J Struct Funct Genomics (2009) 10:255–268
DOI 10.1007/s10969-009-9071-1

Structural genomics target selection for the New York consortium on membrane protein structure

Marco Punta · James Love · Samuel Handelman · John F. Hunt · Lawrence Shapiro · Wayne A. Hendrickson · Burkhard Rost

Received: 22 April 2009 / Accepted: 30 September 2009 / Published online: 27 October 2009

Abstract The New York Consortium on Membrane Protein Structure (NYCOMPS), a part of the Protein Structure Initiative (PSI) in the USA, has as its mission to establish a high-throughput pipeline for determination of novel integral membrane protein structures. Here we describe our current target selection protocol, which applies structural genomics approaches informed by the collective experience of our team of investigators. We first extract all annotated proteins from our reagent genomes, i.e. the 96 fully sequenced prokaryotic genomes from which we clone DNA. We filter this initial pool of sequences and obtain a list of valid targets. NYCOMPS defines *valid targets* as those that, among other features, have at least two predicted transmembrane helices, no predicted long disordered regions and, except for community nominated targets, no significant sequence similarity in the predicted transmembrane region to any known protein structure. Proteins that feed our experimental pipeline are selected by defining a protein seed and searching the set of all valid targets for proteins that are likely to have a transmembrane region structurally similar to that of the seed. We require sequence similarity aligning at least half of the predicted transmembrane region of seed and target. Seeds are selected according to their feasibility and/or biological interest, and they include both centrally selected targets and community nominated targets. As of December 2008, over 6,000 targets have been selected and are currently being processed by the experimental pipeline. We discuss how our target list may impact structural coverage of the membrane protein space.

Keywords Membrane proteins · Target selection · Structural genomics · Structure determination

Abbreviations

αIMP	Alpha helical bundle integral membrane protein
DUF	Domain of unknown function
IMP	Integral membrane protein
NYCOMPS	New York consortium on membrane protein structure
PDB	Protein data bank
PSI	Protein structure initiative

Electronic supplementary material The online version of this article (doi:10.1007/s10969-009-9071-1) contains supplementary material, which is available to authorized users.

M. Punta (✉) · L. Shapiro · W. A. Hendrickson · B. Rost
Department of Biochemistry and Molecular Biophysics, Columbia University, 630 West 168th Street, New York, NY 10032, USA
e-mail: punta@rostlab.org

M. Punta · B. Rost
Columbia University Center for Computational Biology and Bioinformatics (C2B2), 1130 St. Nicholas Ave. Rm. 802, New York, NY 10032, USA

M. Punta · J. Love · J. F. Hunt · L. Shapiro · W. A. Hendrickson · B. Rost
New York Consortium on Membrane Protein Structure, New York Structural Biology Center, 89 Convent Avenue, New York, NY 10027, USA

S. Handelman · J. F. Hunt
Department of Biological Sciences, Columbia University, New York, NY 10032, USA

W. A. Hendrickson
Howard Hughes Medical Institute, Columbia University, New York, NY 10032, USA

M. Punta · B. Rost
Northeast Structural Genomics Consortium (NESG), 1130 St. Nicholas Ave. Rm. 802, New York, NY 10032, USA

RefSeq	NCBI reference sequence
SG	Structural genomics
TM	Transmembrane
TMH	Transmembrane helix
UPF	Uncharacterized protein family

Notations used

Reagent genomes	List of entirely sequenced organisms from which PSI clones its targets

Introduction

NYCOMPS as part of the PSI

The protein structure initiative (PSI) is the leading structural genomics (SG) initiative in the USA; it is funded by the National Institute of General Medical Sciences (NIGMS) at the National Institutes of Health (NIH). The PSI currently supports four large production centers and six specialized centers as well as other activities [1–3]. Two of these specialized centers focus on developing new technologies for membrane protein structure determination: the Center for Structures of Membrane Proteins (CSMP) [4] and the New York Consortium on Membrane Protein Structure (NYCOMPS). At NYCOMPS (http://www.nycomps.org/) we have established a high-throughput pipeline beginning with target selection and further including protein purification, protein expression and scale-up. Scaled-up proteins are sent to individual participating labs (within or outside of the consortium) for structure determination trials. While most of our resources are channeled into X-ray crystallography, we also pursue structure determination by NMR, solid-state NMR and cryo-electron microscopy. Here, we describe target selection, the first stage of the NYCOMPS pipeline.

Structure determination of integral membrane proteins is difficult

Integral membrane proteins (IMPs) are usually classified into two structural classes, according to the secondary structure conformation adopted by their membrane spanning segments [5, 6]: alpha helical bundle integral membrane proteins (αIMPs; estimated to constitute ~25% of an average proteome [7, 8]) and beta barrel IMPs (estimated, for example, to account for ~2–3% of proteins in Gram-negative bacteria [9–11]). These two classes of proteins differ also in their membrane localization: beta barrels are exclusive to the outer membrane of Gram-negative bacteria, atypical Gram-positive bacteria, mitochondria and chloroplasts, while alpha helical IMPs have been observed in all other membranes [12]. Recent structural data demonstrates the existence of at least one additional structural class of IMPs, the alpha barrel [13]. At NYCOMPS, we have so far focused only on αIMPs, because of their abundance in genomes and high biological and medical relevance [6, 14].

αIMPs are among the most difficult proteins for structure determination studies [15–17]. Consequently, they are extremely under-represented in the set of proteins for which we have high-resolution experimental structures. In particular, fewer than 1% of the proteins in the Protein Data Bank (PDB) [18] are IMPs (in March 2009 there are a total of 432 IMPs including both αIMP and beta barrels among the 56,217 PDB structures, see http://blanco.biomol.uci.edu/Membrane_Proteins_xtal.html for up-to-date IMPs statistics [19]) while estimates based on fully sequenced genomes predict that 25% of all annotated proteins are αIMPs [7, 8].

αIMPs present several challenges for successful experimental structure determination. First, high protein yields are often essential for structural studies; unfortunately, αIMPs are generally expressed at naturally low levels and their over-expression is often toxic to the cell [20]. Second, αIMPs are usually insoluble due to the long hydrophobic helices that are needed to span the lipid bilayer of the membrane core (typically around 17 residues long [21, 22]). Detergents are used to disrupt the membrane and prevent nonspecific aggregation. However, the choice of the detergent and the optimization of other buffer components, such as salt and glycerol, are challenging tasks [23, 24]. Finally, solution components that are useful for protein solubilization can interfere with crystallization, and crystallization success is also heavily dependent on the lipid content of the protein–detergent complexes [17]. In essence, each step in the experimental purification and structure determination of an αIMP is extremely demanding; typically, many parameters must be optimized to obtain a high-resolution membrane protein structure [15], usually working with small amounts of proteins.

SG adds large sampling of diversity as a new dimension

Structural genomics (SG) tries to increase the odds of experimentally obtaining high-resolution structures by using a pan-genomic approach that adds homology as a new dimension to the structure determination problem. This approach has been described in numerous publications [25–27]. Typically, all proteins within a given realm are clustered into pan-genomic sequence families, i.e. families with members found in different genomes, and then a set of those proteins are selected for which DNA templates are available from 'reagent' genomes. The set of proteins is tested experimentally for structure determination. An experimental structure for any one member of the family can serve through comparative modeling to inform studies

on any other family member [28]. This rationale is behind the tremendous success and impact of the PSI structural leverage [29]. Here, we take a slightly different approach based on the concept of seed sequences. In brief, given a target protein π^* (the "seed") our goal is to find a protein π, the structure of which is predicted to be similar to π^* and that surrenders a high-resolution structure using available experimental procedures (whereas the seed may fail). As in the conventional approach, the structure of π can provide a comparative structural model for π^*. The advantage of this approach over clustering of the full set of targets at the inception of the project is that we can create our families by expanding promising seeds whenever such seeds become available, instead of having to map these seeds to predefined families. This is likely to increase the 'centrality' of the seed with respect to the cluster.

