Biochemical and Biophysical Research Communications 445 (2014) 724–733

Contents lists available at ScienceDirect

Biochemical and Biophysical Research Communications

journal homepage: www.elsevier.com/locate/ybbrc

Drug screening strategy for human membrane proteins: From NMR protein backbone structure to *in silica*- and NMR-screened hits

Steffen Lindert [a,1], Innokentiy Maslennikov [b,c,1], Ellis J.C. Chiu [b], Levi C. Pierce [a,2], J. Andrew McCammon [a,d,e], Senyon Choe [b,c,f,*]

[a] Department of Chemistry and Biochemistry, Department of Pharmacology, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA
[b] Structural Biology Laboratory, The Salk Institute for Biological Studies, 10010 North Torrey Pines Rd., La Jolla, CA 92037, USA
[c] Joint Center for Biosciences, 301-B, Songdo Smart Valley 214, Songdo-dong, Yeonsu-ku, Incheon 406-840, Republic of Korea
[d] Howard Hughes Medical Institute, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA
[e] NSF Center for Theoretical Biological Physics, National Biomedical Computation Resource, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA
[f] Drug Discovery Collaboratory, Carlsbad, CA 92008, USA

ARTICLE INFO

Article history:
Available online 10 February 2014

Keywords:
Human membrane proteins
NMR screening
Molecular dynamics
Computational docking

ABSTRACT

About 8000 genes encode membrane proteins in the human genome. The information about their druggability will be very useful to facilitate drug discovery and development. The main problem, however, consists of limited structural and functional information about these proteins because they are difficult to produce biochemically and to study. In this paper we describe the strategy that combines Cell-free protein expression, NMR spectroscopy, and molecular DYnamics simulation (CNDY) techniques. Results of a pilot CNDY experiment provide us with a guiding light towards expedited identification of the hit compounds against a new uncharacterized membrane protein as a potentially druggable target. These hits can then be further characterized and optimized to develop the initial lead compound quicker. We illustrate such "*omics*" approach for drug discovery with the CNDY strategy applied to two example proteins: hypoxia-induced genes HIGD1A and HIGD1B.

1. Introduction

Membrane proteins account for approximately 50% of most important targets for pharmaceutical industry [1] because they constitute a key element in cell communication with the environment. With the support of novel techniques of membrane protein production and approaches for advanced labeling, structure determination methods, predominantly X-ray crystallography and NMR spectroscopy, show impressive progress with hundreds of membrane protein structures determined during the last decade [2–6]. However, membrane protein structures still represent less than 1% of all known unique protein structures (http://www.pdb.org). Integral membrane proteins are very difficult to investigate both structurally and functionally in order to understand the signaling mechanisms or for structure-based drug design. Human integral membrane proteins (hIMPs) are the most arduous targets among other membrane proteins: less than 50 structures of unique hIMPs are known (http://www.pdb.org). Therefore, every known spatial structure of hIMP provides valuable experimental information for structure-based drug design.

Over the course of the past few decades, computational methods have become a viable tool in drug discovery efforts. Computational approaches contribute to the search for drugs at different stages of drug discovery, such as target identification, validation of potential hits, and lead design. They can also be used at the preclinical trial stage [7]. Arguably the most profound impact by these computational approaches is made through molecular docking techniques. The most efficient use of the techniques would require knowledge of the receptor protein structure. A wide array of different methods exists which generally account of ligand flexibility [7]. Recent years have also seen the advent of docking methods that account for receptor flexibility [8,9]. The method used in this work – the relaxed

Abbreviations: hIMPs, human integral membrane proteins; CF, cell-free; HI, hypoxia-induced; MD, molecular dynamics; DMPC, 1,2-dimyristoyl-*sn*-glycero-3-phosphocholine; LMPG, 1-myristoyl-2-hydroxy-sn-glycero-3-[phospho-rac-(1-glycerol).

* Corresponding author at: Structural Biology Laboratory, The Salk Institute for Biological Studies, 10010 North Torrey Pines Rd., La Jolla, CA 92037, USA.
E-mail address: choe@salk.edu (S. Choe).

[1] These authors contributed equally to this work.

[2] Present address: Schrödinger, LLC, 120 West 45th Street 7th Floor, Tower 45, New York, NY 10036, USA.

http://dx.doi.org/10.1016/j.bbrc.2014.01.179
0006-291X/

complex scheme [10,11] – utilizes molecular dynamics simulations in combination with docking algorithms to account for receptor flexibility.

