns: Hxx.xxx.xxH (HCOH-t1, HCOH-t2, HCOH-t3, and HCOH-t4), xxH.Hxx.xxx (HCOH-t5 and HCOH-t6) and xxx.xxH.Hxx (HCOH-t7) (Fig. 7). These three patterns can also be related by circular permutations of the three EUs [Fig. 7(G)]. We also identified a small number of archaeal HCOH-t5 proteins (e.g., gi|389847617 in Fig. 2) that contain a combination of patterns of HCOH-t1 and HCOH-t5 and have four histidines in the motif pattern of HxH.Hxx.xxH (Fig. 7). These proteins could thus bind two hemes.

HCOs/NORs follow the histidine pattern of Hxx.xxH.HH.HxH. The HH motif in the seventh TMH (the third TMH in EU2) is a unique feature of HCOs/NORs that is not present in HCOH proteins. Such an addition allows HCOs/NORs to coordinate a copper or non-heme iron in addition to binding two hemes. We also identified several small groups of proteins closely related to HCOs/NORs that have some or all of the conserved histidines deteriorated. Two small groups (light blue dots in Fig. 3) have the HH motif deteriorated. Four small groups (yellow dots in Fig. 3) have all of the six histidines deteriorated. These groups could have evolved from the HCOs/NORs by sequence divergence [Fig. 7(H)].

Although it is likely that three-EU HCOH proteins have evolved by gene duplication and fusion of single-EU HCOH ancestors, the opposite evolutionary scenario, gene split (or fission) of three-EU HCOH genes to generate single-EU HCOH open

reading frames [Fig. 7(I)], is also plausible. For example, HCOH-s1 proteins of clusters H and G, especially cluster H proteins, are close to the HCOs/NORs in the CLANS diagram (Fig. 3). Their top BLAST hits include C-type HCOs. In the genome of *Sideroxydans lithotrophicus* ES-1, one cluster G HCOH-s1 gene and one cluster H HCOH-s1 gene are close to each other and neighboring to a C-type HCO (Fig. 4). It is likely that HCOH-s1 proteins of clusters G and H are derived from C-type HCO proteins by gene fission.

CONCLUSIONS

Comparative sequence-structure analysis revealed novel homology between a number of HCOH proteins and the catalytic subunits of HCOs/NORs. HCOH proteins form groups of four-TMH single-EU proteins (HCOH-s1 and HCOH-s2) and groups of 12-TMH three-EU proteins (HCOH-t1, HCOH-t2, HCOH-t3, HCOH-t4, HCOH-t5, HCOH-t6, and HCOH-t7). Among these groups, only HCOH-s2 and HCOH-t1 correspond to known domains (DUF2871 and NnrS, respectively), while the majority of other HCOH members are currently annotated as hypothetical proteins without known domains. Gene context and domain content analyses, coupled with limited experimental studies of NnrS, suggest that most HCOH proteins are involved in the denitrification process and/or detoxification of reactive small molecules. Based on the structures of HCOs/NORs, single-EU HCOH proteins could form homotrimers or heterotrimers with active sites located at the interfaces between monomers. Conserved histidines in HCOH

proteins indicate that they can bind heme. Strong structural similarity was observed between the homotrimers of the MAPEG family membrane enzymes and the catalytic subunits of HCOs/NORs. Such a similarity, together with the discovery of single-EU HCOH proteins, suggests that HCOs/NORs and three-EU HCOH proteins could have evolved from four-TMH ancestors that form homotrimers similar to MAPEG proteins. Gene duplication, sequence divergence, and gene fusion of ancestral single-EU HCOH proteins could give rise to three-EU HCOH proteins and HCOs/NORs. Conversely, gene fission of three-EU HCOH proteins or HCOs/NORs may have produced some extant single-EU HCOH proteins.

MATERIALS AND METHODS

Sequence similarity searches

PSI-BLAST[26] iterations were conducted to search for homologs of the HCO superfamily proteins starting from one representative with known structure (protein databank (PDB[65]) id: 3o0r, chain B)[16] against a database composed of NCBI non-redundant proteins and environmental sequences with maximal 90% identity (nre90) protein database (e-value inclusion cutoff: 1e-4). To perform transitive searches, PSI-BLAST hits were grouped by BLASTCLUST (with the score coverage threshold [−S, defined as the bit score divided by alignment length) set to 1, length coverage threshold (−L) set to 0.5, and no requirement of length coverage on both sequences (−b F)], and a

representative sequence from each group was used to initiate new PSI-BLAST searches. Such an iterative procedure was repeated until convergence. HHpred[27] was used for profile-profile-based similarity searches to identify distant homologous relationships for (1) HCOs/NORs (gi|315583520), (2) the HCOH-s1 group (gi|499132825), (3) the HCOH-s2 group (gi|316941303), (4) the HCOH-t1 group (gi|110347088), (5) the HCOH-t5 group (gi|292656262), and (6) the HCOH-t6 group (gi|256378768) (profile databases used: Pfam,[66] PDB,[65] and CDD[67]). Detections of conserved domains are performed by the CDD server[67] and the HMMER3 package.[68] We also employed the HorA server[50] to detect structural homologs for the pseudosymmetric units of HCO proteins, using the C terminus of a NOR structure (pdb: 3o0r, chain B, residue 302–458) as input.

Sequence clustering and multiple sequence alignment

Sequence clustering was performed and visualized by the CLANS program.[30] Several cutoffs of *P*-values were tried. The *P*-value cutoff 1e-10 was chosen since it gave the best separation between clusters according to manual inspections. We extracted the sequences in each manually defined group of CLANS results and performed multiple sequence alignments by PROMALS3D[69] for each group. Representative sequences for HCOs/NORs and the newly defined HCOH groups were selected and split into individual EUs of four TMHs. A multiple sequence alignment was constructed for the EUs of these representatives by PROMALS3D. This alignment was manually

adjusted by taking into account sequence conservation, structural superposition of known structures of HCOs/NORs by DaliLite[70] and MUSTANG,[71] hydrophobicity and small residue patterns, and transmembrane regions predictions by Phobius[72] and TMHMM.[73]

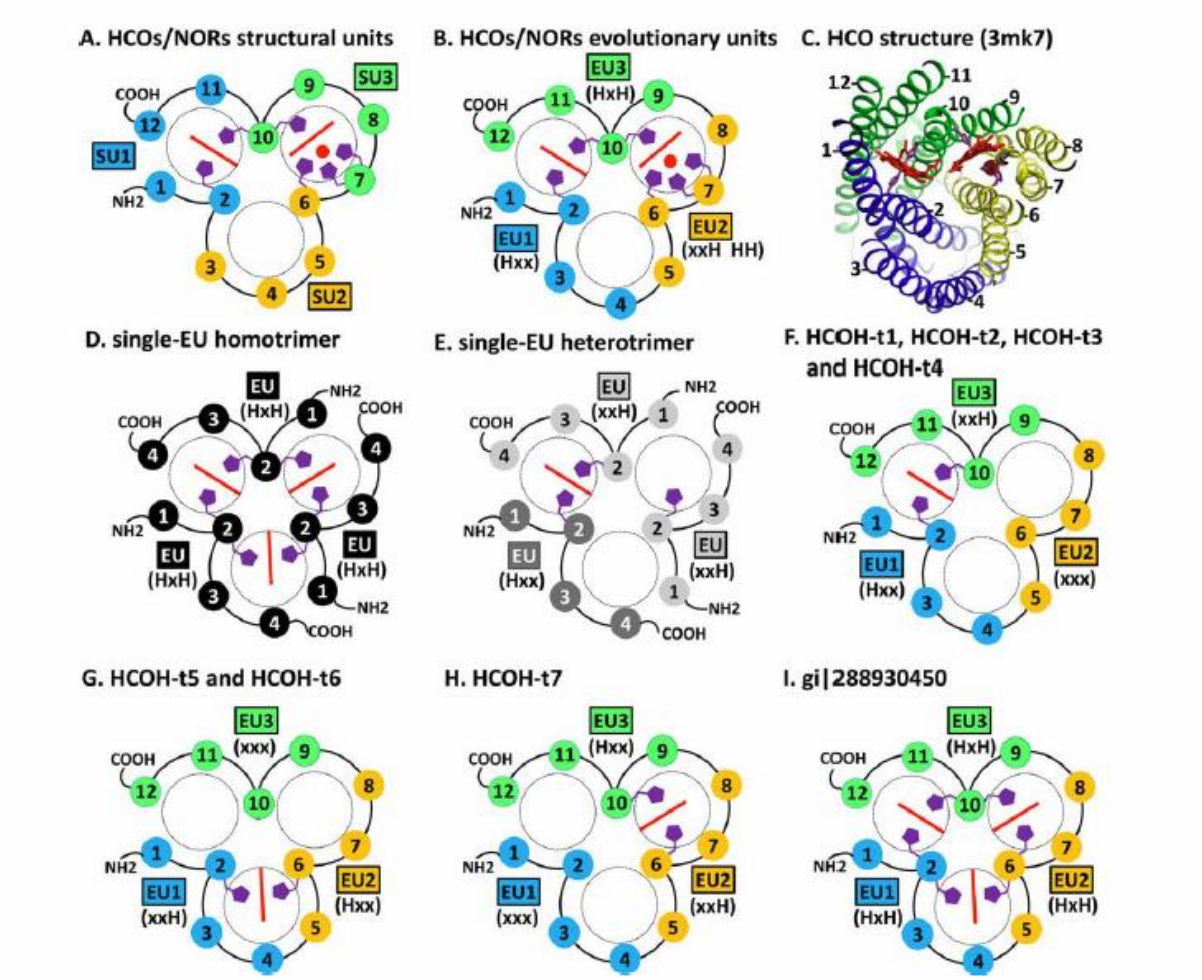


Figure 1. TMH topology diagrams of HCOs/NORs and HCOH proteins. TMHs are shown in numbered circles. TMHs in the same structure unit (marked by SU) or the same evolutionary unit (marked by EU) are filled with the same color. The histidine patterns of the EUs are shown in parentheses. Hemes are shown as red lines. Copper or non-heme irons in HCOs/NORs are shown as red spheres. Conserved histidine sidechains are shown in purple. N- and C-termini of each modeled protein are marked by NH2 and COOH, respectively. **(A)**. Previously defined structural units of HCOs/NORs. **(B)**. Newly defined EUs of HCOs/NORs. **(C)**. Structure of a C-type HCO catalytic subunit (pdb: 3mk7). **(D)**. Model of homotrimer for single-EU proteins with the HxH motif. **(E)**. Model of heterotrimer for single-EU proteins with xxH and Hxx motifs. **(F)**. Model for HCOH-t1, HCOH-t2, HCOH-t3 and HCOH-t4. **(G)**. Model for HCOH-t5 and HCOH-t6. **(H)**. Model for HCOH-t7. **(I)**. Model for the protein (gi|288930450) with the HxH motif in all three EUs.

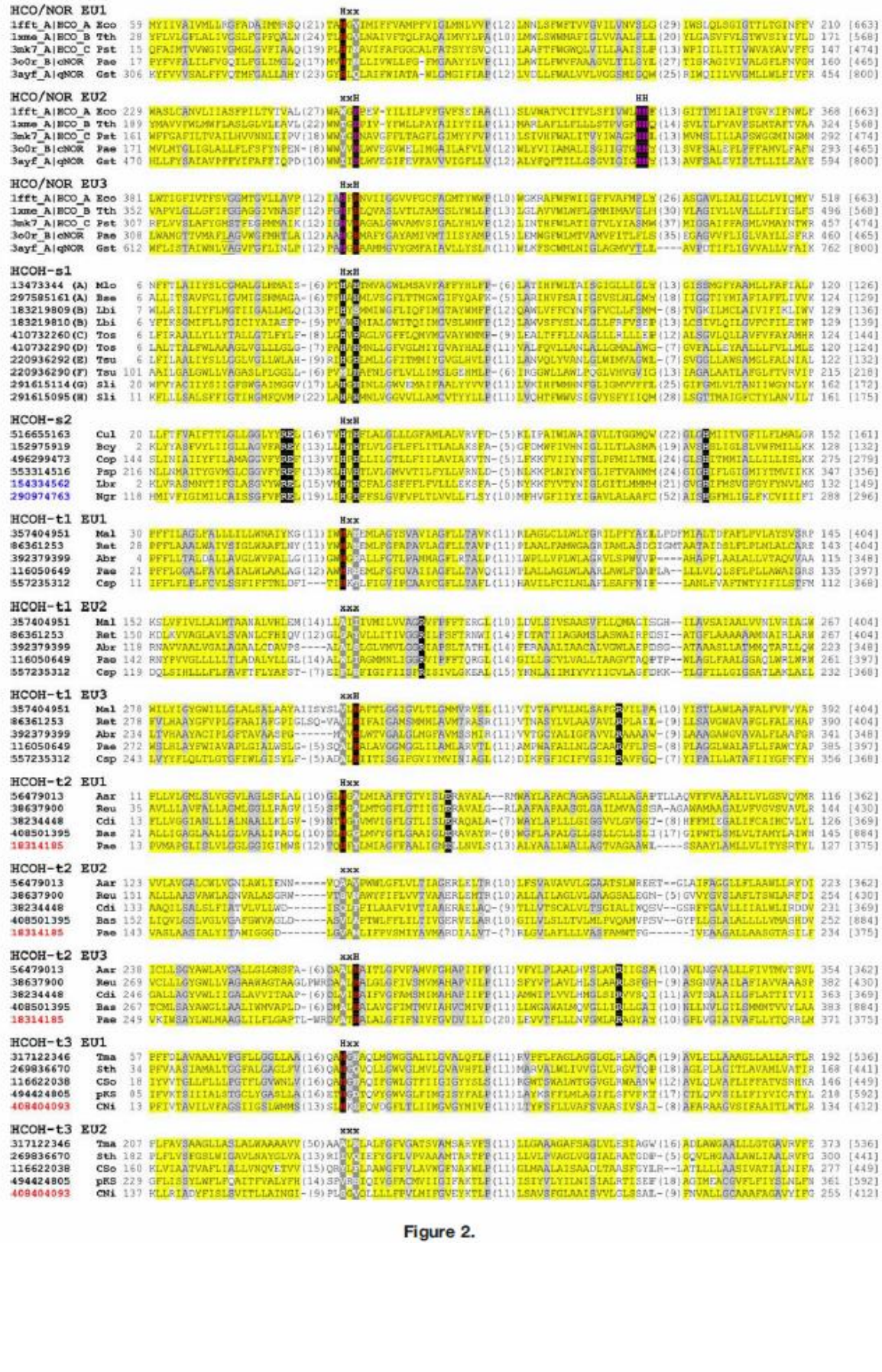


Figure 2.

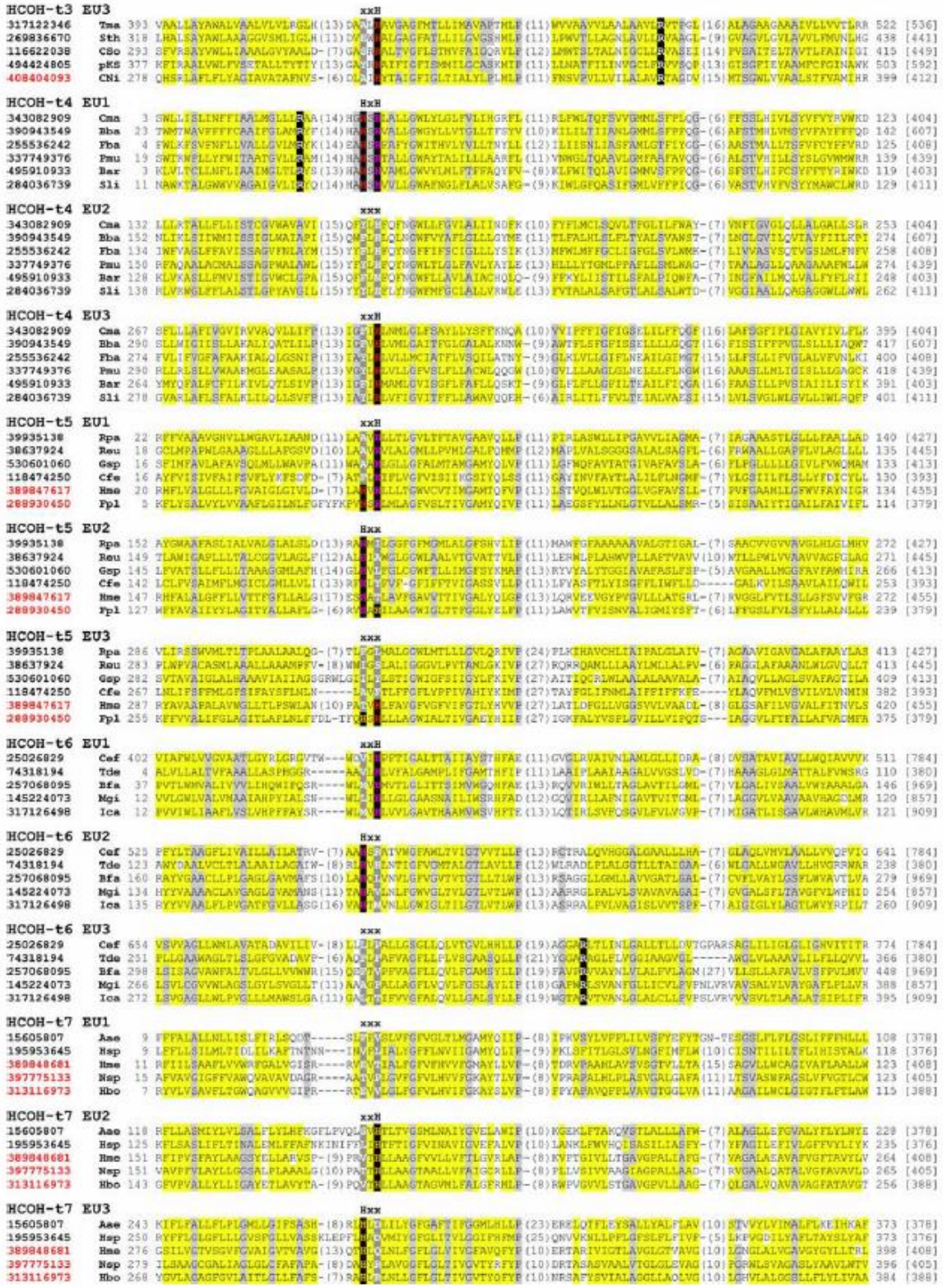


Figure 2. Continued

Figure 2. Multiple sequence alignment of EUs of HCOs/NORs and HCOH proteins. For HCOs/NORs, the sequences are denoted by their pdb id and chain id, followed by type name. For HCOH proteins, each sequence is denoted by its NCBI gene identification (gi) number. The cluster type for HCOH-s1 is shown in parentheses after gi numbers. Starting and ending residues numbers are shown before and after the sequences, respectively. Protein lengths are shown in brackets at the end. Positions with conserved motifs in the second TMHs of these EUs are marked with HxH, xxH, Hxx, or xxx. The conserved HH motif in the second EU of HCOs/NORs is also marked. Conserved histidines in these positions are in black background. For three-EU proteins, conserved histidines at the interface between EU1 and EU3 are colored red, those at the interface between EU2 and EU3 are colored orange, and those at the interface between EU1 and EU2 are colored magenta. Other family-specific conserved residues are also highlighted in black background. Non-charged residues in positions with mainly hydrophobic residues are shaded in yellow. Small residues (G,A,S,C,T,P, and V) in positions with mainly small residues are shaded in grey. Insertion regions are replaced by the number of inserted residues in parentheses or omitted in between underscored letters. Three-letter species name abbreviations shown before the residue starting numbers are as follows: Aae, *Aquifex aeolicus*; Aar, *Aromatoleum aromaticum*; Abr, *Azospirillum brasilense*; Bcy, *Bacillus cytotoxicus*; Bse, *Bacillus selenitireducens*; Bba, *Belliella baltica*; Bas, *Bifidobacterium asteroides*; Bar, *Bizionia argentinensis*; Bfa, *Brachybacterium faecium*; Cfe, *Campylobacter fetus*; Csp, *Campylobacter* sp.; CNi, *Candidatus Nitrososphaera gargensis*; CSo, *Candidatus Solibacter usitatus*; Cop, *Coprobacillus*; Cdi, *Corynebacterium diphtheriae*; Cef, *Corynebacterium efficiens*; Cul, *Corynebacterium ulceribovis*; Cma, *Cyclobacterium marinum*; Eco, *Escherichia coli*; Fpl, *Ferroglobus placidus*; Fba, *Flavobacteriaceae bacterium*; Gsp, *Geobacillus* sp.; Gst, *Geobacillus stearothermophilus*; Hme, *Haloferax mediterranei*; Hbo, *Halogeometricum borinquense*; Hsp, *Hydrogenobaculum* sp.; Ica, *Intrasporangium calvum*; Lbr, *Leishmania braziliensis*; Lbi, *Leptospira biflexa*; Mlo, *Mesorhizobium loti*; Mal, *Methylomicrobium alcaliphilum*; Mgi, *Mycobacterium gilvum*; Ngr, *Naegleria gruberi*; Nsp, *Natrinema* sp.; pKS, *planctomycete* KSU-1; Pmu, *Paenibacillus mucilaginosus*; Psp, *Peptoniphilus* sp.; Pae, *Pseudomonas aeruginosa*; Pst, *Pseudomonas stutzeri*; Pae, *Pyrobaculum aerophilum*; Reu, *Ralstonia eutropha*; Ret, *Rhizobium etli*; Rpa, *Rhodopseudomonas palustris*; Sli, *Sideroxydans lithotrophicus*; Sth, *Sphaerobacter thermophilus*; Sli, *Spirosoma linguale*; Tma, *Thermaerobacter marianensis*; Tos, *Thermus oshimai*; Tth, *Thermus thermophilus*; Tsu, *Thioalkalivibrio sulfidophilus*; Tde, *Thiobacillus denitrificans*. Gi numbers of bacterial, archaeal, and eukaryotic proteins are in black, red, and blue colors, respectively.

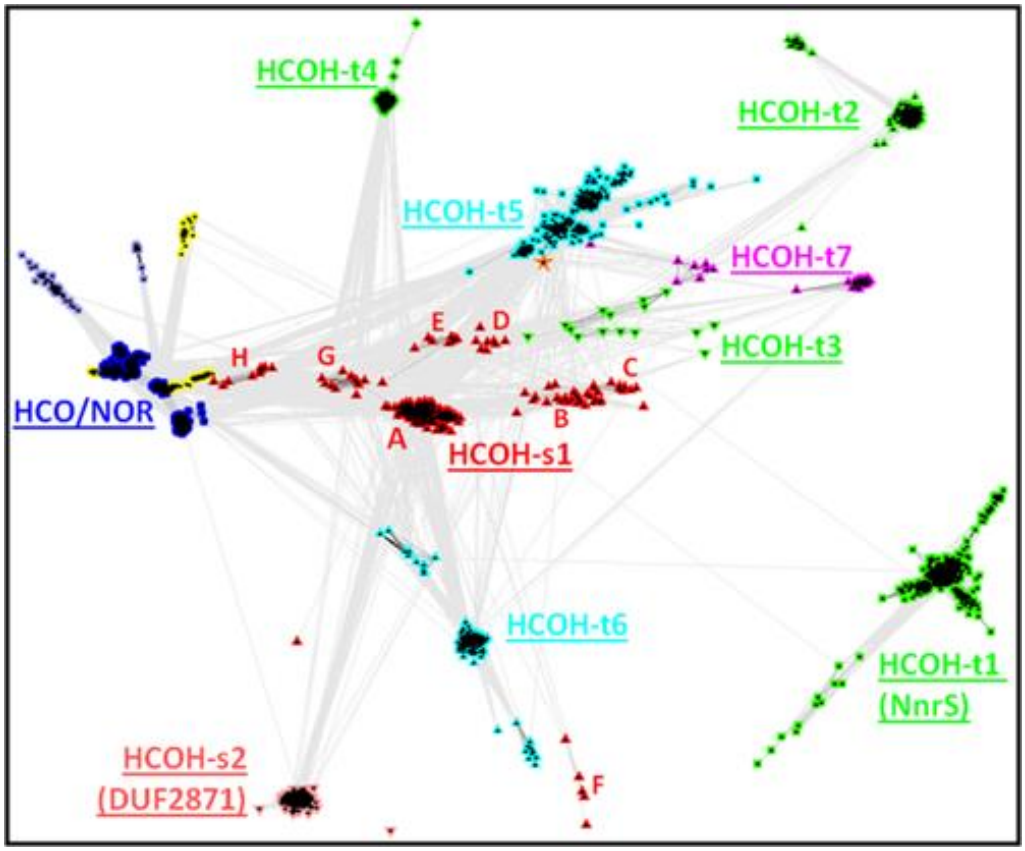


Figure 3. CLANS diagram of HCOs/NORs and HCOH proteins. Connections between proteins indicate BLAST Pvalues less than 1e-10. Single-EU HCOH proteins are shown as red up triangles (HCOH-s1) and pink low down triangles (HCOH-s2). Members of HCOH-t1, HCOH-t2, HCOH-t3, and HCOH-t4, the four groups with the Hxx.xxx.xxH motif pattern, are shown in green squares, green up triangles, green down triangles, and green diamonds, respectively. HCOH-t5 and HCOH-t6 members, both with the xxH.Hxx.xxx motif pattern, are shown in cyan square and cyan up triangles, respectively. The sequence in the HCOH-t5 group with the HxH.HxH.HxH motif pattern is marked by an orange star. HCOH-t7 members with the xxx.xxH.Hxx motif pattern are shown in magenta. Underlined group names are shown. For the HCOH-s1 group, the A to H clusters are marked in red letters. HCOs/NORs are shown as blue dots. Two small groups of proteins shown as light blue dots are closely related to HCOs/NORs. They do not contain the HH motif in the third helix of EU2, while maintaining all the other conserved histidines (Hxx, xxH, and HxH in EU1, EU2, and EU3, respectively). Some HCO-related sequences with all conserved histidines deteriorated are shown as yellow dots.

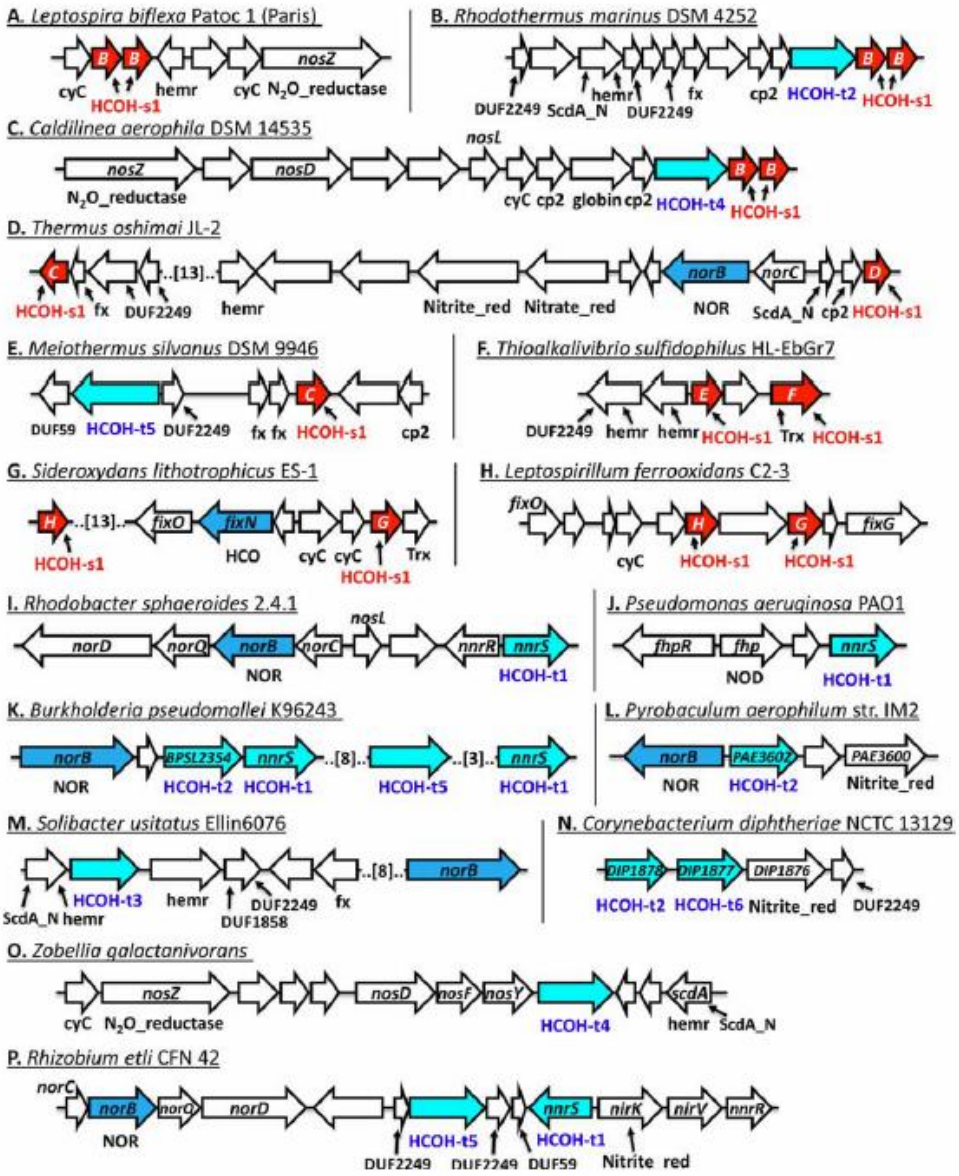


Figure 4. Gene structure diagrams of selected HCOH proteins. Genes are shown as arrows. HCOH-s1 genes, three-EU HCOH genes, and HCO/NOR genes are shown as red arrows, cyan arrows, and blue arrows, respectively. Names of frequently occurring domains are shown below the genes. Numbers of omitted genes are shown in brackets. Domain name abbreviations are as follows: cp2, cupin_2 domain; cyC, cytochrome c; fx, iron-sulfur ferredoxin domain; HCO: heme-copper oxidase; hemr, hemerythrin domain; N_2O_reductase, nitrous oxide reductase; Nitrate_red, nitrate reductase; Nitrite_red, nitrite reductase; NOD, nitric oxide dioxygenase; NOR, nitric oxide reductase; and Trx, thioredoxin.

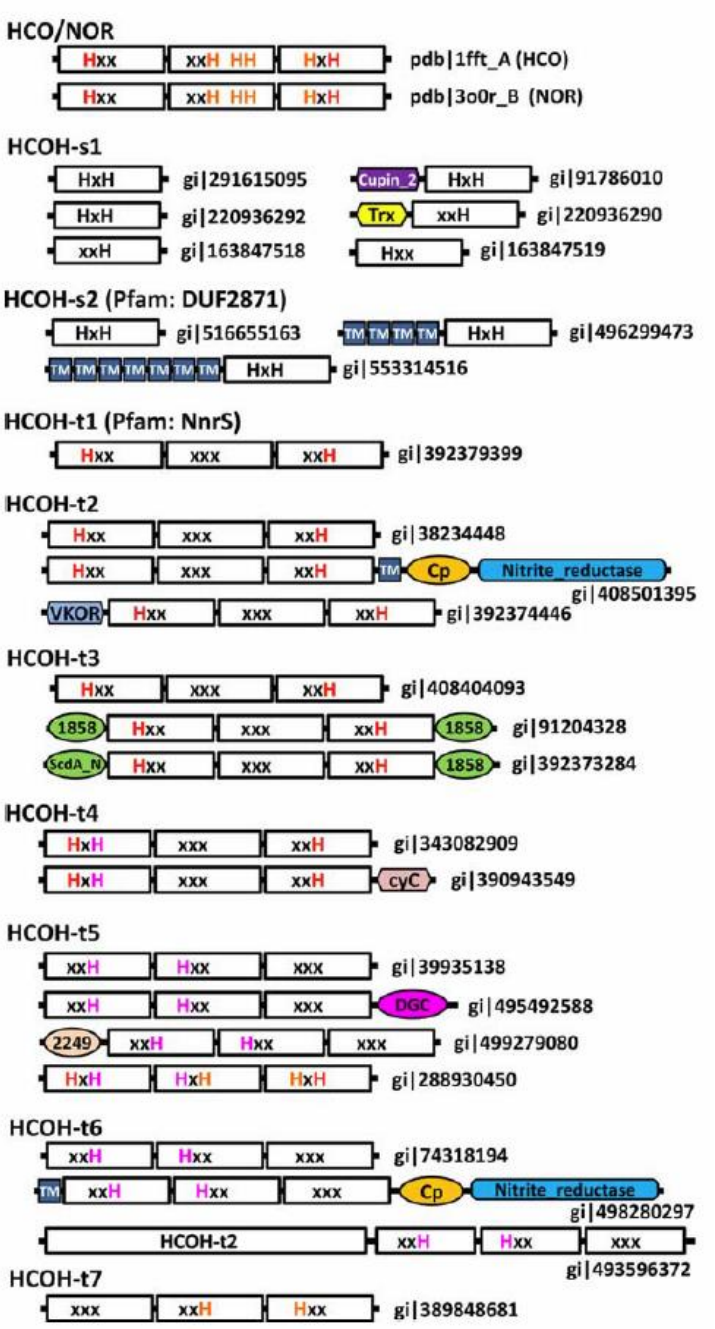


Figure 5. Domain structure diagrams of selected HCO/NOR and HCOH proteins. NCBI gi number or pdb/chain id is shown for each protein. EUs are shown in white rectangular boxes with motifs (HxH, xxH, Hxx, xxx, and HH). For three-EU proteins, conserved histidines at the interface of EU1 and EU3 are shown in red letters, those at the interface of EU2 and EU3 are shown in orange letters, and those at the interface of EU1 and EU2 are shown in magenta letters. Domain or module name abbreviations are as follows: cupin_2, cupin2 domain; cyC, cytochrome c domain, Trx: thioredoxin domain; Cp, cupredoxin domain; 1858, DUF1858; 2249, DUF2249; TM, predicted transmembrane helix; and VKOR, vitamin K epoxide reductase domain.

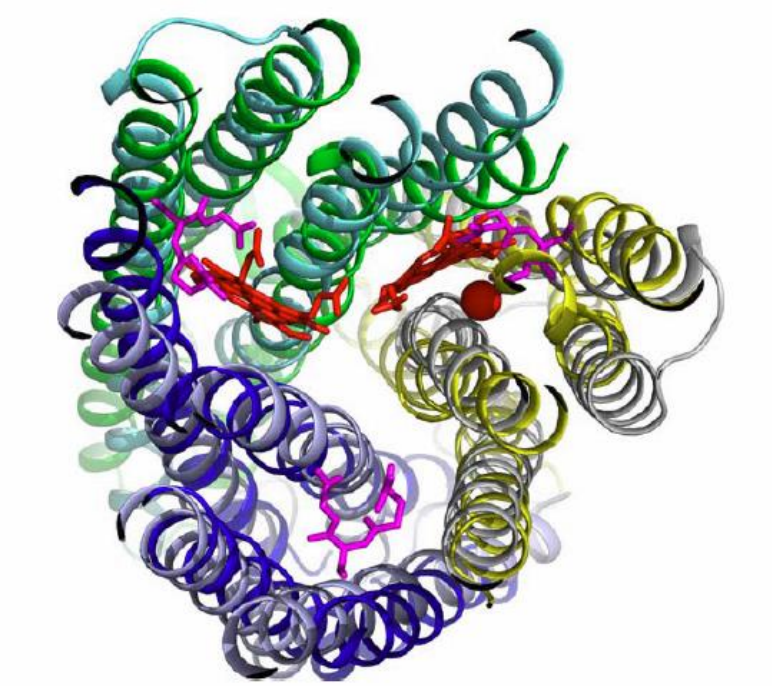


Figure 6. Structure superposition of an HCO superfamily member (pdb id: 3mk7) and a MAPEG member (pdb id: 4al0). EU1, EU2, and EU3 of 3mk7 are shown in blue, yellow, and green, respectively. The heme groups and the non-heme iron are shown as red sticks and a red sphere, respectively. The three chains of 4al0 aligned to EU1, EU2, and EU3 of 3mk7 are shown in light blue, grey, and cyan, respectively. The ligands of 4al0 are shown as magenta sticks.