Goals of target selection at NYCOMPS

The NYCOMPS target selection aims at providing the experimental pipeline with αIMPs that are: (1) novel with respect to what is already in the PDB, (2) diverse, with respect to their known sequence, structural and functional features, and (3) most likely to yield a structure. In order to increase αIMP feasibility we apply several computational filters that eliminate candidates less likely to succeed. These filters range from the exclusion of proteins known to constitute individual subunits of hetero-oligomeric complexes to the removal of proteins predicted to have long regions of disorder. Whatever passes those filters will enter the experimental structure determination pipeline at the New York Structural Biology Center. Targets pursued experimentally by NYCOMPS fall mostly into two broad categories: nominated and centrally selected targets. The way these targets are selected and their potential significance for structural coverage of the αIMP sequence and functional space is the subject of this contribution. We also briefly cover several special case targets that do not fall into either of the previous two categories. The Results section is organized into two parts: target selection (Task I) and target analysis (Task II).

Our target selection protocol has constantly been evolving since the start of NYCOMPS in fall 2005. The protocol has been modified to accommodate comments and suggestions coming from our team of investigators, as well as from the NYCOMPS scientific advisory committee, the NIH review panels and novel data that have appeared in the literature. In the first part of Results (target selection), we describe the protocol as of January 2009, although a fraction of the targets analyzed in the second part of Results (target analysis) were selected according to older criteria that did not include all the filtering steps described here.

Materials and methods

Creation of the NYCOMPS98 dataset

RefSeq target sequences

We downloaded 96 prokaryotic genomes in their amino acid sequence translation from the NCBI Reference Sequence collection (RefSeq [30]; ftp://ftp.ncbi.nih.gov/genomes; Table S1). 82 genomes are *Bacteria* (55 Gram- and 26 Gram+, including 3 Gram+ in the genus *Mycoplasma*, plus one genome that belongs to the phylum of *Cyanobacteria*), 14 are *Archaea*. Note that some of the NYCOMPS targets described here belong to 19 additional genomes that were retired in June 2008 because of: strain mismatch between the RefSeq strain and the one provided by the ATCC® (the source of our genomic DNA), early indications of poor cloning/expression performance from our experimental pipeline or both. Note also that our active list of genomes still comprises seven genomes for which we currently do not have an exact strain match with the ATCC provided genomic DNA or for which a match is uncertain. These genomes were retained based on early indications that, in these cases, strain mismatch was not causing major problems at the cloning level. Overall, the 96 NYCOMPS genomes encode 310,357 protein sequences.

Trasmembrane helix predictions

In order to identify αIMPs in the 96 chosen prokaryotic genomes, we run TMHMM2 [31] on all sequences. While TMHMM2 has been reported to be one of the best transmembrane helix (TMH) prediction programs in more than one independent assessment [32, 33], it is also one of the fastest, i.e. ideal for predicting TMHs on a large number of sequences. TMHMM2 returns the number and location of predicted TMHs in a sequence. In principle, all proteins with at least one predicted TMH are predicted by TMHMM2 to be αIMPs. To remain on the safe side, we considered as valid targets only proteins with two or more predicted TMHs.

Redundancy reduction

We run CD-HIT [34] with a 98% sequence identity threshold (we used parameters: -n 5 -c 0.98 -l 30 -d 30) on all proteins left from the previous step. This ensured that no two proteins in our dataset shared more than 98% sequence identity. When considering proteins sharing more than 98% sequence identity, the decision on which one to retain in our database depended on the genome the sequences belonged to. In particular, we prioritized: (1) sequences

from *Archaea* (following the guess of our team of experimentalists that they may provide more stable proteins) and from a list of best performing genomes ("best" in terms structure yield for globular soluble proteins) provided to us by the Northeast Structural Genomics Consortium; this list was later substituted by a list of genomes with best expression yield based on preliminary data from our experimental pipeline (data not shown); (2) the longest sequence (i.e. as selected by CD-HIT).

Signal peptide predictions

We run SignalP [35] on all sequences from *Bacteria* left in our list (note: no SignalP for sequences in *Archaea* is available [35]), and excluded all sequences predicted to have two TMHs but for which the first predicted TMH started before a predicted cleavage site. For the position of the cleavage site, we took the maximum out of the neural network and the HMM SignalP predictions.

Disorder predictions

We identified disordered residues in our sequences by running IUPred [36] using the option 'glob' that predicts structured domains in a protein. IUPred is one of the best performing programs for prediction of long disordered regions [37, 38] and it is also extremely fast. We discarded all proteins that had more than 15 consecutive residues predicted by IUPred not to be in a structured domain.

The sequences left after running this protocol constitute what we call the NYCOMPS98 dataset (39,037 sequences total).

Criterion for establishing evolutionary relationships between αIMPs

In order to find homologs of an αIMP query sequence, we used sequence similarity. We run three iterations of PSI-BLAST [39]: two profile generating iterations of the query against a large database composed of the sum of UniProtKB [40] and PDB [18], and one final iteration on the αIMP dataset of interest, e.g. TCDB [41] (parameters first two iterations: *-j* 3 *–v* 1,000 *–t* 1 *-h* 1*e*-10 *-e* 0.001 *-F F*; last iteration: *-e* 1 *–t* 1 *–v* 50,000 *-b* 50,000 *-F F*). Note that we input the 'effective length of database' of the first iterations into the last iteration (*-z* option) in order to have an estimate of the alignments' *E* values based on a large database. We first selected as 'homologs' of the query protein all sequences that aligned to it with *E* value $< 10^{-3}$ in the last iteration. Then, we additionally required all retained sequences to align to the query so that the alignment covered at least 50% of the residues predicted to be in a TMH in both query and subject sequence. All proteins that satisfied these constraints were considered part of the same structural family as the query sequence. Except when otherwise indicated, this is the criterion used to establish similarity between αIMPs throughout the paper.

Clustering of proteins in the 'von Heijne list'

We clustered *E. coli* proteins in the 'von Heijne list' (613 protein total) using a variation of the CLUP algorithm [25, 42]. For establishing similarity between proteins, we used the criterion described above. For clustering, we chose as the initial seed the shortest protein in the list and then seeds of increasing length until no sequences were left [25, 42]. Once the *E. coli* proteins were grouped into paralogous clusters, we further merged any two clusters that had at least one common member such that the whole region aligned to the seed of the first cluster was also aligned to the seed of the second, or vice versa.

Post-seed-expansion filtering of targets

Exclusion of protein sequences similar to those in the PDB

We run three iterations of PSI-BLAST [39] using the same databases and parameters used for establishing evolutionary relationships between αIMPs (see above). Only differences were the *E* value threshold we used and the transmembrane (TM) coverage we required. Indeed, we discarded all proteins that at any iteration aligned with *E* value < 1 to a PDB protein and for which the alignment covered at least 25% of the residues predicted to be in a TMH in the target protein. Note that in this case the fraction of a PDB protein TM region aligned to the query was not deemed relevant. Even in cases in which the alignment extended over 100% of the TM region of a PDB protein, the target was not discarded unless the alignment covered more than 25% of the target TM region.

Exclusion of individual subunits of hetero-oligomeric complexes

EcoCyc [43] is an annotated database for *E. coli* strain K-12 MG1655; Swiss-Prot [40, 44] is a general-purpose database including proteins from thousands of different species. To collect information about the possible role of a given protein as constituent subunit of a hetero-oligomeric complex, we queried EcoCyc manually for *E. coli* proteins and Swiss-Prot automatically for both *E. coli* and non *E. coli* proteins. From the Swiss-Prot searches we used information from the "FUNCTION", "INTERACTIONS", "SUBUNITS", "COFACTORS", "SUBCELLULAR LOCATION", "DISEASE", "DOMAIN" and "SIMILARITY" fields. This data was then manually inspected by our team of experimentalists,

who were asked to take a decision on whether or not to approve a given seed family.