Experimental verification of the potential hit compounds is a crucial step in a process of drug design. Modern solution NMR spectroscopy is well suited for monitoring protein–ligand interactions and allows for the rapid screening of compound libraries [12,13]. Protein-based NMR screening provides valuable information about interaction with a ligand by detection of chemical shift perturbation in heteronuclear ^{1}H–^{15}N-HSQC or ^{1}H–^{15}N-TROSY-HSQC experiments [13–15]. While interacting ligands could be identified simply by detection of changes in protein chemical shifts even without knowledge of assignment, in order to locate ligand-binding pocket, the protein-based screening by NMR requires resonance assignment for the target protein. The ability to differentiate binding sites for different ligands or with different affinities is also a great advantage of protein-based NMR screening.

Using an advanced strategy of combining cell-free (CF) membrane protein expression and fast NMR structural analysis we have recently demonstrated a high-speed determination of backbone structures of six small human integral membrane proteins (hIMPs) [16]. Two of these proteins, HIGD1A and HIGD1B, belong to the class of hypoxia-induced (HI) genes [17]. Presumably they are subunits of Cytochrome C oxidase and HIGD1A is important for assembly of the respiratory supercomplexes [18], but actual function of these proteins in human cells, as well as in other mammalian cells, is yet unknown.

Despite the vast genomics data available for hIMPs, we lack information what hIMPs are potentially druggable. Here we extended our strategy by including the structure-based search of protein ligands. We combined Cell-free protein synthesis, NMR spectroscopy, and molecular DYnamics simulation methods into a strategy, we named CNDY, aimed for functional analysis and drug design. To illustrate the potential of such *omics* approach, we used the CNDY strategy for ligand search for HIGD1A and HIGD1B. Starting from the backbone NMR structures of the HI proteins, we performed molecular dynamics (MD) simulations of the structures embedded in a lipid bilayer, computational screening of compounds from the Open Chemical Repository at the National Cancer Institute for binding to the representative structures derived from the MD trajectories, and protein-based NMR screening of the hit compounds with the CF-expressed HI proteins. The results provide us a starting point to identify the hit compounds, which can then be optimized to discover the initial lead molecules.

2. Materials and methods

2.1. Cell-free expression and structure determination by NMR spectroscopy

CF expression of HIGD1A and HIGD1B and high-speed determination of their structures by NMR spectroscopy have been described by Klammt et al. [16].

2.2. System preparation for MD simulation

The backbone NMR structures were used to prepare the membrane systems for simulation of HIGD1A and HIGD1B (PDB codes 2LOM, 2LON). The CHARMM-GUI membrane builder [19,20] was used for the system setup. PDB files were loaded as starting point. The protein segment was specified as all 93 and 99 residues present in the PDB files of HIGD1A and HIGD1B, respectively. No terminal patching was carried out. Protonation and phosphorylation states were built based on standard assumptions. The proteins were oriented by translating it along the z axis by -3 Å with respect to the original PDB coordinates. A rectangular simulation box of 75 Å was chosen. A 15 Å water layer in the z direction was added to both sides of the membrane. The extension of the simulation box in x–y direction was determined based on the ratios of lipid components. A homogenous DMPC lipid bilayer was chosen, resulting in 84 lipid molecules in the upper leaflet and 80 lipid molecules in the lower leaflet for HIGD1A and 86 lipid molecules in the upper leaflet and 78 lipid molecules in the lower leaflet for HIGD1B. The center of the system was placed at $z = 0$. The total system sizes were 93.0 Å in x direction, 98.0 Å in y direction, and 83.2 Å in z direction (HIGD1A) and 95.5 Å in x direction, 94.0 Å in y direction, and 86.5 Å in z direction (HIGD1B). 20 Na^+ and 26 Cl^- ions were added to neutralize the HIGD1A system and 21 Na^+ and 27 Cl^- ions were added to neutralize the HIGD1B system and to obtain a 150 mM ionic strength. TIP3P waters were added.