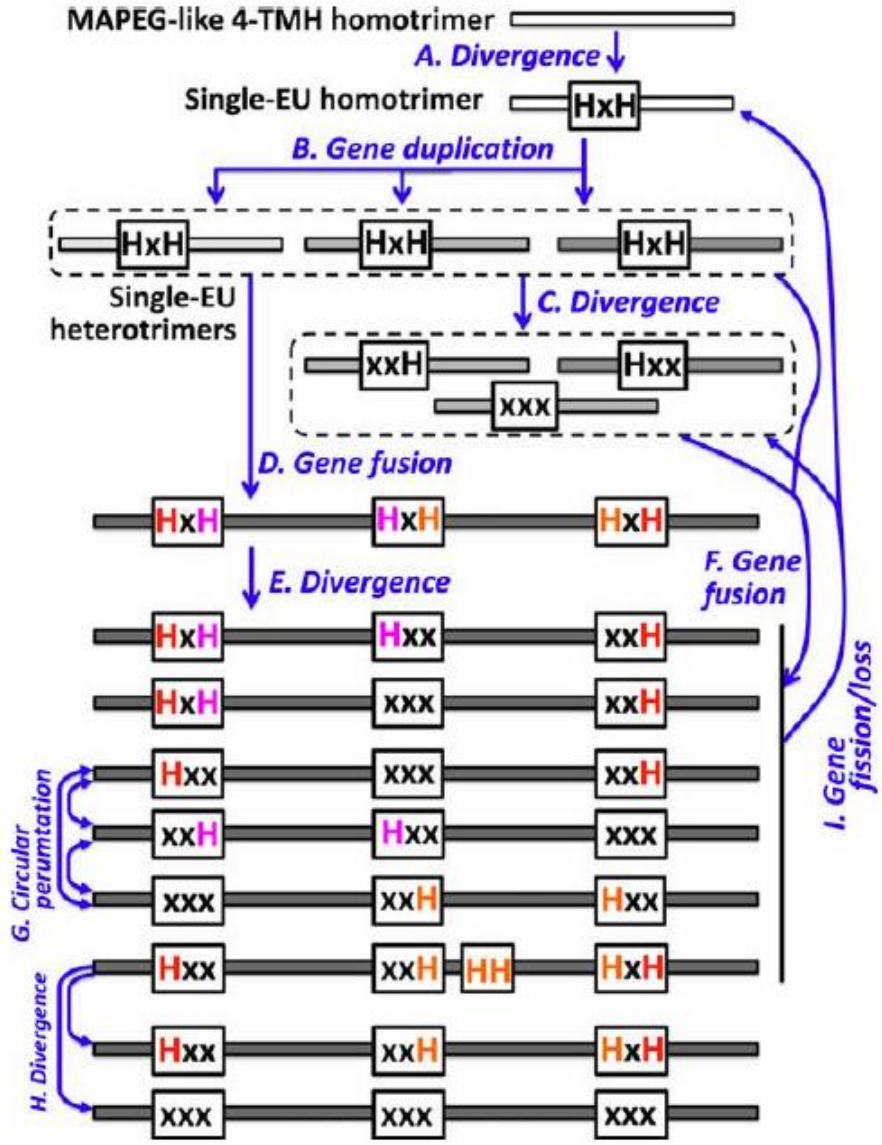


Figure 7. Possible evolutionary events in HCOs/NORs and HCOH proteins.

REFERENCES

1. Catling DC, Glein CR, Zahnle KJ, McKay CP. Why O2 is required by complex life on habitable planets and the concept of planetary "oxygenation time" Astrobiology. 2005;5:415–438.
2. Payne JL, Boyer AG, Brown JH, Finnegan S, Kowalewski M, Krause RA, Jr, Lyons SK, McClain CR, McShea DW, Novack-Gottshall PM, Smith FA, Stempien JA, Wang SC. Two-phase increase in the maximum size of life over 3.5 billion years reflects biological innovation and environmental opportunity. Proc Natl Acad Sci USA. 2009;106:24–27.
3. Dahl TW, Hammarlund EU, Anbar AD, Bond DP, Gill BC, Gordon GW, Knoll AH, Nielsen AT, Schovsbo NH, Canfield DE. Devonian rise in atmospheric oxygen correlated to the radiations of terrestrial plants and large predatory fish. Proc Natl Acad Sci USA. 2010;107:17911–17915.
4. Gribaldo S, Talla E, Brochier-Armanet C. Evolution of the haem copper oxidases superfamily: a rooting tale. Trends Biochem Sci. 2009;34:375–381.
5. Pereira MM, Santana M, Teixeira M. A novel scenario for the evolution of haem-copper oxygen reductases. Biochim Biophys Acta. 2001;1505:185–208.
6. van der Oost J, de Boer AP, de Gier JW, Zumft WG, Stouthamer AH, van Spanning RJ. The heme-copper oxidase family consists of three distinct types of terminal oxidases and is related to nitric oxide reductase. FEMS Microbiol Lett. 121:1–9.

7. Richter OM, Ludwig B. Cytochrome c oxidase—structure, function, and physiology of a redox-driven molecular machine. Rev Physiol Biochem Pharmacol. 2003;147:47–74.

8. Ferguson-Miller S, Babcock GT. Heme/copper terminal oxidases. Chem Rev. 1996;96:2889–2908.

9. Michel H, Behr J, Harrenga A, Kannt A. Cytochrome c oxidase: structure and spectroscopy. Annu Rev Biophys Biomol Struct. 1998;27:329–356.

10. Iwata S, Ostermeier C, Ludwig B, Michel H. Structure at 2.8 A resolution of cytochrome c oxidase from Paracoccus denitrificans. Nature. 1995;376:660–669.

11. Sousa FL, Alves RJ, Ribeiro MA, Pereira-Leal JB, Teixeira M, Pereira MM. The superfamily of heme-copper oxygen reductases: types and evolutionary considerations. Biochim Biophys Acta. 2012;1817:629–637.

12. Brochier-Armanet C, Talla E, Gribaldo S. The multiple evolutionary histories of dioxygen reductases: Implications for the origin and evolution of aerobic respiration. Mol Biol Evol. 2009;26:285–297.

13. Gennis RB. Multiple proton-conducting pathways in cytochrome oxidase and a proposed role for the active-site tyrosine. Biochim Biophys Acta. 1998;1365:241–248.

14. Rauhamaki V, Baumann M, Soliymani R, Puustinen A, Wikstrom M. Identification of a histidine-tyrosine cross-link in the active site of the cbb3-type cytochrome c oxidase from Rhodobacter sphaeroides. Proc Natl Acad Sci USA. 2006;103:16135–16140.

15. Hemp J, Christian C, Barquera B, Gennis RB, Martinez TJ. Helix switching of a key active-site residue in the cytochrome cbb3 oxidases. Biochemistry. 2005;44:10766–10775.
16. Hino T, Matsumoto Y, Nagano S, Sugimoto H, Fukumori Y, Murata T, Iwata S, Shiro Y. Structural basis of biological N2O generation by bacterial nitric oxide reductase. Science. 2010;330:1666–1670.
17. Moenne-Loccoz P, Fee JA. Biochemistry. Catalyzing NO to N2O in the nitrogen cycle. Science. 2010;330:1632–1633.
18. Hendriks J, Gohlke U, Saraste M. From NO to OO: nitric oxide and dioxygen in bacterial respiration. J Bioenerg Biomembr. 1998;30:15–24.
19. Stevanin TM, Laver JR, Poole RK, Moir JW, Read RC. Metabolism of nitric oxide by Neisseria meningitidis modifies release of NO-regulated cytokines and chemokines by human macrophages. Microbes Infect. 2007;9:981–987.
20. Hendriks J, Oubrie A, Castresana J, Urbani A, Gemeinhardt S, Saraste M. Nitric oxide reductases in bacteria. Biochim Biophys Acta. 2000;1459:266–273.
21. Sousa FL, Alves RJ, Pereira-Leal JB, Teixeira M, Pereira MM. A bioinformatics classifier and database for heme-copper oxygen reductases. PLoS One. 2011;6:e19117.
22. Castresana J, Lubben M, Saraste M, Higgins DG. Evolution of cytochrome oxidase, an enzyme older than atmospheric oxygen. EMBO J. 1994;13:2516–2525.
23. de Vries S, Schroder I. Comparison between the nitric oxide reductase family and its aerobic relatives, the cytochrome oxidases. Biochem Soc Trans. 2002;30:662–667.

24. Ducluzeau AL, van Lis R, Duval S, Schoepp-Cothenet B, Russell MJ, Nitschke W. Was nitric oxide the first deep electron sink? Trends Biochem Sci. 2009;34:9–15.

25. Castresana J, Saraste M. Evolution of energetic metabolism: the respiration-early hypothesis. Trends Biochem Sci. 1995;20:443–448.

26. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389–3402.

27. Soding J, Biegert A, Lupas AN. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res. 2005;33:W244–248.

28. Jakobsson PJ, Morgenstern R, Mancini J, Ford-Hutchinson A, Persson B. Common structural features of MAPEG—a widespread superfamily of membrane associated proteins with highly divergent functions in eicosanoid and glutathione metabolism. Protein Sci. 1999;8:689–692.

29. Holm PJ, Bhakat P, Jegerschold C, Gyobu N, Mitsuoka K, Fujiyoshi Y, Morgenstern R, Hebert H. Structural basis for detoxification and oxidative stress protection in membranes. J Mol Biol. 2006;360:934–945.

30. Frickey T, Lupas A. CLANS: a Java application for visualizing protein families based on pairwise similarity. Bioinformatics. 2004;20:3702–3704.

31. Oliveira F, de Carvalho AM, de Oliveira CI. Sand-fly saliva—man: the trigger trio. Frontiers Immunol. 2013;4:375.

32. Vannier-Santos MA, Martiny A, de Souza W. Cell biology of Leishmania spp.: invading and evading. Curr Pharma Des. 2002;8:297–318.

33. Koreny L, Lukes J, Obornik M. Evolution of the haem synthetic pathway in kinetoplastid flagellates: an essential pathway that is not essential after all? Intl J Parasitol. 2010;40:149–156.

34. Campos-Salinas J, Cabello-Donayre M, Garcia-Hernandez R, Perez-Victoria I, Castanys S, Gamarro F, Perez-Victoria JM. A new ATP-binding cassette protein is involved in intracellular haem trafficking in Leishmania. Mol Microbiol. 2011;79:1430–1444.

35. Huynh C, Yuan X, Miguel DC, Renberg RL, Protchenko O, Philpott CC, Hamza I, Andrews NW. Heme uptake by Leishmania amazonensis is mediated by the transmembrane protein LHR1. PLoS Pathog. 2012;8:e1002795.

36. Dutta S, Furuyama K, Sassa S, Chang KP. Leishmania spp.: delta-aminolevulinate-inducible neogenesis of porphyria by genetic complementation of incomplete heme biosynthesis pathway. Exp Parasitol. 2008;118:629–636.

37. Chang CS, Chang KP. Heme requirement and acquisition by extracellular and intracellular stages of Leishmania mexicana amazonensis. Mol Biochem Parasitol. 1985;16:267–276.

38. Anzaldi LL, Skaar EP. Overcoming the heme paradox: heme toxicity and tolerance in bacterial pathogens. Infect Immun. 2010;78:4977–4989.

39. Kwiatkowski AV, Laratta WP, Toffanin A, Shapleigh JP. Analysis of the role of the nnrR gene product in the response of Rhodobacter sphaeroides 2.4.1 to exogenous nitric oxide. J Bacteriol. 1997;179:5618–5620.

40. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez P, Bork P, von Mering C, Jensen LJ. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res. 2013;41:D808–D815.
41. Bartnikas TB, Wang Y, Bobo T, Veselov A, Scholes CP, Shapleigh JP. Characterization of a member of the NnrR regulon in Rhodobacter sphaeroides 2.4.3 encoding a haem-copper protein. Microbiology. 2002;148:825–833.
42. Stern AM, Hay AJ, Liu Z, Desland FA, Zhang J, Zhong Z, Zhu J. The NorR regulon is critical for Vibrio cholerae resistance to nitric oxide and sustained colonization of the intestines. MBio. 2012;3:e00013–00012.
43. Stern AM, Liu B, Bakken LR, Shapleigh JP, Zhu J. A novel protein protects bacterial iron-dependent metabolism from nitric oxide. J Bacteriol. 2013;195:4702–4708.
44. Jamet A, Euphrasie D, Martin P, Nassif X. Identification of genes involved in Neisseria meningitidis colonization. Infect Immun. 2013;81:3375–3381.
45. Cozen AE, Weirauch MT, Pollard KS, Bernick DL, Stuart JM, Lowe TM. Transcriptional map of respiratory versatility in the hyperthermophilic crenarchaeon Pyrobaculum aerophilum. J Bacteriol. 2009;191:782–794.
46. French CE, Bell JM, Ward FB. Diversity and distribution of hemerythrin-like proteins in prokaryotes. FEMS Microbiol Lett. 2008;279:131–145.
47. Overton TW, Justino MC, Li Y, Baptista JM, Melo AM, Cole JA, Saraiva LM. Widespread distribution in pathogenic bacteria of di-iron proteins that repair

oxidative and nitrosative damage to iron-sulfur centers. J Bacteriol. 2008;190:2004–2013.

48. Frey AD, Kallio PT. Nitric oxide detoxification—a new era for bacterial globins in biotechnology? Trends Biotechnol. 2005;23:69–73.
49. Lo Conte L, Ailey B, Hubbard TJ, Brenner SE, Murzin AG, Chothia C. SCOP: a structural classification of proteins database. Nucleic Acids Res. 2000;28:257–259.
50. Kim BH, Cheng H, Grishin NV. HorA web server to infer homology between proteins using sequence and structural similarity. Nucleic Acids Res. 2009;37:W532–W538.
51. Cheng H, Kim BH, Grishin NV. Discrimination between distant homologs and structural analogs: lessons from manually constructed, reliable data sets. J Mol Biol. 2008;377:1265–1278.
52. Martinez Molina D, Wetterholm A, Kohl A, McCarthy AA, Niegowski D, Ohlson E, Hammarberg T, Eshaghi S, Haeggstrom JZ, Nordlund P. Structural basis for synthesis of inflammatory mediators by human leukotriene C4 synthase. Nature. 2007;448:613–616.
53. Martinez Molina D, Eshaghi S, Nordlund P. Catalysis within the lipid bilayer-structure and mechanism of the MAPEG family of integral membrane proteins. Curr Opin Struct Biol. 2008;18:442–449.
54. Ohno S. Evolution by Gene Duplication. New York: Springer-Verlag; 1970.
55. Kummerfeld SK, Teichmann SA. Relative rates of gene fusion and fission in multi-domain proteins. Trends Genet. 2005;21:25–30.

56. Pasek S, Risler JL, Brezellec P. Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins. Bioinformatics. 2006;22:1418–1423.

57. Goodsell DS, Olson AJ. Structural symmetry and protein function. Annu Rev Biophys Biomol Struct. 2000;29:105–153.

58. Levy ED, Boeri Erba E, Robinson CV, Teichmann SA. Assembly reflects evolution of protein complexes. Nature. 2008;453:1262–1265.

59. Lynch M. The evolution of multimeric protein assemblages. Mol Biol Evol. 2012;29:1353–1366.

60. Bouzat JL, McNeil LK, Robertson HM, Solter LF, Nixon JE, Beever JE, Gaskins HR, Olsen G, Subramaniam S, Sogin ML, Lewin HA. Phylogenomic analysis of the alpha proteasome gene family from early-diverging eukaryotes. J Mol Evol. 2000;51:532–543.

61. Archibald JM, Logsdon JM, Jr, Doolittle WF. Origin and evolution of eukaryotic chaperonins: phylogenetic evidence for ancient duplications in CCT genes. Mol Biol Evol. 2000;17:1456–1466.

62. Pereira-Leal JB, Levy ED, Kamp C, Teichmann SA. Evolution of protein complexes by duplication of homomeric interactions. Genome Biol. 2007;8:R51.

63. Choi S, Jeon J, Yang JS, Kim S. Common occurrence of internal repeat symmetry in membrane proteins. Proteins. 2008;71:68–80.

64. Duran AM, Meiler J. Inverted topologies in membrane proteins: a mini-review. Comput Struct Biotechnol J. 2013;8:e201308004.

65. Rose PW, Bi C, Bluhm WF, Christie CH, Dimitropoulos D, Dutta S, Green RK, Goodsell DS, Prlic A, Quesada M, Quinn GB, Ramos AG, Westbrook JD, Young J, Zardecki C, Berman HM, Bourne PE. The RCSB Protein Data Bank: new resources for research and education. Nucleic Acids Res. 2013;41:D475–482.
66. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer EL, Tate J, Punta M. Pfam: the protein families database. Nucleic Acids Res. 2014;42:D222–D230.
67. Marchler-Bauer A, Zheng C, Chitsaz F, Derbyshire MK, Geer LY, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Lanczycki CJ, Lu F, Lu S, Marchler GH, Song JS, Thanki N, Yamashita RA, Zhang D, Bryant SH. CDD: conserved domains and protein three-dimensional structure. Nucleic Acids Res. 2013;41:D348–D352.
68. Eddy SR. Accelerated profile HMM searches. PLoS Comput Biol. 2011;7:e1002195.
69. Pei J, Kim BH, Grishin NV. PROMALS3D: a tool for multiple protein sequence and structure alignments. Nucleic Acids Res. 2008;36:2295–2300.
70. Holm L, Park J. DaliLite workbench for protein structure comparison. Bioinformatics. 2000;16:566–567.
71. Konagurthu AS, Whisstock JC, Stuckey PJ, Lesk AM. MUSTANG: a multiple structural alignment algorithm. Proteins. 2006;64:559–574.
72. Kall L, Krogh A, Sonnhammer EL. Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server. Nucleic Acids Res. 2007;35:W429–432.

73. Sonnhammer EL, von Heijne G, Krogh A. A hidden Markov model for predicting transmembrane helices in protein sequences. Proc Intl Conf Intellig Syst Mol Biol. 1998;6:175–182.

CHAPTER 6 ESTIMATION OF UNCERTAINTIES IN THE GLOBAL DISTANCE TEST (GDT_TS) FOR CASP MODELS[5]

INTRODUCTION

The Critical Assessment of techniques for protein Structure Prediction (or CASP) is a community-wide experiment to establish the capabilities and limitations of structure prediction methods, as well as to determine the progress of modeling methodologies [1]. Since CASP3 in 1998, assessors have been using the Global Distance Test (GDT_TS) score [2,3] in model evaluation due to its tolerance for partial structure segments that could create a large root mean square deviation (RMSD). The GDT algorithm uses the residue correspondence between the model and the target structure to search for optimal superpositions under selected distance cutoffs. The GDT_TS score reports an average of the maximum number of residues that can be superimposed under four distance cutoffs 1Å, 2Å, 4Å, and 8Å. Current GDT_TS comparisons produce a point estimate for structure similarity without confidence intervals. Although the statistical significance of differences in GDT_TS between group performances can be tested in CASP where participating groups submitted a number of predictions [4,5], identifying significant differences between individual models with close structural similarity would be

[5] This chapter was published as:
Li W, Schaeffer RD, Otwinowski Z, Grishin N V. *Estimation of uncertainties in the global distance test (GDT_TS) for CASP Models*. PLoS One 2016;11(5):e0154786.

challenging for GDT_TS point estimates due to the potential underlying structural flexibility of the modeled proteins.

The flexibility of protein structures could add uncertainty to the atomic positions, which subsequently introduces uncertainty to structure comparison by GDT_TS measure. Currently, CASP models are submitted as sets of coordinates representing accurate atom positions. Although efforts to estimate the certainty of atom positions have been made, such estimations are only accurate for a few top performing models [6]. In addition, the estimated values vary dramatically in scale, which limits their utility in estimation of the uncertainty of atom positions. On the other hand, target structures are snapshots of flexible protein molecules that exist as ensembles of conformational states [7–9]; the atomic fluctuations caused by the dynamic properties of target proteins would contribute to the uncertainty of atom positions in their structures. In our study, we derived the GDT_TS uncertainty from simulated fluctuations of target structures. NMR spectroscopy can reveal the functional dynamics of proteins on a wide range of time scales and is used to generate a structure ensemble of (usually 20) conformations [10]. However, the standard X-ray refinement produces the static structure averaged over time and space for the dynamic ensembles contained in crystals [11]. Although B-factors are thought to reflect the conformational diversity of such ensembles [11], insufficient information about collective motions [12] make it intractable to translate the uncertainty of B-factors into that of GDT_TS scores. To re-capitulate the structural heterogeneous ensembles in the crystal lattices, we performed time-averaged refinement [13] for X-ray datasets to generate structural ensembles for our GDT_TS uncertainty analysis.

Here, we utilize structure ensembles either from NMR deposits or generated by time-average refinements from X-ray structures to determine the uncertainty in GDT_TS scores for CASP models. Our results demonstrate that the time-averaged refinements produced structure ensembles in better agreement with the experimental datasets than the averaged X-ray structures, due to the ability to model anharmonic motions. As GDT_TS increases, its standard deviation (SD) also increases, reaching a maximum of 0.3 and 1.23 for X-ray and NMR structures, respectively. To facilitate score comparisons by the community, we developed a user-friendly web server that produces structure ensembles for NMR and X-ray structures and is accessible at http://prodata.swmed.edu/SEnCS. Our work helps to identify the significance of GDT_TS score differences for structures with high similarity, as well as to provide structure ensembles for estimating SDs of any scores.

RESULTS AND DISCUSSIONS

Generation of Structural Ensembles

CASP targets are primarily determined by X-ray crystallography and sometimes by NMR spectroscopy. NMR structures are deposited as ensembles of multiple conformations indicating the variation due to a combination of protein dynamics and uncertainty in NMR refinement. To generate ensembles indicative of the structural heterogeneity of X-ray structures, we performed time-averaged refinements [13] for

crystallographic datasets. Briefly, time-averaged refinement is performed using molecular dynamics simulations with time-averaged constraints on the X-ray dataset. Time-averaged refinement can model anharmonic motions, unlike traditional averaged refinement using B-factors, generating structure ensembles more compatible with the crystallographic data.

In our time-averaged refinement procedure, the global structure flexibility is approximated by the TLS (Translation/Libration/Screw) fitting procedure [16]. This TLS procedure requires a pTLS parameter, which defines the fraction of atoms used in the flexibility approximation and cannot be determined *a priori*. As the authors suggested, we performed simultaneous refinements with an array of pTLS values and observed that the pTLS value controls the amplitude of atomic fluctuations within the produced ensembles. Illustrated by the mean and SDs of selfGDT scores, i.e. the GDT_TS scores comparing models within one ensemble (details in methods), simulations with larger pTLS values produced ensembles of lower structure flexibility exhibiting lower SDs (cyan bars in Fig 1A). More importantly, the time-averaged refinements produced better R-free values only when a sufficient fraction of atoms is included in the flexibility approximation (Fig 1B, R value improvement as 0.01 for pTLS = 1); simulations with pTLS values no more than 0.6 produced worse R-free values (decreasing as much as 0.13) than those of averaged structures and might over-optimize the structure.

As the choice of pTLS value affects the structure flexibility of the generated ensembles (Fig 1), we chose a pTLS value such that the simulated fluctuations were similar to the expected fluctuations in crystal structures of native proteins. In doing so,

we suggest that the observed distribution of GDT_TS scores between members of our simulated ensembles is representative of the true dynamic ensemble of the target protein. To test whether our simulated ensemble was a reasonable model of structural fluctuations, we analyzed cases where the same protein was crystalized in different space groups. Those proteins were experimentally captured in distinct conformational states and were demonstrated to reveal functionally relevant dynamics [17]. A large portion (69%) of these proteins were determined in three distinct space groups (S1 Fig), inhibiting the statistical power of SDs to indicate the structure fluctuations. We used the minimal GDT_TS score (minGDTs) among all scores in an ensemble to indicate the minimal structural similarity of an ensemble. Higher minGDTs implies higher structural similarity and thus lower structure fluctuations. The minGDTs for all proteins lies above 95, with a majority of average minGDT values ranging from 98.9 to 99.5 (Fig 2A). Compared to the majority of such structures (Fig 2B, red dot), time-averaged ensembles exhibit higher fluctuations for all pTLS values (Fig 2B, cyan dots). Therefore, considering both the structure flexibility (indicated by minGDTs) and the compatibility with experimental data (indicated by R-free values), we used the largest possible pTLS value (pTLS = 1).

NMR ensembles showed even higher structure flexibility than X-ray ensembles of pTLS = 1, even when we applied a 3.5Å threshold suggested by the CASP assessors to filter the highly flexible regions (purple bar in Fig 1A). Such differences in structure flexibility were attributed to the discrepancy in environmental influences (such as solvent properties) and experimental interpretation [18]. Most NMR structures are determined in water or organic solvents, whereas proteins in crystallography form a well-ordered crystal

lattice with less solvent between protein molecules. When interpreting the experimental data, NMR spectroscopy determines structures with larger allowance for errors from data misinterpretation [19], compared to the high resolution X-ray structures we included (resolution ≤ 1.8Å). Conclusively, consistent with previous studies, our results suggested that NMR ensembles should be more flexible than X-ray ensembles of high resolutions.

GDT_TS Scores Calculated Using the Ensembles

In CASP evaluations, assessors employed statistical tests, e.g. bootstrapping and Student's t-test, to identify the top-performing groups [4,5]. However, for comparison between individual models, the lack of uncertainty estimation makes it difficult to distinguish the subtle performance differences between models. Comparisons lacking statistical significance might lead to over-aggressive claims about performance improvement, as small gains could be claimed as performance improvement. To solve this problem, we aimed to estimate the uncertainty of GDT_TS scores from our simulated ensembles to provide confidence intervals for statistical significance.

To quantify uncertainty, we computed the standard deviations (SD) of the GDT_TS scores, superimposing models against the generated target ensembles. The mean of such GDT_TS scores would infer the expected value in the canonical comparison between models and a single target structure, as the mean is the most likely value for such GDT_TS scores that follow a normal distribution (refer to S1 File, S2 and S3 Figs for normality test). The SDs in the scatter plots (Fig 3A) exhibited

differing scales for X-ray and NMR structures. For further analysis, we binned the SDs by 10 GDT_TS mean and averaged within each bin (Fig 3B and 3C, red bars). The averaged SDs increase with the GDT_TS means for models of low performances, reaching maximum values of 0.3 and 1.23 for X-ray and NMR structures, respectively. The averaged SDs of X-ray ensembles reach the maximum values in bins of smaller GDT_TS mean than NMR ensembles, likely due to the lower structure flexibility of X-ray ensembles. Interestingly, although similar to the maximum values, the average SDs slightly decrease with the GDT_TS mean for high performance models. We also investigated the structure flexibility of ensembles over the bins and found that the models of high GDT_TS scores were predicting the targets of lower structure flexibility; the SDs of GDT_TS comparison within individual ensembles (selfGDT, Fig 3B and 3C inset red lines) decrease for all NMR ensembles and X-ray ensembles of GDT_TS larger than 60. We speculate that such a correlation between the predictability, approximated by the GDT_TS values, of a target and the stability of a protein fold, indicated by the SDs of GDT_TS scores, could be related to the abundance of structure templates. Presumably, lower structural flexibility would facilitate the determination of experimental structures, which could then serve as modeling templates to boost the performance of prediction methods.

To reduce the bias in the distribution of ensemble flexibility over the bins, we further normalized the SDs by filtering a subset of ensembles of similar flexibility (see Methods). The normalized GDT_TS (Fig 3B and 3C green bars) display similar maximums to the raw data (before normalization) for X-ray ensembles, whereas NMR

ensembles showed an increased SD as 1.49 due to the reduced sample sizes in high GDT_TS score bins. However, those bins that exhibit the maximum value shift to a higher value, likely due to the exclusion of highly flexible ensembles in lower GDT_TS mean bins. Interestingly, neither NMR nor X-ray ensembles show SDs similar to those of the GDT_TS comparison within individual ensembles (Fig 3B and 3Cinset green lines, 1.38 and 1.94 for X-ray and NMR ensembles, respectively), possibly due to superposition optimization. Models of high performance/similarity would potentially superimpose to the conserved core regions of the target structure, leaving the highly flexible loops unaligned and thus reducing the fluctuation of aligned region. On the other hand, low quality models would be aligned over multiple differing regions to individual structures in an ensemble; as a result, the atomic fluctuations in the ensemble are averaged by the superposition optimization.

Uncertainty of Other CASP Scores

CASP targets are classified into two categories, Template-Based Modeling (TBM) and Free-Modeling (FM), based on the template availability and model performance [20,21]. GDT_TS scores are primarily employed in FM assessment, due to their increased capability to identify high performing models in the presence of short regions with large structural dissimilarities. In the TBM category, the high accuracy version of GDT-based scores, i.e. GDT_HA, was used to better recognize local differences between highly accurate models. Compared to the GDT_TS score, GDT_HA uses stricter distance

thresholds for superposition optimization and thus is more sensitive in identifying small improvements in local segments [22]. During CASP11, assessors introduced the superposition-independent Local Distance Difference Test (lDDT) score [23], which is constantly used in Continuous Automated Model EvaluatiOn (CAEMO) [24], to evaluate the local distance difference between structures. In addition to GDT_TS scores, we also evaluated the uncertainty in structure comparison quantified by GDT_HA and lDDT metrics using our generated ensembles.

Of the two GDT scores under consideration, GDT_HA is generally 10–20 less than the GDT_TS scores computed from the same models (Fig 4A and 4B), reflecting its higher stringency. The SDs of GDT_HA (shown in Fig 4C and 4D) are correlated with the SDs of GDT_TS scores, with R^2 of 0.71 and 0.87 for X-ray and NMR ensembles, respectively. Due to increased sensitivity of GDT_HA, we expect that the SDs of GDT_HA would be slightly higher than those of GDT_TS scores; indeed, more than half of GDT_HA scores display higher SDs than those of GDT_TS scores (57.5% for X-ray ensemble and 56.5% for NMR ensembles). GDT_HA, after normalization for structure flexibility, exhibits distributions similar to those of GDT_TS scores (Fig 4E). The SDs of GDT_HA increase with the mean of the scores, reaching a maximum value of 0.45 and 2.36 for X-ray and NMR structures, respectively. Our comparison demonstrates a similar uncertainty distribution for the high accuracy version of GDT-based scores.

In contrast to the strong correlation between GDT_TS and GDT_HA scores (coefficient as 0.98 for both X-ray and NMR structures), lDDT has a weaker correlation to GDT_TS scores (Fig 5A and 5B, coefficient as 0.82 for X-ray and 0.89 for NMR

structures, respectively), which potentially reflects the different evaluation emphasis wherein lDDT scores focus on the preservation of local contacts and GDT_TS highlights the global structure geometry. Consistent with the lower correlations between mean values, the SDs of lDDT and GDT_TS scores have lower R^2 values of 0.54 and 0.72 for X-ray and NMR structures, respectively (Fig 5C and 5D). Notably, the slope of the linear fits for the SDs showed large deviations from the diagonal (lDDT = 0.01 GDT_TS, 1 GDT_TS score is equivalent to 0.01 lDDT score), especially for NMR structures. Some errors from superposition, which are not included for lDDT scores, could potentially explain the larger SD for GDT_TS scores. The lDDT scores, after normalization for structure flexibility, show similar distributions to those of GDT_TS and GDT_HA scores (Fig 5E). The SDs of lDDT scores increase with the mean of the scores, reaching a maximum value of 0.0051 and 0.0131 for X-ray and NMR structures, respectively. However, due to the lack of high performing models (lDDT>0.8), the observed maximum SDs may not necessarily be the theoretical maximums for lDDT scores, as high performing models could continue the increasing trend for SDs. In conclusion, our study reveals the potential of our generated ensembles in evaluating the uncertainty of any structure similarity metrics.

Application and Limitations of Estimated Uncertainty in Model Comparison

In CASP assessments, the performance significance between groups is established by bootstrap and Student's t-test [4,5] statistics. However, comparing individual models

of close structural similarity can be difficult due to the lack of estimation for the score uncertainty induced by the structural flexibility. Here, we utilized the simulated structural flexibility of prediction targets to provide an estimate of uncertainty potentially underlying an individual point estimate score of a single model structure, which may prevent over-aggressive claims of improved performance. For example, two models from group TS410 and TS117 under target T0839 domain 1 have GDT_TS scores 58.20 and 57.72, respectively. The structural comparison between the models (Fig 6) identified very high similarity between secondary structure elements; however, large structural deviations were observed in the flexible loops connecting those secondary elements. The looped regions from both structures show little structural similarity to the respective regions in the target structure; potentially, the model from TS410 received a higher GDT_TS score due to the incidental overlap of some residues in these loops. By using our estimated uncertainty, the difference between this pair of scores is statistically insignificance under the 95% confidence interval (which requires GDT_TS differs at least 0.6). Therefore, our uncertainty estimation can help identify those models that differ by the random fluctuations in the loop region.

As the GDT_TS scores report the percentage of residues aligned under specified distance cutoff, the length of the structure plays a crucial role in the scale of its variations. For example, one misaligned residue in a protein of 50 residues would cause GDT_TS scores differ by 2, whereas one residue difference in a protein of 200 amino acids would contribute to 0.5 GDT_TS difference. We attempted to study the effect of length on the SDs of GDT_TS scores. Although we can see the tendency for shorter proteins to have

larger SDs (S4 Fig), insufficient target numbers for specific protein lengths (S1 Table) hinders the clarification of the quantitative relationship between length and GDT_TS uncertainty. As a single residue misalignment in the shorter protein could potentially create larger score fluctuation that deviates from the most likelihood SDs we concluded, we recommend generating the structure ensembles using our procedure and computing the SDs particularly for short proteins of interest.

Public Availability of Structural Ensemble Generation

To facilitate the SD calculation for any given structure, we implemented our method for generating structure ensembles as a user-friendly web server named SEnCS (Structure Ensemble of Conformational States, available at http://prodata.swmed.edu/wenlin/server/senCS/). The server takes a PDB ID as the input and computes the ensembles based on the type of structures. For NMR structures, it will fetch the ensemble from the PDB database [ref] and process the structure using a 3.5Å threshold to remove highly flexible regions without sufficient NMR constraints. For X-ray structures, it will retrieve the structures and experimental data from the PDB_REDO database [14] and perform time-averaged refinements. By default (fast mode), the time-averaged refinement would use all atoms in flexibility estimation (pTLS = 1) to generate the most conservative ensembles. Alternatively, one can explore an array of atom fractions in flexibility estimation (pTLS value) and generate a series of ensembles (exhaustive mode). The result page (Fig 7) exhibits the structural view of the

ensemble in JSmol [25] and the residue-based fluctuation along the protein sequences. The options are available in the result page to vary the distance threshold for NMR ensembles or to compute X-ray ensembles for more user-specified pTLS values. Once the ensemble is generated, users can download them to perform structure comparisons for uncertainty estimation for their scores.

CONCLUSIONS

Our study utilized structural ensembles either from NMR deposits or generated by time-averaged refinement to estimate the uncertainty of GDT_TS scores for CASP models. We quantified the SDs of GDT_TS scores and found that the SDs increase for low GDT_TS models and decrease for high GDT_TS models in our dataset. The X-ray and NMR structures have a maximum SD of 0.3 and 1.23, respectively. Subsequent application of our method to the high accuracy version of GDT-based scores, i.e. GDT_HA, and superposition-independent lDDT scores demonstrates the potential of our procedure to estimate the uncertainty for any other scores. Particularly, GDT_HA produces slightly higher SDs due to the increased sensitivity of GDT_HA. The SDs from lDDT scores are less correlated with those of GDT_TS scores, possibly due to the different dependency of structure superposition. We have also developed a web server that generates structure ensembles for uncertainty estimations. Our work provided generic SDs for estimating confidence intervals of GDT_TS scores, as well as the web server that provides the structure ensembles for any given protein.

MATERIALS AND METHODS

Proteins in Different Crystal Forms

We downloaded the non-redundant pdbaa database from http://dunbrack.fccc.edu/Guoli/culledpdb/pdbaa.gz and identified 1706 protein sequences associated with more than 2 space groups from the pdbaa database. The structures with the highest resolution for each space group were selected as representatives for the GDT_TS calculation. Briefly, representative structures are superimposed by the sequence-independent LGA structural aligner to generate sequence alignments, which in turn were used to produce sequence-dependent GDT_TS scores. We note that protein segments undergoing dramatic conformational changes do not align in the LGA superposition and thus do not contribute to the GDT_TS score calculation.

Time-Averaged Refinement for X-Ray Structures

We filtered the publically available X-ray structures in CASP9, CASP10, and CASP11 with resolution less than 1.8 Å and obtained 59 high resolution structures. Those structures and their experimental datasets were downloaded from the pdb_redo database [14]. We used the phenix.ensemble_refinement module[13] in the phenix software suit (version 1.9) to perform time-averaged refinement. As the author suggested in the tutorial

(http://www.phenix-online.org/documentation/reference/ensemble_refinement.html), we performed simulations with an array of pTLS values: 0, 0.1, 0.2, 0.4, 0.6, 0.8, 0.9, 1.0. We note that the program would automatically adjust the threshold to include at least 63 non-solvent and non-hydrogen atoms; additionally, it would fail if insufficient atoms were included.

NMR Structure Parsing

33 NMR structures were extracted from CASP9, CASP10, and CASP11 targets and downloaded from the pdb database [15]. To filter flexible regions, we computed the maximum Cα distance deviations of each residue per ensemble. We applied a 3.5 Å maximum Cα threshold, which was used in CASP target processing, to filter flexible residues potentially caused by the insufficient experimental NMR constraints.

Parameters for Ensembles

We first determined the central model of an ensemble as the structure with the largest sum of pairwise GDT_TS scores to other models in the ensemble. Second, we define selfGDT as the sum of GDT_TS scores comparing the central model to other structures in the ensemble. Finally, we computed the means and standard deviations (SDs) of the selfGDT for all targets, and excluded outlier ensembles (3δ away from the average of the means and SDs). To compare proteins with structures of multiple space groups, we

computed the minimal value of all-against-all GDT_TS scores for all models in an ensemble (minGDT) to replace SD as an estimate of ensemble fluctuation. To reduce the sample size of the target ensembles to a number similar to the most prevailing number of crystal forms, we calculated the average minGDT from 1000 random samples of three models from the target ensemble.

Comparison between Models and Target Structures

Sequence-dependent GDT_TS scores were calculated between the models and the individual structures in an ensemble. GDT_TS mean and standard deviation (SD) were calculated from the population of computed GDT_TS scores for each ensemble. As CASP models include partial structures, we filtered models of NMR structures with less than half of the target sequence length and models of X-ray structures with 100 residues less than the target structures. We binned the SDs by their corresponding means and removed outliers (>3σ) in each bin. When normalizing SDs by the structure flexibility of ensembles, we removed the outlier ensembles using 0.5σ as cutoff for the mean and SDs of selfGDTs, and computed SDs of GDT_TS scores comparing the filtered ensembles against the corresponding models.