Exclusion of seed family 'outliers'

In this manual step, we excluded proteins that were very different with respect to the seed in terms of length and number of TMHs. What 'very different' meant depended on the family under consideration but, in general, proteins that had a difference of more than two predicted TMHs or had long (>100 residues) insertions with respect to the seed were prime candidates for exclusion. Also, we generally excluded proteins whose N-terminus we suspected might have been wrongly annotated. To this aim, we built a multiple sequence alignment of the whole family using CLUSTALW [45] and manually inspected all members' N-terminal regions. If a consensus N-terminus could be identified, sequences that aligned with the consensus but displayed extra N-terminal residues were discarded.

Comparing NYCOMPS targets to Pfam-A

Pfam [46] is a large database of protein families. Here we used the Pfam-A collection of "high quality, manually curated families" (10,340 total, version 23.0). These families may or may not correspond to domains [46, 47]. We run the Pfam-provided program pfam_scan.pl on all selected target sequences (using the "-overlap" option to keep all hits within the same clan of families and default parameters). We considered all Pfam family hits below E value 10^{-3}, 10^{-10} or 10^{-50} and additionally required the Pfam families to cover a percentage of the target's TM region equal to or higher than 25, 50, 75 or 100%. If more than one Pfam family matched the target, we considered the overall coverage of the target TM region provided by all matching Pfam families. For E value $< 10^{-3}$ and $\geq$50% coverage, we repeated the same calculation this time requiring that it existed at least one individual Pfam family providing the full 50% coverage of the target TM region. This gave the 70% (of target proteins) figure that we provide in the "Results" section. We additionally calculated annotated-only Pfam family hits by excluding Domains of Unknown Function (DUFs) and Uncharacterized Protein Families (UPFs).

Comparing NYCOMPS targets to TCDB

Transport Classification Database (TCDB) [41] is a membrane transport protein database based on the Transporter Classification system, which is analogous to the enzyme commission (EC) system for enzyme classification. Note that TCDB includes both αIMPs and beta barrel integral membrane proteins. For our analysis, we used the version of the database from July 2008. TCDB is organized into 5 levels (each level identified by a number or a letter), representing the transporter class (7 classes in the version that we used), subclass (24), superfamily/family (557), subfamily (1,320) and substrate transported (3,224) for a total of 5,005 sequences. In order to estimate the fraction of TCDB classes, subclasses and superfamilies/families covered by our targets (first 3 levels of TCDB), we proceeded as follows. We run 1,000 bootstrapping [48] iterations. At each iteration, we picked at random exactly one target for each seed family (to mimic the situation in which we solve only one structure per seed family) and aligned all such targets against TCDB. At the end of each iteration, we calculated the number of different TCDB identifiers covered by our randomly picked targets (i.e. the number of TCDB identifiers corresponding to TCDB proteins that aligned to those targets). Finally, after 1,000 iterations, we calculated the average and standard deviation of the percentage of TCDB numbers covered with respect to the total (again, for the first three levels in the classification). We repeated this operation considering only 25, 50 and 75% of the seed families, selecting the families at random in each iteration, without re-sampling.

Comparing NYCOMPS targets to UniProtKB

UniProtKB [44, 49] is a protein repository composed of the manually annotated Swiss-Prot and of the automatically annotated TrEMBL databases. Here, we only considered UniProtKB proteins with at least 2 predicted TMH (UniProtKB-TMH). Since we wanted to calculate *novel* leverage [50] of the UniProtKB-TMH subset provided by NYCOMPS targets, we first had to calculate the leverage on the same subset provided by αIMPs currently in the PDB. In fact, UniProtKB-TMH proteins that show significant similarity to PDB proteins cannot be claimed as "novel leverage" by NYCOMPS targets. To find αIMPs in the PDB we considered all PDB sequences (February 2009) and run TMHMM2 [31] on the entire set. We discarded all sequences predicted to have less than 2 TMHs and reduced redundancy at 98% sequence identity using CD-HIT [34] (as done for the NYCOMPS targets). This left us with 187 proteins. We preferred this definition of αIMPs rather than using the annotations provided by PDB because it more closely reflected the way our target sequences were annotated as αIMPs (i.e. using TMH prediction). Finally, we aligned these sequences against UniProtKB-TMH and labeled all sequence similar UniProtKB-TMH proteins as 'not-novel'. Note that in this case criterion for similarity to PDB proteins is as in Fig. 2. When calculating the UniProtKB-TMH leverage of our targets, we excluded all proteins labeled as 'not novel' and performed bootstrapping [48] to calculate averages and standard deviations in the same way as described for TCDB. In the same way, we

calculate novel leverage for subsets Swiss-Prot-TMH and UniProtKB-TMH-Human. For comparison, we also calculated ratios between NYCOMPS target leverage and PDB protein leverage by taking average leverage values obtained from bootstrapping for NYCOMPS targets and leverage of the just described 187-protein subset for PDB proteins.

Results and discussion

Task I: target selection

NYCOMPS98: creating a valid target list of predicted alpha helical membrane proteins (Fig. 1a)

First, we chose from the RefSeq collection [30] 96 fully sequenced prokaryotic genomes (Table S1) for which genomic DNA was available from ATCC® (total of 310,357 proteins). This choice was heavily influenced by the experiences from the large-scale SG centers [1], in particular from NESG, the Northeast Structural Genomics consortium [51], and from NYSGXRC, the New York SGX Research Center for Structural Genomics [52]. We predicted TMHs for all genomic regions annotated as proteins. Overall, we found that the average number of predicted TMHs in all 96 genomes was 24%. This confirmed previous findings [7, 47]. Per genome percentages varied from 19% (*Methanocaldococcus jannaschii*) to 30% (*Clostridium perfringens*; Fig. S1). Among all membrane proteins that were predicted to integrate helices into the membrane (referred to as αIMPs), those with a single helix (single span) dominated (7.5% of all proteins) accounting for about one-third of all predicted αIMPs, followed by αIMPs with 2 and 6 TMHs (Fig. S2).

Since most of our targets had no experimental annotation linking them to the membrane, we had to rely on prediction methods. Such methods are estimated to be very accurate [32, 53], however, inevitably they will at times make TMH prediction mistakes. For target selection, we therefore only included proteins with ≥2 predicted TMHs. After this, we reduced redundancy by filtering out targets with exceedingly similar sequences guaranteeing that no two proteins in our dataset shared more than 98% pairwise sequence identity. Indeed, while we welcomed redundancy, we wanted to avoid cloning very similar proteins as very