The fully solvated system for HIGD1A contained 48,423 atoms (including 164 lipid molecules and 9,194 water molecules) and the system for HIGD1B contained 50,086 atoms (including 164 lipid molecules and 9,707 water molecules). The CHARMM27 force field [21] was used for all the simulations. Minimization and equilibration using NAMD 2.9 [22] was performed in six stages. The first and second stages simulated 25 ps in the NVT ensemble with a 1 fs timestep. During the first stage, harmonic force restraints were applied to all system components, i.e. protein (positional restraints), waters (restraint to prevent water from entering the hydrophobic membrane region), lipids (restraints to keep structural integrity of membrane) and ions (positional restraints). In the second stage of equilibration the restraints on the ions were removed and the restraints on the protein backbone and side chains were cut in half. The remaining four stages all simulated in the NPAT ensemble. Stage 3 simulated for 25 ps with a 1 fs timestep, while stages 4–6 simulated for 100 ps with a 2 fs time step. Restraints on all system components are gradually decreased within stages 3–6. Only a 0.1 kcal/mol/Å^2 restraint on the protein backbone remained in stage 6. For a more detailed description of the equilibration protocol see [19].

2.3. Molecular dynamics simulations and trajectory clustering

All simulations were performed under the NPT ensemble at 300 K using NAMD 2.9 [22] and the CHARMM27 force field [21]. Periodic boundary conditions were used along with a non-bonded interaction cutoff of 12 Å. Bonds involving hydrogen atoms were constrained using the SHAKE algorithm [23], allowing for a time step of 2 fs. Structures were saved every 2 ps. All systems were simulated for 100 ns.

Visual inspection of the 100 ns trajectories allowed identification of potential binding sites (pockets) for HIGD1A and HIGD1B. Owing to the different screening strategies applied to HIGD1A and HIGD1B, the molecular dynamics trajectories were processed differently. For HIGD1B 86 equally spaced frames (every ~1.16 ns) were extracted from the 100 ns MD trajectory. For HIGD1A, frames every 10 ps were extracted from the MD trajectory. The extracted frames were subsequently aligned in two separate sets using the C_α atoms in two potential binding site of HIGD1A. Structures in the aligned frame sets were clustered by RMSD using the GROMOS++ conformational clustering [24]. A RMSD cutoffs of 1.9 Å and 2.1 Å was chosen for two HIGD1A potential binding sites. The chosen cutoffs resulted in 6 clusters for each potential binding site that represented at least 90% of the respective trajectories. The central members of each of these clusters were chosen to represent the pocket conformations within the cluster and thereby the conformations sampled most prominently by that pocket over the course of the simulation.

2.4. Virtual screens of hIMPs

Virtual screens were carried out for both HIGD1A and HIGD1B. Different strategies were applied to the two systems in order to test their performance and reliability in identifying possible binders.

2.4.1. HIGD1A

The entire National Cancer Institute (NCI) compound database was used for the virtual screen. The original unprepared dataset contained 265,241 compounds. Ligands were prepared using LigPrep [25], adding missing hydrogen atoms, generating all possible ionization states, as well as tautomers. The final prepared dataset used for virtual screening contained 798,555 compounds. Docking simulations were performed with Glide [26–28]. Due to the high number of test compounds, the Virtual Screening Workflow was utilized. A pre-filter routine decreased the time spent on non-desirable compounds. QikProp [29] was run to efficiently evaluate pharmaceutically relevant properties of the test compounds. A Lipinski's Rule [30] pre-filter was applied. The OPLS2005 force field [31] was used for docking, as well as Epik state penalties [32]. A hierarchical docking scheme was used: 100% of the filtered compounds were docked with Glide HTVS. The 10% best scoring states after HTVS transitioned into Glide SP. Again, the 10% best scoring states after SP were used for docking with Glide XP. All dockings were performed flexibly and a post-docking minimization was applied. The docked XP poses and scores were finally evaluated for compound selection. This docking procedure was performed independently for 12 different HIGD1A structures extracted from the MD trajectory by RMSD clustering (see above) – six cluster centers for each potential binding site.

2.4.2. HIGD1B

The virtual screen was performed using the NCI diversity set III, a subset of the full NCI compound database. Again, ligands were prepared using LigPrep, adding missing hydrogen atoms, generating all possible ionization states, as well as tautomers. The final prepared dataset used for virtual screening contained 1,013 compounds. No pre-filtering was applied for the diversity set. Docking simulations were performed with AutoDock Vina [33]. The entire diversity set was docked into all 86 individual trajectory frames.