Calculations for GDT_HA and lDDT Scores

The high accuracy version of GDT-based score, i.e. GDT_HA, was computed using LGA, which calculates the percentages of correctly aligned residues under four distance cutoffs: 0.5Å, 1Å, 2Å, and 4Å. Calculating the GDT_HA scores by averaging the correct percentage under these cutoffs, we applied the same pipeline as for the GDT_TS scores to compute the SDs of GDT_HA. We performed linear regression (suppressing the intercept term) for the SDs of GDT_HA and GDT_TS. The R^2 of the regression model was calculated using Microsoft Excel. Normalization of structure flexibility was performed using a similar procedure as for GDT_TS, substituting 1σ as cutoff for NMR ensembles (the original 0.5σ cutoff excluded all ensembles with GDT_HA greater than 20). The same procedure was also applied to lDDT score calculation, using 0.5σ as the normalization cutoff.

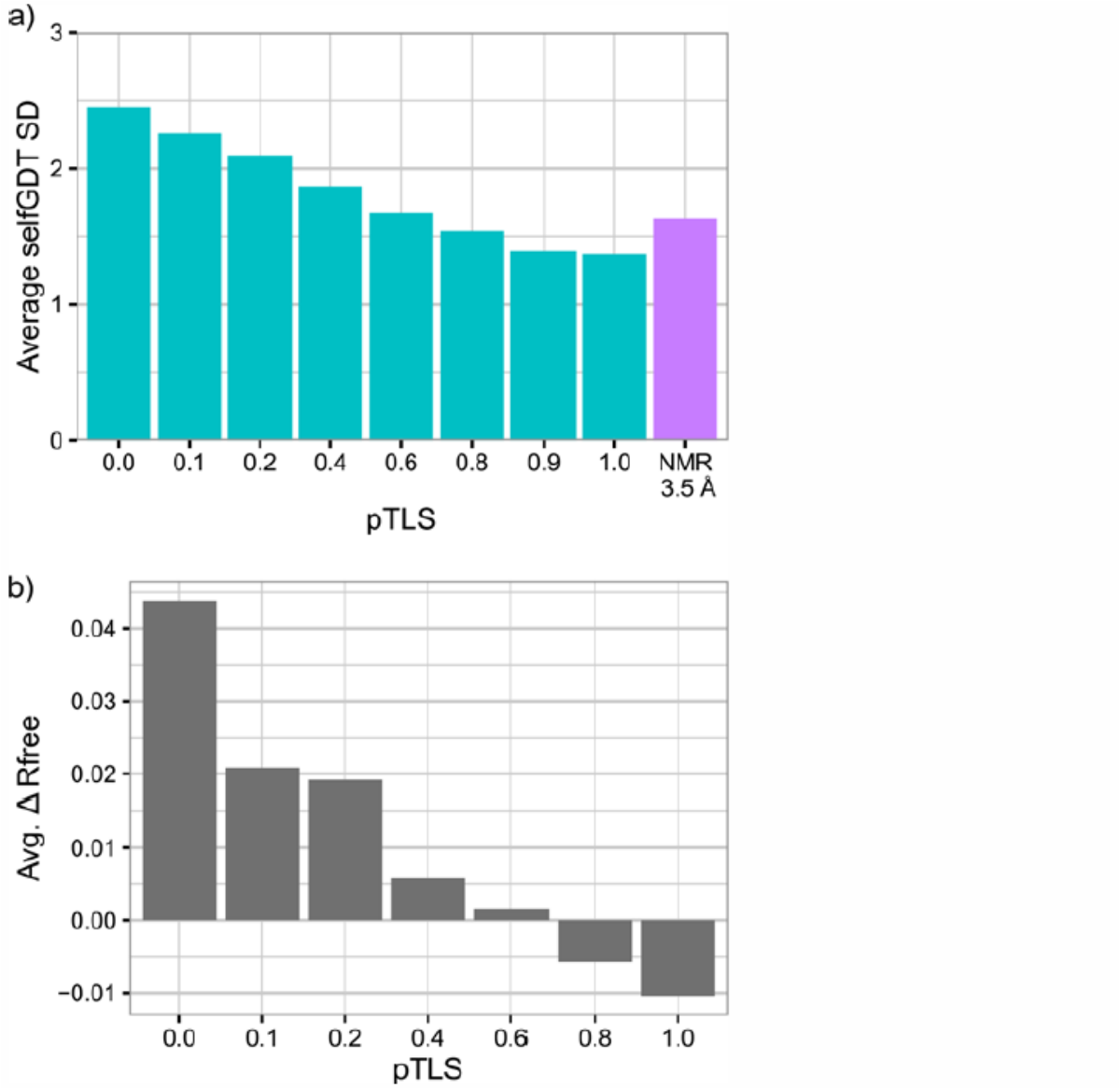


Fig 1. Comparison of structure fluctuations (a) and R-free values (b) for ensembles of different pTLS values. The choice of pTLS influences both the structural variability and the Rfree values of the generated assembles. As pTLS increases, the resulting ensembles show less variability. The structure fluctuations are implied by the SDs of GDT_TS scores calculated between models with an ensemble (selfGDT). NMR (a, purple) ensembles were compared after applying the 3.5Å distance threshold for filtering highly flexible segments. R-free values only improved with respect to experiment when pTLS was greater than 0.6.

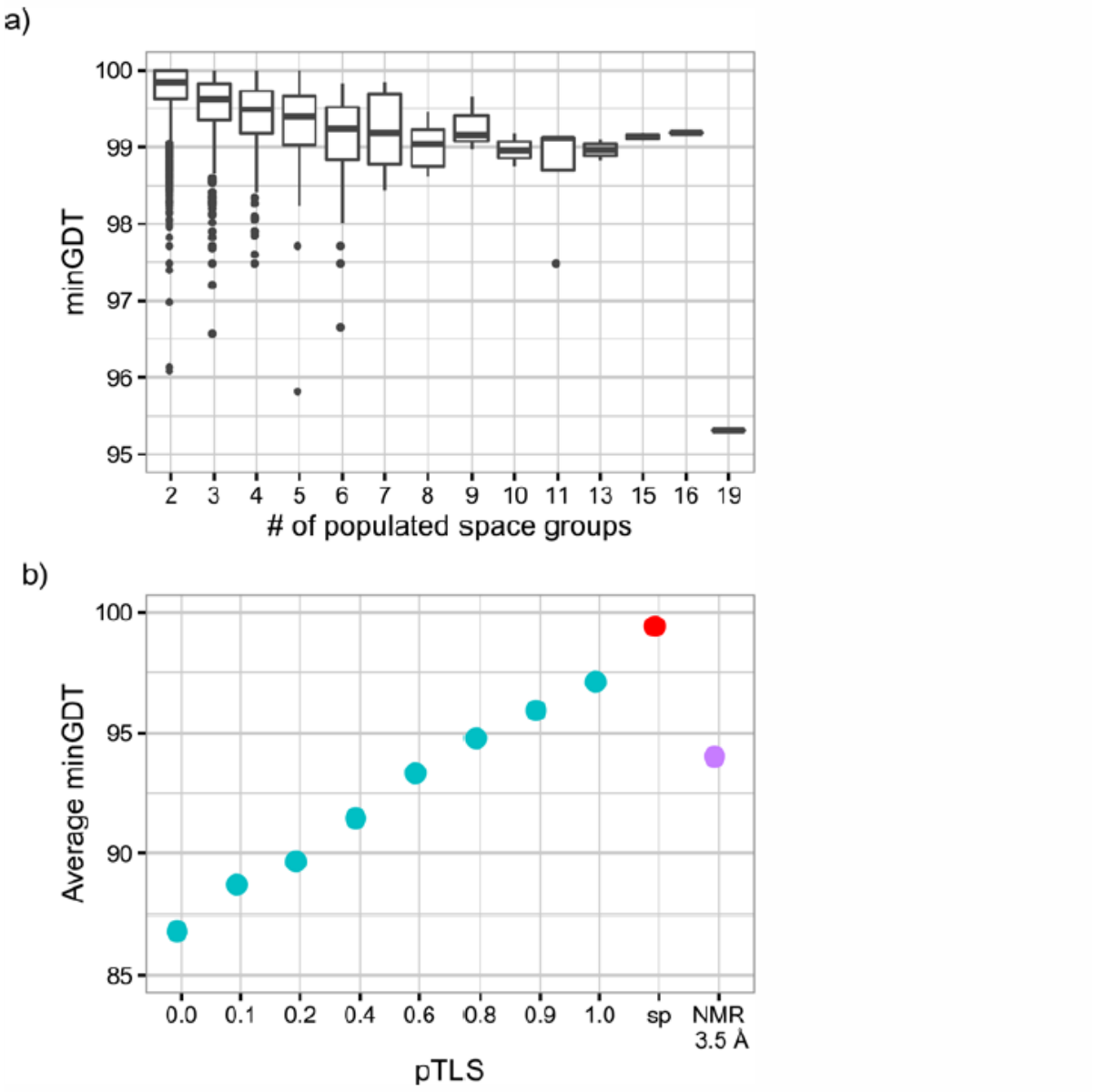


Fig 2. Structure fluctuations of proteins crystallized in different space groups. Structures in different crystal forms serve as a control for the expected structural flexibility. (a) We compared the distribution of minGDTs depending on the number of crystal forms (space groups). (b) The average minGDTs are displayed for the proteins of three space groups ('SP', red), time-averaged refinements (cyan), and NMR ensembles (purple). The minGDTs of the latter two (cyan and purple) were computed from resampled ensembles of three random structures. Generated ensembles approached the structural variability seen between different space groups of the same protein at pTLS = 1.

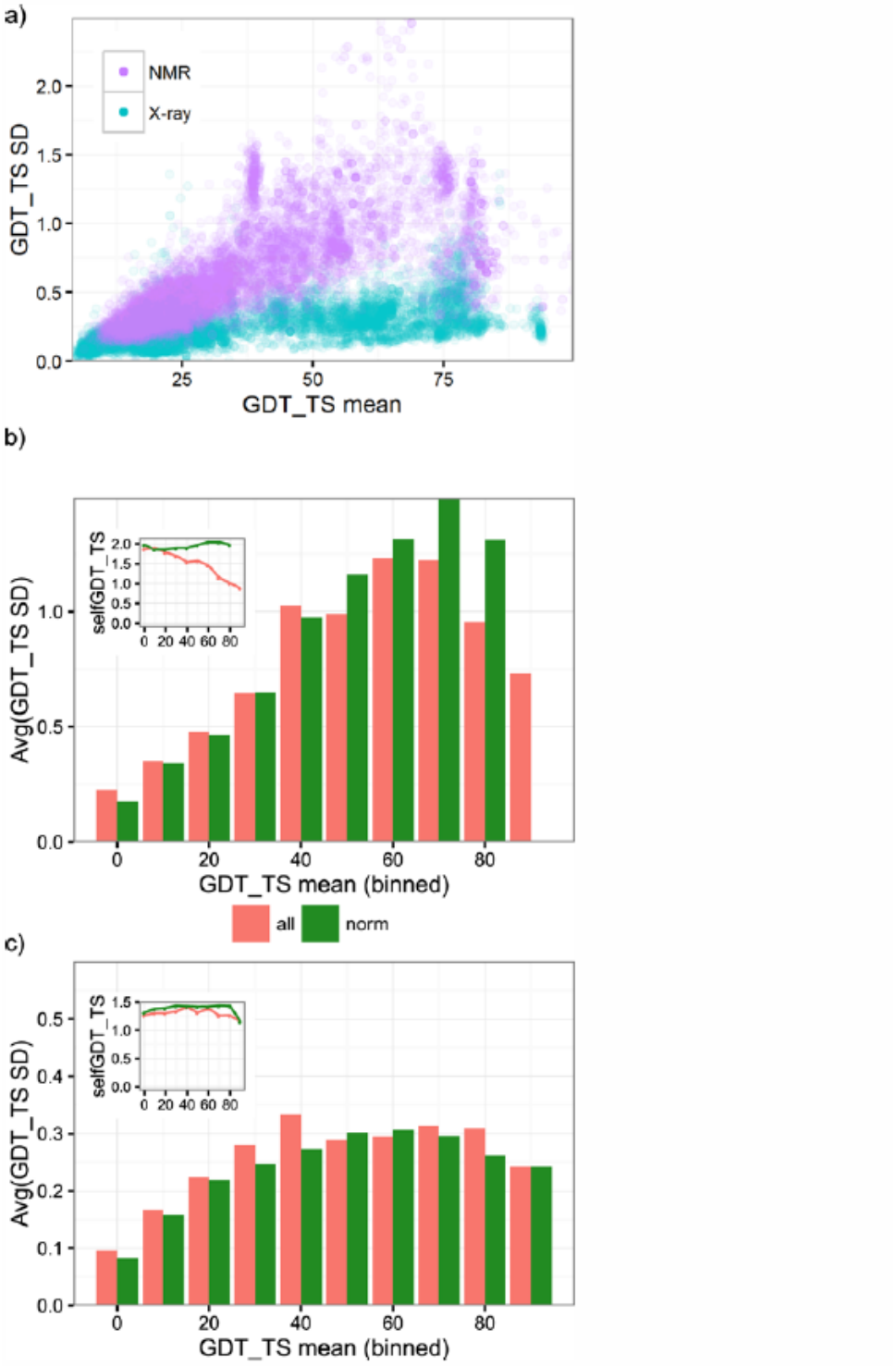


Fig 3. Uncertainty of GDT_TS scores (quantified by the SDs) against the mean of GDT_TS scores. (a) GD_TS scores show a close-to linear relationship between the mean of a GDT_TS score and its standard deviation. The SDs of high scoring NMR models are generally greater than those of X-ray scores in the same regime. We binned SDs for all ensembles before (red bars) and after (green bars) normalization for selfGDTs (shown in insets) for NMR (b) and X-ray (c) ensembles, respectively. Bins are labeled by their left edge. Bins with no models are not shown.

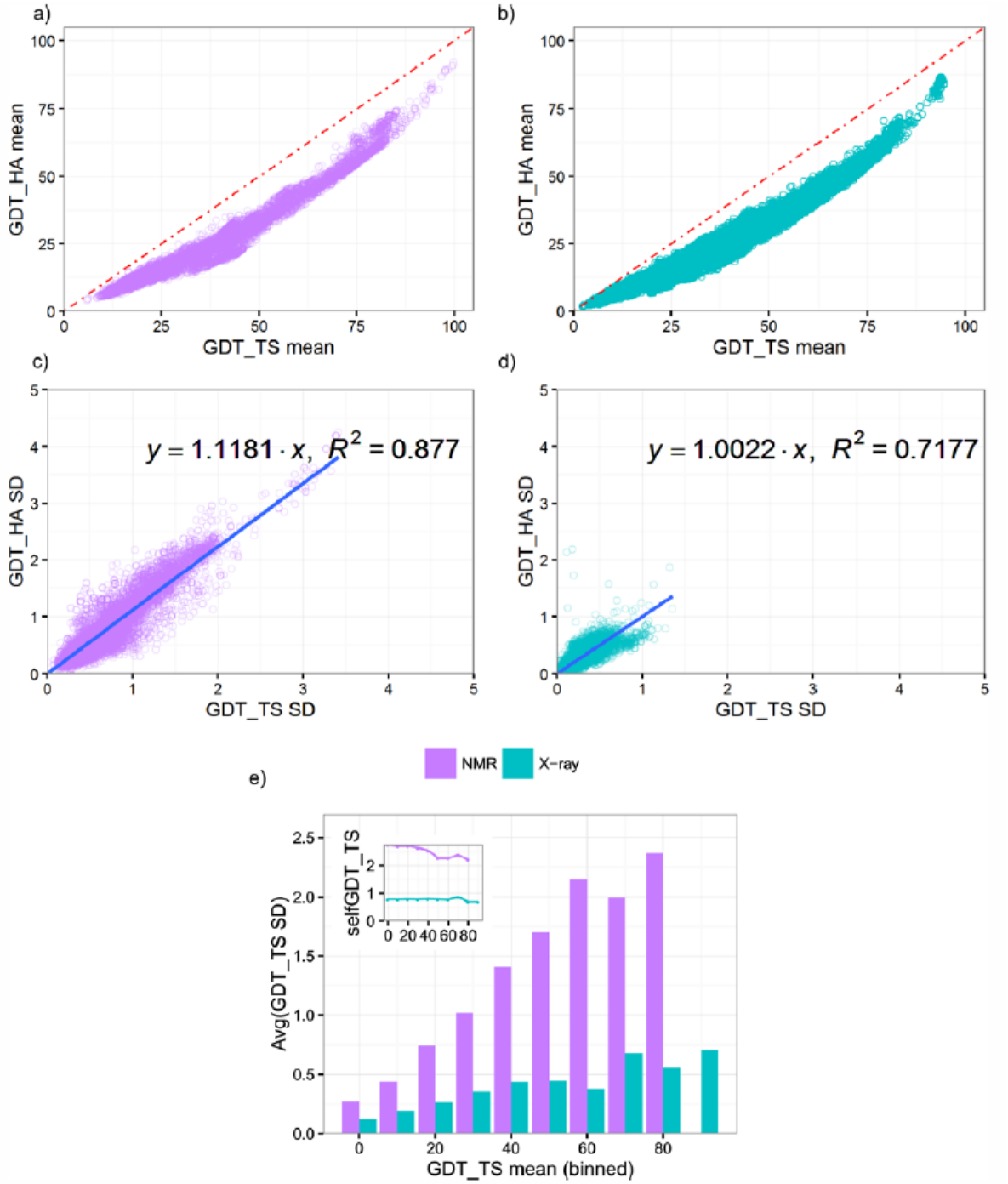


Fig 4. Comparison between GDT_HA and GDT_TS scores on generated ensembles. We compared the relationship between the high-accuracy GDT-based score (GDT_HA) and GDT_TS. Panels a-d display the means and SDs of GDT_HA versus GDT_TS computed from X-ray

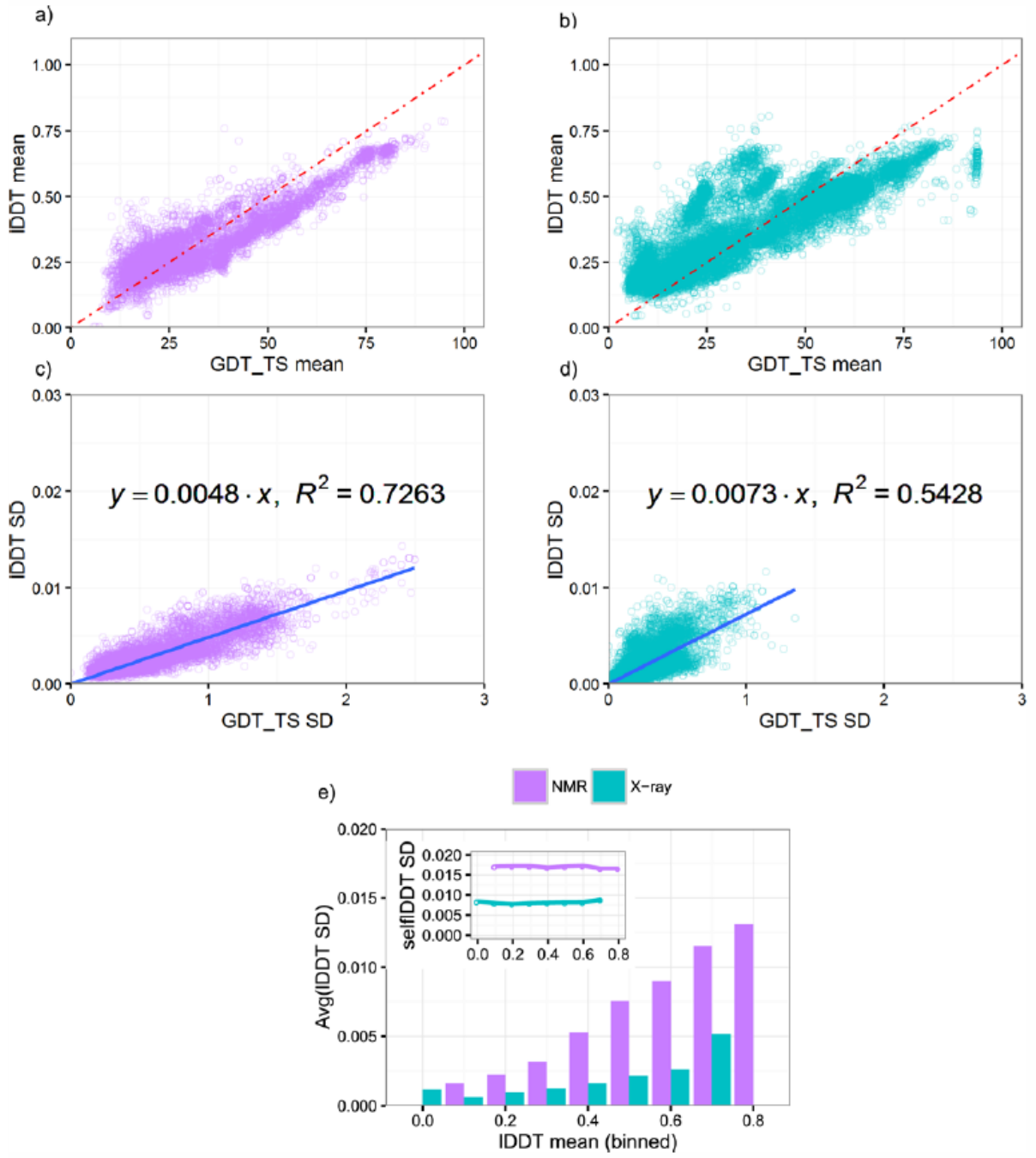


Fig 5. Comparison between IDDT and GDT_TS scores on generated ensembles. Comparisons of mean and SDs were shown for X-ray (cyan) and NMR (purple) ensembles, respectively. Panels a-d display the means and SDs of IDDT versus GDT_TS computed from same models, respectively. Red dash line denote the diagonal. Linear regression (blue line) of IDDT SDs with respect to GDT_TS SDs showed deviations from the 1:1 ratio between scores over the observed range. Panel e illustrated the binned IDDT SDs after normalization, which shows a similar trend to GDT_TS and GDT_HA SDs. The X-axis of the panels is labeled by the left edge of the bin. Bins with 0 models are not shown.

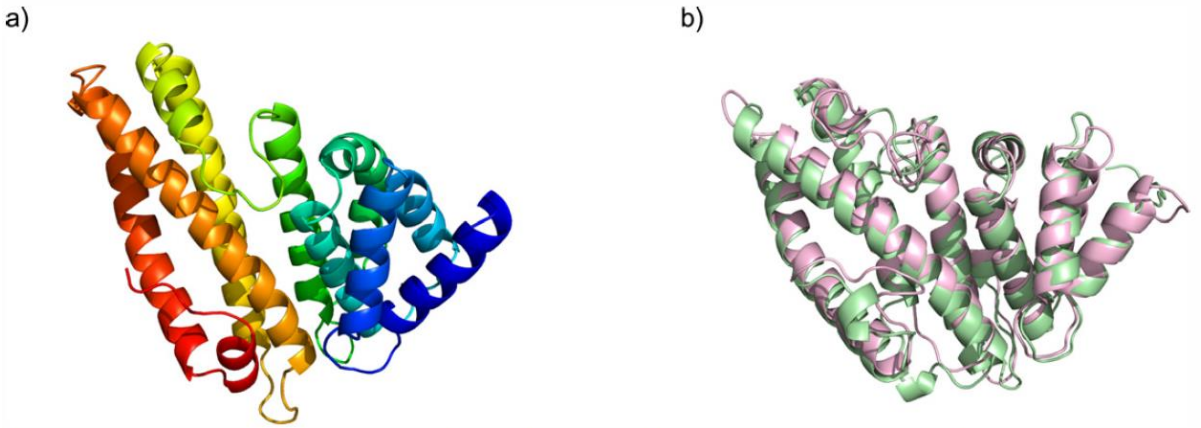


Fig 6. Structure comparison for highly similar models. Two CASP 11 models of TS0839 showed small differences in GDT_TS score, primarily due to differences in their modeled loop regions. Application of our uncertainty estimation reveals that the structural differences between these two models are insignificant and that the models are of similar quality. The model structure colored rainbow was displayed in panel a; the models TS0839TS410_1-D1 (pale-green, GDT_TS = 58.20) and TS839TS117_1-D1 (pink, GDT_TS = 57.72) are superimposed in panel b.

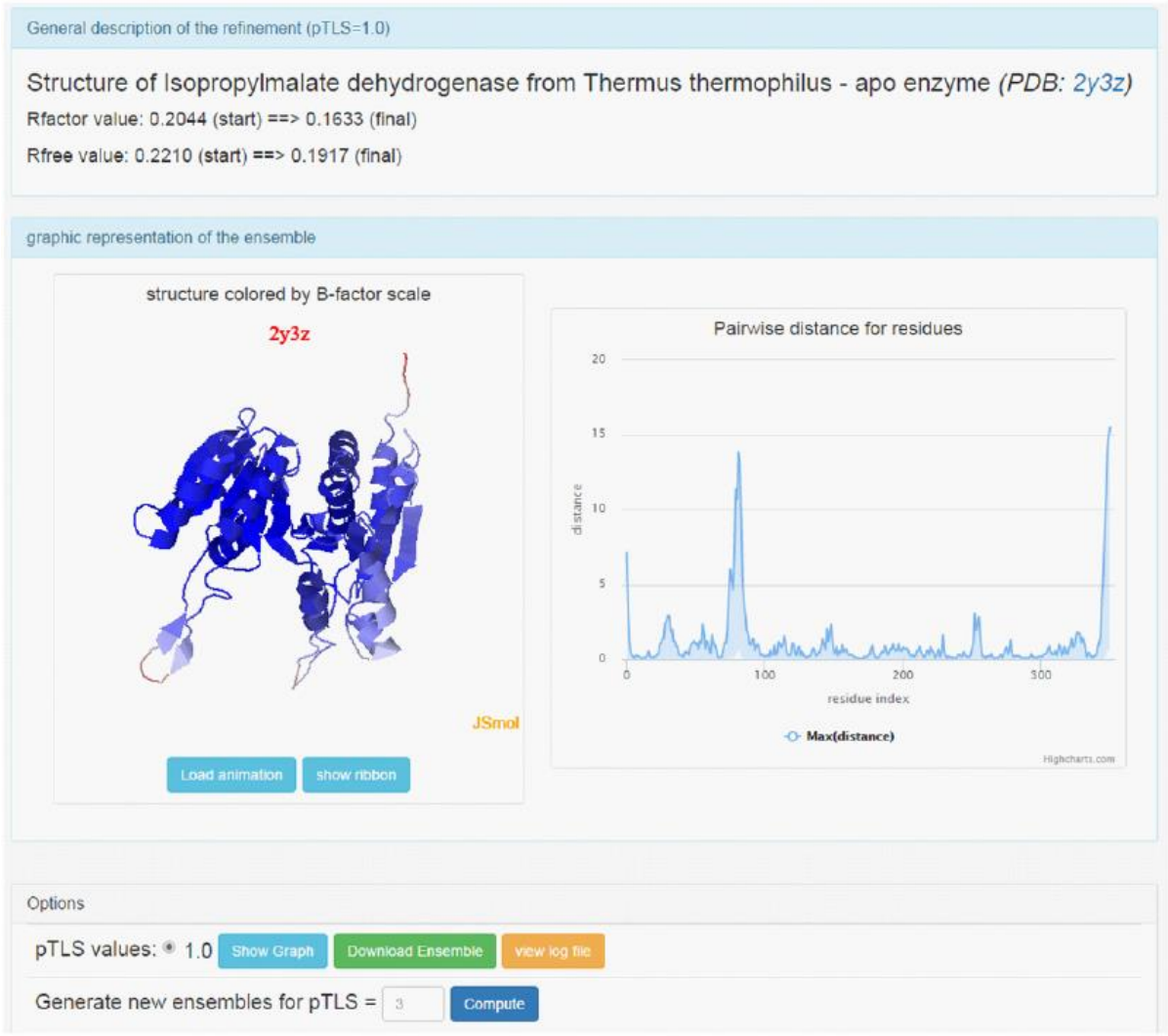


Fig 7. Snapshots of a result page from SEnCS server. This webpage is available via the link: http://prodata.swmed.edu/wenlin/server/enGen/result.php?pdb=2y3z.

REFERENCES

1. Moult J, Fidelis K, Zemla A, Hubbard T. Critical assessment of methods of protein structure prediction (CASP)-round V. Proteins. 2003;53 Suppl 6: 334–9. doi:10.1002/prot.10556
2. Zemla A. LGA: A method for finding 3D similarities in protein structures. Nucleic Acids Res. 2003;31: 3370–4. Available: http://www.ncbi.nlm.nih.gov/pubmed/12824330
3. Zemla A, Venclovas C, Moult J, Fidelis K. Processing and analysis of CASP3 protein structure predictions. Proteins. 1999;Suppl 3: 22–9. Available: http://www.ncbi.nlm.nih.gov/pubmed/10526349
4. Kinch L, Yong Shi S, Cong Q, Cheng H, Liao Y, Grishin N V. CASP9 assessment of free modeling target predictions. Proteins. 2011;79 Suppl 1: 59–73. doi:10.1002/prot.23181
5. Kinch LN, Li W, Monastyrskyy B, Kryshtafovych A, Grishin N V. Assessment of CASP11 Contact-Assisted Predictions. Proteins. 2016; doi:10.1002/prot.25020
6. Kryshtafovych A, Barbato A, Fidelis K, Monastyrskyy B, Schwede T, Tramontano A. Assessment of the assessment: evaluation of the model quality estimates in CASP10. Proteins. 2014;82 Suppl 2: 112–26. doi:10.1002/prot.24347
7. Henzler-Wildman K, Kern D. Dynamic personalities of proteins. Nature. 2007;450: 964–72. doi:10.1038/nature06522
8. Bernadó P, Blackledge M. Structural biology: Proteins in dynamic equilibrium.

Nature. 2010;468: 1046–8. doi:10.1038/4681046a

9. Wrabl JO, Gu J, Liu T, Schrank TP, Whitten ST, Hilser VJ. The role of protein conformational fluctuations in allostery, function, and evolution. Biophys Chem. 2011;159: 129–41. doi:10.1016/j.bpc.2011.05.020

10. Osawa M, Takeuchi K, Ueda T, Nishida N, Shimada I. Functional dynamics of proteins revealed by solution NMR. Curr Opin Struct Biol. 2012;22: 660–9. doi:10.1016/j.sbi.2012.08.007

11. Kuzmanic A, Pannu NS, Zagrovic B. X-ray refinement significantly underestimates the level of microscopic heterogeneity in biomolecular crystals. Nat Commun. 2014;5: 3220. doi:10.1038/ncomms4220

12. Berendsen HJ, Hayward S. Collective protein dynamics in relation to function. Curr Opin Struct Biol. 2000;10: 165–9. Available: http://www.ncbi.nlm.nih.gov/pubmed/10753809

13. Burnley BT, Afonine P V, Adams PD, Gros P. Modelling dynamics in protein crystal structures by ensemble refinement. Elife. 2012;1: e00311. doi:10.7554/eLife.00311

14. Huang YJ, Mao B, Aramini JM, Montelione GT. Assessment of template-based protein structure predictions in CASP10. Proteins. 2014;82 Suppl 2: 43–56. doi:10.1002/prot.24488

15. Joosten RP, Long F, Murshudov GN, Perrakis A. The PDB_REDO server for macromolecular structure model optimization. IUCrJ. 2014;1: 213–20. doi:10.1107/S2052252514009324

16. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000;28: 235–42. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=102472&tool=pmcentrez&rendertype=abstract

17. Winn MD, Isupov MN, Murshudov GN. Use of TLS parameters to model anisotropic displacements in macromolecular refinement. Acta Crystallogr D Biol Crystallogr. 2001;57: 122–33. Available: http://www.ncbi.nlm.nih.gov/pubmed/11134934

18. Kohn JE, Afonine P V, Ruscio JZ, Adams PD, Head-Gordon T. Evidence of functional protein dynamics from X-ray crystallographic ensembles. PLoS Comput Biol. 2010;6. doi:10.1371/journal.pcbi.1000911

19. Wallace BA, Janes RW, Bassolino DA, Krystek SR. A comparison of X-ray and NMR structures for human endothelin-1. Protein Sci. 1995;4: 75–83. doi:10.1002/pro.5560040110

20. Nabuurs SB, Spronk CAEM, Vuister GW, Vriend G. Traditional biomolecular structure determination by NMR spectroscopy allows for major errors. PLoS Comput Biol. 2006;2: e9. doi:10.1371/journal.pcbi.0020009

21. Kinch LN, Li W, Schaeffer RD, Dunbrack RL, Monastyrskyy B, Kryshtafovych A, et al. CASP 11 Target Classification. Proteins. 2016; doi:10.1002/prot.24982

22. Read RJ, Chavali G. Assessment of CASP7 predictions in the high accuracy template-based modeling category. Proteins. 2007;69 Suppl 8: 27–37. doi:10.1002/prot.21662

23. Mariani V, Biasini M, Barbato A, Schwede T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics. 2013;29: 2722–8. doi:10.1093/bioinformatics/btt473

24. Haas J, Roth S, Arnold K, Kiefer F, Schmidt T, Bordoli L, et al. The Protein Model Portal--a comprehensive resource for protein structure and model information. Database (Oxford). 2013;2013: bat031. doi:10.1093/database/bat031

25. Jmol: an open-source Java viewer for chemical structures in 3D. Available: http://www.jmol.org/

CHAPTER 7 CHSEQ: A DATABASE OF CHAMELEON SEQUENCES[6]

INTRODUCTION

Protein secondary structure elements have been viewed as the fundamental building blocks of protein tertiary structures.[1-3] The formation of α-helical and β-strand elements is induced by the interplay between local amino acid propensities and global interactions.[4-6] To investigate the influence of global interactions on the formation of secondary structures, researchers have discovered stretches of identical amino acid sequences that adopt distinct conformations, also called as chameleon sequences (ChSeqs).[7] Further studies[8] revealed the importance of such structural ambiguity in ChSeqs for a better understanding of amyloid diseases,[9-11] where native proteins can refold into β-strands to stabilize the pathogenic assemblies. Additionally, ChSeqs are reported to contribute to functional diversity described in alternatively spliced isoforms.[12]

The first search for ChSeqs in proteins was carried out by Kabsch and Sander.[13] They reported 25 chameleon pentapeptides from 62 protein structures. From then on, researchers have shown increased interest in the detection of ChSeqs.[12, 14-19] Besides analyzing the amino acid properties of ChSeqs, scientists have used ChSeqs to

---

[6] This Chapter was published as:
Li W, Kinch LN, Karplus PA, Grishin N V. *ChSeq: a database of chameleon sequences*. Protein Sci 2015;24(7):1075–1086.

evaluate the performance of secondary structure predictors.[12, 20, 21] Collectively, such evaluation studies showed that methods based on sequence profiles outperformed methods based on single sequences.[22] Surprisingly, the evaluations of neural network-based secondary structure predictors have shown that profile-based methods predict ChSeqs with similar efficiency as on sequences where alternative conformations are never observed.[21, 23]

To better understand the principles of protein structure changes, aided with increasing numbers of available Protein Data Bank (PDB)[24] structures, we searched for ChSeqs and identified a large set ranging from 6 to 10 in residue length. ChSeqs found in homologous structures tend to reveal conformational changes involved in switching protein functional states. Alternatively, the different environments surrounding ChSeqs from unrelated structures tend to dictate their conformation. We found that the evolutionary information provided by the sequence profiles can successfully predict the secondary structure feature that prevails in a given protein family. We present our dataset in a user-friendly web interface available at prodata.swmed.edu/chseq, as well as in csv format at http://prodata.swmed.edu/chseq/downloads/.

RESULTS AND DISCUSSION

Our comprehensive search for ChSeqs identified 19,603 (20 homologous and 19,583 unrelated) ChSeqs of entirely helix-to-strand transitions (Fig. 1) in the current nonredundant PDB database. For a fair comparison with the latest study,[18] which

detected ChSeqs with any secondary structure difference in the sequence strings, we also loosened our criteria and detected 128,703 ChSeqs in unrelated proteins with any helix-to-strand transition in the middle two residues of the sequence strings.

ChSeqs in homologous structures highlight dramatic conformational changes

We detected 20 ChSeqs that undergo complete helix-to-strand transitions in homologous structures. We found 12 of the 20 ChSeqs to be associated with biological functions (Table 1). Based on their experimental studies, the biological processes of the 12 ChSeqs can be classified into four types. First, the conformational changes upon activation (6 ChSeqs); these include the fusion protein of respiratory syncytial virus (2 ChSeqs),[25-27] the fusion protein of paramyxovirus (2 ChSeqs),[28, 29] the 50S ribosomal protein L24,[30, 31] and a cysteine proteinase.[32, 33] Second, the changes upon substrate binding (3 ChSeqs); these include the transcription factor Rfah (2 ChSeqs)[34, 35] and the 4Fe–4S cluster domain of human DNA primase.[36, 37] Third, the changes resulting from cleavage or insertion of a peptide (2 ChSeqs); these include the serine protease inhibitor ovalbumin[38, 39] and the cell surface adhesion molecule neurexin 1β.[40, 41] Fourth, the changes upon oligomerization (1 ChSeq); this includes a tubulin acetyltransferase.[42, 43]

The fusion protein in respiratory syncytial virus[25-27] contains one of the longest ChSeqs (10 residues), as well as another ChSeq of six residues (Fig. 2). In the prefusion structure (pdb: 4jhw, Chain F), the two ChSeqs together form a $\beta3_{176–181}/\beta4_{185–194}$ hairpin that packs against the "fusion peptide."[27] In the profusion structure (pdb: 3rki, Chain A),

each of the ChSeq strands transforms into a helical conformation, extending the “fusion peptide” helix and packing with the C-terminal helix to form a coiled coil stalk for membrane insertion.[26] As illustrated in this example, the ChSeqs undergo dramatic conformational changes and participate in the transition between the protein's inactive and active states.