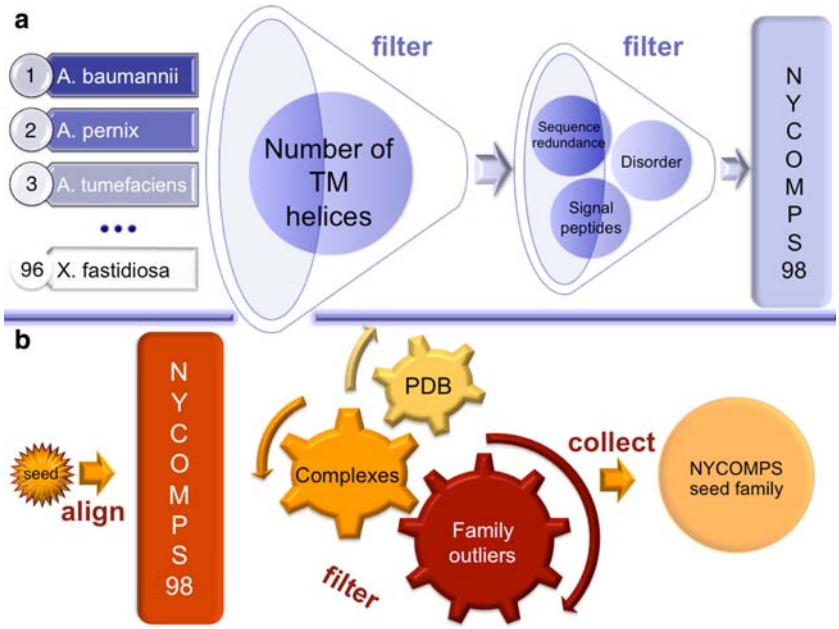

Fig. 1 Target selection protocol at NYCOMPS. **a** *Building the NYCOMPS98 dataset of valid targets.* We selected targets from 96 fully sequenced prokaryotic genomes. We used TMHMM2 [31] to predict TMHs in this set and retained only sequences with ≥2 TMHs. Finally, we applied a series of additional filters: we reduced redundancy at 98% using CD-HIT [34], we removed all sequences with 2 predicted TMHs for which the first TMH overlapped with a predicted signal peptide (using SignalP [35]) and we discarded sequences with at least 15 consecutive residues predicted to be disordered (using IUPred [36]). All sequences left constitute our set of valid targets, which we call NYCOMPS98. **b** *Expanding a protein seed into a family of related proteins within NYCOMPS98.* The seed is aligned against the whole NYCOMPS98 dataset using PSI-BLAST [39]. Retained sequences are those that satisfy our similarity criterion (Fig. 2). From this list we eliminate: sequences that are significantly similar to PDB proteins (filter is not applied to nominated targets), sequences known to constitute individual subunits of hetero-oligomeric complexes and sequences that differ significantly from the seed with respect to sequence length and number of predicted TMHs. We also discard proteins that align well with the family N-terminus consensus sequence (if any such consensus can be identified) but add some extra N-terminal residues, i.e. are possibly mis-annotated (Fig. S3). All remaining sequences are finally sent to cloning

close homologs have been shown to have similar crystallization propensities [54].

In order to further decrease the chance of introducing water-soluble non-αIMPs into our pipeline, we also excluded sequences with two predicted TMHs for which the position of the most N-terminal TMH overlapped with a predicted signal peptide (the most common mistake of TMH prediction programs is to predict an N-terminal TMH in place of a signal peptide [53]). Finally, we filtered out proteins that were predicted to have more than 15 consecutive disordered residues and hence might be problematic for crystallization [55]. The remaining sequences constitute what we refer to as the NYCOMPS98 dataset (39,037 αIMPs).

Cloning families (Fig. 1b)

The two principal steps in target selection are: (1) the identification of targets that constitute promising candidates for structure determination and/or are of utmost biological interest; we refer to these as to the "seeds". (2) The *expansion* of the seeds into families of—usually homologous—proteins likely to have membrane structures similar to the seed; we expand the seeds considering only sequences that are part of the NYCOMPS98 set of valid targets. NYCOMPS seeds are chosen from two distinct tracks that we refer to as "central selection" and "nomination". Centrally selected seeds (139 proteins) have so far been selected according to prior indications of successful over-expression in our *E. coli* host [56] (details below). In contrast, nominated seeds (35 proteins) have been handpicked by participating and adjoined laboratories; most of these seeds are well-studied proteins of known function. Note that seeds do not need to be proteins within NYCOMPS98 and not even within our collection of genomes. A protein is a valid seed as long as we can find homologs in the NYCOMPS98 dataset and as long as these homologs pass all the additional filters described below.

Central seed selection

In central seed selection our main concern was to pick membrane proteins that were more likely to readily provide high-resolution structures and that differed substantially in sequence within the TM region from proteins for which structures had already been determined experimentally. Given the small number of membrane protein structures available in the PDB, predictions as to which membrane proteins constitute structure-prone targets can at the moment rather be based on a misunderstanding of statistics than on sustainable science. We therefore scaled down our ambitions and started with proteins that were likely to give high yields of expression in our *E. coli* host (eventually increasing the odds to determine their structure). Such a list of proteins became available at the outset of the project thanks to a published genome-wide expression study performed on *E. coli* proteins [56]. Although the main goal of the work by von Heijne and coworkers was to determine the localization of the C-terminus of *E. coli* αIMPs (either inside the cell or in the periplasm), an important side effect was to provide us with a list of proteins that were successfully over-expressed in *E. coli*. In fact, all 139 NYCOMPS seeds that have so far been centrally selected have been extracted from a list of 613 proteins provided to us by Erik Granseth (Stockholm University, Sweden). NYCOMPS refers to this set of proteins as "the von Heijne list" (www.rostlab.org/punta/vonHeijnelist.txt). All proteins in this list were successfully over-expressed in *E. coli* in fusion with GFP or phosphatase A [56]. Our hope had therefore been that expanding these proteins into our 96 reagent genomes would provide a longer list of "good expressers", eventually increasing the odds for obtaining a structure for each seed family.

The initial *von Heijne list* was not unique at the sequence level; it included paralogs. Hence, we first clustered the proteins in the *von Heijne list* according to sequence similarity (the clustering procedure was adapted from methods described in references [25, 42], see "Methods" for more details). This resulted in 268 *E. coli* paralogy groups. The protein with the highest fluorescence level in each group was then selected as seed for the expansion into the NYCOMPS98 dataset (following prior indications that fluorescence levels correlated with expression levels [57]). Several of these initial families were subsequently excluded due to one or several of the following reasons: (1) they were similar to membrane proteins of known structure (i.e. in the PDB; note, however, that for a handful of centrally selected seeds now exists a structure of a homolog in the PDB that was deposited after the seed was selected), (2) they represented isolated subunits of hetero-oligomeric complexes, (3) they had less than 5 homologs in the entire UniProtKB [44, 49]; i.e. they provided low structural leverage). Some of the largest families (hundreds of homologs in NYCOMPS98) have also been held back, waiting for a data-driven criterion that could allow us to select only a fraction of homologs among all those available for the family; a criterion that would, for example, allow us to select proteins that we predict to express well under our experimental protocol. To date, 139 seeds from the *von Heijne list* have been selected for cloning.

Nominated seed selection

Nominated seeds (handpicked by participating groups) are special in many respects. For one, novelty with respect to PDB proteins is not enforced. Instead, we simply report the

observed similarities to the nominating group, which then takes the final decision on whether or not to pursue that specific target. Indeed, according to our criterion for novelty (E value < 1, alignment extending on at least 25% of the target TM region, see "Methods" for more details), 14 (or 40%) of our nominated seeds have significant sequence similarity to a least one PDB protein. There are various reasons why nominated targets are sometimes selected disregarding similarity to PDB proteins. Technology development projects often need to work on well-characterized test cases. One such example was the nomination of the KcsA channel [58] by the solid state NMR group. Also, there are cases in which sequence similarity as detected by our automatic protocol may not capture important structural and/or functional differences between the nominated seed and protein(s) already in the PDB. These differences may mean that the seed is a very valuable target, notwithstanding the presence of a homolog in the PDB. Again, this type of evaluation is left to the nominating group. Another procedural difference between nomination and central selection pertains to redundancy: nominated seeds do not need to be non-redundant. The same group can nominate seeds that are so sequence similar that they expand into the same family. Despite their similarity, they will be processed as separate seeds by our pipeline (with overlapping members being assigned to only one of the resulting seed families). If different groups nominate similar seeds though, we ask that they reach an agreement on how the resulting targets will be distributed.

Seed expansion

The seed expansion procedure is the same for centrally selected and nominated targets and it is based on reciprocal sequence similarity in the predicted TM region between seed and NYCOMPS98 proteins (Fig. 2). In particular, given an alignment between a seed and a NYCOMPS98 protein with PSI-BLAST [39] E value $< 10^{-3}$, we require that $\geq$50% of the residues predicted to be in TMHs in both proteins are found in the aligned region. The rationale of the TM region constraint on the alignment is to avoid association of a NYCOMSP98 protein to a seed based merely on the presence of a common water-soluble domain.

Filtering of targets

After a seed is expanded into a family of proteins predicted to have similar membrane cores, all family members are subjected to additional filters. Since these filters depend on information that may quickly change, they are best applied at the moment the targets have to be submitted to the experimental pipeline rather than when creating the NYCOMPS98 dataset.