2.5. Screening by NMR

2.5.1. NMR screening approach

The top 80 available hit compounds for each protein with the maximal docking scores were obtained from NCI database depository (The Open Chemical Repository Collection). The compounds were ordered in a list according to the computed binding score. Each hit compound was assigned an unique within the list seven-digits code. An X^{th} digit in the code defines the presence (1) or absence (0) of the compound in a sample X. The compounds with lower predicted docking score were assigned with the codes, corresponding to infrequent presence of these compounds in the samples. Among the 80 compounds in each list, 7 were present in one sample, 21 in two samples, 35 in three samples, and 17 in four samples. In turn, samples 1–7 contained 33, 32, 31, 31, 31, 32, and 32 compounds, accordingly.

2.5.2. Preparation of the proteins

Both proteins, HIGD1A and HIGD1B, were prepared uniformly ^{15}N-labeled using CF system, as described in [16]. The proteins were expressed as a precipitant [34,35] in 6 ml CF reaction, which was enough for 8 NMR samples with a protein concentration of approximately 150–250 μM. The expressed proteins were solubilized in 2% (HIGD1B) or 3% (HIGD1A) 1-myristoyl-2-hydroxy-*sn*-glycero-3-[phospho-rac-(1-glycerol)] (LMPG) buffered with 20 mM MES–Bis–Tris, pH 6.0. The solubilized proteins were divided into 8 equivalent samples and the sample volumes were adjusted to 330 μl with 20 mM MES–Bis–Tris, pH 6.0, 2% (HIGD1B) or 3% (HIGD1A) LMPG buffer, contained also D_2O (the final D_2O concentration in the samples were 5%). One sample for each protein was used as a control without the compounds, and seven samples for each protein were used for solubilization of the different compound mixtures.

2.5.3. Preparation of the ligands

All hit compounds were suspended in 4:4:1 chloroform:methanol:water mixture at a concentration close to 10 mg/ml. Aliquots (5 μl) of the compound solutions or suspensions were combined in 7 different mixtures for each list, according to the protocol described above and air-dried. The NMR samples of the detergent-solubilized protein were added to the vials, which contained dried compound mixtures, vortexed, and left overnight at RT. This procedure results in almost 100% solubilization of all compounds in the NMR samples. The samples were flush-frozen in a liquid N_2 and stored at −20 °C. For NMR experiments the samples were thawed at 37 °C, transferred into Shigemi NMR tube, and degased under low vacuum in a sonication bath for 3 min.

2.5.4. NMR experiments and analysis

The ^{1}H–^{15}N-TROSY-HSQC spectra (256 t_2 increments, 16 scans) were measured for all samples at 310 K on a 700 MHz Bruker NMR spectrometer equipped with a cryogenic probe. The spectra were transformed using Bruker NMR software, Topspin. The transformed spectra were analyzed in CARA [36]. The known assignment of HIGD1A and HIGD1B [16] was used in the analysis of chemical shifts changes in ligand-containing samples. Weighted-average chemical shift differences (WCS) were calculated for each residue as described [37]. Molecular graphics and analysis were performed using UCSF Chimera program [38].

3. Results

3.1. MD simulation of backbone NMR structures in explicit bilayer

One of the important contributions that MD can make to the structural elucidation of membrane proteins is a simulation in a more natural environment than that which the experiments were performed in. NMR measurements were done in micelles, which shape and surface properties differ from the lipid bilayer membrane environment of HIGD1A and HIGD1B. Experimentally determined structures for HIGD1A and HIGD1B were embedded in a homogenous DMPC lipid bilayer, minimized and equilibrated. Subsequently, 100 ns MD simulations were performed on both proteins in the lipid bilayer.

Structural stability of the proteins with respect to the equilibrated starting structure was monitored. Fig. 1 shows the root mean square deviation (RMSD) from the starting structure as a function of simulation time as well as the root mean square fluctuations (RMSF) as a function of residue numbers. RMSDs were calculated over heavy backbone atoms in the regions of interest. The RMSD plot, which is a measure of overall protein stability, shows a fast increase of the RMSD within the first 10 ns of simulation time. Then the protein remains more or less stable (at around 4. . .5 Å RMSD) for the rest of the simulation. All proteins, including soluble proteins, exhibit an initial increase in RMSD upon start of the MD simulation. This increase is generally associated with full equilibration of the structure in the molecular mechanics force field. For a protein of less than 100 amino acids, one would expect an increase between 2 and 3 Å associated with equilibration. The higher increase in the case of HIGD1A sug-