The remaining eight ChSeq examples lack experimentally verified functions. Five of them come from structures of substantially different lengths (Table 1). The longer length structures form complete protein domains (determined by X-ray crystallography), whereas the shorter length structures are limited to several secondary structure elements (solved by NMR). As exemplified by the DH domains of Dab2 (illustrated in Fig. 3),[44, 45] we found that all the ChSeqs from truncated structures exhibit helical conformation. Alternately, the ChSeqs from the complete domains form β-strands. For example, in the complete DH domain (Fig. 3, pdb: 1p3r), the ChSeq β-strand (magenta) integrates into the center of an open β-barrel, forming a hydrogen bonding network with two neighboring β-strands (residues 92–97 and 145–151) that are missing in the truncated structure (Fig. 3, pdb: 2lsw). In the absence of the stabilizing hydrogen bonding network provided by the β-barrel, the single β-strand transforms into an α-helix in the shorter length structures. All five ChSeqs from truncated domains exhibit similar conformational transitions, suggesting that the helical conformations resulting from truncations are nonphysiological and caused by the lack of sufficient hydrogen bonding networks.

Two of the remaining three ChSeq examples include unpublished structures. For one, an unpublished NMR structure (pdb: 2mdk) of a human major sperm protein (MSP) domain contains an α-helix, whereas the crystal structure (pdb: 3ikk)[46] contains a β-strand. For the other, an unpublished crystal structure (pdb: 3lru) of a truncated human pre-mRNA processing factor 8 (Prp8) RNase H-like domain[47] exhibits a β-strand in a sheet formed by a swapped dimer, whereas a crystal structure of the complete domain (pdb: 4jke) has an α-helix. The last homologous ChSeq (sequence: AKEEAIKE) is from two engineered proteins designed to explore the mutation pathways for a single mutation to switch from an IgG-binding fold (α + β topology) into an albumin-binding fold (all-α topology).[48, 49]

Previous searches for ChSeqs either did not distinguish homologies of the ChSeqs[12, 16] or focused their searches on unrelated ChSeqs.[13, 15-19] However, some studies have investigated conformational diversity and structural motions present in the structures.[50-56]We examined whether our ChSeqs are also present in these studies. Although these studies collected redundant chains of close homologs (and we removed redundancy), five of the homologous ChSeqs we identified have been recorded in the "dynamic domains" (DynDom) database.[54] Recently, the database of conformational diversity in the native state of proteins (CoDNaS)[56] characterized structures of 100% sequence identity. The database for protein structural change upon ligand binding (PSCDB)[55] concentrated on the conformational changes on binding small molecules. As we used nonredundant structures and no conformational changes induced by binding small molecule were detected, none of our ChSeqs were reported in these two

most recent databases. We attempted to compare our ChSeqs with the database of protein conformational diversity (PCDB)[50]; however, the dataset seems to be no longer accessible through its website.

ChSeqs in unrelated structures illustrate the interplay between local and nonlocal interactions

We detected ChSeqs in unrelated structures using two different criteria. The more stringent search aims to detect entirely helix-to-strand transitions and detected 19,583 ChSeqs. However, the results using this set of criteria are not suitable for direct comparison with previous works. Therefore, we also searched with a looser criteria that allows shorter secondary structural element transitions; the detected ChSeqs increased to 128,703. When compared with previous studies (see Table 2), this search identified approximately 20-fold more ChSeqs. This increase corresponds well with the approximately 20-fold growth in the data size of nonredundant PDB structures (from 3214 to 67,589). The large number of hexamers detected is more than double the pentamer count in the most recent study.[18] We also increase the length of the longest ChSeqs identified from 8 to 10 (with four 10-mers seen here).[18]

ChSeqs that form different secondary structures in unrelated proteins were used to analyze the interplay between local and nonlocal interactions.[16-18] Such interactions can be illustrated in one of the 10-residue ChSeqs detected by loose criterion (Fig. 4). This ChSeq (sequence: QGTAVVVSAA) is found in an immunoglobulin fold (ECOD domain

ID: e4jb9H4) and a Rossmann fold (ECOD domain ID: e1vl6A1). In the immunoglobulin structure (pdb: 4jb9), the ChSeq forms a β-strand (residues 105–114) embedded in a β-sandwich; in the Rossmann-fold structure (pdb: 1v16), it forms a helix (residues 157–166). In this example, the ChSeq sequence includes a number of strong α-helix formers (e.g., A) and strong β-strand formers (e.g., V), as measured by Chou-Fasman parameters.60 This mix of strong but ambiguous α-helical and β-strand propensities is similar to that observed in a previous study of helix-to-strand transitions.16 In the immunoglobulin structure, nearby β-strands form a hydrogen-bonding network with the ChSeq to stabilize the extended conformation; in the Rossmann fold, the lack of surrounding hydrogen bonding partners allows the ChSeq to form a helix induced by strong α-helix propensity of its sequence [Fig. 4(b)]. Therefore, in this example, the global interactions impose constraints on the sequences of ambiguous secondary structure propensity, guiding local interactions to stabilize the respective secondary structures.

In the above example (Fig. 4), the ChSeq has a mixture of amino acids with ambiguous secondary structure preferences. We compared the amino acid frequencies of all detected ChSeqs (under the stringent criterion) with the amino acid frequencies of proteins in the Swiss-Prot database (Fig. 5). When compared with the frequencies in Swiss-Prot (green line in Fig. 5), the residues Ile, Val, Ala, and Leu are overrepresented in ChSeqs. As pointed out in previous analyses,12, 16 these residues have strong propensities in forming either α-helix (residues) or β-strand (residues). Alternately, Pro is underrepresented in ChSeqs consistent with its tendency to be both a helix and a strand breaker. Other residues with low Chou-Fasman60 helical or strand propensities, that is,

Gly, Ser, Asp, and Asn, also show low frequencies in ChSeqs. The low frequency of Cys can be explained by its potential to reduce structural flexibility through forming disulfide bonds.[12, 16] The low frequencies of Trp, His, Met, and Gln were also observed previously.[12, 16]

As has been noted[15] and was seen in the examples in Figure 4, ChSeqs tend to be largely buried in the protein core, forming interactions with surrounding secondary structure elements. To study the solvent exposure of residues in ChSeqs, we calculated the relative solvent accessibility (RSA), which indicates the percentage of surface area exposed to the solvent for a residue (Fig. 6). In general, when compared with residues in proteins, the distribution of RSAs in ChSeqs shows more fully buried residues (<5% RSA) and many fewer highly exposed residues (>85% RSA). However, when compared with residues contained in β-strands and α-helices, the distribution of RSAs in ChSeqs is comparable (green), indicating that the RSA decrease may be simply a result of the constraints of being in secondary structures.

Evaluation of secondary structure predictors on ChSeqs highlights the advantage of profile-based predictors

ChSeqs may be the most stringent test set for secondary structure predictors.[20, 21]Previous studies have applied profile-based secondary structure prediction methods to unrelated ChSeqs and have shown their high accuracy in predicting ChSeq secondary structures.[12, 21, 23] To study the influence of the evolutionary

information on the success of profile-based predictors, we applied both a profile-based predictor, here called psiP (for PSIPRED using sequence profile), and a single sequence-based predictor, here called psiS (for PSIPRED using single sequence), to the set of 655 ChSeqs with more than six residues. Consistent with previous evaluations, for the overwhelming majority, 92% (605/655) of ChSeqs, the profile-based psiP predicted correct secondary structures for both forms. Influenced by flanking residues, single-sequence-based psiS is in principle able to produce distinct predictions for sequences in a ChSeq pair; however, correct psiS secondary structure predictions for both forms are obtained for fewer than half, 42% (274/655), of the ChSeqs. Among the 58% ChSeqs that had incorrect predictions, for 96% (i.e., 56% of the 655 ChSeqs), the correct secondary structure is obtained for one of the families but not the other.

As was seen for the example ChSeq shown in Figure 4, psiS produced mainly β-strand predictions for both structures, whereas psiP could successfully distinguish the secondary structures from different protein structures. As shown in the secondary structure predictions for the ChSeq helix in the Rossmann fold [Fig. 4(c)], psiS predicts the “AVVV” stretch as a strand. However, the family profile includes alternate residues that allow psiP to correctly predict the AVVV as a helix. To quantify the prevalence of this type of alternate single-sequence-based prediction, we computed a prediction *P*-value (PPV) to indicate the probability of observing a given psiS prediction based on the psiS predictions carried out for every sequence in a given protein family. A lower PPV means the single-sequence prediction is more dissimilar to the prevailing psiS prediction among members of a protein family. The PPV distribution of incorrect psiS predictions for

ChSeqs is different from the distribution of psiS predictions for random sequences without observed helix-to-strand transitions (green line in Fig. 7). For incorrect psiS ChSeq predictions [blue bars in Fig. 7(a)], about one-third of the PPVs are below 0.05, indicating that the predictions significantly deviate from the prevailing predictions of family members. On the other hand, the distribution of ChSeqs with correct psiS prediction closely approximates the random distribution except at PPVs < 0.15 [Fig. 7(b)].

To study the influence of secondary structure type on the PPV distributions, we separately analyzed the helix and strand conformations. The PPV distributions [Fig. 8(a)] show that ChSeqs adopting strands have significantly lower PPVs than ChSeqs adopting helices, with a two-sided Kolmogorov–Smirnov (K-S) test *P*-value of $1.36\,e-06$. This indicates that psiS predictions for β-strands tend to deviate more from their prevailing family predictions than do the predictions for α-helices. This explains a clear asymmetry in the predictability of helices and strands in that, among all the ChSeqs, 42% had both α-helices and β-strands predicted correctly, 40% had only the α-helix predicted correctly, 16% had only the β-strand predicted correctly, and 2% had neither predicted correctly. Interestingly, if we further divide each conformation into those having correct versus incorrect psiS predictions, the PPV distributions are not distinguishable for either the correctly [Fig. 8(b)] or the incorrectly [Fig. 8(c)] predicted ChSeqs, with the K-S test *P*-values to be 0.21 and 0.39, respectively. Correctly predicted ChSeqs of both conformations tend to have higher PPVs [Fig. 8(b)], and incorrectly predicted ChSeqs of both conformations show a trend for lower PPVs [Fig. 8(c)]. Therefore, psiS predictions

from α-helices tend to match the prevailing family prediction more than β-strands, consistent with the higher fraction of correct predictions for α-helices.

Cross-validation of homologies by ECOD identified ChSeqs in unrelated regions of homologous protein folds

ECOD is an evolutionary classification of protein domains based on structural and sequence similarity, where structures within the same H-group are considered homologs.[59] As a cross-check of our homology assignments, we applied the ECOD classification to our BLAST-based ChSeq homologs. ECOD allowed us to correct classifications of three ChSeqs that are falsely found as homologs by BLAST due to multidomain problem.[61] Additionally, ECOD helped us to filter 65 ChSeqs that were in homologous proteins but did not represent homologous parts of the proteins. For example, the ChSeq shown in Figure 9 (with sequence: AIVLSKY) is from two structures classified by ECOD as homologous Rossmann folds (pdb: 3id6 and 4lg1); however, the ChSeq is in the N-terminal helix in one structure (4lg1) but in the C-terminal strand in another structure (3id6). The pairwise alignment of these two structure sequences is only limited to the ChSeq region (*E*-value 0.12), which is not sufficient to support their homology. Examples of unrelated ChSeqs in homologous folds are mainly concentrated in three large H-groups: the Rossmann fold (20 ChSeqs), the TIM barrel (16 ChSeqs), and the P-loop domain (8 ChSeqs).

We did not include 400 ChSeqs (1.8% of total ChSeqs) in our final dataset, as they could not be mapped to current ECOD domains. Those sequences include (i) 257 ChSeqs mapped to ECOD as peptides, coiled coils, fragments, and artificial sequences, for which homology cannot be inferred with confidence; and (ii) 143 ChSeqs mapped to the protein regions not covered by ECOD domains due to ECOD domain parsing limitations. These 400 ChSeqs are available at http://prodata.swmed.edu/wenlin/pdb_survey2/index.cgi/artifacts/.

A user-friendly web interface to the ChSeq database integrates a wide range of relevant information

For making this information accessible, we imported our dataset into a web interface (Fig. 10) that integrates structural and sequence information relevant for a ChSeq analysis. For efficiency, the default display includes only a single representative PDB entry for each form of a ChSeq, with a "show all PDB chains for this group" option to display all relevant PDB entries. Cross-database information, including protein names from PDB[24] and H-groups from ECOD,[59] is provided at the top of each panel. For more in-depth study of the structures, one can load the structure in JSmol[62] or download PyMol[63] session files (having a white protein chain with magenta ChSeq). In addition, below each image, the secondary structure (from PDB and psiS and psiP predictions) and sequence information (including gap fraction) are given along with a weblogo[64] visualization of the sequence profile of the ChSeq region. The full alignment

of the protein family is accessible via the link on the right of the weblogo image. This web interface to the ChSeq database is available through a portal at prodata.swmed.edu/chseq.

CONCLUSIONS

We have developed a rather comprehensive, updated dataset of ChSeqs. Interestingly, among the 20 examples of homologous ChSeqs that undergo helix-to-strand conformational changes, 12 were found to be involved in biological function. When compared with the most comprehensive previous study, we achieved a roughly 20-fold increase in detected unrelated ChSeqs (similar to the growth of the nonredundant PDB database in the relevant timeframe) and increased the length of the longest ChSeq from 8 to 10 residues. We find that for the ~56% of ChSeqs, for which a prediction based on single sequences is correct for only one of the families, there is a strong tendency for the sequence to be an "outlier" sequence for the other family. Its presence as a minority type of sequence in the family explains why it does not negatively impact the success of profile-based secondary structure predictions, which effectively capture the information present in the prevailing sequence patterns present in the family. A user-friendly web interface to the ChSeq database (available at prodata.swmed.edu/chseq) will facilitate future studies of ChSeqs and the gleaning of insights they can provide into the interplay between the influences of local and nonlocal interactions on protein structures.

MATERIALS AND METHODS

Detection of ChSeqs

The nonredundant PDB database, which combines structures of an identical sequence into one record, was downloaded on February 14, 2014, from ftp://ftp.ncbi.nih.gov/blast/db/FASTA/pdbaa.gz. The structures with Cα-atoms only were filtered. To select representative structures for each record, we prioritized crystal structures with the best resolution, followed by NMR structures, and then EM structures. We used a sliding window ranging from 6 to 40 to detect identical sequence strings. We further filtered out sequence strings contained in a longer sequence. The DSSP software65 was used to define ChSeq secondary structures from representative PDBs. We followed the DSSP nomenclature66 and reduced the eight DSSP secondary structure states into three: (1) "H," "G," and "I" as "H," (2) "E" and "B" as "E," and (3) others as "C." As a stringent criterion, we define ChSeqs1 as sequence strings with transitions between α-helices (H) and β-strands (E) in every position. To make our statistics comparable with previous studies, we also applied a looser criterion to define ChSeqs2 as segments for which helix-to-strand transitions occurred for the middle two residues of identical segments from unrelated proteins (for how relatedness was defined, see the next section).

Classification of ChSeqs by protein homology

We ran BLAST against the nonredundant PDB database to identify homologs for each structure. BLAST hits with an *E*-value better than 1 *e*−5 were considered homologs. As a cross-check, we also applied the ECOD59 classification to our dataset using H-groups (similar to SCOP67 superfamily) to define homologs. We manually inspected all the homologous ChSeqs detected by BLAST and ECOD to make sure that (1) structures of a homologous ChSeq are from only one ECOD H-group and that (2) homologous ChSeqs are aligned in the BLAST alignment with confident statistics.

Evaluation of PSIPRED prediction on ChSeqs

By default, the PSIPRED68 program runs PSIBLAST69 and uses the statistics from the sequence profile to perform prediction (denoted as psiP for "*P*rofile"). To study the influence of the sequence profile, we tweaked PSIPRED to use the statistics from the input sequence alone without running PSIBLAST (denoted as psiS for "*S*ingle" sequence). To evaluate the performance of psiP and psiS, we compared the secondary structure prediction with that found in the representative structures. The DSSP program has relatively strict criteria in defining α-helices and β-strands. As "C" might contain atypical helices or strands, we allowed mismatches against Cs and only penalized incorrect predictions between Es and Hs. We also allowed errors in defining the secondary structure boundary and only penalized the E and H mismatches in the middle

four residues of a ChSeq. Therefore, a correct prediction is defined as a prediction with no H versus E mismatches in the middle four residues of a ChSeq. To quantify the magnitude of the difference between the psiS and psiP predictions for a given sequence, we extracted the multiple sequence alignments (MSAs) used in psiP and calculated the prediction distance (D_p) for each sequence in the MSA using the following equation:

$$D_p = \sum_{i=1}^{n} ||V^i_{psiSS} - V^i_{psiSP}||,$$

where n is the length of the ChSeq, || || is the operator to calculate a Euclidean distance, and V^i_{psiSP} and V^i_{psiSS} are the probability vectors of secondary structure predictions for position i from psiP and psiS, respectively.

To indicate the extent to which the psiS of a sequence diverges from those that would be predicted by single sequences within its family, we estimated a PPV using the following equation:

$$PPV = \frac{N_{tail}}{N_{all}},$$

where N_{tail} is the number of D_ps larger than the D_p of the sequence, and N_{all} is the number of proteins in the MSA. To ensure the statistical significance of the PPVs, we filtered out protein families with $N_{all} < 150$.

Calculation of amino acid frequency and solvent accessibility

For the sequences of unrelated ChSeqs1 (i.e., those stringently defined), we calculated the frequencies of the 20 amino acid types. A set of reference frequencies of

amino acids was obtained by the amino acid frequencies of proteins in the Swiss-Prot[70] database available at http://web.expasy.org/protscale/pscale/A.A.Swiss-Prot.html. RSA was calculated as dividing the solvent accessibility (in $Å^2$) observed for each residue in a protein of interest (from DSSP) by the total surface area of the residue.[71] To estimate the RSA distribution in proteins, we sampled 1000 proteins from ChSeq-containing structures and calculated the RSA for every residue. To estimate the RSA distribution of α-helices and β-strands of length *N* (for comparison with ChSeqs of length *N*), we randomly selected a segment of N residues from the secondary structure elements (excluding coils) of ChSeq-containing structures and calculated the RSA for every residue.

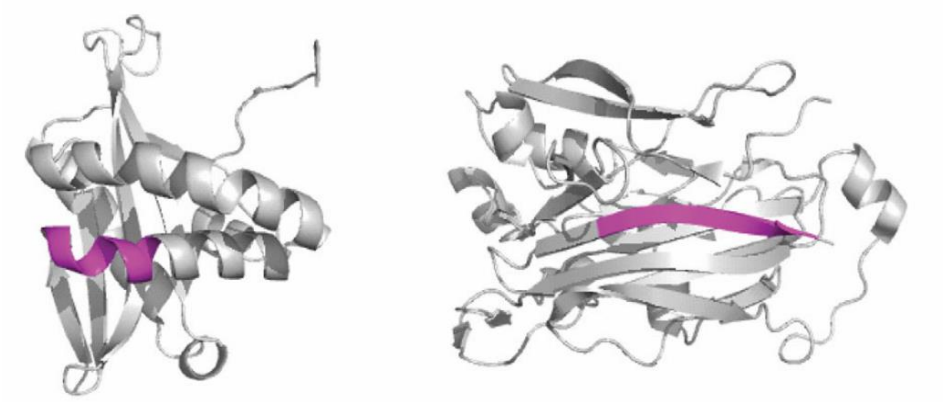
Filtering ambiguous and non-native sequences

We used the PDBx/mmCIF file of each structure in the PDB database to convert modified residues to their original (parent) residues. After our conversion, sequences containing unknown residues remained (e.g., the unknown residues in Chain D of pdb: 4hu6), which hindered our definition of identical sequence strings. Additionally, we detected protein expression tags near the termini by checking sequence conservation. Homologous sequences were retrieved by PSI-BLAST with three iterations against the UniRef90 database. The results were filtered to include sequences with *E*-value better than 0.001, identity larger than 30%, and gap positions smaller than 50% of the sequence length. The resulting positional gap fractions were calculated and rescaled to 0–9 (9 is

more gapped). If positions within 20 residues of either terminus had an average positional gap fraction larger than 6, we categorized the termini as protein expression tags. These ambiguous and non-native sequences (8.5% of total ChSeqs) can be found at http://prodata.swmed.edu/wenlin/pdb_survey2/index.cgi/artifacts/.

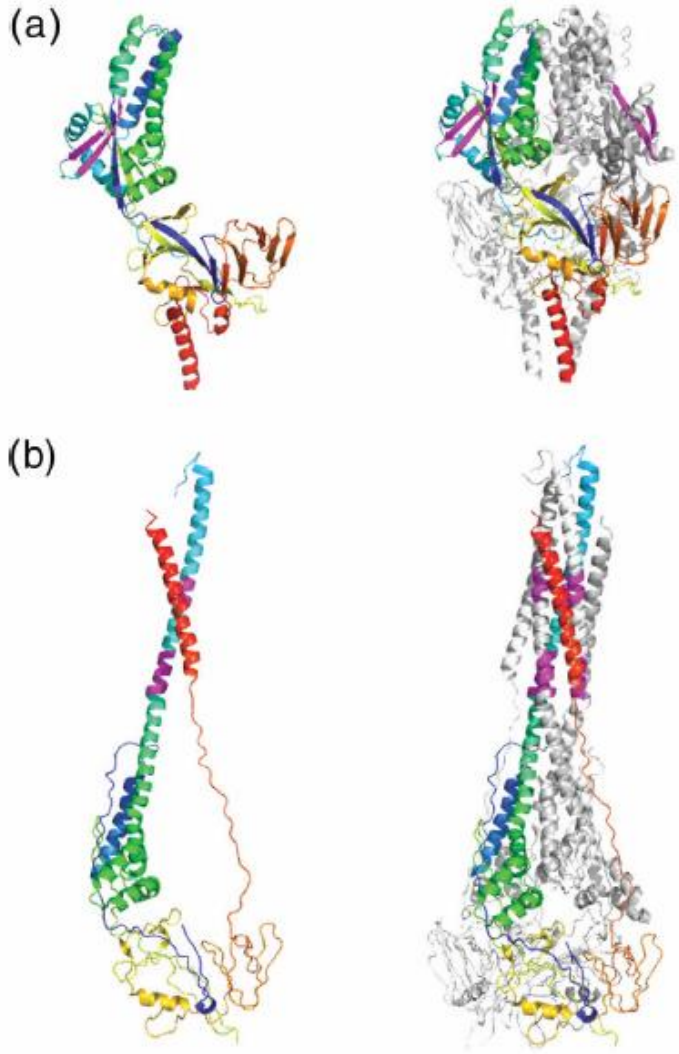
Preparation of the web interface

To reduce redundancy for web visualization, we clustered the ChSeqs by their secondary structure elements such that each cluster contains ChSeqs of identical secondary structures. For unrelated ChSeqs, these clusters were further split according to ECOD H-groups. By default, we show the most diverse representative pair on top. In the downloadable PyMol63sessions of the structures, we limit to unique chains containing ChSeqs to reduce the file size. The MSAs used in detecting protein expression tags are included in the web interface.



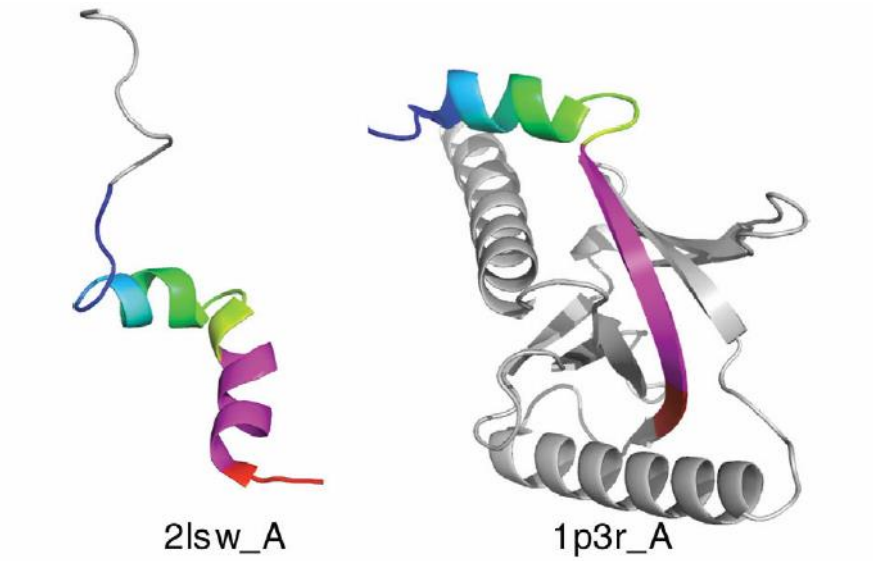
```
2Q0Y_A 1   GMECRPLCIDDLELVCRHREA 21
3S30_A 334 GAPGSTLLIDDLELVCKQPLR 354
```

Figure 1. Chameleon sequences (ChSeqs) and their distributions in homologous and unrelated proteins. A ChSeq adopting different conformations. The pdb codes are 2Q0Y (left) and 3S30 (right), respectively. ChSeqs are colored magenta in both the structure and sequence.



(c)
```
Sequence: STNKAVVSLSNGVSVLTSKVLDLKNY
4jhw_F(a):CCCEEEEEECCCEEEEEEEEEHHHH
3rki_A(b):HHHHHHHHHHHHHHHHHHHHHHHHH
```

Figure 2. Conformational changes in Type I fusion protein of respiratory syncytial virus. (a) ChSeqs (colored magenta) between residues 185–194 and 176–181 (pdb: 4jhw, Chain F) form a β-hairpin in the prefusion complex (pdb: 4jhw) illustrated in rainbow as monomeric (left panel) and trimeric (right panel). (b) The ChSeqs form helical conformations in the profusion complex (pdb: 3rki, Chain A) illustrated as above. (c) The sequence and the corresponding secondary structures of the ChSeq segments in prefusion (Line 2: 4jhw) and profusion (Line 3: 3rki) complexes.




```
2LSW_A 33 EKTDEYLLARFKGDGVKYKAKLIGID 58
dssp      CCCCCCHHHHHCCCHHHHHHHHHCCC
1P3R_A 33 EKTDEYLLARFKGDGVKYKAKLIGID 58
dssp      CCCHHHHHHHHCCCCEEEEEEEEEE
```

Figure 3. ChSeqs in proteins of different lengths. The region of identical sequences is shown in the alignment and colored rainbow in the structures. ChSeqs are colored magenta.

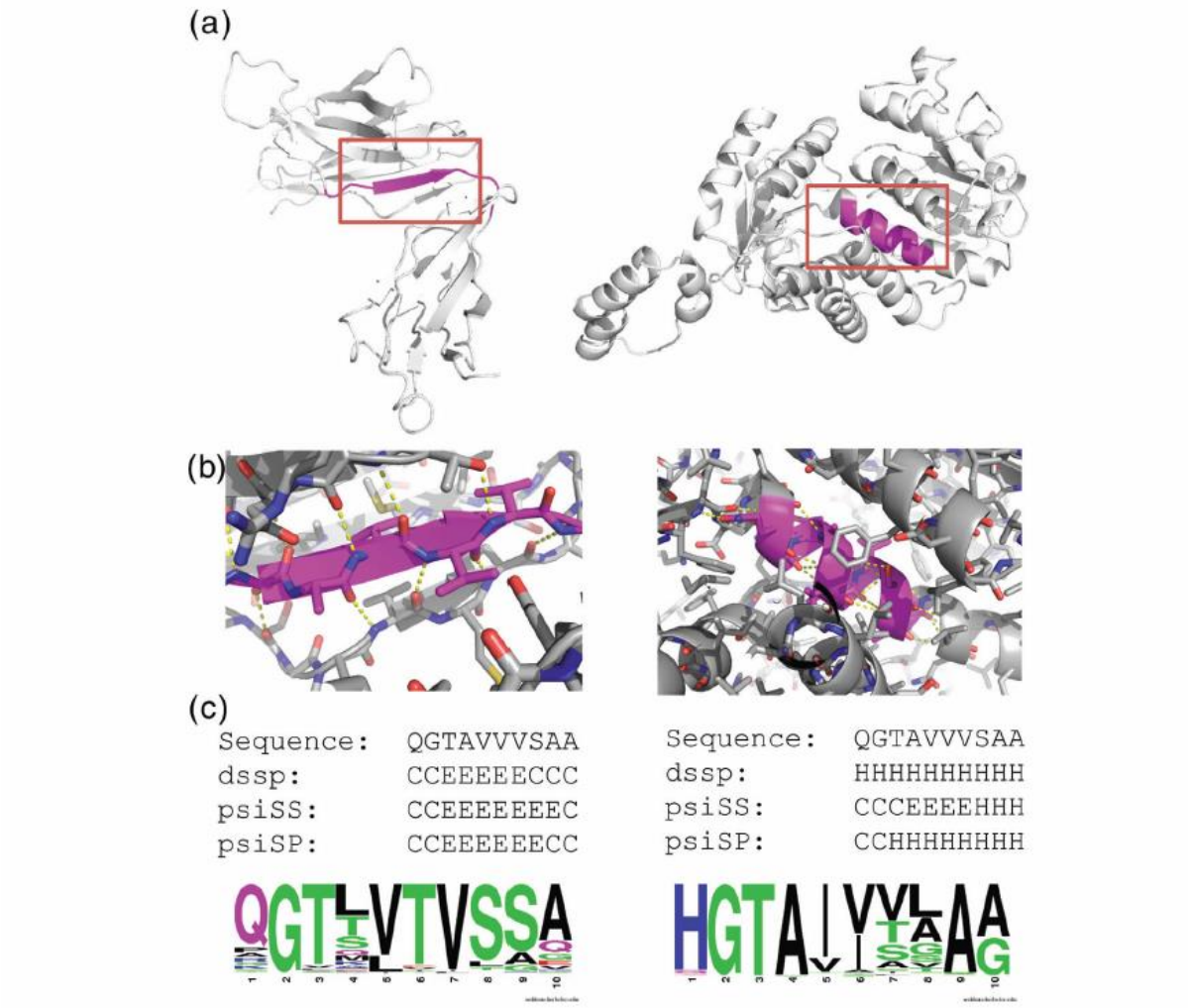


Figure 4. Example of a 10-residue ChSeq in unrelated proteins. (a) ChSeqs (magenta) in the structures 4JB9 (left) and 1VL6 (right). (b) Close-ups of red box regions of panel (a) with some backbone hydrogen bonds (dashed yellow lines) shown. (c) Sequence, observed secondary structure, and psiS- and psiP-predicted secondary structure are shown along with weblogo pictures visualizing the sequence profiles in each protein family.

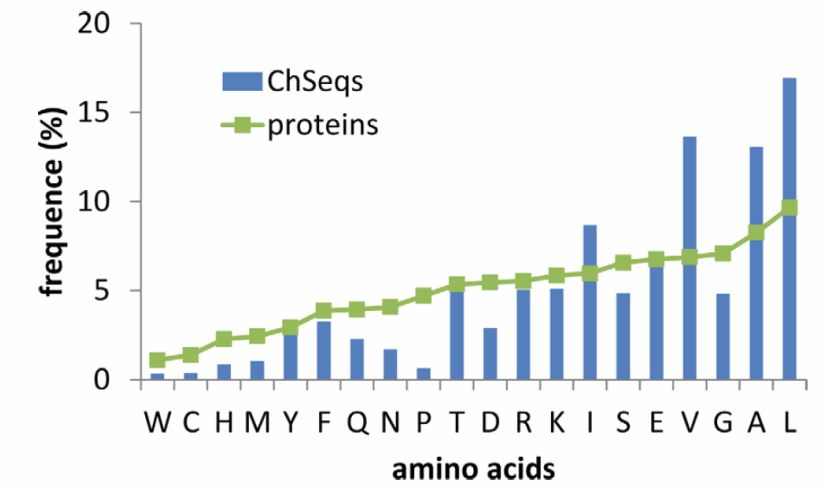


Figure 5. Amino acid composition of ChSeqs. Amino acid frequencies in ChSeqs (blue) are compared with the frequencies seen in proteins from the Swiss-Prot database (green).

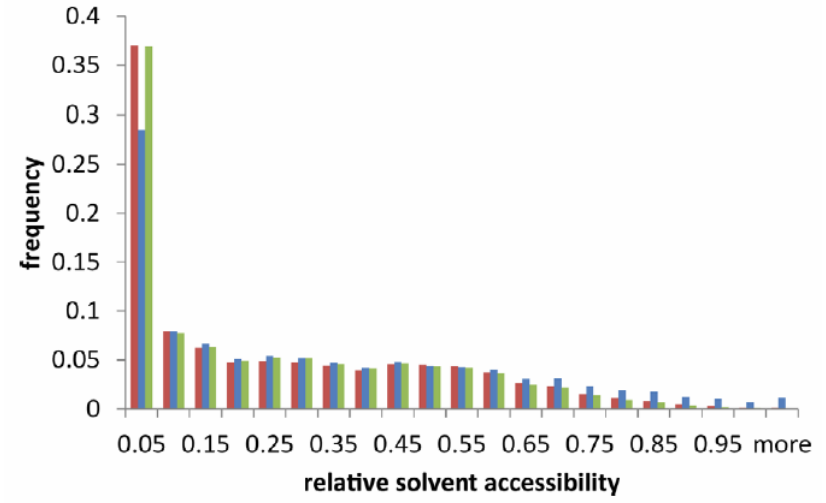


Figure 6. ChSeqs are similarly buried as residues in strands and helices. Histogram of the RSA distribution of residues in "stringent" ChSeqs (red), in a set of 1000 random proteins (blue), and in a set of "random" β-strands and α-helices (green).

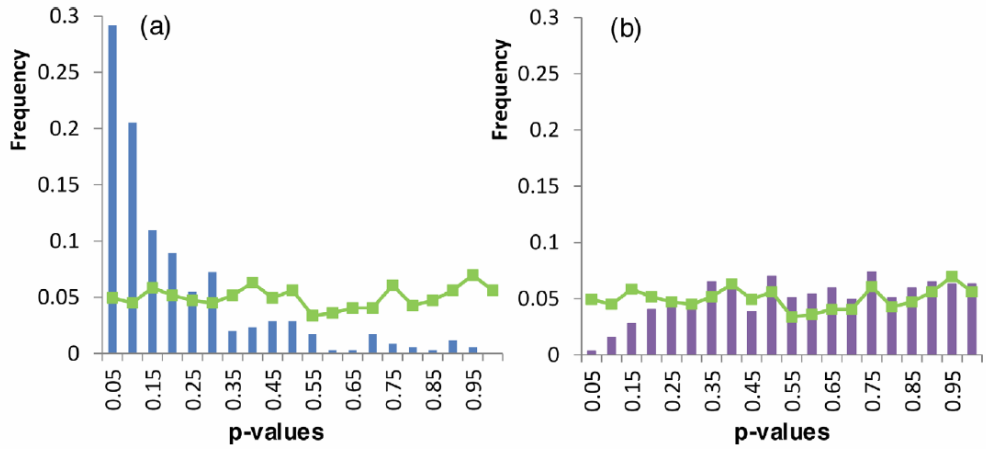


Figure 7. Histograms of prediction *P*-values (PPVs) for ChSeqs with (a) incorrect psiS predictions and (b) correct psiS predictions. Green lines represent the PPVs for controls computed from a random sequence from the family.

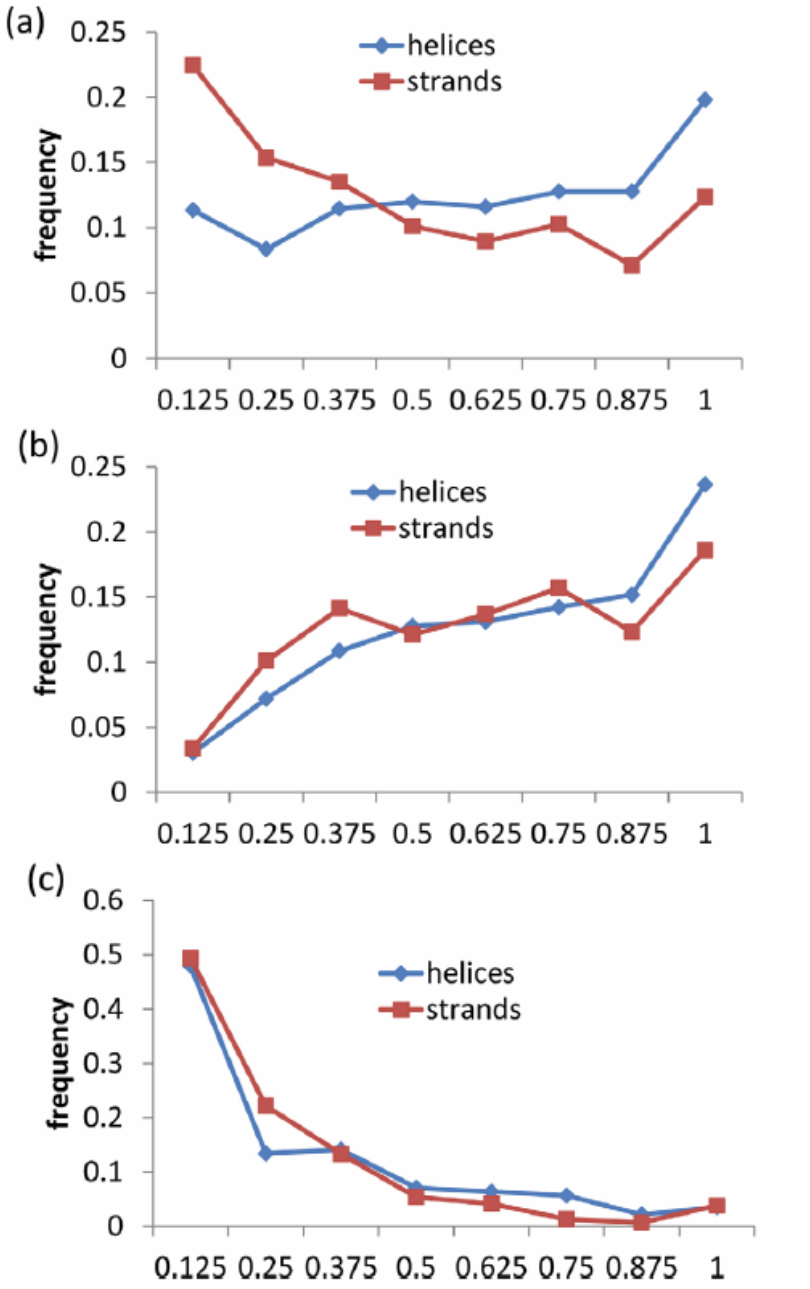


Figure 8. Histograms of PPVs for ChSeqs with helical (red) and stranded (blue) conformations. All studied ChSeqs (a) are further divided into those with correct psiS predictions (b) and incorrect predictions (c).