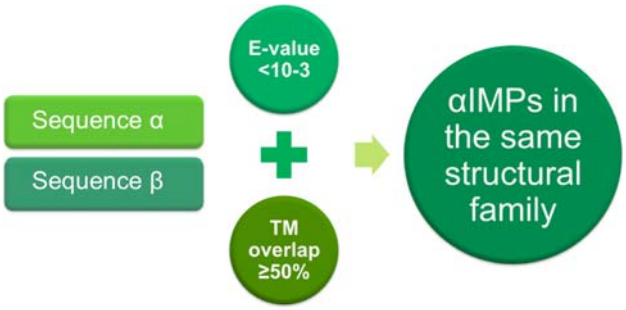

Fig. 2 αIMPs similarity criterion. We align sequence α to sequence β (both αIMPs) using PSI-BLAST [39]. If the alignment has E value $<10^{-3}$ and it extends over $\geq$50% of the predicted TM regions of both proteins, than we consider β similar to α. This criterion is used throughout the paper to establish similarity between αIMPs, e.g. similarity between a seed protein and proteins in the NYCOMPS98 dataset

Novelty: exclusion of PDB homologs

The first filtering step is meant to ensure that target proteins provide novel coverage of the protein universe [50]. Again, this does not apply to nominated targets. We filter out any centrally selected target with significant similarity in the predicted TM region to any protein in the PDB. This excludes all proteins aligning to a PDB protein with PSI-BLAST E values < 1 and for which the alignment extends over more than 25% of the predicted TM region ("Methods"). Note that while our E value cut-off ensures selection of targets whose TM domains are more novel than the average domain selected by other PSI high-throughput consortia [29], we do allow some overall similarity to PDB proteins to occur. Indeed, while we want to select novel targets, we do not want to exclude sequences simply based on the presence of a soluble domain that has a homolog in the PDB or on the fact that at most 25% of its TM region can be modeled based on a protein in the PDB. Finally, note also that our E value cutoff for avoiding similarity to PDB proteins is stricter (i.e. requires less similarity for exclusion of a protein) than the one we used, for example, for seed expansion (E value < 1 vs. $<10^{-3}$). This reflects our different goals in the two situations. Whereas in seed expansion we try to minimize the number of false positives (proteins not evolutionary related to the seed), in the comparison with PDB proteins we want to minimize the number of false negatives (in other words, we want to exclude as many targets related to PDB proteins as possible even at the cost of excluding a number of targets that are not in fact related to any PDB protein). We run this PDB filter both at the seed and at the single target level. Seeds with significant similarity to a PDB protein are not considered for further processing. When a seed passes this filter, each individual target in the family into which the

seed expands is still subjected to the same filter and discarded if it matches our criteria for exclusion.

Exclusion of isolated subunits of hetero-oligomeric complexes

Occasionally, protein subunits that natively are parts of larger hetero-oligomeric complexes are structurally stable even when expressed in isolation (e.g. homotetrameric A subunit cyclic nucleotide-gated ion channels [59]). However, in general, when expressed in isolation they are expected to be less likely to yield experimental structures. Therefore, our second filter removes all candidates for which we have evidence that they might constitute individual subunits of hetero-oligomeric complexes. Our identification of such subunits mostly relies on information extracted from EcoCyc [43] and Swiss-Prot [44]. Additionally, we seek input from our team of experimentalists. When we find evidence that one or more members of a family may constitute a subunit of a larger hetero-oligomeric complex, we usually discard the entire family. The final decision is taken after consulting with our team of experimental experts. An alternative way to confront such cases might be to clone and co-express them with all components of the complex. However, except for a few special cases (see below), our experimental pipeline is currently not set up to perform such operations in a high-throughput manner.

Removal of outliers

As a final step, we try to correct for "inconsistencies" in our families. We usually discard proteins for which the number of predicted TMHs differs greatly from that of the seed, as well as, proteins that differ significantly in their length with respect to the seed ("Methods"). Also, we try to exclude proteins that align well with the consensus N-terminus of the family (when any such consensus can be identified) but that feature additional N-terminal residues, because they constitute cases of proteins that may have been mis-annotated (Fig. S3).

Other NYCOMPS targets

Several NYCOMPS targets were selected according to other criteria. For instance, biological-theme targets are individual proteins handled by participating laboratories in the usual style of hypothesis-driven rather than hypothesis-generating structural biology. Such targets typically do not enter our pipeline at any stage. Another set of examples is constituted by 230 nominated histidine kinase targets that were hand-picked by one of our participating groups based on functional annotations. Some of those have <2 predicted TMHs (either 1 or 0). Additionally, we cloned 18 constructs that consist of co-cistronic (i.e. localized in neighboring regions of the genome) subunits of hetero-oligomeric complexes. This particular set of complexes can enter the existing experimental pipeline without modifications, because we can clone the full DNA stretch spanning the genes of all involved subunits.

Task II: target analysis

NYCOMPS targets diverse in terms of number of TMHs and length

In the following, we analyze 6,118 targets (from 174 seed families) that have been submitted to our experimental pipeline. 79% of these originated from 139 centrally selected seeds; the other 21% from 35 nominated seeds (complete list available at www.rostlab.org/punta/NYCOMPS-targets.txt). NYCOMPS targets are diverse in terms of their sequences as well as their predicted structural and functional features. First, the targets selected so far sample a wide range in terms of sequence length and of the number of TMHs (Fig. 3a, b, respectively). Proteins with 4 and 10 predicted TMHs are the most frequent among our selected targets. Proteins with four predicted TMHs include, for example, the centrally selected *E. coli* seeds HtpX, a family of heat shock proteins that may participate in degradation of misfolded proteins, and KdpD, the sensor member of a two-component signal transduction system responding to changes in potassium ion concentration. Proteins with 10 predicted TMHs include RhtA, a threonine/homoserine exporter, and WecH, an *O*-acetyltransferase (functional information for these seeds extracted from EcoCyc [43]). RhtA, in particular, constitutes our largest seed family comprising 552 proteins (largest by far, the second largest family has 200 members). This underlines the fact that the fraction of proteins with a given number of TMHs in our list of selected targets does not translate in general into the frequency of seeds with the same number of TMHs. In fact, a given seed might get expanded into targets with a (usually slightly) different number of predicted TMHs and different seeds might generate a very different number of targets. 64% of the selected targets are predicted to have the N-terminus inside, the other 36% have the N-terminus outside. In the next paragraphs, we analyze the sequence and functional diversity of NYCOMPS targets by mapping them to well-annotated databases such as Pfam [60], the TCDB [41] and UniProtKB [44, 49]. Note that excluding our largest families (such as RhtA) from the analysis reported hereafter would slightly change the numbers but lead to the same overall conclusions.

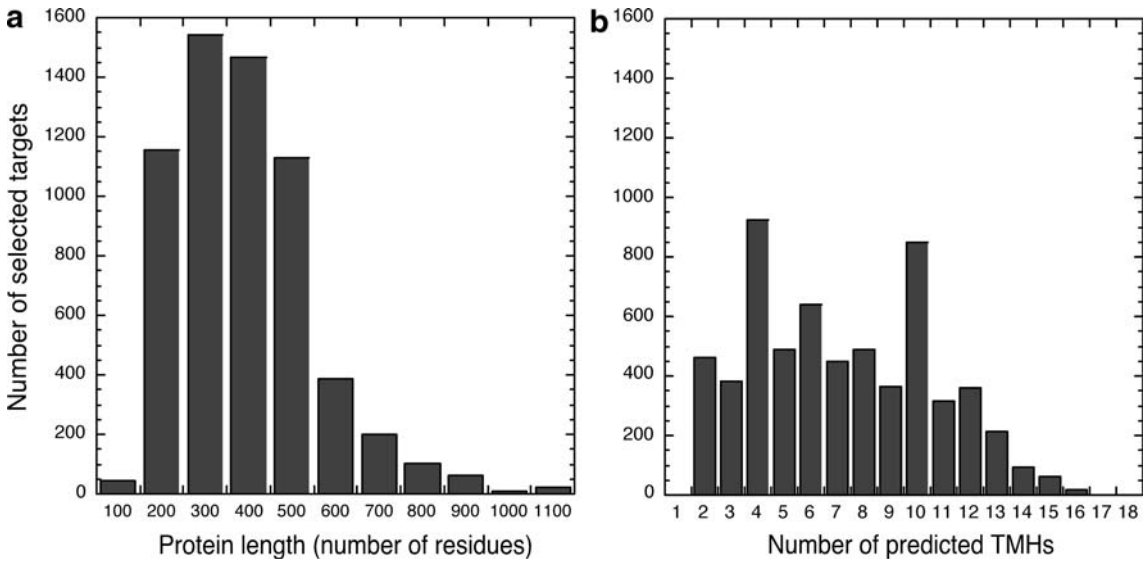

Fig. 3 Diversity of NYCOMPS targets. **a** Distribution of sequence lengths. *x*-axis tick labels represent ranges, e.g. 100 means between 0 and 100 residues. The last bin (1,100) includes all proteins longer than 1,000 residues. **b** Distribution of number of TMHs predicted by TMHMM2 in all selected targets