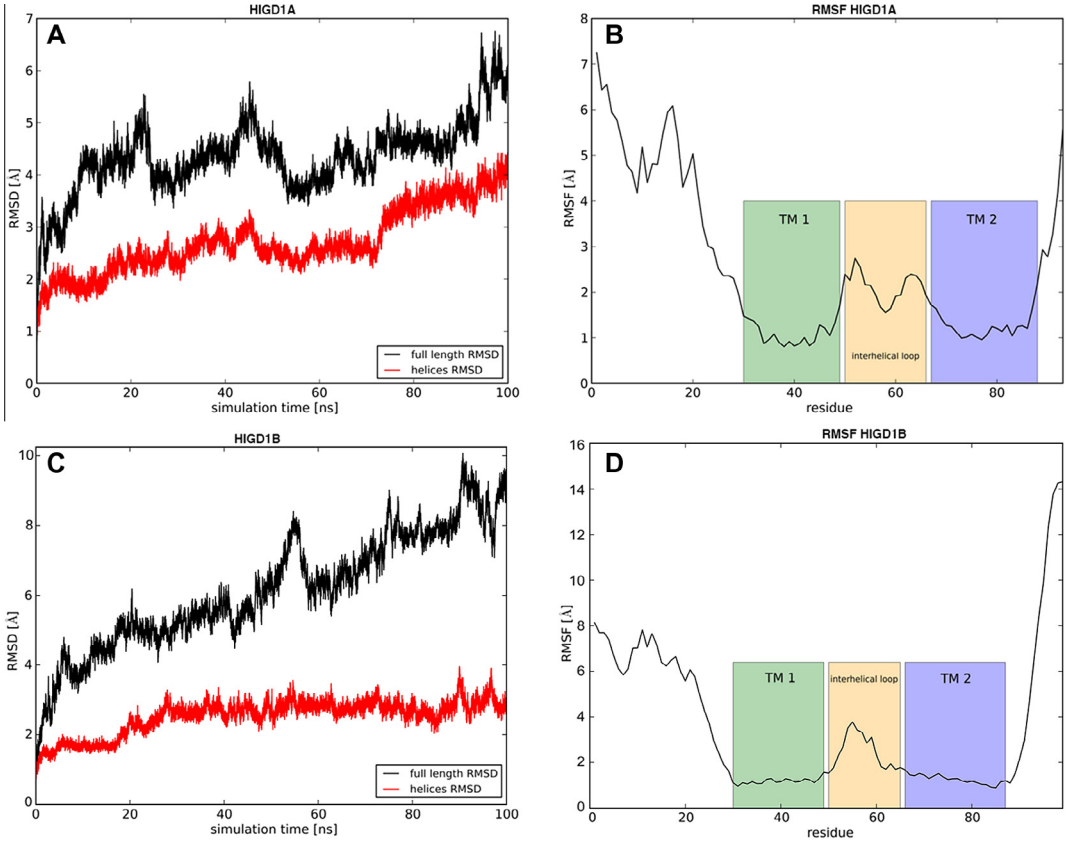

Fig 1. Conformational flexibility of the HIGD1A (A and B) and HIGD1B (C and D) structures during 100 ns MD simulation. (A and C) Root mean square deviation (RMSD) of all backbone heavy atoms (black) and backbone heavy atoms of TM helices (residues 30–49 and 67–88 for both proteins, red) with respect to the starting structure as a function of simulation time. (B and D) Root mean square fluctuations (RMSF) as a function of residue numbers. The two TM regions and the interhelical loop are labeled.