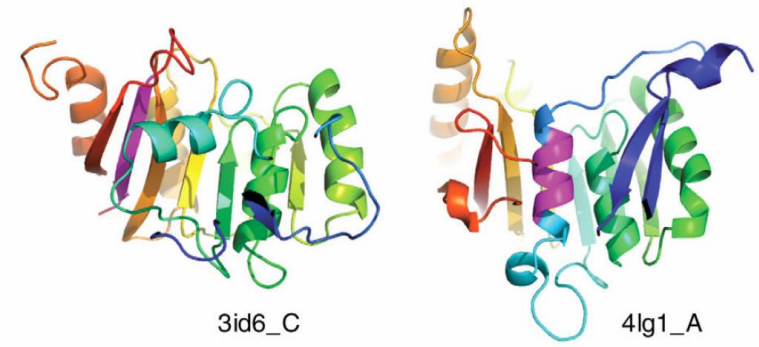


Figure 9. Nonhomologous ChSeq in homologous proteins. The ChSeq (purple) is highlighted in the two ribbon diagrams, and the BLAST alignment is shown.

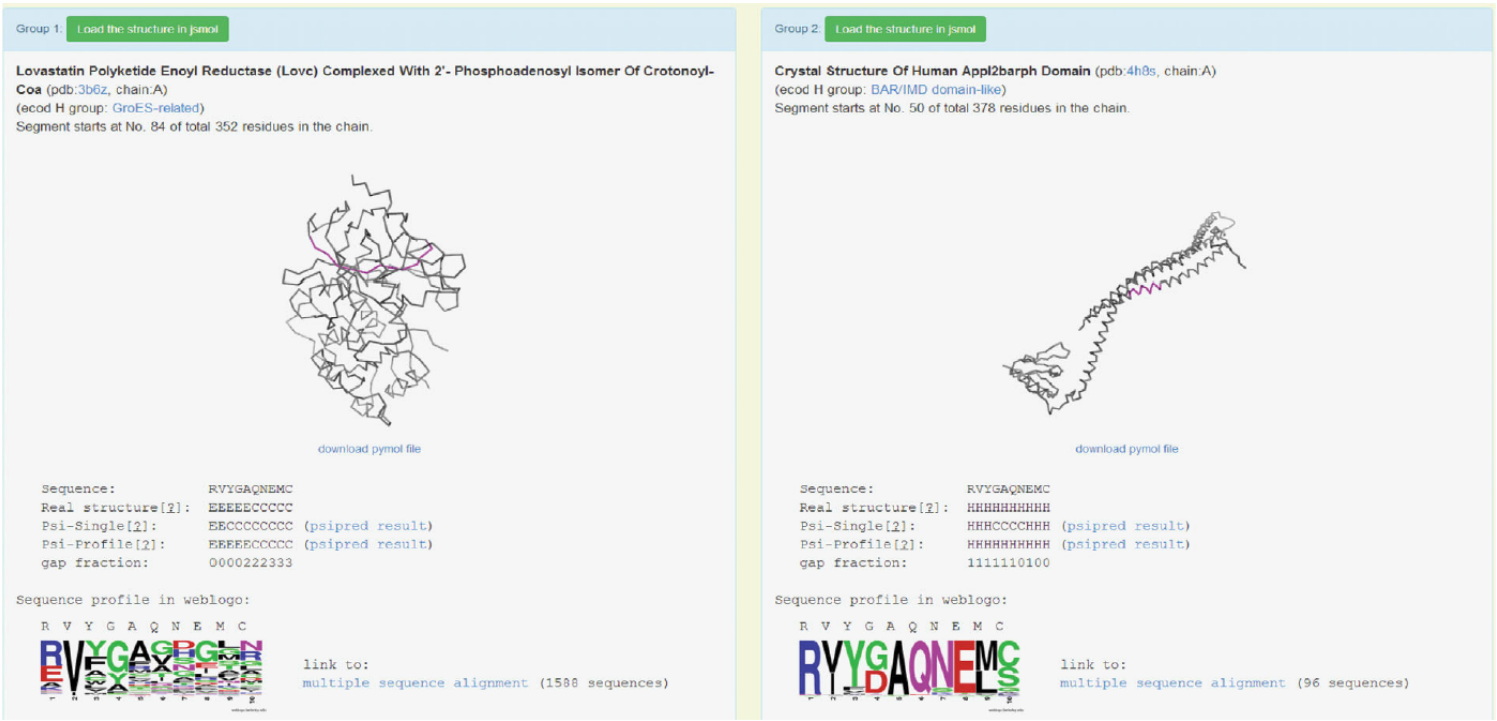


Figure 10. An example web interface. This shows a ChSeq that occurs in unrelated proteins (accessible at http://prodata.swmed.edu/wenlin/pdb_survey2/index.cgi/new_dssp/middle-match/RVYGAQNEMC/).

Table I. *ChSeqs in Homologous Proteins*

Sequence	pdb1	length1	pdb2	length2	Alignment length[a]	Alignment fraction[b] (%)	Annotation[c]	Protein name[d]
VSVLTSKVLD	4jhwF	498	3rkiA	528	454	86	Functional	Fusion protein of respiratory syncytial virus
MDSKLRCVFE	3ikkA	127	2mdkA	125	124	98	Unpublished	hVAPB MSP domain
IKASQELV	3n4pA	279	2kn8A	68	68	24	Fragment	Human cytomegalovirus terminase nuclease domain
SAEAGVDA[f]	1jtiA	385	1ovaD	386	383	99	Functional	Serine protease inhibitor ovalbumin
AKEEAIKE	2kdmA	56	2jwsA	56	51	91	Engineer	GA95 and GB95
VKYKAKLI[e]	1p3rA	160	2lswA	40	26	16	Fragment	Phosphotyrosin binding domain (Ptb) of mouse disabled 2
EIKHSVK	2ldA	66	2ougA	162	62	38	Functional	Transcription factor Rfah
RSMLLLN	2ldA	66	2ougA	162	62	38	Functional	Transcription factor Rfah
LGRVVDE	3mw3A	208	2r1bA	220	168	76	Functional	Cell surface adhesion molecule neurexin 1β
LDPLEVH	3lruA	160	4jkeA	222	160	72	Unpublished	Human Prp8 Rnase H-like domain
QSLGTAV[e]	4gipD	409	1svfA	64	63	15	Functional	Fusion protein of paramyxovirus
FKKIKVL	2rfeA	324	1z9iA	53	20	6	Fragment	Epidermal growth factor receptor
KILVQA[e]	1p3hA	99	1p82A	25	24	24	Fragment	Mycobacterium tuberculosis chaperonin 10
RLFQVK	3ffnA	782	1solA	20	20	3	Fragment	Calcium-free human gelsolin
VADVVQ[e]	4gipD	409	1svfA	64	63	15	Functional	Fusion protein of paramyxovirus
KKVRFF	3r8sU	102	2gyaS	99	99	97	Functional	50S ribosomal protein L24
LIEYFR	3ly6A	697	2q3zA	687	683	98	Functional	Cysteine proteinase
SYNIRH	3l9qA	195	3q36A	192	186	95	Functional	4Fe–4S cluster domain of human DNA primase
KAVVSL	4jhwF	498	3rkiA	528	454	86	Functional	Fusion protein of respiratory syncytial virus
TVIDEL	4h6zA	190	4hkfA	191	186	97	Functional	Tubulin acetyltransferase

[a] Alignment length: the length of the alignments between pdb1 and pdb2.
[b] Alignment fraction: the alignment length divided by the maximum of length1 and length2.
[c] Annotation: categorization of conformational differences in ChSeqs, including conformational changes (i) with associated function (functional), (ii) in protein of diverse lengths (fragment), (iii) involving unpublished structures (unpublished), and (iv) in engineered proteins (engineer).
[d] Protein name: the protein name summarized from the pdb entries.
[e] One structure of the ChSeqs recorded in the DynDom database.
[f] Both structures of the ChSeqs recorded in the DynDom database.

REFERENCES

1. Ballew RM, Sabelko J, Gruebele M. Direct observation of fast protein folding: the initial collapse of apomyoglobin. Proc Natl Acad Sci U S A 1996;93(12):5759–5764.
2. Freund SM, Wong KB, Fersht AR. Initiation sites of protein folding by NMR analysis. Proc Natl Acad Sci U S A 1996;93(20):10600–10603.
3. Han KF, Baker D. Global properties of the mapping between local amino acid sequence and local structure in proteins. Proc Natl Acad Sci U S A 1996;93(12):5814–5818.
4. Socci ND, Onuchic JN, Wolynes PG. Protein folding mechanisms and the multidimensional folding funnel. Proteins 1998;32(2):136–158.
5. Dill KA. Polymer principles and protein folding. Protein Sci 1999;8(6):1166–1180.
6. Gross M. Protein folding: think globally, (inter)act locally. Curr Biol 1998;8(9):R308–R309.
7. Minor DL, Kim PS. Context-dependent secondary structure formation of a designed protein sequence. Nature 1996;380(6576):730–734.
8. Gendoo DMA, Harrison PM. Discordant and chameleon sequences: their distribution and implications for amyloidogenicity. Protein Sci 2011;20(3):567–579.
9. Aguzzi A, Sigurdson C, Heikenwaelder M. Molecular mechanisms of prion pathogenesis. Annu Rev Pathol 2008;3:11–40.

10. Caughey B, Lansbury PT. Protofibrils, pores, fibrils, and neurodegeneration: separating the responsible protein aggregates from the innocent bystanders. Annu Rev Neurosci 2003;26:267–298.
11. Chiti F, Dobson CM. Protein misfolding, functional amyloid, and human disease. Annu Rev Biochem 2006;75:333–366.
12. Guo J-T, Jaromczyk JW, Xu Y. Analysis of chameleon sequences and their implications in biological processes. Proteins 2007;67(3):548–558.
13. Kabsch W, Sander C. On the use of sequence homologies to predict protein structure: identical pentapeptides can have completely different conformations. Proc Natl Acad Sci U S A 1984;81(4):1075–1078.
14. Wilson IA, Haft DH, Getzoff ED, Tainer JA, Lerner RA, Brenner S. Identical short peptide sequences in unrelated proteins can have different conformations: a testing ground for theories of immune recognition. Proc Natl Acad Sci U S A 1985;82(16):5255–5259.
15. Cohen BI, Presnell SR, Cohen FE. Origins of structural diversity within sequentially identical hexapeptides. Protein Sci 1993;2(12):2134–2145.
16. Zhou X, Alber F, Folkers G, Gonnet GH, Chelvanayagam G. An analysis of the helix-to-strand transition between peptides with identical sequence. Proteins 2000;41(2):248–256.
17. Kuznetsov IB, Rackovsky S. On the properties and sequence context of structurally ambivalent fragments in proteins. Protein Sci 2003;12(11):2420–2433.

18. M Saravanan K, Selvaraj S. Search for identical octapeptides in unrelated proteins: Structural plasticity revisited. Biopolymers 2012;98(1):11–26.

19. Ghozlane A, Joseph AP, Bornot A, Brevern AG de. Analysis of protein chameleon sequence characteristics. Bioinformation 2009;3(9):367–369.

20. Saravanan KM, Selvaraj S. Performance of secondary structure prediction methods on proteins containing structurally ambivalent sequence fragments. Biopolymers 2013;100(2):148–153.

21. Jacoboni I, Martelli PL, Fariselli P, Compiani M, Casadio R. Predictions of protein segments with the same aminoacid sequence and different secondary structure: a benchmark for predictive methods. Proteins 2000;41(4):535–544.

22. Rost B, Sander C. Third generation prediction of secondary structures. Methods Mol Biol 2000;143:71–95.

23. Rost B. Review: protein secondary structure prediction continues to rise. J Struct Biol;134(2-3):204–218.

24. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28(1):235–242.

25. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22(12):2577–2637.

26. Eyrich VA, Przybylski D, Koh IYY, Grana O, Pazos F, Valencia A, Rost B. CAFASP3 in the spotlight of EVA. Proteins 2003;53 Suppl 6:548–560.

27. Cheng H, Schaeffer RD, Liao Y, Kinch LN, Pei J, Shi S, Kim B-H, Grishin N V. ECOD: An Evolutionary Classification of Protein Domains. PLoS Comput Biol 2014;10(12):e1003926.

28. Conte L Lo, Ailey B, Hubbard TJ, Brenner SE, Murzin AG, Chothia C. SCOP: a structural classification of proteins database. Nucleic Acids Res 2000;28(1):257–259.

29. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292(2):195–202.

30. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25(17):3389–3402.

31. Boutet E, Lieberherr D, Tognolli M, Schneider M, Bairoch A. UniProtKB/Swiss-Prot. Methods Mol Biol 2007;406:89–112.

32. Zamyatnin AA. Protein volume in solution. Prog Biophys Mol Biol 1972;24:107–123.

33. The PyMOL Molecular Graphics System, Version 1.5.0.4 Schrödinger, LLC. .

34. Heerens AT, Marshall DD, Bose CL. Nosocomial respiratory syncytial virus: a threat in the modern neonatal intensive care unit. J Perinatol 2002;22(4):306–307.

35. Swanson KA, Settembre EC, Shaw CA, Dey AK, Rappuoli R, Mandl CW, Dormitzer PR, Carfi A. Structural basis for immunization with postfusion respiratory syncytial virus fusion F glycoprotein (RSV F) to elicit high neutralizing antibody titers. Proc Natl Acad Sci U S A 2011;108(23):9619–9624.

36. McLellan JS, Chen M, Leung S, Graepel KW, Du X, Yang Y, Zhou T, Baxa U, Yasuda E, Beaumont T, Kumar A, Modjarrad K, Zheng Z, Zhao M, Xia N, Kwong PD, Graham BS. Structure of RSV fusion glycoprotein trimer bound to a prefusion-specific neutralizing antibody. Science 2013;340(6136):1113–1117.

37. Baker KA, Dutch RE, Lamb RA, Jardetzky TS. Structural basis for paramyxovirus-mediated membrane fusion. Mol Cell 1999;3(3):309–319.

38. Yin H-S, Wen X, Paterson RG, Lamb RA, Jardetzky TS. Structure of the parainfluenza virus 5 F protein in its metastable, prefusion conformation. Nature 2006;439(7072):38–44.

39. Mitra K, Schaffitzel C, Fabiola F, Chapman MS, Ban N, Frank J. Elongation arrest by SecM via a cascade of ribosomal RNA rearrangements. Mol Cell 2006;22(4):533–543.

40. Dunkle JA, Wang L, Feldman MB, Pulk A, Chen VB, Kapral GJ, Noeske J, Richardson JS, Blanchard SC, Cate JHD. Structures of the bacterial ribosome in classical and hybrid states of tRNA binding. Science 2011;332(6032):981–984.

41. Han B-G, Cho J-W, Cho YD, Jeong K-C, Kim S-Y, Lee B Il. Crystal structure of human transglutaminase 2 in complex with adenosine triphosphate. Int J Biol Macromol 2010;47(2):190–195.

42. Pinkas DM, Strop P, Brunger AT, Khosla C. Transglutaminase 2 undergoes a large conformational change upon activation. PLoS Biol 2007;5(12):e327.

43. Burmann BM, Knauer SH, Sevostyanova A, Schweimer K, Mooney RA, Landick R, Artsimovitch I, Rösch P. An α helix to β barrel domain switch transforms the transcription factor RfaH into a translation factor. Cell 2012;150(2):291–303.

44. Belogurov GA, Vassylyeva MN, Svetlov V, Klyuyev S, Grishin N V, Vassylyev DG, Artsimovitch I. Structural basis for converting a general transcription factor into an operon-specific virulence regulator. Mol Cell 2007;26(1):117–129.

45. Agarkar VB, Babayeva ND, Pavlov YI, Tahirov TH. Crystal structure of the C-terminal domain of human DNA primase large subunit: implications for the mechanism of the primase-polymerase α switch. Cell Cycle 2011;10(6):926–931.

46. Vaithiyalingam S, Warren EM, Eichman BF, Chazin WJ. Insights into eukaryotic DNA priming from the structure and functional interactions of the 4Fe-4S cluster domain of human DNA primase. Proc Natl Acad Sci U S A 2010;107(31):13684–13689.

47. Yamasaki M, Arii Y, Mikami B, Hirose M. Loop-inserted and thermostabilized structure of P1-P1' cleaved ovalbumin mutant R339T. J Mol Biol 2002;315(2):113–120.

48. Stein PE, Leslie AG, Finch JT, Carrell RW. Crystal structure of uncleaved ovalbumin at 1.95 A resolution. J Mol Biol 1991;221(3):941–959.

49. Shen KC, Kuczynska DA, Wu IJ, Murray BH, Sheckler LR, Rudenko G. Regulation of neurexin 1beta tertiary structure and ligand binding through alternative splicing. Structure 2008;16(3):422–431.

50. Koehnke J, Katsamba PS, Ahlsen G, Bahna F, Vendome J, Honig B, Shapiro L, Jin X. Splice form dependence of beta-neurexin/neuroligin binding interactions. Neuron 2010;67(1):61–74.

51. Kormendi V, Szyk A, Piszczek G, Roll-Mecak A. Crystal structures of tubulin acetyltransferase reveal a conserved catalytic core and the plasticity of the essential N terminus. J Biol Chem 2012;287(50):41569–41575.

52. Li W, Zhong C, Li L, Sun B, Wang W, Xu S, Zhang T, Wang C, Bao L, Ding J. Molecular basis of the acetyltransferase activity of MEC-17 towards α-tubulin. Cell Res 2012;22(12):1707–1711.

53. Yun M, Keshvara L, Park C-G, Zhang Y-M, Dickerson JB, Zheng J, Rock CO, Curran T, Park H-W. Crystal structures of the Dab homology domains of mouse disabled 1 and 2. J Biol Chem 2003;278(38):36572–36581.

54. Xiao S, Charonko JJ, Fu X, Salmanzadeh A, Davalos R V, Vlachos PP, Finkielstein C V, Capelluto DGS. Structure, sulfatide binding properties, and inhibition of platelet aggregation by a disabled-2 protein-derived peptide. J Biol Chem 2012;287(45):37691–37702.

55. Shi J, Lua S, Tong JS, Song J. Elimination of the native structure and solubility of the hVAPB MSP domain by the Pro56Ser mutation that causes amyotrophic lateral sclerosis. Biochemistry 2010;49(18):3887–3897.

56. Schellenberg MJ, Wu T, Ritchie DB, Fica S, Staley JP, Atta KA, LaPointe P, MacMillan AM. A conformational switch in PRP8 mediates metal ion

coordination that promotes pre-mRNA exon ligation. Nat Struct Mol Biol 2013;20(6):728–734.

57. Alexander PA, He Y, Chen Y, Orban J, Bryan PN. A minimal sequence code for switching protein structure and function. Proc Natl Acad Sci U S A 2009;106(50):21149–21154.

58. He Y, Chen Y, Alexander P, Bryan PN, Orban J. NMR structures of two designed proteins with high sequence identity but different fold and function. Proc Natl Acad Sci U S A 2008;105(38):14412–14417.

59. Juritz EI, Alberti SF, Parisi GD. PCDB: a database of protein conformational diversity. Nucleic Acids Res 2011;39(Database issue):D475–D479.

60. Gerstein M, Krebs W. A database of macromolecular motions. Nucleic Acids Res 1998;26(18):4280–4290.

61. Flores S, Echols N, Milburn D, Hespenheide B, Keating K, Lu J, Wells S, Yu EZ, Thorpe M, Gerstein M. The Database of Macromolecular Motions: new features added at the decade mark. Nucleic Acids Res 2006;34(Database issue):D296–D301.

62. Lee RA, Razaz M, Hayward S. The DynDom database of protein domain motions. Bioinformatics 2003;19(10):1290–1291.

63. Qi G, Lee R, Hayward S. A comprehensive and non-redundant database of protein domain movements. Bioinformatics 2005;21(12):2832–2838.

64. Amemiya T, Koike R, Kidera A, Ota M. PSCDB: a database for protein structural change upon ligand binding. Nucleic Acids Res 2012;40(Database issue):D554–D558.

65. Monzon AM, Juritz E, Fornasari MS, Parisi G. CoDNaS: a database of conformational diversity in the native state of proteins. Bioinformatics 2013;29(19):2512–2514.

66. Chou PY, Fasman GD. Prediction of the secondary structure of proteins from their amino acid sequence. Adv Enzymol Relat Areas Mol Biol 1978;47:45–148.

67. Kim B-H, Cong Q, Grishin N V. HangOut: generating clean PSI-BLAST profiles for domains with long insertions. Bioinformatics 2010;26(12):1564–1565.

68. Jmol: an open-source Java viewer for chemical structures in 3D. .

69. Crooks GE, Hon G, Chandonia J-M, Brenner SE. WebLogo: a sequence logo generator. Genome Res 2004;14(6):1188–1190.

CHAPTER 8 ASSESSMENT OF CASP11 CONTACT-ASSISTED PREDICTIONS[7]

INTRODUCTION

The CASP11 contact-assisted structure modeling categories intend to learn how knowledge of long-range contacts improved the quality of tertiary structure prediction models provided by so-called hybrid prediction methods.[1-3] For a selection of more challenging tertiary structure prediction targets (T0), contact-assisted data were distributed to the CASP community subsequent to the release of the target structure and collection of the initial predictions, but prior to the public release of the experimental coordinates. Four types of contact-assisted data (abbreviated T*) were provided: predicted three-dimensional contacts gathered from the contact prediction category of CASP11 (Tp, subscript 'p' for predicted), selected subsets of correct contacts from the contact prediction category (Tc, 'c' for correct), simulated sparse NMR contacts (Ts, 's' for simulated), and contacts obtained from cross-linking mass spectroscopy studies (Tx, 'x' for crosslinked). These categories expanded on the promising results observed in the CASP10 contact-assisted assessment,[3] which evaluated only correct contacts (Tc).

---

[7] This Chapter was published as:
Kinch LN*, **Li W***, Monastyrskyy B, Kryshtafovych A, Grishin N V. *Assessment of CASP11 Contact-Assisted Predictions*. Proteins 2016.
* Authors contributed equally

An overview of the experimental setup for the CASP 11 contact assisted categories is illustrated in Figure 1. The Prediction Center chose sets of pairwise contacts for the predicted Tp and correct Tc contact-assisted categories from long-range contacts collected in CASP's Residue-Residue Contact Prediction (RR) category. The lists of submitted contacts in the RR category (both true and false positive) were filtered to retain only long-range contacts (separation along the sequence >23 residues), sorted according to the submitted probability, and truncated to the first L/5 contacts if necessary (L- target length in residues). For each predicted Tp target, the processed lists were released for ten CASP11 RR groups that were among the best performers in the previous CASP.4 For the correct Tc category contacts, the lists of predicted contacts in the RR category were pre-filtered for correctness by measuring the contact distances in the native structure. Correct contacts were defined as distance between C_β from each residue of the given pair being <8 Å. The correct Tc pairs were then subjected to the procedure used in the predicted Tp category, usually limiting to L/5 contacts, with the number being sometimes smaller (if not enough long-range contacts existed) or larger (to include all contacts with the same probability as that of the bottom, L/5-th contact).

The simulated NMR Ts and crosslinked Tx contact data were generated by the Montelione and Rappsilber labs, respectively. CASP organizers provided coordinates of crystal structures of the selected simulated NMR Ts targets to the Montelione group (Rutgers). These coordinates were used to mimic the data available in the initial stage of an NMR study. First, NOESY cross peaks were assigned to targets using a simulation procedure [G.Montelione, this issue], and then ambiguous distance restraints from these

peaks were generated using the Automated Structure Determination Platform ASDP.[5] CASP organizers arranged for shipment of biological material from CASP target providers to the Rappsilber lab (Technical University of Berlin). The target proteins were cross-linked and distance restraints were obtained using mass spectrometry [J.Rappsilber, this issue].

A total of 27 targets were selected by the Prediction Center for contact-assisted predictions in CASP11 (Table 1). The targets were divided into the following categories: 24 in the predicted Tp set, 19 in the simulated NMR Ts set, 24 in the correct Tc set, and 4 in the crosslinked Tx set. The targets were designated according to the category abbreviation (Tp, Ts, Tc, or Tx) followed by the three-digit T0 target number (that is, 761 from T0761-D0). One target (Tp826) was omitted from evaluation because the simulated NMR Ts contacts were released prior to the predicted Tp contacts.

The CASP11 contact-assisted targets included 17 that were evaluated in the tertiary structure prediction category as single domains, with 12 categorized as FM, two categorized as TBM, and three categorized as TBM-Hard.[6] The remaining ten targets are multidomain, with four exhibiting duplications of the same domain and one exhibiting a triplication. The multidomain targets were categorized as all TBM (1 target), all FM (3 targets), a combination of TBM and FM (4 targets), and a combination of TBM-Hard and FM (2 targets).

A number of groups participated in the contact-assisted categories in CASP11, including six servers and 23 human groups (Table 2). Only 10 groups contributed models for nearly all targets in all of the contact-assisted categories. Five additional groups

contributed models for nearly all targets in three of the four categories while one group contributed in two of the four categories. Three groups concentrated on the crosslinked Tx category with the smallest number of targets.

RESULTS AND DISCUSSION

Target-based performance improvements

We used performance improvement measures developed in the previous CASP evaluation3to assess the CASP community's ability to use contact information to improve tertiary structure predictions (T0). The first measure, individual performance improvement, represents the difference between the contact assisted (T*) scores and the score of the best unassisted T0 prediction from the same group. If the corresponding unassisted prediction T0 was missing, we used the average GDT_TS score from all unassisted predictions submitted on the target in place of the reference score. The second measure, absolute performance improvement, compares scores of assisted T* models and a gold standard unassisted T0 model (the best among all participating predictors in the specific contact-assisted category). However, the absolute performance unrealistically assumes that each group started with the same best unassisted T0 model. Despite the drawbacks of these measures, the difference distributions for best GDT_TS models on each assisted target (Fig.2) provide insight into the performance improvements of the CASP community as a whole using various types of contact information. Predicted Tp

and crosslinked Tx targets exhibited a relatively poor overall performance, with broadly negative absolute improvement values and relatively lower individual improvement values than those calculated for correct Tc and simulated NMR Ts targets, which tended to display positive improvements on most targets.

For the predicted Tp and crosslinked Tx categories, the absolute performance is overwhelmingly negative (Fig. 2, most red bars in the left panel representing predicted Tp scores and lower part of the right panel representing crosslinked Tx scores are below 0). The average absolute performance difference of best predicted Tp models over all targets was negative (−10.86 GDT_TS), with only 8% of the best models showing positive absolute performance improvement. The individual predicted Tp performance on average differed by −0.19 GDT_TS, and approximately half (51%) of the best predicted Tp models exhibited positive individual performance improvement (Fig. 2, blue bars above 0). Similarly, the best crosslinked Tx models had negative averages of −10. 2 GDT_TS (absolute) and −1.9 GDT_TS (individual), beating their unassisted models in 9% (absolute) and 36% (individual) of the cases. The discrepancy between some of the absolute and individual performance improvements suggested that positive individual performance scores might simply reflect poor initial models. Despite this potential caveat, community-wide T-tests as performed in the previous CASP contact-assisted evaluation3 (Table 3) showed marginal, yet significant improvements for 4 of the 23 predicted Tp targets: Tp767-D0, Tp804-D0, Tp806-D1, and Tp834-D0. At the same time, the predictions showed significant deteriorations with respect to their unassisted models on seven of the 23 predicted Tp targets. Only three predicted Tp targets (Tp763, Tp804

and Tp827) included promising absolute group performance (GDT_TS improvement > 10). Three of four crosslinked Tx targets showed average deterioration in model quality using assisted information, with one (Tx808-D0) being significantly worse.

In contrast to the poor performance on predicted Tp and crosslinked Tx targets, CASP11 predictors achieved good results modeling correct Tc and simulated NMR Ts targets (Fig. 2, center panel for correct Tc and upper left panel for Ts). The average GDT_TS improvement of the best correct Tc models was 12.1 GDT_TS for absolute performance, with the top score improvement approaching 69.2 GDT_TS for Tc763. For individual performance, the average of all best correct Tc models over all targets was 22.9 GDT_TS, with the top score improvement approaching 72.1 GDT_TS for target Tc763. All but one correct Tc target showed significant improvements using community-wide t-tests (Table 3). Similarly, the average performance improvements for best simulated NMR Ts models were both positive (1.5 GDT_TS for absolute and 11.7 for individual), with all but two of the targets (Ts794 and Ts835) showing significant improvements in average model quality by the community-wide t-tests (Table 3).

The correct Tc and simulated NMR Ts score distributions highlight another drawback of comparing assisted scores to initial unassisted T0 scores. Several of the targets exhibited negative absolute performance differences, yet the individual performance differences were generally positive (i.e. Tc/Ts806, Tc/Ts824, and Tc/Ts827). These discrepancies suggested that the gold standard best unassisted T0 models used for calculating absolute performance had unusually high scores. Indeed, one of the manual groups participating in the contact-assisted predictions (Baker, CASP group number 064

– see Table 2 for CASP11 group name-number correspondence) provided outstanding "unassisted" T0 predictions for two of these targets (T0806, Fig. 4, and T0824). We learned that the Baker group had successfully incorporated co-evolution based contact predictions into their T0 tertiary structure predictions.17 As such, the top T0 GDT_TS scores did not fairly reflect those of unassisted models, and this incorrect basis for comparison resulted in unusually low community-wide absolute performance scores (and penalized the individual performance scores for group 64 on these two targets).

Group-based performance improvements

We used the same absolute and individual performance improvement measures (with slight alterations) to understand how each group used contact information to improve unassisted T0 models. For the group-based performance improvement evaluation (Fig. 3 and Table 4), we considered all assisted models in calculating averages so that the most information possible was included for statistical evaluation, and we compared these models to either the top group unassisted T0 (individual) or the gold standard unassisted T0 (absolute). Most of the groups' individual and absolute average performance differences were negative for predicted Tp (blue) and crosslinked Tx (orange) targets [Fig. 3(A)]. In contrast, the average individual performance differences for both correct Tc and simulated NMR Ts were above 30 GDT_TS for six of the participating groups (three groups from Jooyoung Lee's lab: Lee, LeeR and NNS server; the Baker group; the Wiskers group; and the Laufer group), with similar trends in the

absolute performance [Fig. 3(B)]. While the Wiskers group showed one of the most promising GDT_TS difference score trends, they contributed models for only 2 of the 24 correct Tc targets and 2 of the 19 simulated NMR Ts targets (Table 4). In fact, six of the groups contributed models for <10 of the 70 total targets in all of the assisted categories (indicated by grey group labels in Fig. 3) and were ultimately excluded from rankings.

According to pairwise Student's *t*-tests evaluating the individual and absolute GDT_TS performance improvements for the groups participating in the CASP11 contact assisted categories, only three groups (NNS, Fusion, and Stap) showed significantly positive individual average performances on predicted Tp targets, whereas one additional group (Baker) showed a positive, but insignificant average performance (Table 4A). Fifteen of the twenty participating groups in the predicted Tp category significantly declined as measured by individual performance differences, and all were significantly worse using absolute performance differences. In the crosslinked Tx category, two groups (Meiler Lab and Stap) showed significant positive individual average performance, one group (Baker) showed positive, but insignificant performance, and the rest showed significantly negative individual average performance (Table 4D]).

In the correct Tc and the simulated NMR Ts categories, individual and average performance measures showed significant (by Student's *t*-test) improvement over initial models for five groups (Fig. 2: Lee, LeeR, NNS, Baker and Laufer). Four additional groups (Floudas, Anthropic Dreams, Multicom-cluster, and Foldit) significantly improved in both individual and average measures for the correct Tc category, and one group (Floudas) showed significant improvements in both measures for the simulated

NMR Ts category (Table 4 B,C). The top five performing groups had higher scores on the correct Tc targets than both their individual unassisted T0 scores (average increase of 43.0 GDT_TS) and the gold standard unassisted T0 scores (average increase of 35.4 GDT_TS). They also showed similar average improvements in the simulated NMR Ts category (35.3 GDT_TS for individual and 28.5 GDT_TS for absolute).

Two of the top-performing groups in the contact-assisted prediction (Baker and LeeR) also performed well in the FM tertiary structure prediction evaluation of unassisted T0 models.[18]Since most (21 out of 27) of the contact-assisted targets belong at least in part to the FM category (Table 1), above the average GDT_TS scores of these two groups on unassisted T0 targets could introduce negative bias in difference scores. Thus in theory, evaluation of groups that outperform on T0 targets by their individual GDT_TS difference tests might be unfair. Indeed, the average best T0 GDT_TS score (29.5) on all contact assisted targets for the Baker and LeeR groups was significantly different than the average best T0 GDT_TS score for the remaining groups (22.5) using a two-sample, one-tailed *t* test. Given these drawbacks to the performance improvement scores, we chose to rank groups using alternate scores (see Performance evaluation section below).

Examples of top assisted target predictions from top-performing groups

Target Tp806 exhibited the highest overall significant mean difference (4.3 GDT_TS) reflecting performance improvement for the predicted Tp category (Table 3). The FM-categorized T0806 target protein [Fig. 4(A)] adopts an α/β three-layered

sandwich architecture in the Evolutionary Classification Of protein Domains (ECOD) database[19] that is distantly related by structure (top LGA_S 25.0 to 2q07A) to folds in the X-group "other Rossmann-like structures with the crossover". The Rossmann-like domain in the target is interrupted by a unique 3-helix insertion that is not present in any structurally related templates. The relatively high GDT_TS score of 60.7 for this target's top T0 model (64_1, by the Baker group) reflected a correct overall topology for the prediction [Fig. 4(B)] that was significantly closer to the target than the top templates. Despite this impressive top T0 prediction, the mean GDT_TS was much lower (16.56) for T0 models from groups participating in the contact-assisted categories. The best model for this target in the predicted Tp category [also the Baker's group model 64_5, Fig. 4(C)] slightly improved the GDT_TS score (to 62.5). The next best group prediction [38_3 by the NNS server, Fig. 4(D)] retained the correct topology of the Rossmann fold, but incorrectly oriented the helical insertion with respect to the β-sheet.

Target Tc810-D1 exhibited the highest overall significant mean difference (30.4 GDT_TS) reflecting performance improvement for the correct Tc category, and Ts810-D1 exhibited the third highest mean difference (22.3 GDT_TS) for the simulated NMR Ts category (Table 3). The ECOD database[19] classifies the FM-categorized target T0810-D1 as an α-superhelices architecture with a somewhat irregular ARM-repeat fold [Fig. 4(E)]. This target domain is fused to a C-terminal domain exhibiting an α/β-barrel architecture fold that is homologous to a TIM barrel in ECOD. This C-terminal domain was categorized as TBM and was excluded from the contact-assisted predictions. The top unassisted prediction model among contact-assisted predictors for this domain (TS162_3,

from McGuffin group) displayed a roughly similar topology (GDT_TS 40.5), except the N-terminal helices did not pack against the subdomain formed by the C-terminal helices [Fig. 4(F)]. The two top Tc prediction models (44_1 and 169_1 from J. Lee's lab) were identical and improved over the top T0 model by 45.8 GDT_TS [Fig. 4(G)], while the top simulated NMR Ts prediction model by another group [Laufer, 428_4, Fig. 4(H)] improved over the top T0 model by 38.7 GDT_TS. The top correct Tc and simulated NMR Ts prediction models for T0810-D1 adopted the correct overall topology of the ARM-repeat fold, with the main differences stemming from an extended C-terminal linker sequence with no secondary structure.

The single-domain target T0812-D1 [Fig. 4(I)] was categorized as TBM-hard, and displayed a β-sandwiches ECOD architecture that is homologous to Concanavalin A-like folds. The top T0 prediction model [64_3 from the Baker group, Fig. 4(J)] retained the same overall fold as the target domain, except for the N-terminal residues (5–56) corresponding to the first three β-strands. The overall mean difference for the target T0812-D1 was negative (−2.1 GDT_TS), yet the top performing crosslinked Tx model improved over the T0 model by 3.2 GDT_TS [64_3 from Baker, Fig. 4(K)]. The next best group prediction model [42_1 from the Tasser group, Fig. 4 (L)] decreased by 4 GDT_TS, as compared to the T0 model. While the top performing crosslinked Tx model only improved by 3.2 GDT_TS, it correctly placed the three N-terminal β-strands and attained the entire fold topology. The next best group model also predicted the correct overall fold topology, but the model exhibited gaps and incorrectly structured β-strands.