81% of the selected targets map to Pfam-A families, 63% to TCDB proteins

How "relevant" are NYCOMPS targets to the today's biology? We try to address this point by mapping our target list to two manually curated protein databases that carry functional annotations: the general purpose Pfam-A domain database [60] and the IMP-specific TCDB, a comprehensive database of membrane transport proteins [41]. Pfam-A families are manually curated from sequence-based alignments and they often do not span entire structural domains [46, 47]. This means that the TM region of a target may align to more than one Pfam-A family and one or more Pfam-A families may not cover it entirely. On the other hand, some of the TM regions in our targets do not represent single structural domains, e.g. the TM region of 2-hydroxycarboxylate transporters [61]. For these reasons, we evaluate similarity to Pfam-A families in two ways. First, we calculate the fraction of our targets that align (HMMER *E* value $< 10^{-3}$) to one or more Pfam-A families, collectively covering more than half of the target TM region. This fraction amounts to 81% of all targets (89% of nominated and 79% of centrally selected targets, respectively; see Table 1 for different *E* value thresholds and different fraction of TM region coverage). Second, we consider only targets for which it exists at least one Pfam-A family that by itself aligns over more than half of the predicted TM region. The percentage of targets satisfying this condition is equal to 70%. Overall, these latter targets map to 142 different Pfam-A families, offering additional evidence of NYCOMPS targets diversity at the sequence level. Finally, since not all Pfam-A families carry a functional annotation, we additionally calculate the percentage of matches after excluding DUFs and UPFs. In this case, the previous values become 61% (one or more families, Table S2) and 58% (single family).

Table 1 Percentage of NYCOMPS selected targets that map to at least one Pfam-A family

E value\ TM coverage	0.25	0.5	0.75	1.0
$<10^{-3}$	88%	81%	68%	25%
$<10^{-10}$	81%	75%	62%	24%
$<10^{-50}$	42%	41%	35%	16%

Note that if a target matches more than one Pfam-A family (with HMMER *E* value lower than the given threshold), TM region coverage is calculated as the sum of TM region coverage for the different families. Example: 81% of our targets match one or more Pfam-A families with *E* value(s) $<10^{-3}$ and alignment(s) covering at least half of the target TM region (second row "$<10^{-3}$", third column "0.5")

We next compared our list of selected targets to TCDB proteins. We found 63% of all targets to have significant similarity (defined as in Fig. 2, "Methods") to at least one transport protein in TCDB (nominated targets: 91%; centrally selected targets: 55%). While TCDB classification is organized into five levels, hereafter we considered only the first three, namely: *class*, *subclass* and *superfamily/family*. In particular, we investigated the percentage of TCDB classes, subclasses and superfamilies/families for which we could provide a structural model if we determined one structure for 25–100% of our seed families (Table 2).

If we solved one structure for each seed family, 12% of TCDB superfamilies/families distributed over 45% of all

Table 2 Target leverage: TCDB

Fraction of target families solved out of 174	TCDB Level 1 (7) (%)	TCDB Level 2 (24) (%)	TCDB Level 3 (557) (%)
1	86±0	45±2	12±0
0.75	82±6	40±4	9±1
0.50	77±8	35±5	7±1
0.25	66±11	25±5	3±0

We run PSI-BLAST against all proteins in TCDB using all our targets as queries. We then use bootstrapping to calculate the percentage of TCDB classification numbers we could provide a structural model for (leverage) if we solved one target per each seed family ("Methods"). This is done for the first three levels in the TCDB classification. Levels 1–3 correspond to class, subclass, superfamily/family. Example: if we solved a structure for each one of our seed families, we could provide a structural model for at least one protein in 45% of the 24 TCDB subclasses (second row "1", third column "TCDB Level 2 (24)")

TCDB subclasses could, on the average, be modeled (considering our criterion for similarity: at least 50% of their TM region could be modeled). If instead we solved one structure for every fourth seed family, we could provide models for 3% of TCDB superfamilies/families distributed over 25% of all TCDB subclasses. When considering these numbers one has to take into account that TCDB also contains beta barrel integral membrane proteins that we currently do not taken into consideration as targets for NYCOMPS and that not all membrane proteins are transporters (i.e. TCDB does not cover the entire universe of annotated αIMPs).

In conclusion, comparison to Pfam-A (61%) and TCDB (63%) shows that a little less than two-thirds of our targets can at least partially be mapped to functionally annotated proteins found in manually curated public databases. On the other hand, we also see that achieving a comprehensive structural coverage of IMP databases such as TCDB will require scaling-up considerably the number of selected seeds and targets.

UniProtKB novel leverage provided by NYCOMPS targets

One important aspect of any SG effort is novel leverage, i.e. the degree to which each experimental structure enables the generation of comparative models that were not available prior to its determination [50]. PSI has overall performed extremely well by this criterion [29]. In the context of NYCOMPS target selection we estimated the novel leverage that would result if we obtained structures for all, or a fraction of, our families. Note that here we apply a slightly more restrictive criterion for novel leverage with respect to its original definition [50], that is, not novel leverage with respect to the time targets were selected but novel leverage as of February 2009. We defined the leverage by aligning all our targets against UniProtKB but discarding aligned UniProtKB sequences predicted to have <2 TMHs (we call this subset *UniProtKB-TMH*).

If we solved one structure for each seed family, we could model at least 50% of the TM region of ~130,000 novel proteins in UniProtKB-TMH (Fig. 4a, blank circles and full line); if we determined one in every fourth family, we could still model on the average about 33,000 UniProtKB proteins (Fig. 4a); finally, should we solve a structure for only 10 of the targeted families, we could model on average close to 8,000 proteins (although for such a number of solved structures actual coverage would crucially depend on what these proteins are, see standard deviation Fig. 4a). Looking at leverage for each family, we see that about two-thirds of the seed families have over 200 novel UniProtKB-TMH homologs; only 18 families have novel leverage <50. UniProtKB is an ever-increasing database of annotated protein sequences (mostly open reading frames). As the number of sequenced proteins and organisms increases, UniProtKB-TMH leverage for a given set of targets will also increase. This means that the naked numbers we just reported may not be very meaningful. In other words, while it is true that we can model a large number of proteins if we solve the structure of at least some of the NYCOMPS targets, it is also true that this may simply reflect the sheer size of UniProtKB and the fact that, as more data become available, proteins cluster into increasingly large homologous families. For this reason it is probably more interesting to compare our projected UniProtKB-TMH leverage with the UniProtKB-TMH leverage obtained by using a non-redundant set of PDB proteins predicted to have ≥2 TMHs (187 proteins total, see "Methods"). This is a more direct measure of the impact NYCOMPS could have on our knowledge of the αIMPs structural universe. To identify PDB αIMPs we use predictions instead of annotations to be consistent with the way we picked our targets and with the way we identified αIMPs in UniProtKB. If we do this (Fig. 4b), we see that while solving one structure for each of our selected seeds would provide novel leverage equal to about 43% of what currently possible with available structures, solving only 10 structures would provide leverage equal to 2 to 3% of what is already possible.

Next, we investigated the leverage with respect to Swiss-Prot-TMH, i.e. the manually annotated subset of UniProtKB-TMH (Fig. 4a, filled diamonds and long-dashes). In this case, NYCOMPS targets would allow to model between 300 and 4,700 novel Swiss-Prot-TMH proteins, depending on the number of seed families for which we can solve at least one structure. The fraction of novel leverage with respect to what currently possible using PDB proteins follows a very similar trend with respect to what seen for the entire UniProtKB-TMH (Fig. 4b).