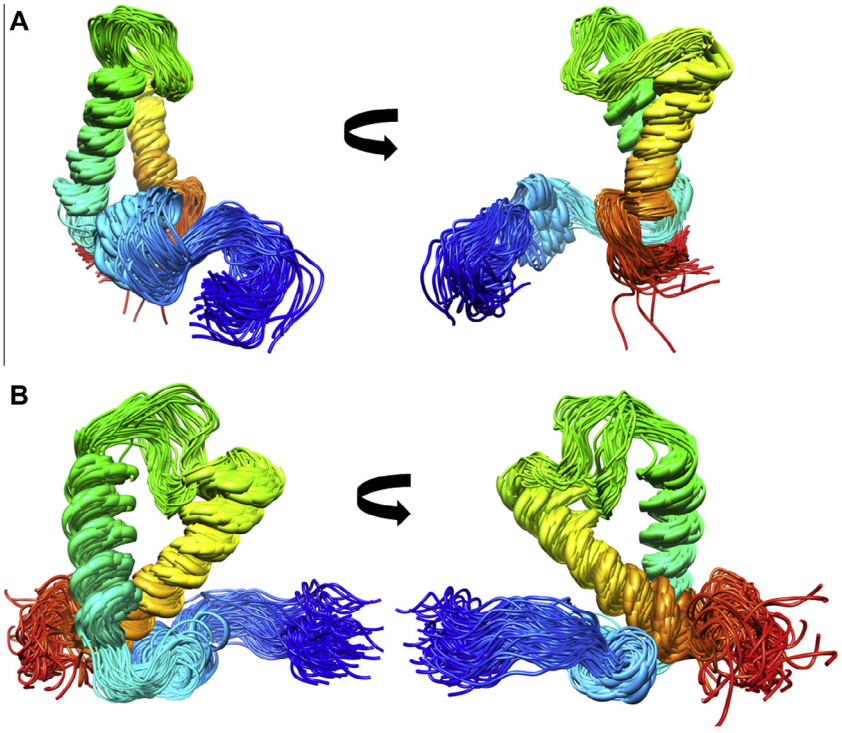

Fig. 2. The ensemble of conformations sampled by the interhelical loop and the N- and C-termini of (A) HIGD1A and (B) HIGD1B. The figure was prepared using Chimera program [38].

gests a structural reorganization due to the transition from micelle to lipid bilayer environment. The RMSF plot depicts the per-residue fluctuations during the course of the simulation. Different regions of the protein are subject to very different fluctuations. While the two transmembrane helices are very stable (RMSF values below 2 Å), there are significant fluctuations and rearrangements at both termini. Interestingly, the interhelical loop exhibits a high degree of stability, probably since it is anchored into the membrane. Fig. 2 depicts the ensemble of conformations sampled by the interhelical loop and the N- and C-termini. The first 10 ns of the simulation have been deemed equilibration time and have thus been removed for all subsequent analysis.

The position of the proteins in the membrane has been monitored over the course of the simulations. No significant changes in position or orientation relative to the membrane could be discerned.

Conformational analysis of the structural changes during MD simulation and comparison of the dihedral angles with the ones in the set of 20 NMR backbone structures show that in both proteins the TM helices keep their length, but became more relaxed (the ranges of backbone angles increased by 10–15 degrees, see Supplemental Figs. 1 and 2). Most pronounced adjustments occurred at the terminal turns of the TM helices, located in the bilayer/water interface. Also the inter-helical loops and short regions between N-terminal amphiphilic helix and first TM helix were significantly refined and their conformational space was restricted during MD simulation. The short amphiphilic helices at the N-terminal tails of both proteins remain folded, located on a bilayer surface and oriented parallel to the surface (Fig. 2), while C-terminal tails remain unfolded and show significant fluctuations in their orientation. During MD simulation side chain angles became clustered around three favorable rotamers (−60/180/+60) in both proteins. On the other hand, there is no substantial restriction in the distribution of side chain angles in helical regions.

3.2. Pocket identification and clustering

Visual inspection of the 100 ns HIGD1A trajectory data revealed two areas of the protein that exhibited pocket-like character for a significant amount of the simulation time and thus may function as interesting targets for virtual screening of possible binders. We identified a top pocket (from here on referred to as pocket 1) and