Performance evaluation without unassisted models: Combining scores for ranks

Due to the potential biases of using unassisted models for the contact-assisted evaluation, we chose to assess group performance using similar score combinations as were used in the FM (see Kinch, this issue) and TBM (see Roland, this issue) evaluations. We generated *Z*-score sums and averages over all contact-assisted (T*=Tp, Tc, Ts, or Tx) targets for the combined scores on each group's best or first submitted models. We evaluated all categories using the FM-style combined scores (GDT_TS, ContS, QCS, TenS, lDDT, and MolProb). However, the relative high performance of groups in the correct Tc and simulated NMR Ts categories prompted additional evaluation using TBM-style score combinations to better distinguish models that are closer to their targets (GDT_TS > 50).

Group performance was ordered by best FM-style *Z*-score sum (Table 5, includes also FM-style average, first models and win/loss counts). All groups that could not be distinguished from the top ranked group according to *t* test and bootstrap significance (for FM-style *Z*-score sum) are bolded. The top-performing groups in the contact-assisted categories according to the FM-style and win/loss scoring schemes (Lee, LeeR, NNS, and Baker) were similar to those that outperformed in performance improvement scores (Fig. 3). As three of these groups correspond to a single CASP11 participant (Jooyoung Lee - groups 38, 44, and 169), we investigated whether having multiple groups (i.e. submitting as multiple groups) tended to alter the *Z*-score ranks or significance scores of the participant when compared to having a single group (i.e. submitting as a single group).

To check for this case, we omitted two of the three J. Lee's groups in turn, and recalculated all the relative scores for all the participating groups in these three scenarios. With the exception of the crosslinked Tx category, which had too few targets, the ranks and significance estimates of any single group from the same CASP11 participant did not change, although the absolute values of the Z-scores did (See prodata.swmed.edu/casp11/contact for tables).

When compared to group performance ranks determined by the GDT_TS Z-score sums, the FM-style *Z*-score sums produced the same ranks for the four top-performing groups in the predicted Tp category (Lee, NNS, McGuffin, and Fusion, in ranked order). However, tests of statistical significance in the predicted Tp category suggested that one of the groups (Baker) that predicted significantly fewer targets (10 out of 23) tied with the two top-performing groups (Lee and NNS). In win/loss counts, the same four groups rank at the top, with the Baker group holding 3rd place.

For the correct Tc category, all scoring methods (GDT_TS, FM-style, and win/loss counts) rank groups LeeR and Lee as first and second, correspondingly. Because the top prediction models in this category were similar to the target (GDT_TS score >50), we also examined TBM-style scoring and significance estimates that were designed to evaluate such similarities. TBM-style scoring ranked the same two groups at the top. These two groups tied in many of the head-to-head trials (10 out of 24 targets), and the performance of the two groups could not be distinguished by significance estimates of TBM-style scoring. The third-place group (Baker) tied with the top-performing group

according to significance of FM-style scores, but not TBM-style scores or GDT_TS only scores.

For the simulated NMR Ts category, the same group (Baker) placed as first for all three *Z*-score-style scoring methods (GDT_TS, FM, and TBM). Two additional groups (LeeR and Lee) tied for top-performance by all statistical measures. The fourth ranked group, NNS server (as well as the Laufer group that predicted less targets), tied with the top groups only using significance from T-tests on TBM-style scoring. Interestingly, win/loss counts with GDT_TS, FM-style, and TBM-style scoring placed the Lee and LeeR groups above the top-ranked Baker group. The cause of this apparent discrepancy in rankings is discussed in the following section (Head-to-Head Comparisons).

For the crosslinked Tx category, the top-performing Baker group was ranked first by GDT_TS and FM-style scoring methods, as well as in win/loss counts. The top group tied with Lee and NNS groups using *t* test significance estimates, while it significantly outperformed by FM-style bootstraps. The differences in significance likely originated from the low number of targets in this category (4 targets).

Head-to-head comparisons of top-performing groups

To help clarify the performance of the top ranked groups in each category that tied by any of the significance estimates, we plotted their head-to-head GDT_TS scores (Fig. 5). For these head-to-head comparisons, we chose the top performing Lee lab group (among Lee, LeeR and NNS) according to FM-style *Z*-score ranks for each assisted

category. For illustrative purposes, we combined the head-to-head results from the Baker and Lee groups for the predicted Tp and crosslinked Tx categories into a single graph [Fig. 5(A)]. The predicted Tp targets were limited to only 10 of the 23, since the Baker group did not predict the remaining targets. Most of the predicted Tp targets clustered near the identity line below 40 GDT_TS. However, the Baker group submitted three predicted Tp prediction models above GDT_TS 40 that outperformed (Tp806, Tp818, and Tp827), while the Lee group submitted one (Tp825) that outperformed. This relative outperformance of the Baker group on the reduced target subset likely explains their elevated performance according to significance estimates and their win/loss rank just under the top-performing Lee and NNS server groups (Table 5). Similarly, three out of the four targets in the crosslinked Tx category clustered near the identity line below 25 GDT_TS. The Baker group outperformed on a single crosslinked Tx target (Tx812), while the Lee group outperformed marginally on two of the crosslinked Tx targets. Thus, the outperformance of group Baker on a single target Tx812 established their position at the top of all ranking methods for the crosslinked Tx category (Table 5).

The correct Tc category head-to-head plot highlights a cluster of 23 targets above 48 GDT_TS, with the LeeR group outperforming on most (16 targets). The Baker group appeared to excel at the assisted prediction of target Tc812, while the LeeR group excelled at target Tc794, among a few others. This relative outperformance by the LeeR group on most of the targets resulted in their top ranking by all methods. Their top ranking was also justified by significance tests using the TBM-style scoring scheme, which was chosen by the TBM assessor as distinguishing models that were generally

closer to the template (above 50 GDT_TS). The bootstrap and *t* test significance estimates using TBM-style scoring suggested the performance of the LeeR group was not distinguishable from the alternate prediction group from the same participants (Lee), yet it was distinguishable from the Baker group (confidence level 0.916).

The simulated NMR Ts category plot comparing LeeR with Baker highlights three outlier targets where Baker outperformed LeeR (Ts761, Ts777, and Ts827), and two targets (Ts794 and Ts826) where LeeR outperformed Baker. Performance scores on the remaining targets clustered closely to the equivalence line, with more favoring the LeeR group, which wins on 10 of 14 remaining targets. Comparison of the Baker group with the Lee group (ranked 2 by GDT_TS *Z*-score sums) yielded similar results (not shown). *Z*-score sums tended to emphasize the magnitude of improvements while win/loss counts evaluated the generalization of the methodology on various targets. Therefore, the apparent discrepancy in rankings by the two methods was caused by the Baker group providing more significantly better outlier targets (top Z-score ranking), whereas the LeeR group provided more subtly better winning targets (12 out of 19 targets). Statistical tests, including bootstrap and *t* test, suggested that the differences between these two groups were statistically insignificant.

In our above analyses, we treated multidomain assisted targets as single evaluation units. Besides this treatment, we also calculated scores, rankings, and significance estimates for first model predictions and domain-based predictions (i.e, predictions on multidomain targets were split and evaluated separately). Group performance using first models resembled that of best models with a few exceptions,

including (1) LeeR significantly outperformed the other groups on correct Tc targets, and (2) nns tied with the top groups on simulated NMR Ts targets using FM-style scores. The top performing groups performed similarly using best models on a per-domain basis, with a few exceptions. The Baker group tied with the Lee and LeeR groups in the correct Tc category by all significance tests and the NNS server no longer tied with the top performing groups (Baker and Lee) in the simulated NMR Ts category using TBM-style scoring. For first models, Baker TS064 tied with the LeeR group on correct Tc targets by TBM-style scores and Laufer, who predicted less than half (11) targets, tied with the four top groups on simulated NMR Ts targets by TBM- and GDT_TS- style scores. All the evaluation tables are accessible via http://prodata.swmed.edu/casp11/contact.

Performance comparisons to previous contact-assisted predictions

The contact-assisted component of CASP11 included several new categories (predicted Tp, simulated NMR Ts, and crosslinked Tx) that had no basis for comparison to the previous assessment. The input data in the only comparable category (designated correct Tc in both CASP10 and CASP11) had some significant differences in both the number and type of provided contacts. The number of provided contacts for CASP10 were restricted to roughly one tenth of the number of residues, and the contacts were only selected if they were present in <15% of the unassisted predictions in CASP10.[3] In contrast, in CASP11 the Prediction Center provided a significantly larger number (~10

fold) of correct Tc contacts that were selected among top contact predictors regardless of the contact coverage in the submitted 3D models.

The previous CASP10 contact-assisted correct Tc category showed significant improvements in mean correct Tc GDT_TS scores when compared to mean T0 scores for each target, with the best absolute improvement approaching 40 GDT_TS. The best absolute improvement for CASP11 correct Tc targets was even higher (70 GDT_TS). Even though it is hard to bring the different types of contacts in two different CASPs to the same frame of reference, the data allowed us to notice similar trends in both CASPs, namely improved average performance with increased number of contacts per residue. A scatter plot of CASP11 target-based best absolute GDT_TS improvement against number of unique provided contacts per target residue (ranged from 0.432 to 1.11) highlighted an overall trend of improving performance with enriching contact information [Fig. 6(A)]. Although the data showed a relatively low goodness of fit ($R^2 = 0.09$), extension of the linear fit line ($Y = 28.20*X + 22.18$) to the number of contacts released in CASP10 (25.6 GDT_TS difference at 0.12 contacts per residue) suggests a similar trend in CASP10 and CASP11. This extrapolation implied that the apparent CASP11 performance "improvement" stemmed from an increase in the number of given contacts.

Two of the correct Tc targets with high outlier T0 predictions (T0806 and T0824, discussed in Target-based performance improvement section above) should have displayed lower than expected best absolute improvements, skewing the trends highlighted in Figure 6(A). Indeed, omitting these two targets from linear fit calculations slightly improved the goodness of fit ($R^2 = 0.11$) and resulted in a somewhat larger slope

of the line: $Y = 30.09*X + 22.31$, which corresponds to a similar number extended to CASP10 levels (25.9 GDT_TS difference at 0.12 contacts per residue).

Given the relatively high number of correct Tc targets, we examined the performance of predictions on different fold types. We considered the ECOD architecture for each correct Tc target, combining the target architectures into broad categories including α/β, $\alpha+\beta$, all-α, all-β, and mixed resulting from the presence of multiple domains. We then plotted the best absolute performance of targets clustered into each category [Fig. 6(B)]. Because the targets displayed a trend in performance based on given contacts per residue, we normalized the best absolute performance by averaging it with an estimate of the best absolute performance (Y) based on the given contacts per residue (X) according to the Figure 6(A) linear fit formula. The results suggest that the provided contacts helped modestly for all-α targets (average normalized performance improvement 35.5 GDT_TS). Only a single target (T0806) populated the α/β category. This target represented an outlier and exhibited a lower than expected absolute difference (33.3 GDT_TS) because of unusually high T0 model quality discussed previously. Indeed, when we used the next-best group T0 target to calculate normalized best absolute performance on the singleton α/β target, the recalculated value (49.9 GDT_TS) exceeded the normalized average best absolute performance value [Fig. 6(B), dotted line, 43.3 GDT_TS]. One possible explanation for the relative contact-assisted outperformance on β-strand-containing targets might involve their more regular interaction in β-sheets dictated by non-local backbone hydrogen bonds. Thus, a single contact provides the correct register for the β-strand with its neighboring β-strands. Alternatively, interactions

between α-helices can occur at different angles, requiring more than one contact pair to define their placement.

CONCLUSIONS: PERFORMANCE INSIGHTS AND SUGGESTIONS

Two research labs significantly outperformed the rest using all types of contact-assisted information to enhance prediction model quality: the Lee lab represented by a server NNS, and two manual groups LeeR, and Lee; and the Baker lab with the same-named prediction group. Using contact-assisted information from two different categories, correct Tc and simulated NMR Ts, these top-performing groups provided significantly improved structure predictions. On the other hand, information provided in the predicted Tp and crosslinked Tx categories yielded marginal improvements, despite the success of the Baker group in utilizing contact predictions to significantly improve structure models for several targets (i.e. T0806 and T0824) in the template free modeling category of CASP11 (Baker, personal communication). Unfortunately, the Baker group did not participate in the RR category, from which the assisted Tp category contact data was selected. Thus, the benefit of depth of alignment and improved co-variation methods that led to Baker's success in residue-residue contact and tertiary structure prediction[17, 18] could not be evaluated for other groups participating in the predicted Tp category. Moreover, we could not clearly separate the contributions of provided contacts from those embedded in the Baker prediction methodology to their success in the contact-assisted categories. The observation that the Baker group best contact-assisted Tp

model (GDT_TS 62.50) was only marginally better than their best unassisted T0 model (GDT_TS 60.65) suggests that the contribution of predicted Tp data from other groups was limited.

Perhaps the most encouraging prediction models came from the simulated NMR Ts category, which aimed to mimic contact information provided by experimental NMR data. The quality of models produced using this information, which albeit only represents a model of real NMR data, approached that of the artificial correct Tc category.

Given the relative outperformance of the Baker and J. Lee's groups on the contact-assisted categories, we decided to use their average GDT-TS scores for all targets in a given category to represent top performance. We then examined why the predicted Tp and crosslinked Tx categories were much more difficult than the correct Tc and simulated NMR Ts categories [Fig. 7(A)]. First, we considered a term that evaluated the quality of provided contacts for each assisted category: the correct contact percentage (CCP). As expected, outperformance in the correct Tc category arose from the high percentage of correct contacts given (100% by definition), with the other three categories having <15% of the provided contacts being correct. Interestingly, the CCP average for the simulated NMR Ts category was almost the same as for the predicted Tp category, for which performance was significantly lower. Thus, CCP alone could not account for performance. The given simulated NMR Ts data included far more contacts than in any of the other categories (see paragraph below), so we also calculated the correct contact coverage (CCC) of the target to see if this property could compensate for a lack of correct provided contacts. Indeed, the simulated NMR Ts category displayed a higher CCC

average (2.5-fold coverage) than the other three categories (Tp 0.19-fold, Tc 0.7-fold, and Tx 0.16-fold coverage). Thus, the outperformance on the correct Tc targets stemmed from the high percentage of correct contacts, whereas the outperformance in the simulated NMR Ts category stemmed from a reduced percentage of correct contacts that was offset by a much higher coverage of correct contacts. A number of possible explanations for the relatively poor performance in the crosslinked Tx category exist. From our evaluation of contact quality [Fig. 7(A)], the contacts provided by the crosslinked Tx data were only 10.8% correct on average when defined by the 8 Å distance cutoff in the experimental structures. Such poorly defined contacts likely result from the cross-linking agent being too long to represent interacting residues. Additionally, the nature of the crosslinking agents could result in an uneven distribution on the structures. This notion might lead to the relatively low average coverage of the correct contacts noted for the category (Fig. 7, crosslinked Tx CCC is 0.16). Thus, the crosslinked Tx category experiment provided a fundamentally different type of contact information, as residues must be accessible to the crosslinking reagent (i.e. relatively exposed) and might be more distant (> 8 Å) than the traditional concept of contacting residues. Perhaps including such restrictions in methodology for using crosslinked Tx contacts would improve the quality of structure models.

To gain further insights into the quality of Ts predictions, we compared Ts models generated by predictors to ‘dummy models’ generated by us using standard NMR structure determination software. To generate dummy models, we used one of most cited NMR packages,16 the NMR routines in the Crystallography and NMR System (CNS).

The CNS package utilizes the distance restraints in simulated annealing protocol to produce a model most compatible with these restraints. The average number of contacts per target given to predictors in Ts category was 14724 hydrogen pairs, corresponding to 9283 residue pairs [Fig. 7(B), dark and light cyan bars]. This number far exceeds that given in other contact-assisted categories. For instance, the largest number of contacts per target from any of the other three categories is only 673 residue pairs (Tp814). However, the overwhelming majority (about 98.5%) of these contacts is "ambiguous", and the NMP peak is usually assigned to multiple atom pairs. When all given Ts contacts (ambiguous and unambiguous) are used as input, CNS package generated dummy models with approximately random GDT_TS scores for each Ts target [average GDT_TS = 13.56, Fig. 7(C) cyan line], close to some of the worst predictions. Apparently, the ambiguity of the contacts hindered the reconstruction of the structures by CNS, and most predictors found a more clever way to deal with ambiguities.

We next attempted to reduce the ambiguity provided to the CNS software. As the first step, we used only unambiguous contacts, that is, those for which distance constraint corresponded to a single given pair of atoms. While this method of contact selection does not require the knowledge of the target structure and could have been used by predictors, it comes at the cost of losing most of Ts contact information, because unambiguous assignments corresponded to an average of 1.2% (by atom)/1.7% (by residue) of the total Ts contacts [Fig. 7(B), dark and pale purple bars]. With unambiguous contacts being the only input, the CNS package generated dummy models with 29.3 GDT_TS score on average [Fig. 7(C), purple line]. Dummy models from five targets predicted the correct

fold and achieved GDT_TS above 40 (maximal GDT_TS = 53.8 for target Ts812). Therefore, although the number of unambiguous contacts was limited, those contacts were mostly correct (98.6% of unambiguous atom pairs are correct) and could be used to generate reasonable seed structures for further refinement. Interestingly, many of the CASP simulated NMR Ts predictions [Fig. 7(C), blue dots] had GDT_TS scores lower than the dummy structures generated from unambiguous contacts by CNS, suggesting that these groups could have benefitted from including standard NMR structure determination software in their methodologies.

Because assessors are granted access to the target structures, we further attempted to disambiguate ambiguous contacts using the knowledge of the target structure. We selected all the correct constraints in the provided simulated NMR Ts contacts to evaluate the theoretical upper limit of the CNS performance. For the purpose of cross-category comparison in previous section calculating CCP and CCC [Fig. 7(A)], the correct contacts were defined as those with Cβ distance no >8 Å. Here, we extracted the cutoff for the 'Ts-specific' true contacts from the upper limit (UPL) of the atomic distance for individual atom pairs provided by the simulated NMR data, resulting in an average of 1041 correct atom pairs in 625 correct residue pairs [Fig. 7(B), dark and pale green bars]. This definition was slightly higher than the number of correct contacts computed in the cross-category comparison [586 residue pairs, Fig. 7(B), medium green bar]. The dummy models generated by CNS using those 'Ts-specific' true contacts produce GDT_TS scores ranging from 43 to 75, with an overall average of 58 [Fig. 7(C), green line]. Impressively, many predictions achieved better performance than the structures built

from the true distance constraints selected with the knowledge of the target structure. The best predictions for every target outperform the dummy models obtained by CNS using true contacts. Although the lack of chemical shifts in Ts contacts provided to predictors limits the utilization of the NMR package to its full potential, the structure prediction methods seemed to utilize additional information to push the limit of the NMR methods based purely on the distance constraints. These best prediction methods should be useful for NMR researchers in protein structure determination and may have some advantages over the CNS package.

CASP11 exhibited a number of significant differences in the implementation of the contact-assisted category experiment when compared to the previous CASP10. These differences made evaluation of performance improvement difficult. Performance of the correct Tc categories from both CASPs was roughly dependent on the number contacts given per residue [Fig. 6(A)]. Given the artificial nature of the correct Tc category, perhaps future contact-assisted experiments could explore the correlation between given contacts per residue and top structure prediction performance by incrementally providing sets of correct Tc contact pairs over time. At the very least, this category should include more consistently defined contact pairs between CASP experiments to allow methods performance comparisons over time.

MATERIALS AND METHODS

Improvements over unassisted T0 models

We evaluated the community-wide improvement in performance quality by comparing the contact-assisted models (all T*: Tp, Ts, Tc, and Tx) to the unassisted (T0) models, using the GDT_TS score[7] that has been used in CASP assessments for over a decade.[8-13] We considered the differences in both individual performance and absolute performance on a target-wide basis similar to the evaluation of the CASP10 contact-assisted category.[3] For comparing overall performance improvements on each of the assisted targets, the best unassisted T0 GDT_TS from the group (individual performance) or the best overall unassisted GDT_TS among all groups (absolute performance) was subtracted from the group's T*model GDT_TS. To include individual performance scores for those groups that did not provide T0 models, the average T0 GDT_TS for all groups participating in the contact-assisted category for that target substituted for the missing T0s. To be consistent with the previous CASP10 assisted evaluation, we estimated the significance of community-wide performance improvement for each target using one-tailed *t*-tests that compared all assisted T* model GDT_TS scores to all T0 model GDT_TS scores (not only best T0's). We used one-tailed paired *t*-tests to evaluate the significance of each group's performance improvements (absolute and individual) over their unassisted T0 targets. The *t*-tests compared all of the group's assisted T* model GDT_TS scores to either the group's best T0 model scores (substituting missing T0 scores with the average GDT_TS for the corresponding target) or the overall maximum T0 model GDT_TS scores among all participating groups, respectively.

Group performance using combined scores, win/loss counting, and head-to-head trials

We calculated *Z*-score sums (and averages) over all the targets in each category for several different scores. *Z*-scores were calculated as in previous CASPs[10, 11] using first and best GDT_TS scores, as well as the combined score used to evaluate CASP 11 tertiary structure predictions (Kinch *et al,*. Evaluation of CASP11 free modeling targets and CASP ROLL in this issue). Briefly, we calculated *Z*-scores over each target for first and best GDT_TS, FM-style combined score (GDT_TS, TenS, QCS, ContS, lDDT, and Molprb), and TBM-style combined score (GDT_HA, GDC_ALL, lDDT, SG, and 0.2 x Molprb); and summed (or averaged) the *Z*-scores for all targets in each contact-assisted category.

The statistical significance of whether each group's performance differed from that of the other groups was inferred from one-tailed paired t-tests and bootstrap tests[10, 14, 15] on GDT_TS, FM-style, and TBM-style scoring schemes. We also carried out a pairwise comparison (head-to-head trials) of the group results, as well as the CASP10-style overall win/loss counts for all-against-all pairwise comparisons.[3] In head-to-head trails, for each pair of groups, we calculated the fraction of common targets/domains for which one group outperformed the other according to the selected score. In win/loss counts, we performed all-against-all pairwise prediction model comparisons on the selected scores for each target and summed the numbers of win/loss

cases for each group. The groups were ranked primarily by the probability that a win/loss record was equal to or better than the observed record that could have been obtained by chance, and secondarily by the fraction of winning comparisons. In GDT_TS comparisons for both head-to-head trials and win/lose count, we extended our comparison to consider models within both 1 and 2 GDT_TS score units as ties to address models with insignificant differences. Due to the registration of multiple groups by a single participant, we studied whether registering multiple groups (as opposed to having a single group) would provide an advantage or disadvantage to the participant's *Z*-score and ranking. To address this question, we compared original *Z*-scores, *t* test probabilities, and ranks to those calculated using only one of the multiple groups from the same participant.

Calculating correct contact percentage and correct contact coverage for contact assisted targets

The correct Tc and predicted Tp categories included some duplicated residue pairs that stemmed from overlapping predicted contacts provided by multiple prediction groups. Simulated NMR Ts target contacts included hydrogen atom pairs (as opposed to residue pairs), with some having multiple peak assignments as well as multiple atom counts for some residue pairs. Additionally, contacts in the simulated NMR Ts category and for the cross-linking target Tx781 included residue pairs limited to the same residue (noted as self-contacts). We filtered out duplications and self-contacts, using the numbers

for unique and non-self contact pairs. The correct contact percentage (CCP) was calculated as the number of correct residue pairs divided by the number of total residue pairs (times 100 to convert to percentage), with correct contact pairs defined as having C_β atoms in the target structure no >8 Å apart. We also computed the correct contact coverage (CCC) as the correct residue pair count divided by the target length.

Production of dummy structure models using simulated NMR Ts contact restraints

The simulated NMR Ts contacts represent hydrogen pairs from simulated NMR peak assignments, with an indicated distance upper limit (UPL) and its corresponding peak. Due to the ambiguity of the NMR assignments, peaks could be assigned to multiple hydrogen pairs. We produced dummy structure models with the CNS package using different distance restraint sets from the simulated NMR Ts contacts: (1) all contacts, (2) unambiguous contacts, and (3) true contacts. Unambiguous contacts were generated by taking those peaks with only one contact pair. As the UPLs for hydrogen pairs vary, we defined 'Ts-specific' contacts as those with distances lower than the corresponding UPLs. Note that the 'Ts-specific' contact threshold differs from the contact threshold used in comparison across categories (C_β atoms within 8 Å).

The simulated annealing protocol of the CNS package[16] was used to calculate structures based on provided distance restraints. As these restraints were limited to hydrogen atoms, we assigned the lower limit for distance constraints as 1.5 Å and the upper limit as the UPL given in the contact information. Simulations were performed

from both an extended chain (‘anneal.inp’ template option) and an embedded substructure starting model generated for H_N, N, CO, Cα, Cβ, and Cγ atoms by distance geometry calculations based on the Nuclear Overhauser Effect (NOE) restraints (‘dg_sa.inp’ template option) and ‘sum’ mode for NOE averaging. Simulations were complete after generating 10 accepted structures or reaching a 48-hour time limit. All simulations using unambiguous contacts, and 7 out of 19 simulations using correct contacts produced 10 accepted structures before reaching the time limit. The simulations generated from 634 to 10815 NMR structure solutions for each target, due to variations in protein length and provided contact numbers. We reported the best GDT_TS score among all the trial structures for each simulated NMR Ts target.

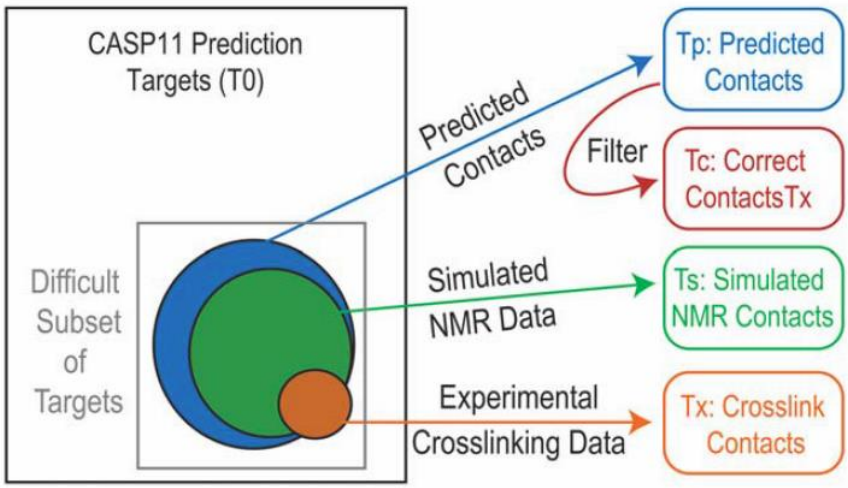


Figure 1

Contact Assisted Category Dataset Schematic. The leftmost outer box represents all CASP11 T0 targets, with a relatively smaller subset of the difficult targets selected for various contact-assisted categories (circles). The data for predicted Tp contact targets (blue) were selected from the predicted contacts provided for the CASP11 RR category and were filtered for close contacts for data provided in the correct Tc category (red). The data for the simulated NMR Ts targets (green) were provided by the Montelione lab, and the data for the crosslinked Tx targets (orange) were generated experimentally by the Rappsilber lab.

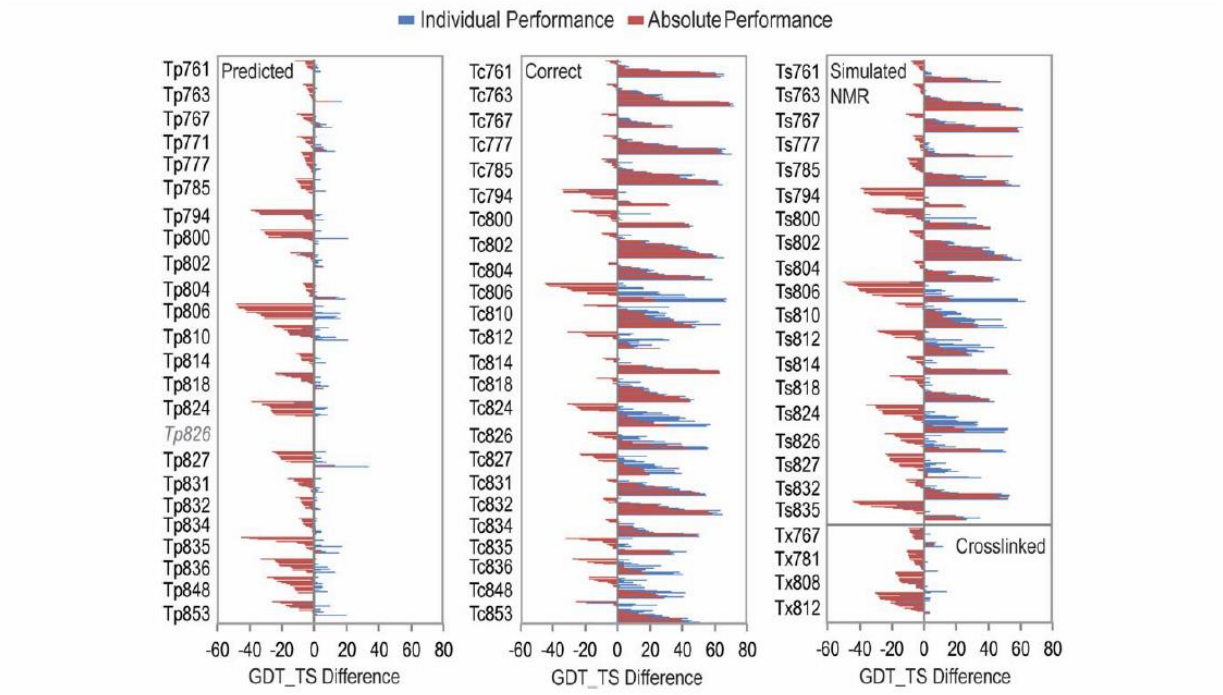


Figure 2

Absolute and individual performance for assisted targets. Each group has one bar that corresponds to absolute performance (red), measured by subtracting the gold standard T0 model's GDT_TS from the best T* model's GDT_TS for the group; and one bar that corresponds to individual performance (blue), measured by subtracting the best T0 model GDT_TS for each group from their best T* model GDT_TS. The value of the GDT_TS performance difference is indicated below the bar graph, with a grey line drawn through 0. Overall performance on predicted Tp targets (left panel) and crosslinked Tx targets (lower right panel) was worse than on correct Tc targets (middle panel) and simulated NMR Ts targets (upper right panel).

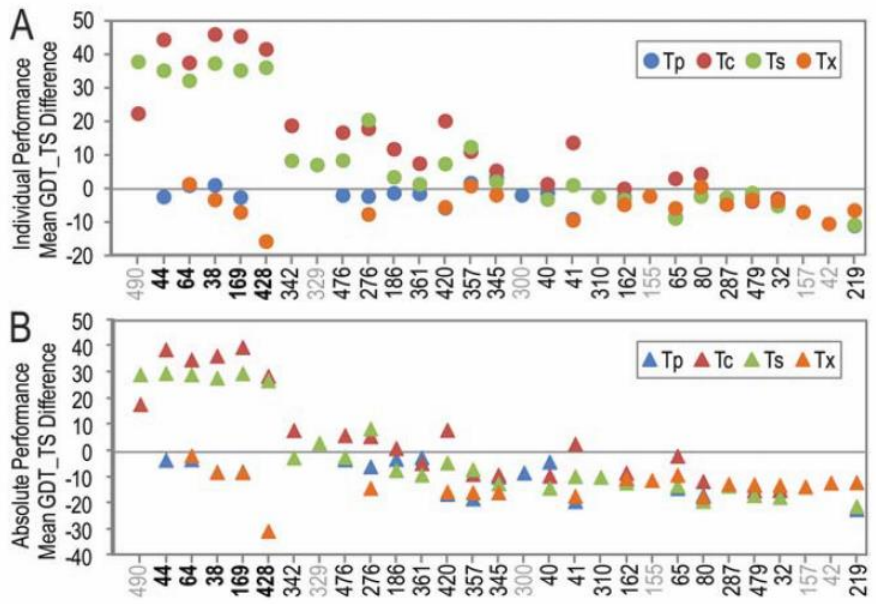


Figure 3

Groupwise performance improvements on assisted targets. The individual performance (**A**) and absolute performance (**B**) averages of each indicated group (X coordinate) are plotted for predicted Tp (blue), correct Tc (red), simulated NMR Ts (green), and crosslinked Tx (purple) targets. Groups are ordered from left (highest) to right (lowest) based on the sum of averages over all categories. Groups with <5 total predictions (out of 70 possible) are in labeled in grey.

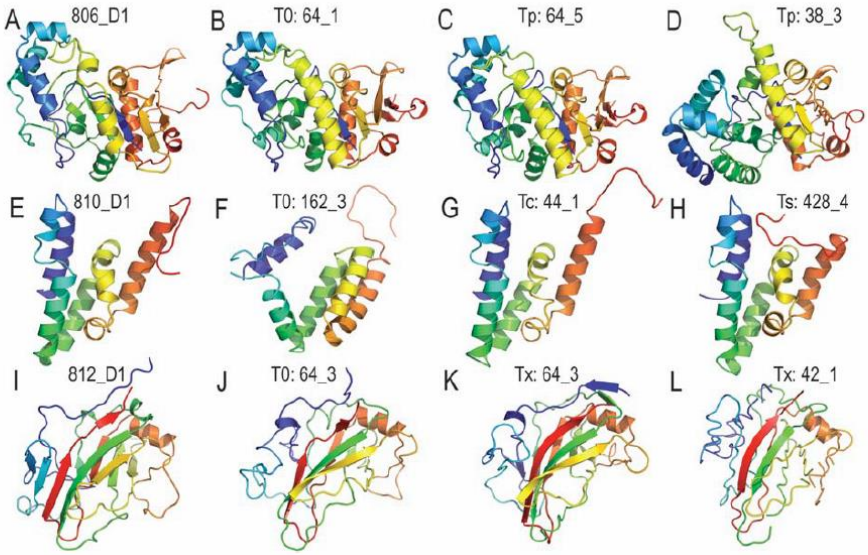


Figure 4

Examples of improved assisted prediction models. Targets and models are illustrated in cartoon and colored in rainbow from blue (N-terminus) to red (C-terminus). Target 806 D1 (**A**) is compared to the top unassisted T0 model 64_1 (**B**) having a GDT_TS of 60.7. The top predicted Tp model 64_5 (**C**) improves the GDT_TS slightly to 62.5, while the next best predicted Tp group model 38_3 (**D**) decreases the GDT_TS to 29.5. Target 810 D1 (**E**) is compared with the top unassisted T0 model 162_3 (**F**) having a GDT_TS of 40.5. Two identical top correct Tc models 44_1and 169_1 (**G**) improve the GDT_TS significantly to 86.3, while one top simulated NMR Ts model 428_4 (**H**) improves the GDT_TS significantly to 79.2. Target 812 D1 (**I**) is compared to the top unassisted T0 model 64_3 (**J**) having a GDT_TS of 44.2. The top crosslinked Tx model 64_3 (**K**) improves the GDT_TS slightly to 47.4, while the next best group model 42_1 (**L**) decreases the GDT_TS to 40.2.

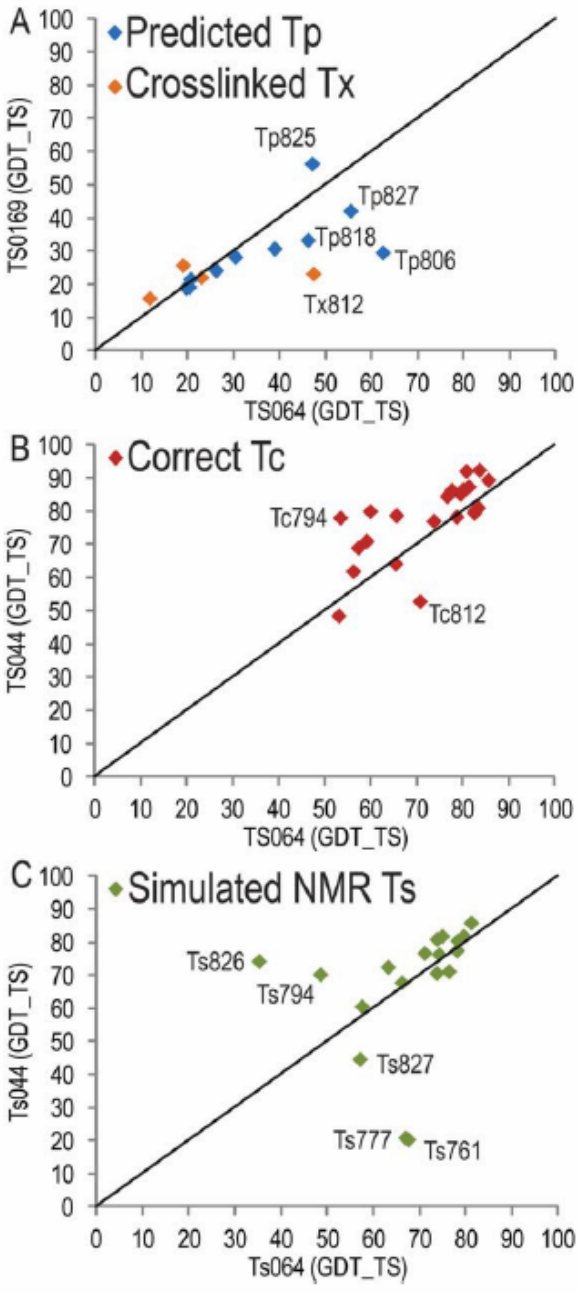


Figure 5

Head-to-head plots for top-performing groups. Top-performing groups according to significance tests were chosen for comparison. FM-style *Z*-scores were used to select the top group number among multiple submissions from the same prediction team. GDT_TS scores were plotted for (**A**) Baker Group 64 against Lee group 169 for the predicted Tp (blue) and the crosslinked Tx categories (orange), (**B**) Baker group 64 against LeeR group 44 for the correct Tc category, and (**C**) Baker group 64 against LeeR group 44 for the simulated NMR Ts category.