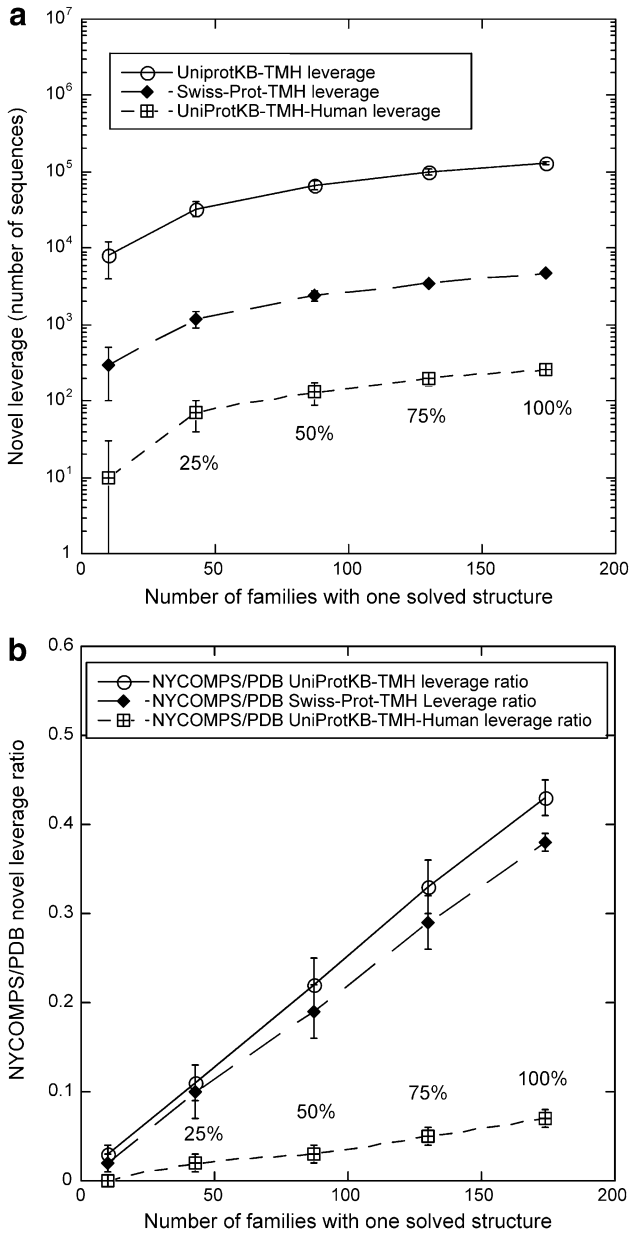

Fig. 4 Potential novel αIMP leverage provided by NYCOMPS targets. **a** The *x*-axis gives the number of seed families for which we hypothetically determine a structure (corresponding to 10 seed families or to 25–100% of all seed families; e.g. 25% corresponds to 43 seed families and 100% to 174 seed families); the *y*-axis reports the number of predicted αIMPs with more than 2 TMHs for which more than 50% of the residues in the TM region could be modeled using the NYCOMPS targets on the *x*-axis as templates (leverage). *Numbers* on the *y*-axis are for proteins in: UniProtKB-TMH (i.e. all predicted αIMPs in UniProtKB with more than 2 TMHs, see "Methods"; *blank circles* and *continuous line*), Swiss-Prot-TMH (*filled diamonds* and *long-dash line*) and UniProtKB-TMH-Human (i.e. human proteins in UniProtKB-TMH, *crossed squares* and *short-dash line*). *Error bars* are obtained by bootstrapping [48] ("Methods"). **b** Comparison between NYCOMPS target and PDB protein leverage. On the *y*-axis we report the ratio between the respective leverage values. Notations are as in (**a**). See "Methods" for the way UniProtKB-TMH leverage by PDB proteins is calculated

Finally, if we consider only human sequences within UniProtKB-TMH, we find that novel structural information could be obtained for 10–260 proteins (Fig. 4a, crossed squares and short-dash line). This time, the ratio to current leverage is markedly smaller (Fig. 4b). This is not very surprising given that all of our targets come from prokaryotic organisms.

Conclusions

New York Consortium On Membrane Protein Structure (NYCOMPS), targets alpha helical bundle integral membrane proteins, adopting a strategy that seeks to optimize success while maintaining the commitment to novelty, target relevance and leverage. In this paper, we have shown that the selected targets cover a wide range of protein lengths, TM topologies and functions. We have also demonstrated that the experimental determination of representative structures for these targets would allow transfer of structural information to a large number of known, but structurally uncharacterized, proteins. In the near future, we plan to expand the list of valid targets by introducing new genomes and by targeting eukaryotic proteins, non-cocistronic complexes and beta barrel integral membrane proteins.

Acknowledgments This work was supported by the grant U54-GM75026-01 to the NYCOMPS from PSI of the NIH. Thanks to all NYCOMPS collaborators who contribute to making NYCOMPS a wonderful experience, in particular thanks to Ann McDermott, Filippo Mancia, Ming Zhou, Francesca Gubellini (all Columbia), Da-Neng Wang (New York University), Mark Girvin (Albert Einstein), Guy Montelione (Rutgers), Renato Bruni and Brian Kloss (both NYCOMPS). Specific thanks to Jinfeng Liu (Genentech), Jessica Locke (Rutgers) and Rajesh Nair (Food and Drug Administration) for providing preliminary information and programs; to Ta-tsen Soong (Columbia), Henry Bigelow and Dariusz Przybylski (both Broad Institute), Kaz Wrzeszczynski (Cold Spring Harbor Laboratory), Zsuzsanna Dosztanyi (Hungarian Academy of Sciences, Budapest) and Zsolt Zolnai (University of Wisconsin—Madison) for useful discussions; to Guy Yachdav and Laszlo Kajan (both Columbia) for computer assistance and the collection of genome data sets. Last, not least, thanks to all those who deposit their experimental data in public databases, and to those who maintain these databases.

References

1. Burley SK, Joachimiak A, Montelione GT, Wilson IA (2008) Contributions to the NIH-NIGMS protein structure initiative from the PSI production centers. Structure 16:5–11