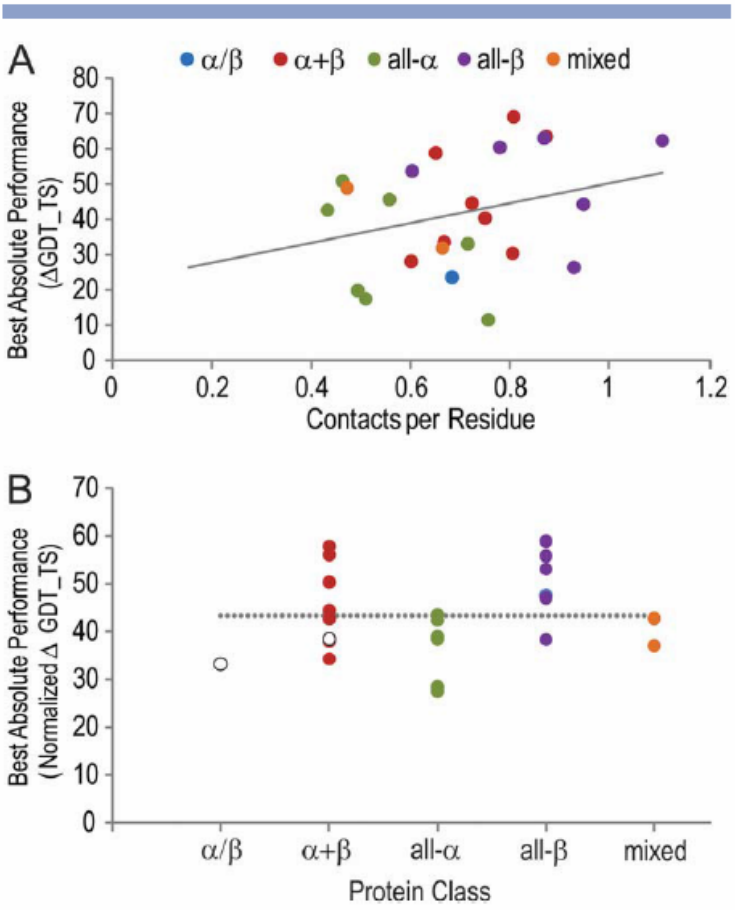


Figure 6

Absolute correct Tc performance improves with increasing provided contacts. The best absolute performance (top correct Tc GDT_TS – T0max GDT_TS, *y* axis) is plotted against the number of contacts provided per residue in the target (x axis) in panel (**A**) and colored according to protein class: α/β (blue), α + β (red), all-α (green), all-β (purple), and mixed (orange). A linear fit to the data has a relatively low goodness of fit $R^2 = 0.09$. In panel (**B**) the best absolute performance is normalized by averaging the absolute performance with the expected absolute performance according to the contacts per residue given the linear fit in A. The normalized performance is separated in panel according to protein classes as in panel A. White markers represent data for targets T0806 and T0824 with expected bias in T0. A dashed line indicates the average best absolute performance on all targets.

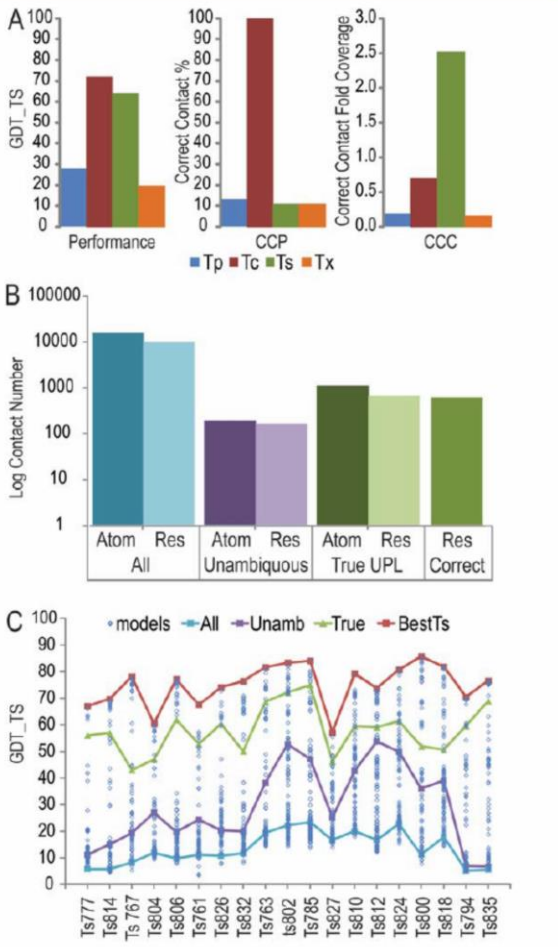


Figure 7

Dissecting prediction quality. A: The performance of the top groups (averaged GDT_TS, left panel) is dictated by two components of the provided contact information: the percentage of correct contacts (those within 8 Å in the target structure) over all given contacts (CCP, middle panel) and the fold coverage of correct contacts over the target structure (CCC, right panel). The bars represent the averages over targets from each contact assisted category: predicted Tp (blue), correct Tc (red), simulated NMR Ts (green), and crosslinked Tx (orange). B: The contacts provided for the simulated NMR Ts category can be subdivided into several classes of given information: all provided contacts that include both single peaks and multiple peaks for certain atom pairs (cyan), unambiguous contacts that correspond to given atom pairs with a single peaks (purple), 'Ts-specific' true contacts defined as pairs with atomic distance in the target structure within the given upper distance limit (UPL) (green), and correct contacts defined as pairs within 8 Å in the target structure (medium green). Contacts of each subcategory are shown in logarithmic scales and counted by atom (dark colors, labeled 'Atom') or by residue (light colors, labeled 'Res'). C) The various classes of simulated NMR Ts information lead to different levels of performance measured for "dummy" models generated by us using standard NMR structure determination techniques (see Matherials and Methods for details). The GDT_TS scores of these dummy models produced with all contacts, unambiguous contacts, and 'Ts-specific' true contacts are colored (from bottom to top) cyan, purple, and green, respectively, and shown as solid lines to aid visualization. Dummy model performance (colored lines) is compared to prediction model performance (GDT_TS) for all groups (blue circles), with the top simulated NMR Ts prediction models (solid red line) outperforming the top dummy models for all targets.

Table I
Summary of Contact-Assisted Targets

Target ID	Range	Dom	Class	Category	ECOD architecture
T0761-D0	49-285	2	FM	Tp, Ts, Tc	α+β two layers; duplication
T0763-D1	31-160	1	FM	Tp, Ts, Tc	α+β two layers
T0767-D0	39-312	2	TBM;FM	Tp, Ts, Tc, Tx	α+β two layers; a+b two layers
T0771-D0	26-203	1	FM	Tp	α+β two layers
T0777-D1	18-362	1	FM	Tp, Ts, Tc	*α complex topology*
T0781-D0	34-415	2	FM;TBM-H	Tx	α+β two layers; duplication
T0785-D1	3-114	1	FM	Tp, Ts, Tc	β-sandwiches
T0794-D0	1-462	2	TBM;FM	Tp, Ts, Tc	α+β four layers; beta sandwiches
T0800-D1	36-247	1	TBM-H	Tp, Ts, Tc	β duplicates or obligate multimers
T0802-D0	5-122	1	FM	Tp, Ts, Tc	β-sandwiches
T0804-D0	9-202	2	FM	Tp, Ts, Tc	β duplicates or obligate multimers; beta sandwiches
T0806-D1	1-256	1	FM	Tp, Ts, Tc	α/β three-layered sandwiches
T0808-D0	19-418	2	TBM;FM	Tx	β sandwiches; duplication
T0810-D1	24-136	1	FM	Tp, Ts, Tc	α superhelices
T0812-D1	5-187	1	TBM-H	Ts, Tc, Tx	β-sandwiches
T0814-D0	23-419	3	FM;TBM-H	Tp, Ts, Tc	β-sandwiches; triplication
T0818-D1	30-163	1	TBM	Tp, Ts, Tc	α+β two layers
T0824-D1	2-109	1	FM	Tp, Ts, Tc	few SS elements
T0826-D1	11-211	1	FM	Tp, Ts, Tc	*α bundles*
T0827-D2	212-369	1	FM	Tp, Ts, Tc	*α complex topology*
T0831-D2	109-352	1	FM	Tp, Tc	α bundles
T0832-D1	10-218	1	FM	Tp, Ts, Tc	α+β complex topology
T0834-D0	2-215	2	FM	Tp, Tc	α+β three layers; alpha bundles
T0835-D1	21-424	1	TBM	Tp, Ts, Tc	α superhelices
T0836-D1	1-204	1	FM	Tp, Tc	α bundles
T0848-D2	172-354	1	TBM-H	Tp, Tc	α+β two layers
T0853-D12	5-152	2	TBM	Tp, Tc	α+β two layers; duplication

Table II
Summary of Group Participation in Contact-Assisted Categories

Group Num	Group name	Type	T0 (27)	Tp (23)	Ts (19)	Tc (24)	Tx (4)
32	Legato	Human	27	21	19	21	4
38	nns	Server	27	23	19	24	4
40	GoScience	Human	1	1	8	9	0
41	MULTICOM-NOVEL	Server	27	23	19	24	4
42	TASSER	Human	27	0	0	0	2
44	LEER	Human	27	3	19	24	0
64	BAKER	Human	27	10	19	23	4
65	Jones-UCL	Human	26	21	19	23	4
80	MeilerLab	Human	26	23	19	24	3
155	Cornell-Gdansk	Human	27	0	0	0	1
157	FLOUDAS_A1	Human	27	0	0	0	4
162	McGuffin	Human	27	23	19	24	4
169	LEE	Human	27	23	19	24	4
186	Void_Crushers	Human	1	1	8	9	0
219	Sternberg	Human	0	17	19	0	3
276	FLOUDAS_A4	Human	27	1	18	22	4
287	RBO-human	Human	0	0	11	0	4
300	PhyreX	Server	27	4	0	0	0
310	MUFOLD-R	Human	25	0	15	0	0
329	NMR-I-TASSER	Human	0	0	4	0	0
342	Anthropic_Dreams	Human	1	0	8	9	4
345	FUSION	Server	27	22	19	24	4
357	STAP	Human	27	21	19	21	4
361	Contenders	Human	1	1	6	7	1
420	MULTICOM-CLUSTER	Server	27	23	19	24	0
428	Laufer	Human	1	0	10	8	0
476	Foldit	Human	1	1	8	9	0
479	RBO_Aleph	Server	27	21	9	21	4
490	Wiskers	Human	1	0	2	2	0

Table III
Significance of Target Improvement Using Assisted Information

Target ID	T0 Num	T* Num	MeanT0 GDT_TS	MeanT* GDT_TS	Mean Diff	P values	Target ID	T0 Num	T* Num	MeanT0 GDT_TS	MeanT* GDT_TS	Mean Diff	*P* values
A: Predicted Contacts Tp							**Tc814-D0**	80	61	10.25	35.77	25.52	2.74E-13
Tp761-D0	79	53	13.69	14.08	0.39	1.77E-01	**Tc818-D1**	67	96	32.99	52.5	19.51	2.53E-12
Tp763-D1	99	75	18.02	18.19	0.17	3.41E-01	**Tc824-D1**	74	93	26.5	49.04	22.54	1.72E-14
Tp767-D0	71	53	12.18	13.89	1.71	4.73E-03	**Tc826-D1**	66	68	18.59	41.49	22.9	2.22E-13
Tp771-D0	56	59	14.87	15.58	0.71	1.33E-01	**Tc827-D2**	68	85	21.26	37.82	16.56	4.02E-15
Tp777-D1	71	65	11.09	11.51	0.42	1.39E-01	**Tc831-D2**	65	62	16.88	44.18	27.3	5.08E-19
Tp785-D1	71	59	21.09	19.56	−1.53	5.93E-03	**Tc832-D1**	66	69	16.28	45.17	28.89	7.89E-16
Tp794-D0	77	65	33.71	25.51	−8.2	1.11E-03	**Tc834-D0**	67	66	13.26	26.22	12.96	9.67E-11
Tp800-D1	60	55	33.4	22.79	−10.61	1.81E-06	**Tc835-D1**	67	66	37.33	50.18	12.85	5.70E-05
Tp802-D0	71	60	20.3	21.55	1.25	5.41E-02	**Tc836-D1**	65	61	22.58	39.33	16.75	1.23E-11
Tp804-D0	66	65	12.35	13.61	1.26	4.83E-02	**Tc848-D2**	66	81	19.23	39.59	20.36	1.36E-18
Tp806-D1	65	62	16.56	20.89	4.33	2.19E-02	**Tc853-D0**	66	84	24.27	43.34	19.07	2.17E-11
Tp810-D1	66	60	23.3	23.5	0.2	4.50E-01	**C: Simulated NMR Contacts Ts**						
Tp814-D0	80	58	10.25	8.94	−1.31	2.13E-02	**Ts761-D0**	79	86	13.69	21.78	8.09	8.01E-07
Tp818-D1	67	60	32.99	28.62	−4.37	1.12E-03	**Ts763-D1**	99	106	18.02	35.65	17.63	3.89E-12
Tp824-D1	74	58	26.5	24.7	−1.8	3.75E-02	**Ts767-D0**	71	71	12.18	34.97	22.79	4.26E-12
Tp827-D2	68	60	21.26	23.5	2.24	6.61E-02	**Ts777-D1**	71	80	11.09	17.88	6.79	5.60E-05
Tp831-D2	65	65	16.88	16.29	−0.59	2.14E-01	**Ts785-D1**	71	111	21.09	36.62	15.53	2.72E-08
Tp832-D1	66	54	16.28	15.32	−0.96	2.52E-02	Ts794-D0	77	83	33.71	26.7	−7.01	1.07E-02
Tp834-D0	67	64	13.26	14.33	1.07	2.46E-02	**Ts800-D1**	60	85	33.4	41.61	8.21	1.34E-02
Tp835-D1	67	61	37.33	34.26	−3.07	1.07E-01	**Ts802-D0**	71	116	20.3	43.92	23.62	2.99E-16
Tp836-D1	65	62	22.58	23.53	0.95	2.55E-01	**Ts804-D0**	66	83	12.35	28.17	15.82	1.25E-10
Tp848-D2	66	61	19.23	20.96	1.73	7.22E-02	**Ts806-D1**	65	80	16.56	31.13	14.57	1.02E-05
Tp853-D0	66	55	24.27	23.42	−0.85	2.50E-01	**Ts810-D1**	66	99	23.3	45.64	22.34	8.81E-19
B: Correct Contacts Tc							**Ts812-D1**	81	100	22.88	38.89	16.01	3.14E-09
Tc761-D0	79	70	13.69	39.34	25.65	1.26E-13	**Ts814-D0**	80	73	10.25	25.98	15.73	3.64E-08
Tc763-D1	99	86	18.02	40.57	22.55	7.04E-14	**Ts818-D1**	67	104	32.99	41.58	8.59	3.82E-04
Tc767-D0	71	59	12.18	26.17	13.99	7.98E-16	**Ts824-D1**	74	115	26.5	40.22	13.72	2.69E-09
Tc777-D1	71	70	11.09	38.68	27.59	2.48E-14	**Ts826-D1**	66	80	18.59	30.72	12.13	4.43E-07
Tc785-D1	71	101	21.09	46.68	25.59	5.70E-14	**Ts827-D2**	68	100	21.26	27.77	6.51	5.78E-06
Tc794-D0	77	68	33.71	36.66	2.95	1.85E-01	**Ts832-D1**	66	80	16.28	32.94	16.66	1.48E-08
Tc800-D1	60	69	33.4	42.63	9.23	8.58E-03	Ts835-D1	67	77	37.33	36.9	−0.43	4.51E-01
Tc802-D0	71	100	20.3	49.51	29.21	1.90E-18	**D: Crosslinked Contacts Tx**						
Tc804-D0	66	68	12.35	34.76	22.41	3.73E-13	Tx767-D0	71	72	12.18	12.37	0.19	3.70E-01
Tc806-D1	65	71	16.56	43.7	27.14	8.74E-13	Tx781-D0	75	85	9.89	9.5	−0.39	2.24E-01
Tc810-D1	66	94	23.3	53.69	30.39	1.47E-22	Tx808-D0	78	84	12.81	10.02	−2.79	1.98E-03
Tc812-D1	81	65	22.88	36.89	14.01	6.99E-09	Tx812-D1	81	80	22.88	20.8	−2.08	1.19E-01

Positive mean differences and significant P-values are shaded.

Table IV
Significance of Group Performance Improvement Using Assisted Information

	Individual Performance			Absolute Performance				Individual Performance			Absolute Performance		
Group Num	T* Num	Mean Diff	P values	T* Num	Mean Diff	P values	Group ID	T* Num	Mean Diff	P values	T* Num	Mean Diff	P values
A: Predicted Contacts Tp													
32	86	−4.24	7.20E-09	86	−15.12	3.53E-26	38	95	37.38	3.42E-40	95	27.69	4.25E-28
38	115	1.21	1.88E-02	115	−8.20	2.04E-15	40	40	−3.03	1.51E-02	40	−14.33	2.83E-10
40	5	−1.12	7.93E-02	5	−4.38	1.22E-03	41	95	1.20	2.67E-01	95	−9.89	1.25E-06
41	115	−8.88	3.02E-15	115	−19.42	3.35E-31	44	95	35.24	1.62E-33	95	29.5	4.54E-28
44	15	−2.28	2.57E-03	15	−3.69	4.90E-06	64	95	32.24	1.22E-31	95	28.97	4.61E-27
64	47	1.09	2.38E-01	47	−3.48	1.71E-04	65	91	−8.65	2.76E-07	91	−13.39	5.64E-13
65	106	−8.32	5.69E-12	106	−14.44	2.00E-23	80	95	−2.14	5.24E-04	95	−19.33	4.41E-24
80	115	−1.96	2.22E-04	115	−17.04	2.13E-28	162	95	−3.13	7.48E-05	95	−12.34	6.25E-18
162	115	−2.41	6.49E-04	115	−10.89	2.70E-20	169	95	35.29	1.24E-33	95	29.43	5.02E-28
169	115	−2.45	3.48E-06	115	−8.08	4.50E-15	186	40	3.57	1.42E-02	40	−7.59	5.72E-06
186	5	−1.19	4.34E-02	5	−3.31	1.66E-03	219	93	−10.60	1.05E-19	93	−21.35	2.11E-27
219	84	−10.85	1.74E-19	84	−22.54	5.93E-28	276	90	20.57	1.85E-22	90	8.3	1.33E-04
276	5	−2.12	1.23E-03	5	−6.17	1.92E-05	287	55	−2.50	1.87E-02	55	−13.57	4.05E-09
300	17	−1.78	7.33E-03	17	−8.54	8.71E-10	310	75	−2.34	6.90E-10	75	−10.17	6.01E-15
345	110	3.32	4.05E-07	110	−10.10	4.95E-17	329	4	7.21	1.52E-01	4	2.83	3.27E-01
357	104	1.83	1.20E-09	104	−18.42	2.10E-29	342	40	8.49	5.25E-04	40	−2.76	1.68E-01
361	5	−1.38	8.48E-03	5	−2.73	7.40E-04	345	95	2.22	2.33E-02	95	−12.72	2.15E-12
420	115	−5.60	9.96E-07	115	−16.72	3.00E-26	357	95	12.53	1.61E-08	95	−7.28	3.12E-04
476	5	−1.81	1.74E-02	5	−3.54	1.78E-03	361	30	1.50	2.21E-01	30	−9.38	8.76E-04
479	105	−1.83	4.86E-03	105	−13.32	3.31E-21	420	95	7.51	7.41E-05	95	−4.69	5.97E-03
B: Correct Contacts Tc							428	50	36.16	5.61E-22	50	26.67	7.44E-14
32	85	−2.77	8.82E-06	85	−15.24	1.04E-21	476	40	8.58	2.94E-04	40	−2.53	1.71E-01
38	120	45.98	2.47E-54	120	36.11	4.93E-40	479	45	−1.14	2.61E-03	45	−17.18	3.07E-10
40	45	1.50	1.45E-01	45	−9.74	4.38E-07	490	2	37.85	5.21E-02	2	29	1.08E-01
41	120	13.80	4.17E-15	120	2.47	5.58E-02	**D: Crosslinked Contacts Tx**						
44	120	44.42	8.45E-50	120	38.53	4.07E-44	32	15	−3.55	4.57E-03	15	−13.34	1.33E-07
64	115	37.56	1.35E-47	115	34.66	8.52E-45	38	20	−3.13	3.79E-02	20	−8.35	2.29E-03
65	87	3.13	4.61E-02	87	−2.12	1.17E-01	41	20	−9.26	1.74E-06	20	−17.44	5.64E-09
80	120	4.48	8.86E-09	120	−11.82	7.81E-15	42	10	−10.26	3.13E-03	10	−12.37	9.97E-04
162	120	0.14	3.95E-01	120	−8.58	5.77E-16	64	20	1.50	2.04E-01	20	−1.97	1.06E-02
169	120	45.44	1.87E-55	120	39.43	2.85E-48	65	17	−5.66	2.54E-06	17	−9.41	5.70E-09
186	45	11.93	2.07E-07	45	0.82	3.52E-01	80	15	0.69	2.33E-02	15	−17.8	1.26E-06
276	106	18.00	4.16E-27	106	5.38	1.57E-04	155	5	−2.09	1.65E-03	5	−11.39	2.20E-06
342	45	18.88	3.23E-11	45	7.68	3.69E-03	157	19	−6.82	3.30E-04	19	−13.86	1.22E-09
345	120	5.48	5.16E-08	120	−9.34	3.06E-10	162	20	−4.60	7.72E-03	20	−10.6	1.59E-07
357	98	11.20	1.03E-20	98	−9.07	5.31E-09	169	20	−6.85	4.18E-03	20	−8.4	2.15E-03
361	35	7.62	1.38E-03	35	−4.89	6.11E-02	219	15	−6.29	2.27E-07	15	−12.2	1.37E-09
420	120	20.28	6.30E-23	120	7.79	3.56E-06	276	20	−7.45	1.94E-06	20	−14.4	2.72E-09
428	40	41.58	8.22E-28	40	28.37	5.50E-13	287	20	−4.57	2.96E-03	20	−12.83	9.52E-06
476	45	16.85	3.67E-09	45	5.78	1.94E-02	345	20	−1.74	8.17E-05	20	−16.09	5.90E-08
479	105	−3.58	4.72E-09	105	−15.39	2.35E-24	357	20	1.00	2.59E-03	20	−16.07	2.18E-06
490	2	22.46	2.41E-01	2	17.60	2.79E-01	420	20	−5.37	1.86E-04	20	−15.79	5.75E-08
C: Simulated NMR Contacts Ts							428	5	−15.51	7.33E-06	5	−30.79	4.76E-07
32	84	−4.95	2.37E-08	84	−17.91	8.15E-24	479	20	−3.08	2.45E-05	20	−13.03	1.11E-05

Positive mean differences and significant P-values are shaded.

Table V
Group Ranks and Significance

Grp	Group name	Num Tar-get	Best FM-Style Scoring				First FM-Style Scoring				Win/Loss Scoring		
			Sum *Z*	Sum *R*	Avg *Z*	Avg *R*	Sum *Z*	Sum *R*	Avg *Z*	Avg *R*	Win *F*	*P* values	Win *R*
A	Tp												
169	**LEE**[*a,c]	23	16.54	1	0.72	2	14.85	2	0.68	6	0.86	1.3E-32	1
38	**nns**[*a,c]	23	15.98	2	0.69	3	15.64	1	0.68	5	0.86	7.8E-32	2
162	McGuffin[a]	23	8.85	3	0.38	7	−0.93	5	−0.04	12	0.67	2.5E-08	4
345	FUSION[a]	22	3.25	4	0.24	10	7.44	3	0.43	7	0.59	3.1E-03	5
420	MULTICOM-CLUSTER	23	−0.24	5	−0.01	12	−2.42	6	−0.11	13	0.48	6.7E-01	8
479	RBO_Aleph	21	−2.58	6	0.07	11	2.33	4	0.30	9	0.56	4.3E-02	6
65	Jones-UCL	21	−5.27	7	−0.06	13	−2.96	7	0.05	10	0.52	2.5E-01	7
80	MeilerLab	23	−5.73	8	−0.25	14	−5.41	8	−0.24	14	0.43	9.8E-01	9
64	**BAKER**[b,c]	10	−13.73	9	1.23	1	−14.99	9	1.10	1	0.96	1.4E-26	3
357	STAP	21	−16.97	10	−0.62	15	−15.40	10	−0.54	15	0.24	1.0E+00	10
32	Legato	21	−17.24	11	−0.63	16	−17.90	11	−0.66	17	0.24	1.0E+00	11
41	MULTICOM-NOVEL	23	−21.00	12	−0.91	19	−19.92	12	−0.87	19	0.16	1.0E+00	12
219	Sternberg	17	−27.11	13	−0.89	18	−25.97	13	−0.82	18	0.14	1.0E+00	13
B	Tc												
44	**LEER**[*a,b,c,d]	24	29.94	1	1.25	1	30.37	1	1.27	1	0.95	7.8E-79	1
169	**LEE**[*c,d]	24	28.47	2	1.19	3	28.51	2	1.19	2	0.92	7.6E-66	2
64	**BAKER**[b,c]	23	25.30	3	1.19	2	24.68	4	1.16	3	0.88	4.0E-48	3
38	nns	24	24.70	4	1.03	4	25.15	3	1.05	4	0.83	2.8E-38	4
420	MULTICOM-CLUSTER	24	3.57	5	0.15	8	3.73	5	0.16	8	0.62	6.9E-06	8
41	MULTICOM-NOVEL	24	−10.29	6	−0.43	13	−10.35	6	−0.43	14	0.38	1.0E+00	12
276	FLOUDAS_A4	22	−10.36	7	−0.29	11	−10.67	8	−0.30	11	0.42	1.0E+00	10
65	Jones-UCL	23	−11.37	8	−0.41	12	−10.36	7	−0.36	12	0.39	1.0E+00	11
80	MeilerLab	24	−15.92	9	−0.66	17	−15.83	9	−0.66	17	0.24	1.0E+00	18
345	FUSION	24	−16.69	10	−0.70	18	−16.82	10	−0.70	18	0.26	1.0E+00	17
162	McGuffin	24	−16.87	11	−0.70	19	−19.21	12	−0.80	20	0.27	1.0E+00	16
357	STAP	21	−17.74	12	−0.56	16	−17.30	11	−0.54	15	0.35	1.0E+00	13
479	RBO_Aleph	21	−23.16	13	−0.82	20	−21.05	13	−0.75	19	0.20	1.0E+00	19
428	Laufer	8	−26.15	14	0.73	5	−26.30	16	0.71	5	0.74	1.3E-08	5
476	Foldit	9	−26.28	15	0.41	6	−26.06	14	0.44	6	0.70	1.3E-07	6
342	Anthropic_Dreams	9	−26.66	16	0.37	7	−26.25	15	0.42	7	0.69	1.7E-06	7
186	Void_Crushers	9	−28.97	17	0.11	9	−30.68	17	−0.08	10	0.58	2.8E-02	9
32	Legato	21	−31.74	18	−1.23	21	−30.83	18	−1.18	21	0.05	1.0E+00	20
40	GoScience	9	−34.98	19	−0.55	15	−35.79	19	−0.64	16	0.33	1.0E+00	14
361	Contenders	7	−37.33	20	−0.48	14	−36.66	20	−0.38	13	0.33	1.0E+00	15
C	Ts												
64	**BAKER**[*a,b,c,d]	19	25.96	1	1.37	1	25.21	1	1.33	1	0.90	2.9E-55	3
44	**LEER**[b,e]	19	22.24	2	1.17	2	22.33	3	1.18	3	0.92	2.8E-60	1
169	**LEE**[a,b,c,d]	19	22.17	3	1.17	3	22.52	2	1.19	2	0.92	2.8E-60	2
38	**nns**[d]	19	21.11	4	1.11	4	21.69	4	1.14	4	0.85	1.6E-41	4
420	MULTICOM-CLUSTER	19	1.55	5	0.08	10	0.82	6	0.05	10	0.62	1.3E-05	7
276	FLOUDAS_A4	18	0.98	6	0.17	9	1.41	5	0.19	7	0.65	1.3E-07	6
357	STAP	19	0.33	7	0.02	11	−0.25	7	−0.01	11	0.60	3.0E-04	8
428	**Laufer**[d]	10	−7.75	8	1.02	5	−8.04	8	1.00	5	0.87	1.2E-26	5
345	FUSION	19	−10.38	9	−0.55	18	−10.28	10	−0.54	20	0.32	1.0E+00	18
41	MULTICOM-NOVEL	19	−11.26	10	−0.59	20	−10.18	9	−0.54	19	0.34	1.0E+00	17
65	Jones-UCL	19	−12.33	11	−0.65	21	−11.60	11	−0.61	21	0.25	1.0E+00	19
162	McGuffin	19	−12.65	12	−0.67	22	−16.73	14	−0.88	23	0.25	1.0E+00	20
80	MeilerLab	19	−13.55	13	−0.71	23	−13.24	12	−0.70	22	0.24	1.0E+00	21
310	MUFOLD-R	15	−13.92	14	−0.39	16	−13.60	13	−0.37	17	0.39	1.0E+00	13
32	Legato	19	−18.52	15	−0.97	24	−18.44	16	−0.97	24	0.12	1.0E+00	22
287	RBO-Human	11	−18.80	16	−0.25	13	−18.38	15	−0.22	13	0.49	5.6E-01	12
219	Sternberg	19	−20.19	17	−1.06	25	−20.23	17	−1.06	25	0.08	1.0E+00	23
476	Foldit	8	−20.39	18	0.20	7	−20.81	19	0.15	9	0.59	1.4E-02	10
342	Anthropic_Dreams	8	−20.53	19	0.18	8	−20.73	18	0.16	8	0.61	3.8E-03	9
186	Void_Crushers	8	−22.71	20	−0.09	12	−22.77	20	−0.10	12	0.50	5.0E-01	11
479	RBO_Aleph	9	−24.98	21	−0.55	19	−24.47	21	−0.50	18	0.35	1.0E+00	16
40	GoScience	8	−25.06	22	−0.38	14	−24.81	22	−0.35	16	0.36	1.0E+00	14
361	Contenders	6	−28.70	23	−0.45	17	−28.10	23	−0.35	15	0.35	1.0E+00	15
D	Tx												
64	**BAKER**[*a,c]	4	6.20	1	1.55	1	5.85	1	1.46	1	0.95	3.1E-14	1
287	RBO-Human	4	3.31	2	0.83	2	3.07	2	0.77	2	0.78	6.1E-06	3
169	**LEE**[b]	4	3.06	3	0.76	3	2.91	4	0.73	4	0.82	3.8E-07	2

Table V
(Continued)

Grp	Group name	Num Tar-get	Best FM-Style Scoring				First FM-Style Scoring				Win/Loss Scoring		
			Sum *Z*	Sum *R*	Avg *Z*	Avg *R*	Sum *Z*	Sum *R*	Avg *Z*	Avg *R*	Win *F*	*P* values	Win *R*
38	**nns**[b]	4	3.01	4	0.75	4	2.95	3	0.74	3	0.75	6.7E-05	4
162	McGuffin	4	1.96	5	0.49	5	−0.74	7	−0.19	9	0.65	1.4E-02	6
479	RBO_Aleph	4	0.76	6	0.19	7	1.45	5	0.36	5	0.53	3.5E-01	8
65	Jones-UCL	4	0.49	7	0.12	8	0.21	6	0.05	7	0.58	1.2E-01	7
420	MULTICOM-CLUSTER	4	−1.45	8	−0.36	10	−1.35	9	−0.34	11	0.42	8.8E-01	10
357	STAP	4	−1.55	9	−0.39	11	−0.96	8	−0.24	10	0.30	1.0E+00	14
276	FLOUDAS_A4	4	−1.58	10	−0.40	12	−1.46	10	−0.37	12	0.38	9.5E-01	11
345	FUSION	4	−1.86	11	−0.47	13	−1.77	12	−0.44	15	0.35	9.9E-01	13
32	Legato	4	−1.91	12	−0.48	14	−1.70	11	−0.43	13	0.38	9.5E-01	12
157	FLOUDAS_A1	4	−2.33	13	−0.58	16	−2.97	14	−0.74	17	0.27	1.0E+00	15
80	MeilerLab	3	−2.55	14	−0.18	9	−1.89	13	0.04	8	0.44	7.7E-01	9
42	TASSER	2	−3.49	15	0.25	6	−3.33	15	0.33	6	0.73	8.1E-03	5
219	Sternberg	3	−4.01	16	−0.67	17	−3.93	16	−0.64	16	0.23	1.0E+00	16
41	MULTICOM-NOVEL	4	−4.70	17	−1.18	18	−3.95	17	−0.99	19	0.00	1.0E+00	17

[a]Same ranking by GDT_TS.
[b]Significant Best FM score by *t* test.
[c]Significant Best FM score by Bootstrap.
[d]Significant Best TBM score by *t* test.
[e]Significant Best TBM score by Bootstrap.
Positive FM-style scores and win/loss fraction >=0.5 are shaded; top-ranked groups by best model scores (and any groups not significantly different using Bootstraps and T-tests on FM-style and TBM-style scores for Tc and Ts) are bolded.

REFERENCES

1. Bowers PM, Strauss CE, Baker D. De novo protein structure determination using sparse NMR data. J Biomol NMR. 2000;18(4):311–318.
2. Kim DE, Dimaio F, Yu-Ruei Wang R, Song Y, Baker D. One contact for every twelve residues allows robust and accurate topology-level protein structure modeling. Proteins. 2014;82(Suppl 2):208–218.
3. Taylor TJ, Bai H, Tai CH, Lee B. Assessment of CASP10 contact-assisted predictions. Proteins. 2014;82(Suppl 2):84–97.
4. Monastyrskyy B, D'Andrea D, Fidelis K, Tramontano A, Kryshtafovych A. Evaluation of residue-residue contact prediction in CASP10. Proteins. 2014;82(Suppl 2):138–153.
5. Chance MR, Bresnick AR, Burley SK, Jiang JS, Lima CD, Sali A, Almo SC, Bonanno JB, Buglino JA, Boulton S, Chen H, Eswar N, He G, Huang R, Ilyin V, McMahan L, Pieper U, Ray S, Vidal M, Wang LK. Structural genomics: a pipeline for providing structures for the biologist. Protein Sci. 2002;11(4):723–738.
6. Kinch LN, Li W, Schaeffer RD, Dunbrack RL, Monastyrskyy B, Kryshtafovych A, Grishin IV. CASP 11 Target Classification. Proteins. 2016
7. Zemla A. LGA: A method for finding 3D similarities in protein structures. Nucleic Acids Res. 2003;31(13):3370–3374.
8. Ben-David M, Noivirt-Brik O, Paz A, Prilusky J, Sussman JL, Levy Y. Assessment of CASP8 structure predictions for template free targets. Proteins. 2009;77(Suppl 9):50–65.

9. Jauch R, Yeo HC, Kolatkar PR, Clarke ND. Assessment of CASP7 structure predictions for template free targets. Proteins. 2007;69(Suppl 8):57–67.
10. Kinch L, Yong Shi S, Cong Q, Cheng H, Liao Y, Grishin NV. CASP9 assessment of free modeling target predictions. Proteins. 2011;79(Suppl 10):59–73.
11. Kinch LN, Wrabl JO, Krishna SS, Majumdar I, Sadreyev RI, Qi Y, Pei J, Cheng H, Grishin NV. CASP5 assessment of fold recognition target predictions. Proteins. 2003;53(Suppl 6):395–409.
12. Kopp J, Bordoli L, Battey JN, Kiefer F, Schwede T. Assessment of CASP7 predictions for template-based modeling targets. Proteins. 2007;69(Suppl 8):38–56.
13. Vincent JJ, Tai CH, Sathyanarayana BK, Lee B. Assessment of CASP6 predictions for new and nearly new fold targets. Proteins. 2005;61(Suppl 7):67–83.
14. Cozzetto D, Kryshtafovych A, Fidelis K, Moult J, Rost B, Tramontano A. Evaluation of template-based models in CASP8 with standard measures. Proteins. 2009;77(Suppl 9):18–28.
15. Tramontano A, Morea V. Assessment of homology-based predictions in CASP5. Proteins. 2003;53(Suppl 6):352–368.
16. Brunger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, Read RJ, Rice LM, Simonson T, Warren GL. Crystallography & NMR system: A new software suite for macromolecular structure determination. Acta Crystallogr D Biol Crystallogr. 1998;54(Pt 5):905–921.

17. Monastyrskyy B, D'Andrea D, Fidelis K, Tramontano A, Kryshtafovych A. New encouraging developments in contact prediction: Assessment of the CASP11 results. Proteins. 2015

18. Kinch LN, Li W, Monastyrskyy B, Kryshtafovych A, Grishin IV. Evaluation of free modeling targets in CASP11 and ROLL. Proteins. 2015

19. Cheng H, Schaeffer RD, Liao Y, Kinch LN, Pei J, Shi S, Kim BH, Grishin NV. ECOD: an evolutionary classification of protein domains. PLoS Comput Biol. 2014;10(12):e1003926.