2. Norvell JC, Berg JM (2007) Update on the protein structure initiative. Structure 15:1519–1522
3. Norvell JC, Machalek AZ (2000) Structural genomics programs at the US national institute of general medical sciences. Nat Struct Biol 7 Suppl:931
4. Stroud RM, Choe S, Holton J, Kaback HR, Kwiatkowski W, Minor DL, Riek R, Sali A, Stahlberg H, Harries W (2009) 2007 Annual progress report synopsis of the center for structures of membrane proteins. J Struct Funct Genomics 10:193–208
5. Punta M, Forrest LR, Bigelow H, Kernytsky A, Liu J, Rost B (2007) Membrane protein prediction methods. Methods 41:460–474
6. von Heijne G (2007) The membrane protein universe: what's out there and why bother? J Intern Med 261:543–557
7. Knight CG, Kassen R, Hebestreit H, Rainey PB (2004) Global analysis of predicted proteomes: functional adaptation of physical properties. Proc Natl Acad Sci U S A 101:8390–8395
8. Liu J, Rost B (2001) Comparing function and structure between entire proteomes. Protein Sci 10:1970–1979
9. Bigelow H, Petrey D, Liu J, Przybylski D, Rost B (2004) Predicting transmembrane beta-barrels in proteomes. Nucleic Acids Res 32:2566–2577
10. Bigelow H, Rost B (2006) PROFtmb: a web server for predicting bacterial transmembrane beta barrel proteins. Nucleic Acids Res 34:W186–W188
11. Wimley WC (2003) The versatile beta-barrel membrane protein. Curr Opin Struct Biol 13:404–411
12. Bigelow H, Rost B (2009) Online tools for predicting integral membrane proteins. Methods Mol Biol 528:3–23
13. Dong C, Beis K, Nesper J, Brunkan-Lamontagne AL, Clarke BR, Whitfield C, Naismith JH (2006) Wza the translocon for *E. coli* capsular polysaccharides defines a new class of membrane protein. Nature 444:226–229
14. Overington JP, Al-Lazikani B, Hopkins AL (2006) How many drug targets are there? Nat Rev Drug Discov 5:993–996
15. Carpenter EP, Beis K, Cameron AD, Iwata S (2008) Overcoming the challenges of membrane protein crystallography. Curr Opin Struct Biol 18:581–586
16. Wang G (2008) NMR of membrane-associated peptides and proteins. Curr Protein Pept Sci 9:50–69
17. Wiener MC (2004) A pedestrian guide to membrane protein crystallization. Methods 34:364–372
18. Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, Iype L, Jain S, Fagan P, Marvin J, Padilla D, Ravichandran V, Schneider B, Thanki N, Weissig H, Westbrook JD, Zardecki C (2002) The protein data bank. Acta Crystallogr D Biol Crystallogr 58:899–907
19. White SH (2004) The progress of membrane protein structure determination. Protein Sci 13:1948–1949
20. Wagner S, Baars L, Ytterberg AJ, Klussmeier A, Wagner CS, Nord O, Nygren PA, van Wijk KJ, de Gier JW (2007) Consequences of membrane protein overexpression in *Escherichia coli*. Mol Cell Proteomics 6:1527–1550
21. Chen CP, Rost B (2002) Long membrane helices and short loops predicted less accurately. Protein Sci 11:2766–2773
22. von Heijne G (1994) Membrane proteins: from sequence to structure. Annu Rev Biophys Biomol Struct 23:167–192
23. Eshaghi S (2009) High-throughput expression and detergent screening of integral membrane proteins. Methods Mol Biol 498:265–271
24. Gutmann DA, Mizohata E, Newstead S, Ferrandon S, Postis V, Xia X, Henderson PJ, van Veen HW, Byrne B (2007) A high-throughput method for membrane protein solubility screening: the ultracentrifugation dispersity sedimentation assay. Protein Sci 16:1422–1428
25. Liu J, Hegyi H, Acton TB, Montelione GT, Rost B (2004) Automatic target selection for structural genomics on eukaryotes. Proteins 56:188–200
26. Montelione GT, Anderson S (1999) Structural genomics: keystone for a human proteome project. Nat Struct Biol 6:11–12
27. Rost B (1998) Marrying structure and genomics. Structure 6: 259–263
28. Forrest LR, Tang CL, Honig B (2006) On the accuracy of homology modeling and sequence alignment methods applied to membrane proteins. Biophys J 91:508–517
29. Nair R, Liu J, Soong TT, Acton TB, Everett JK, Kouranov A, Fiser A, Godzik A, Jaroszewski L, Orengo C, Montelione GT, Rost B (2009) Structural genomics is the largest contributor of novel structural leverage. J Struct Funct Genomics 10:181–191
30. Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 35:D61–D65
31. Krogh A, Larsson B, von Heijne G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305:567–580
32. Chen CP, Kernytsky A, Rost B (2002) Transmembrane helix predictions revisited. Protein Sci 11:2774–2791
33. Cuthbertson JM, Doyle DA, Sansom MS (2005) Transmembrane helix prediction: a comparative evaluation and analysis. Protein Eng Des Sel 18:295–308
34. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659
35. Emanuelsson O, Brunak S, von Heijne G, Nielsen H (2007) Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc 2:953–971
36. Dosztanyi Z, Csizmok V, Tompa P, Simon I (2005) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. J Mol Biol 347:827–839
37. Schlessinger A, Punta M, Rost B (2007) Natively unstructured regions in proteins identified from contact predictions. Bioinformatics 23:2376–2384
38. Schlessinger A, Punta M, Yachdav G, Kajan L, Rost B (2009) Improved disorder prediction by combination of orthogonal approaches. PLoS One 4:e4433
39. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402
40. UniProt (2009) The universal protein resource (UniProt) 2009. Nucleic Acids Res 37:D169–D174
41. Saier MH Jr, Tran CV, Barabote RD (2006) TCDB: the transporter classification database for membrane transport protein analyses and information. Nucleic Acids Res 34:D181–D186
42. Liu J, Rost B (2004) CHOP proteins into structural domain-like fragments. Proteins 55:678–688
43. Keseler IM, Bonavides-Martinez C, Collado-Vides J, Gama-Castro S, Gunsalus RP, Johnson DA, Krummenacker M, Nolan LM, Paley S, Paulsen IT, Peralta-Gil M, Santos-Zavaleta A, Shearer AG, Karp PD (2009) EcoCyc: a comprehensive view of *Escherichia coli* biology. Nucleic Acids Res 37:D464–D470
44. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N and Yeh LS (2005) The universal protein resource (UniProt), *Nucleic Acids Res*, 33 Database Issue, D154–D159
45. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680
46. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer EL, Bateman A (2008) The Pfam protein families database. Nucleic Acids Res 36:D281–D288

47. Liu J, Rost B (2003) Domains, motifs, and clusters in the protein universe. Curr Opin Chem Biol 7:5–11
48. Efron B, Tibshirani RJ (1993) An introduction to the bootstrap. Chapman & Hall/CRC, Boca Raton
49. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS (2004) UniProt: the universal protein knowledgebase. Nucleic Acids Res 32:D115–D119
50. Liu J, Montelione GT, Rost B (2007) Novel leverage of structural genomics. Nat Biotechnol 25:849–851
51. Acton TB, Gunsalus KC, Xiao R, Ma LC, Aramini J, Baran MC, Chiang YW, Climent T, Cooper B, Denissova NG, Douglas SM, Everett JK, Ho CK, Macapagal D, Rajan PK, Shastry R, Shih LY, Swapna GV, Wilson M, Wu M, Gerstein M, Inouye M, Hunt JF, Montelione GT (2005) Robotic cloning and protein production platform of the northeast structural genomics consortium. Methods Enzymol 394:210–243
52. Sauder MJ, Rutter ME, Bain K, Rooney I, Gheyi T, Atwell S, Thompson DA, Emtage S, Burley SK (2008) High throughput protein production and crystallization at NYSGXRC. Methods Mol Biol 426:561–575
53. Moller S, Croning MD, Apweiler R (2001) Evaluation of methods for the prediction of membrane spanning regions. Bioinformatics 17:646–653
54. Jaroszewski L, Slabinski L, Wooley J, Deacon AM, Lesley SA, Wilson IA, Godzik A (2008) Genome pool strategy for structural coverage of protein families. Structure 16:1659–1667
55. Esnouf RM, Hamer R, Sussman JL, Silman I, Trudgian D, Yang ZR, Prilusky J (2006) Honing the in silico toolkit for detecting protein disorder. Acta Crystallogr D Biol Crystallogr 62:1260–1266
56. Daley DO, Rapp M, Granseth E, Melen K, Drew D, von Heijne G (2005) Global topology analysis of the *Escherichia coli* inner membrane proteome. Science 308:1321–1323
57. Drew D, Slotboom DJ, Friso G, Reda T, Genevaux P, Rapp M, Meindl-Beinker NM, Lambert W, Lerch M, Daley DO, Van Wijk KJ, Hirst J, Kunji E, De Gier JW (2005) A scalable, GFP-based pipeline for membrane protein overexpression screening and purification. Protein Sci 14:2011–2017
58. Doyle DA, Morais Cabral J, Pfuetzner RA, Kuo A, Gulbis JM, Cohen SL, Chait BT, MacKinnon R (1998) The structure of the potassium channel: molecular basis of K+ conduction and selectivity. Science 280:69–77
59. Biel M (2009) Cyclic nucleotide-regulated cation channels. J Biol Chem 284:9017–9021
60. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR (2004) The Pfam protein families database. Nucleic Acids Res 32:D138–D141
61. Lolkema JS (2006) Domain structure and pore loops in the 2-hydroxycarboxylate transporter family. J Mol Microbiol Biotechnol 11:318–325