CHAPTER 9 GENOMES OF 250 SKIPPER BUTTERFLIES REVEAL RAMPANT CONVERGENCE IN WING PATTERNS[8]

INTRODUCTION

when an animal interacts with its environment, it is phenotype that matters. Phenotypic traits define the place of an organism within its ecosystem and within a human-made classification system as perceived by zoologists for two centuries. Due to evolutionary plasticity of phenotypes riddled by adaptive convergence, it is challenging to decipher phylogeny, and thus deduce phylogenetic classification from morphology. The phenotype is encoded by the genotype, which equally bears the footprint of evolution. DNA sequences are prime for evolutionary studies. Sanger-sequencing of select gene markers revolutionized phylogenetic research and refined morphology-based classification. However, frequent homoplasies and in-ability to perfectly model DNA sequence changes are serious obstacles to phylogeny reconstruction from thousands of base pairs. Complete genomes are composed of millions of base pairs. The knowledge of complete genotypes is expected to confidently resolve many outstanding questions. While large-scale genomic studies are still scarce and are mostly performed by hefty consortia, they are most enlightening [1].

[8] This Chapter was submitted as:
Li W, Shen J, Cong Q, Zhang J, Grishin NV. *Genomes of 250 skipper butterflies reveal rampant convergence in wing patterns*

To exemplify a large-scale genomics project possible within a single small lab, we have chosen Skipper butterflies (Hesperiidae). This family is very diverse in wing patterns and shapes and comprises about 4000 species worldwide, but has not received as much attention as other butterflies. Nevertheless, a number of recent studies focused on Skipper phylogeny from DNA sequences of several genes [2–5]. Pioneering work by Warren et al. [2, 3] revealed many surprising phylogenetic relationships among Hesperiidae compared to the last comprehensive morphological treatment [6–11]. However, many questions about Skipper phylogeny remain unanswered. In particular, the Eudaminae subfamily that includes most diverse, large and colorful Hesperiidae was not analyzed in detail.

Here, we obtained and analyzed shotgun genomic reads of 250 species of Skippers, covering all genera in the Eudaminae subfamily. As a result, we constructed genome-level phylogeny of Hesperiidae and found that many Eudamine skippers have been attributed to evolutionary groups they do not belong. The convergence in wing patterns and shapes is rampant and is possibly mimetic. Interestingly, we also found a group of close relatives with disparate morphology. Furthermore, sequence data suggest possible genomic determinants of morphological traits and indicate that our approach is promising to revamp the way biodiversity research is carried out.

RESULTS AND DISCUSSION

To better understand evolution and biology of the family Hesperiidae with the emphasis on the poorly understood subfamily Eudaminae, we sequenced genomes from all its major phylogenetic lineages and all genera of Eudaminae were covered.

Genomes of 250 skipper butterflies.

We selected 250 species of Hesperiidae from all subfamilies and tribes, and 110 were from the subfamily Eudaminae. Only 36 specimens were collected past 2012 and preserved in a DNA-friendly manner. Others were stored pinned in collections and some were collected over a century ago. The oldest was the holotype of *Pseudodrephalys atinas* collected prior to its description in 1888. Nevertheless, shotgun genomic reads targeting 10x coverage of the genome allowed us to assemble genomic regions with the emphasis on protein-coding genes. Out of nearly 12618 genes from the two reference genomes of Hesperiidae we obtained previously [12, 13], more than 75% were more than 50% complete in 85 specimens. The total length of aligned regions in each specimen was 8,100,834+/-2,387,641 positions. In addition to nuclear genomes, we assembled mitogenomes that covered 10,663+/-674 positions and were >90% complete in 232 specimens.

Timed genomic tree of skippers and subfamilies.

We constructed genomic trees from concatenated alignments of coding regions (Fig. 1) and from introns, and both trees had identical topologies. Moreover, to probe problems with incomplete lineage sorting we constructed gene trees and combined them using ASTRAL [14]. Topology of this tree was also the same, giving confidence in the results. The tree from combined coding regions and introns was timed. The major phylogenetic groups in the timed tree constructed from the genomic data agree with the groundbreaking works by Warren et al. [2, 3] and follow-up publications that employed larger set of species and genes [4, 5]. Importantly, the Awls (Coeliadinae) is the subfamily that is sister to the rest of Hesperiidae (Fig. 1). Australian endemic *Euschemon*, as suggested by several studies [2, 3, 15], is unique and forms its own subfamily, sister to all other Hesperiidae with exclusion of Coeliadinae. The next split in the genomic tree is different from that proposed previously based on DNA studies, but in agreement with morphological view. The Spreadwing skippers (mostly dicot feeders) and Grass skippers (mostly monocot feeders) are sisters. Each of these groups were previously divided into several subfamilies. Genomic data very strongly support (bootstrap above 99%) monophyly of these subfamilies. The two latest diverging subfamilies (Hesperiinae and Trapezitinae) have split about 50Mya. Pyrginae, as defined by Warren et al. (2009) diversified from their common ancestor before that time (>55Mya). Moreover, while other subfamilies except Pyrginae (i.e., Eudaminae, Heteropterinae and Hesperiinae) are well-separated from each other, Pyrginae sensu Warren split within a short time into three compact groups, prior to divergence of Grass skippers into subfamilies. Thus these three groups within Pyrginae are more similar to other subfamilies in Hesperiidae and we treat

them as such. Accordingly, the sister tribes of Warren [3] Tagiadini and Celaenorrhinini are unified in a subfamily Tagiadinae, and the subfamily Pyrrhopyginae is re-instated. The Firetips (Pyrrhopyginae) are strikingly distinct in appearance from other skippers, have been traditionally considered a subfamily and they diverged from their common ancestor with Pyrginae about 55Mya, prior to the divergence of Grass Skippers. Thus, genomic data suggest that the family Hesperiidae consists of 9 subfamilies (Fig. 1) and diversification into subfamilies occurred about 50Mya.

Mitogenomes and COI barcodes.

In addition to nuclear genome tree, we constructed a tree from mitogenomes. The resulting tree recapitulates major phylogenetic groupings of the nuclear genome tree, but with weaker support. All the subfamilies and tribes are composed of the same species in mitogenome or nuclear genome phylogenies. Next, using complete mitogenomes of 250 specimens as a backbone, we increased taxonomic coverage of Eudaminae by adding 290 specimens with COI barcodes only. These specimens confidently grouped with mitogenomes of their expected phylogenetic group. Most of these barcode-only specimens were placed consistently with their current classification with several notable exceptions. We used this mitogenome + barcode tree together with the nuclear genome tree (Fig. 1) as the basis for our proposed classification of Hesperiidae. Differences in barcodes suggested that a number of subspecies as defined by Evans (1952, 1953, 1955)

are more likely to be species, and all such cases were analyzed in detail for wing pattern and genitalic differences. We concluded that 27 subspecies should be treated as species.

Eudaminae tribes and subtribes.

Previous studies refrained from defining tribes in Eudaminae subfamily due to small number of species used in the analysis based on a limited number of genes. Therefore, we focused on this subfamily and attempted to delineate the tribes consistently with their definition in Pyrginae (sensu stricto). The four tribes of Pyrginae (Carcharodini, Achlyodini, Erynnini and Pyrgini) outlined by Warren et al. [3] diverged around 42Mya. Genomic tree is consistent with the Warren et al. definitions with some exceptions: we transfer *Grais*, *Tosta*, *Morvina*, *Myrinia*, *Xispia*, *Pseudodrephalys*, *Mimia*, *Eracon*, and *Spioniades* to Achlyodini, *Cornuphallus* to Carcharodini, *Clito* to Erynnini, and *Jera* to Pyrrhopyginae. Moreover, *Cabirus* does not belong to Eudaminae (as also suggested by [4]), and we confidently place it in Achlyodini. On the other hand, *Emmelus* is transferred to Eudaminae. All Pyrginae tribes received 100% bootstrap support and indeed represent major groupings in the subfamily. Cutting through the genomic tree around the level of Pyrginae divergence into tribes results in 4 Eudaminae phylogenetic groups supported by 100% bootstrap that we define as tribes. Two of the tribes that form best-separated groups are described in Table 1 as Oileidini and Entheini. The two others are closely related sisters Eudamini and Phocidini that diverged about 40Mya, and we give them a tribal rank due to their morphological distinction.

The four Eudaminae tribes defined by genomic divergence correspond to groups with similar morphology [9, 10]. For instance, Entheini is basically "B. Augiades group" of Evans that he defined by the divergent 3rd segment of palpi that is set on the outer edge of the 2nd segment. Inconsistently, Evans included three genera with the central 3rd segment in this group: *Phocides*, *Hypocryptothrix* and *Cabirus* that do not belong to Entheini, and *Cabirus* does not even belong to Eudaminae (Fig. 1). Interestingly, *Phareas* that has divergent 3rd segment (and was included in Auguades group) does not belong to Entheini and males possess tufts of long hair-like scales in the groove along hindwing vein 1A+2A, not found in Entheini. We confidently place this uniquely patterned skipper in Phocidini, some of which have tufts of similar scales on wings. The structure of palpi in *Phareas* is likely convergent.

In the genomic tree, Oileidini is sister to the rest of Eudaminae. Genera in this tribe were grouped together with some Pyrginae genera by Evans [10], suggesting that the tribe may be intermediate in morphology between Pyrginae and Eudaminae. This is the smallest tribe (6 genera) and is characterized by tufts of hair-like scales in the groove along hindwing vein 1A+2A in males, either below (*Oileides*) or above (the rest). Similar structures are present in *Phareas* from the Phocidini tribe, but on both sides of hindwing.

The sisters Eudamini and Phocidini are separated from each other by a short branch and could be treated as one tribe. However, both are strongly monophyletic within (100% bootstrap), and Phocidini stand out morphologically and ecologically: forewing veins R4 and R5 are approximate at their origins, hindwing tornus usually produced (not lobed), and skippers hold wings spread flat in resting pose, are crepuscular and come to

light, and many species have striking sexual dimorphism. Our Phocidini is the "D. Celaenorrhinus group" of Evans [9] after removal of the following genera: *Cephise*, which has strongly lobed or tailed hindwing tornus, *Celaenorrhinus*, which males have a tuft of long scales on hind tibiae fitting in a thoracic pouch, a feature not present in Eudaminae, and *Lobocla*, the only Old World Eudaminae. Genomic tree suggests that *Oileides* is polyphyletic, and only one species (together with *Aurina*) belongs to Phocidini.

Eudamini is the largest and most diverse tribe that encompasses more than half of the subfamily. Genomic tree reveals meaningful groupings within the tribe that are described here as subtribes (Table 1). One of them, substribe Cephisina is monotypic including a single genus *Cephise*, which diverged from its sister tribe Telemiadina 35Mya and is unique in its morphological features [16]. Telemiadina includes three genera: *Telemiades* with its close sister *Polygonus* and *Ectomis*, in which we subsume genera *Hypocryptothrix*, *Heronia*, *Polythrix*, and *Chrysoplectrum*. Along with *Lobocla*, Loboclina unifies genera with the arcuate antennal club from "C. Urbanus group" of Evans plus *Venada*, *Aguna* and a new genus *Zeutus*. The rest belong to the "crown" group of Eudaminae and includes the most interesting array of skippers that have been largely misclassified previously due to rampant and possibly mimetic convergence in wing patterns as detailed below.

Eudaminae genera and subgenera.

While there is no accepted universal criteria for the definition of a genus, it has been proposed that major phylogenetic clusters of species with common ancestors existing within a certain time-frame could correspond to genera [17]. Looking for a consistent definition, we cut a timed phylogenetic tree at a time-point to maximize agreement with the current classification (i.e., most genera that are well-defined by morphological features are neither split not merged) and call genera the groups supported by the cut branches. 15Mya corresponds to such point, that on the one hand keeps well-known genera *Aguna*, *Udranomia* and *Urbanus proteus* group unsplit and, on the other hand, separates traditional and morphologically distinct pairs of sister genera such as *Epargyreus* and *Chioides*. Moreover, we attempt to reduce the number of monotypic genera, unless the genus is truly distinct, because it seems more instructive to indicate relationships to other species by the generic name. As a result (Fig. 1), we outlined 50 Eudaminae genera, 4 of which are described as new in Table 2. The number of monotypic genera decreased from 10 to 4: *Pseudonascus*, *Nicephellus*, *Spathilepia* and *Zeutus* g. nov. [18] (plus *Emmelus* transferred from Pyrginae), and these three diverged from their sister taxa at least 18Mya and are morphologically distinct.

Within some genera, we see informative phylogenetic groups of species that are meaningful to define as subgenera, and 3 new subgenera are described (Table 2). Some of the subgenera, such as *Thorybes* have been used as genera for decades, but their genomic and morphological distinctness is insufficient compared to other genera.

Rampant convergence in wing patterns and shapes.

Arguably the most unexpected result of this study is the astounding number of misclassifications, when species were attributed to genera they do not belong to. The genera themselves proposed over the years of classic entomological studies mostly stood the test of genomic data: i.e., 55 genera were recognized prior to our work and we revise them to 50. We eliminated several monotypic genera for which visual morphological differences hindered close relationships with other species, and merged several phenotypically diverse but genotypically close genera. Apparently, morphological distinction between certain kinds of phenotypes (e.g. wing shape, such as tailed hindwing, or wing pattern, such as a pale stripe across forewing) may be indicative of genomic divergence and thus the time since the taxa split from their common ancestor. However, attribution of a species to a genus by its dominant to human eye phenotypic feature is more problematic. For instance, many tailed skippers have been placed in the genus *Urbanus* based on the tail. However, genomic data imply that half of them do not belong there and they were transferred to 3 other genera, one is newly proposed while others didn't include tailed skippers. Moreover, we transferred some skippers with short tails and without tails to *Urbanus* from *Astraptes*. They were previously misclassified due to wing patterns consisting of shiny metallic-cyan wing bases and/or white forewing spots.

We found this situation to be rampant across Eudaminae and attribute it to mimetic convergence. This convergence is not confined to one or two basic patterns to mimic an un-palatable model, but is significantly more diverse. In the most illustrious

example (Fig. 2), we see five different phenotypes that parallel each other in two genera (*Telegonus* and *Cecropterus*) and outgroups (several genera): (1) greenish bases of wings, white stripe of spots on the forewing, hindwing with white tail and margins; (2) metallic-blue wing bases, forewing with a stripe of spots; (3) forewing with a yellow stripe across and apical white spots; (4) chocolate-brown wings, hindwing with yellow tornus; (5) cream-white, semitranslucent spots on the forewing arranged in a typical for Eudaminae pattern.

At least 4 of these phenotypes are not ancestral and thus are convergent. DNA-based phylogeny reveals monophyly of these two proposed genera (Fig. 1). Amazingly, every single species out of 10 shown here was previously misclassfied by visual appearance into a wrong genus (given after "not" by each specimen in Fig. 2). In retrospect, correct assignment to genus could have been achieved through careful inspection of male genitalia (Fig. 2). In *Telegonus*, dorsal side of valva is concave in the middle and forms a mouth-like (in profile) structure with two "kissing lips". In *Cecropterus*, valva is dorsally and terminally convex without a "kiss", but may have sharp "hooks" instead. These genitalic features agree with genomic phylogeny and reinforce our conclusions. Four genera were chosen to represent these phenotypes in the outgroups (Fig. 1). Since the outgroups are more divergent phylogenetically, their previous attribution to genera was mostly correct.

Uncanny divergence within a genus

Prior to our work, each of the three genera *Ectomis*, *Hypocryptothrix* and *Heronia* included but a single skipper species of unique appearance. Never before they were looked at together. To our surprise, all phylogenetic trees we obtained (even including COI barcodes only) revealed but a slight divergence among these three and two other genera, *Polythrix* and *Chrysoplectrum*, suggesting that it is best to place all these skippers in a single genus *Ectomis*. Moreover, their most divergent phylogenetic cluster was part of *Polythrix*, and is described here as a subgenus *Asina* subgen. n. (Table 2). COI Barcode divergence among the subgenus *Ectomis* is within 10% (e.g., 9.5% between *Ectomis cythna* and *"Hypocryptothrix" teutas*) and is less than within the genus of swallowtail butterflies *Pterourus*, which some researchers consider a subgenus of *Papilio*.

Despite limited genetic divergence, the expanded *Ectomis* contains species of uncanny phenotypic divergence. All skippers in the former genus *Polythrix* (subsumed by *Ectomis*) are tailed. All the remaining congeners are not, but their hindwing is usually lobed at tornus. While most *Ectomis* are brown skippers with a pale forewing band frequently divided into spots, some vary from solid dark brown to dark metallic green with forewing central spot, or tawny with many white spots. Some are even part of the mimetic complex with brilliant-blue wing bases and body and white stripe across forewing. Males of some species possess tufts of hair-like scales on hindwing

below, while others have a double row of yellow spines on hind tibiae. Male genitalia are as diverse as wing shapes and patterns, and the phylogenetic closeness is not apparent from genitalia. E.g., valva varies from simple curved plate without elaborations

(in *E. cythna*) to amazingly complex with several processes (in *"Heronia" labriaris*). It would be worthwhile to investigate genetic mechanisms for such a rapid phenotypic divergence within *Ectomis*.

DISCUSSION: A BROAD PICTURE.

Today, arguably the most efficient and cost-effective way to gain rapid insights about biodiversity is to genome-sequence it. Comparative approach when a group of organisms, e.g., a family, is chosen and all its key representatives are sequences and analyzed, along with their morphology and ecology, is rich in discoveries. Even now such work can be done within a small lab and not a large genomic center. We illustrate how such project can be accomplished and results that can be expected from it, taking a butterfly family Hesperiidae as an example. Not only because this family is interesting in its own right, but rather to emphasize several general points. First, while several reference genomes require freshly collected specimens, the bulk of the project can be done using museum samples. Even specimens collected a century ago yield usable genomic data. Second, we provide the ultimate genome-based phylogeny of the group and reclassify it taxonomically. Unexpectedly, we find that many species were classified incorrectly, suggesting that many other families will not be an exception to this rule. Third, we find amazing examples of phenotypic convergence and divergence, and mine genomic data for the links between genotype and phenotype. With the ever-decreasing cost of

sequencing, we expect that soon any phylogenetic project will start from sequencing and analysis of complete genomes.

MATERIALS AND METHODS

For freshly collected specimens, DNA was extracted from a piece of tissue of a specimen (minus wings and genitalia that were stored in an envelope) field-stored in a vial with RNAlater. For pinned and dry specimens from museum collections, DNA was extracted either from a whole abdomen prior to genitalia dissection or from a leg. Details of methods for DNA extraction, genomic library preparation, DNA barcoding, next-generation sequencing and computational analysis of complete nuclear and mitochondrial genomes have been reported previously [12, 20]. Phylogenetic trees were constructed with RAxML and ASTRAL.

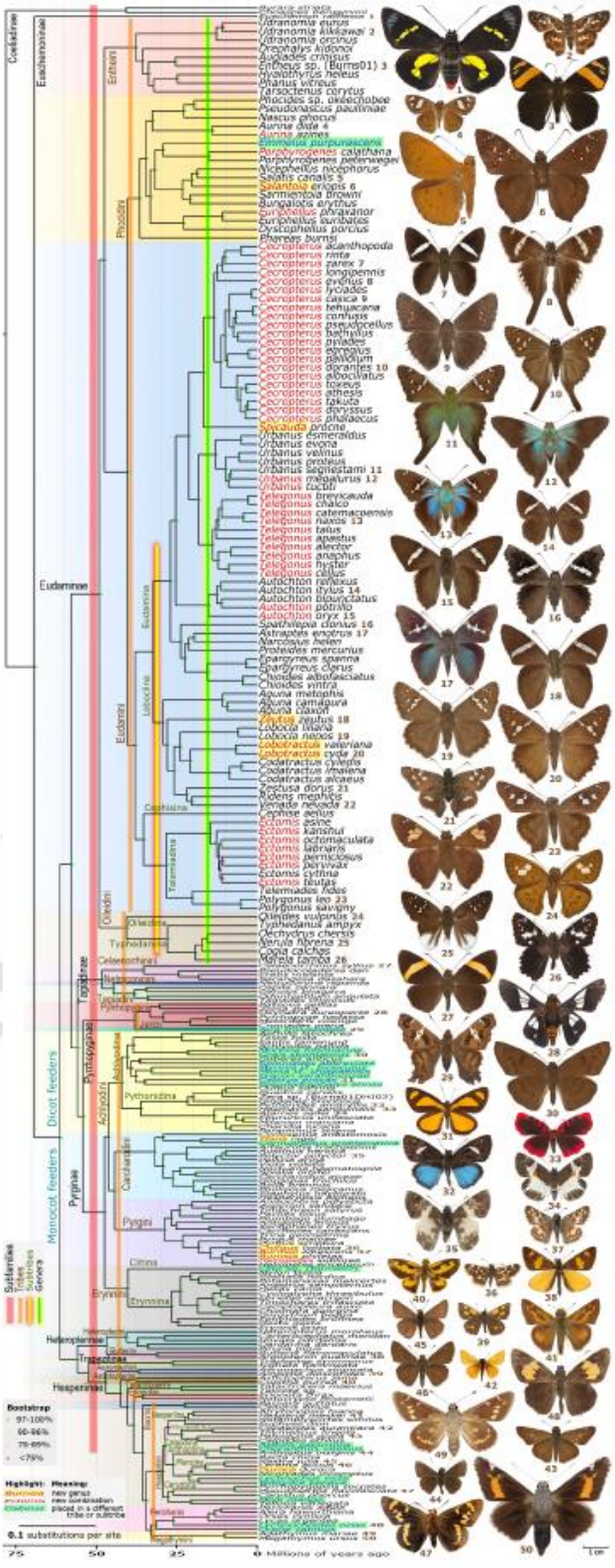


Fig. 1. Timed genomic tree of Hesperiidae. See SI Appendix for details.

Table 1. Description of new tribes and subtribes of Hesperiidae[a]

	Entheini, trib. n.	Oileidini, trib. n.	Loboclina, subtr. n.	Cephisina, subtr. n.
Type genus	*Entheus* Hübner, [1819]	*Oileides* Hübner, [1825]	*Lobocla* Moore, 1884	*Cephise* Evans, 1952
Diagnosic characters[b]	276665.26:A192G; 85.28:C176T; 378.19:G1099C; 374.14:G1169T; keys to B.3a in Evans (1952:6), but exclude B.9	1139.19:T562A; 851.8:C423A,G443A; 11945.11:G391A; 65.4:C330A; tuft of scales by anal fold on hindwing, either above or below	208.2:G145T,T146G; 18312.8:A619C,G620A; keys to C.5, C.10a, C.15.2 or C.18 in Evans (1953)	COI.bc:A44T,C84G,T479A; given by Burns (1996: 182-183) for *Cephise*: genitalia and palpi
	Telemiadina, subtr. n.	Typhedanina, subtr. n.	Pythonidina, subtr. n.	Clitina, subtr. n.
Type genus	*Telemiades* Hübner, [1819]	*Typhedanus* A. Butler, 1877	*Pythonides* Hübner, [1819]	*Clito* Evans, 1953
Diagnosic characters	536.149:G1488C; 997.8:G514T; 860.7:A748G; 3001.3:C1773T; keys to A.2, C.3, C.7a, E.6 or E.9 in Evans (1952, 1953)	1341.12:T25841C; 489.5:G307T; 3446.8:T2308A,C2309G,A2500C; tuft of scales by anal fold on hindwing above, but not below	274.29:G397A; 3478.6:T116C; 7985.5:G916A; 925.10:G199C; in Evans (1953), keys to E.44a or if uncus undivided then E.37a or 40d	COI.bc:G29T,81A,169A, 266A,302T,A353T,A521T; keys to E.52 or E.13.8 in Evans (1953:16,37)
	Netrocorynini, trib. n.	Jerini, trib. n.	Butlerini, trib. n.	Pericharini, trib. n.
Type genus	*Netrocoryne* C. & R. Felder, 1867	*Jera* Lindsey, 1925	*Butleria* Kirby, 1871	*Perichares* Scudder, 1872
Diagnosic characters	2284.30:399A(not T); 904.14:T439G; 275215.7:C925G; 998.8:G308A; 214.24:3520C(not A); keys to C.1 in Evans (1949:12)	103.23:796A(not G); 420.27:G308A; 671.27:935C(not T); 425.5:G1558T; keys to E.3 in Evans (1953:6); forewing cell > 3/5 of costa	2627.8:A1459T; 141.4:C104A; 37338.38:G133T, G134C; keys to H.4 & 5 in Evans (1955:10)	596.8:C1601G; 144.41:G201C; 83.15.e48:G8658A,T8657G; keys to K.27a in Evans (1955:207)

[a]See SI Appendix for Zoobank registration, lists of genera included in each taxon and sequences of exons with diagnostic characters. [b]272.1:A192G means position 192 in exon 1 of the gene 272 is G, changed from A in the ancestor; 169A: ancestral state is unclear; COI.bc is COI barcode region.

Table 2. Description of new genera and subgenera of Hesperiidae[a]

	Salantoia, gen. n.	*Spicauda*, gen. n.	*Zeutus*, gen. n.	*Lobotractus*, gen. n.
Type sp.	*Eudamus eriopis* Hewitson, 1867	*Goniurus procne* Plötz, 1881	*Cecropterus zeutus* Möschler, 1879	*Eudamus valeriana* Plötz, 1881
Diagnosic characters[b]	59C, A79T, T163A, 530T, 598A, T637A; D.3.2 & 3 in Evans (1952: 144); harpe flat, not hook-shaped	307T, T349A, 424(not T), G506A, T562A; keys to C.13.13c in Evans (1952:93); harpe dorsally spiked	A22T, C271A, T278A, A526T, T548C, A607T; genitalia as for *zeutus* by Williams & Bell (1934:27)	49A, T400A, 401T, A477G 517T, C542T, T619A; "*cyda* group" of Burns (1996:196)
Derivation of name[c]	Feminine, a blend of *Sala*[tis] and [Sarmie]*ntoia*	Feminine, a blend of spica and cauda	Masculine, echoes type species name	Masculine, a blend of *Lobo*[cla] and [Coda]*tractus*
	Burnsia, gen. n.	*Urbanoides*, subgen. n.	*Caudatractus*, subgen. n.	*Asina*, subgen. n.
Type sp.	*Syricthus communis* Grote, 1872	*Goniurus esmeraldus* A. Butler, 1877	*Eudamus alcaeus* Hewitson, 1867	*Eudamus asine* Hewitson, 1867
Diagnosic characters	205T, 223A, 241T, 263(not C), T277A, 479T(not A); keys to G.1.5, 8, or 9a in Evans (1953)	T49A, A85T or C, T212A, C542T, T544A, A607C or T, T619A or G; keys to C.13.6a in Evans (1952:86)	355A(not T or C), T556A, A592T; *Codatractus* (C.11 in Evans [1952]) with tailed hindwing	T70A, T127A, T197C, 206T, 208A, A256T, T346A, 373A; keys to C.7.2a in Evans (1952:68)
Derivation of name	Feminine, honors Skipper taxonomist John M. Burns	Masculine, similar to *Urbanus*	Masculine, includes tailed species of *Codatractus*	Feminine, derived from the type species name
	Tiana, gen. n.	*Chirgus*, gen. n.	*Duroca*, gen. n.	*Tekliades*, gen. n.
Type sp.	*Anastrus platypterus* Mabille, 1895	*Hesperia limbata* Erschoff, 1876	*Hesperia duroca* Plötz, 1882	*Thymele ramanatek* Boisduval, 1833
Diagnosic characters	T16C, A43T, G86A, T142C, A196G, T278A, 283C; keys to F.7.3, and F.7.4 in Evans (1953:187)	85A, 205A, 223A, 241A, 263(not C), T277A, A415T, 479T(not A), T574A; keys to G.1.2e in Evans (1953:215)	127T(not A), 163C, 349C, T424C; keys to J.39.5a in Evans (1955: 165), harpe broad, hook-shaped	68A(not T), C90T, T145C, 412T, 553A(not T or C), 583T(not A); keys to I.1.9 in Evans (1937:10)
Derivation of name	Feminine, a blend of *T*[osta] and [li]*iana*	Masculine, a blend of *Chi*[lean] and [Py]*rgus*	Feminine, echoes type species name	Masculine, a blend of [ramana]*Tek* and [Coe]*liades*

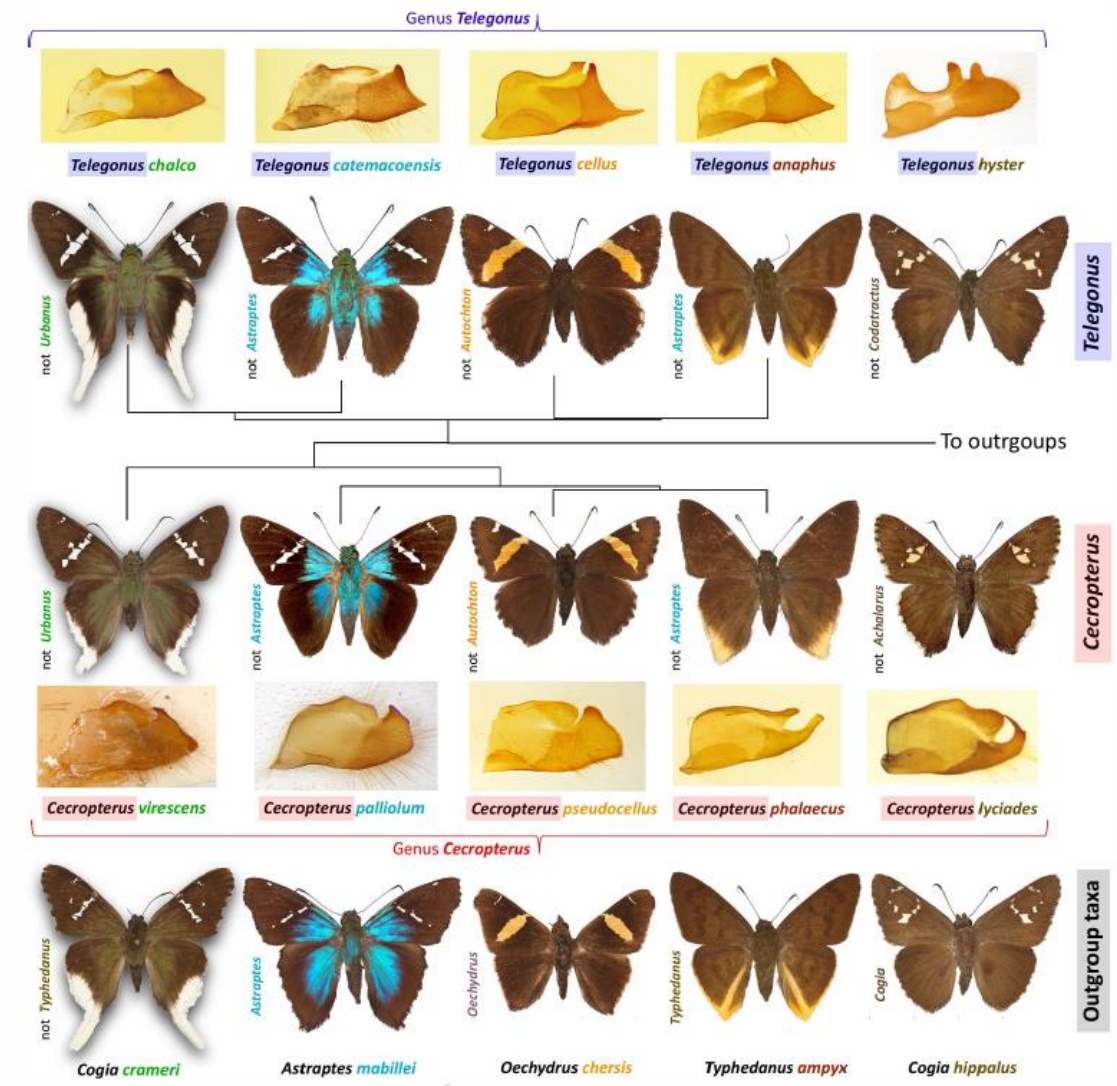


Fig. 2. Four convergent wing pattern groups between Eudaminae genera *Telegonus*, *Cecropterus* and outgroups. Genitalic valvae are shown for ingroups.

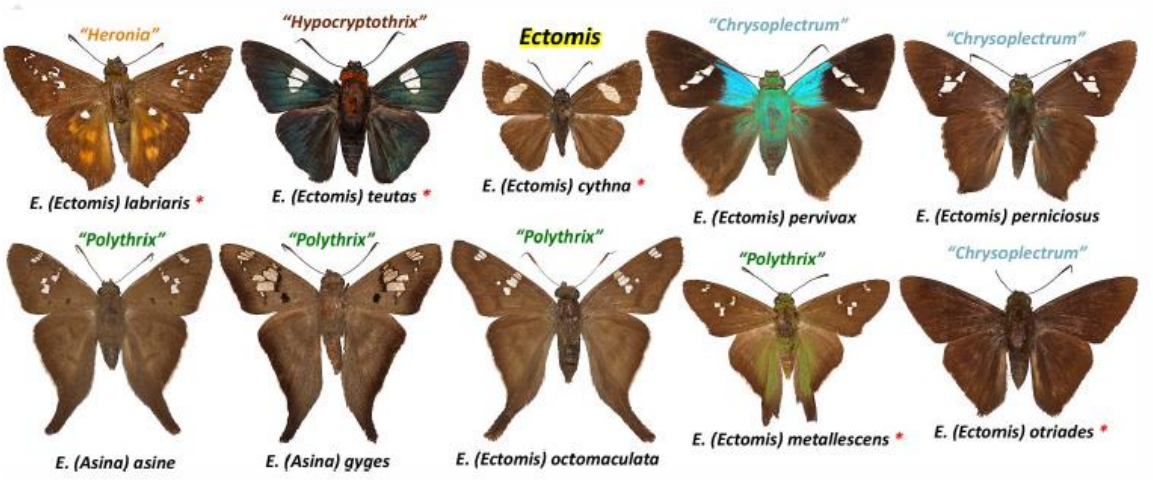


Fig. 3. Uncanny divergence within *Ectomis*. These skippers belonged to 5 genera listed above each image. Type species for these genera are marked with red asterisks.

REFERENCES

1. Jarvis ED, et al. (2014) Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346(6215):1320–1331.
2. Warren AD, Ogawa JR, Brower AVZ (2008) Phylogenetic relationships of subfamilies and circumscription of tribes in the family Hesperiidae (Lepidoptera: Hesperioidea). Cladistics 24(5):642–676.
3. Warren AD, Ogawa JR, Brower AVZ (2009) Revised classification of the family Hesperiidae (Lepidoptera: Hesperioidea) based on combined molecular and morphological data. Systematic Entomology 34(3):467–523.
4. Sahoo RK, et al. (2016) Ten genes and two topologies: an exploration of higher relationships in skipper butterflies (Hesperiidae). PeerJ 4:e2653.
5. Sahoo RK, Warren AD, Collins SC, Kodandaramaiah U (2017) Hostplant change and paleo-climatic events explain diversification shifts in skipper butterflies (Family: Hesperiidae). BMC Evolutionary Biology 17(1):174.
6. Evans WH (1937) A catalogue of the African Hesperiidae indicating the classification and nomenclature adopted in the British Museum. (British Museum (Natural History), London), pp. xii + 212, 30 pls.
7. Evans WH (1949) A Catalogue of the Hesperiidae from Europe, Asia, and Australia in the British Museum (Natural History). (British Museum (Natural History), London), pp. xix + 502, 53 pls.

8. Evans WH (1951) A catalogue of the American Hesperiidae indicating the classification and nomenclature adopted in the British Museum (Natural History). Part I. Introduction and Group A Pyrrhopyginae. (British Museum (Natural History), London), pp. x+92, pls. 1–9.
9. Evans WH (1952) A catalogue of the American Hesperiidae indicating the classification and nomenclature adopted in the British Museum (Natural History). Part II (Groups B, C, D) Pyrginae. Section I. (British Museum (Natural History), London), pp. v + 178, pls. 10–25.
10. Evans WH (1953) A catalogue of the American Hesperiidae indicating the classification and nomenclature adopted in the British Museum (Natural History). Part III (Groups E, F, G) Pyrginae. Section 2. (British Museum (Natural History), London), pp. v + 178, pls. 26–53.
11. Evans WH (1955) A catalogue of the American Hesperiidae indicating the classification and nomenclature adopted in the British Museum (Natural History). Part IV (Groups H to P) Hesperiinae and Megathyminae. (British Museum (Natural History), London), pp. v + 499, pls. 54–88.
12. Cong Q, Borek D, Otwinowski Z, Grishin NV (2015) Skipper genome sheds light on unique phenotypic traits and phylogeny. BMC Genomics 16(1):639.
13. Shen J, Cong Q, Borek D, Otwinowski Z, Grishin NV (2017) Complete Genome of *Achalarus lyciades*, The First Representative of the Eudaminae Subfamily of Skippers. Current genomics 18(4):366–374.

14. Mirarab S, et al. (2014) ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics 30(17):i541–i548.
15. Zhang J, et al. (2017) The complete mitogenome of *Euschemon rafflesia* (Lepidoptera: Hesperiidae). Mitochondrial DNA Part B 2(1):136–138.
16. Burns JM (1996) Genitalia and the proper genus: *Codatractus* gets *mysie* and *uvydixa* - in a compact *cyda* group - as well as a *Hyster*ectomy, while *Cephise* gets part of *Polythrix* (Hesperiidae: Pyrginae). Journal of the Lepidopterists' Society 50(3):173–216.
17. Talavera G, Lukhtanov VA, Pierce NE, Vila R (2012) Establishing criteria for higher-level classification using molecular data: the systematics of *Polyommatus* blue butterflies (Lepidoptera, Lycaenidae). Cladistics 29(2):166–192.
18. Williams RC, Bell EL (1934) Studies in the American Hesperioidea. Paper II (Lepidoptera). Transactions of the American Entomological Society 60:17–30.
19. Watt WB (1972) Xanthine dehydrogenase and pteridine metabolism in *Colias* butterflies. Jour-nal of Biological Chemistry 247(5):1445–1451.
20. Cong Q, Grishin NV (2016) The complete mitochondrial genome of *Lerema accius* and its phylogenetic implications. PeerJ 4:e1546.