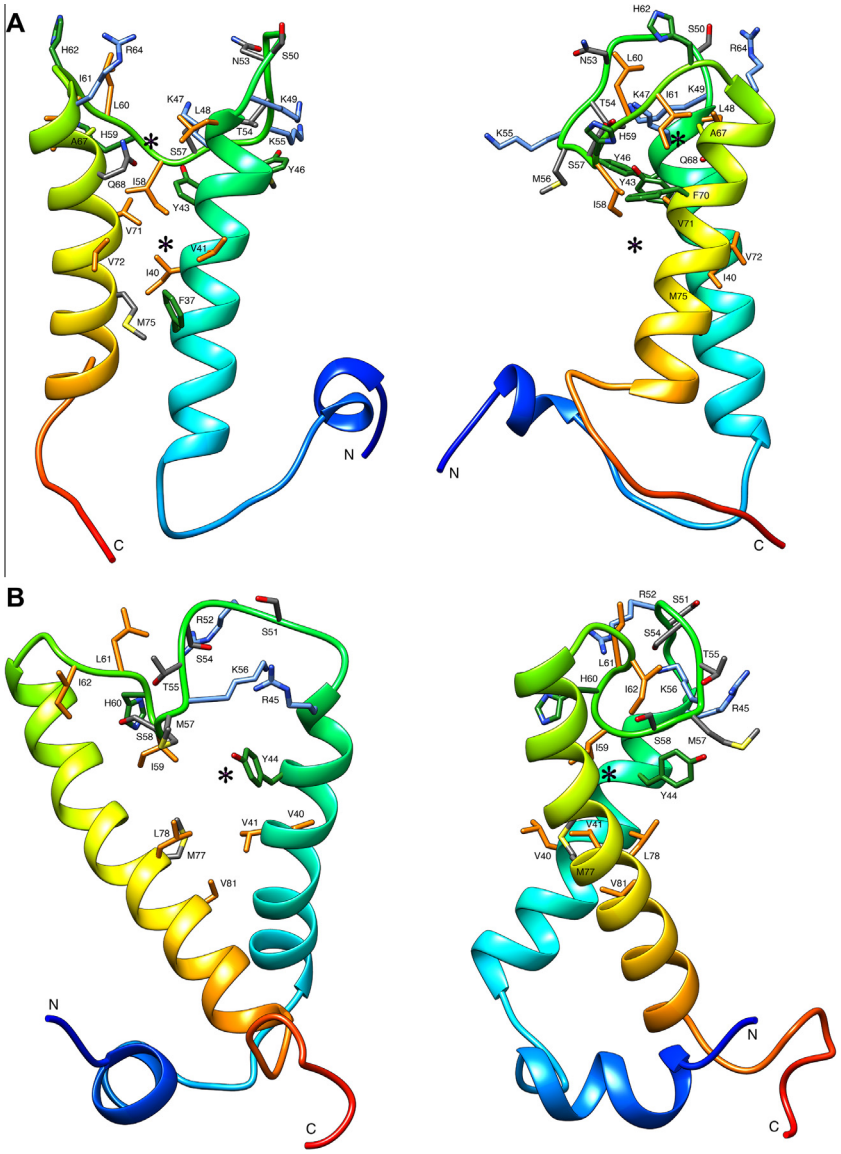

Fig. 3. Representative structures of (A) HIGD1A and (B) HIGD1B from the 100 ns MD simulation. The asterisks show locations of centers of binding pockets used in computational docking. The pocket-forming residues are labeled. The figure was prepared using Chimera program [38].

a side pocket (from here on referred to as pocket 2). The pockets are separated by the helix-helix interface and by the inter-helical loop in such a way, that pocket 1 is easily accessible for hydrophilic compounds, while pocket 2 is buried inside the TM region (Fig. 3A). Pocket 1 is formed by residues Y43, G44, K47, and L48 in helix 1, residues S57, I58, and L60 in the interhelical loop, and residues R64, A67, Q68, F70, and V71 in helix 2. Pocket 2 is formed by residues A39, I40, and Y43 in helix 1, by residue I58 in the interhelical loop, and by residues V71 and M75 in helix 2. Residues Y43 (helix 1), I58 (interhelical lop), V71 (helix 2), and several others separate the pockets and therefore they belong to both pockets (Fig. 3A). Coordinates of these pockets were the basis for trajectory clustering for the virtual screening protocol. The chosen clustering parameters resulted in 6 clusters (per pocket) that represented at least 90% of the respective trajectories. Visual inspection of the 100 ns HIGD1B trajectory data revealed one possible binding pocket. The pocket is formed by residues V40, V41, Y44, R45 in helix 1, residues M57, S58, and I59 in interhelical loop, and residues A70, C71, M77, and L78 in helix 2 (Fig. 3B). No trajectory clustering was performed for HIGD1B. Rather frames every ~1.16 ns were extracted from the simulation and used as receptor structures.

3.3. Computational docking

Two separate pockets on HIGD1A were identified as possible binding sites and targeted in two separate virtual screens. For each of the two pockets six cluster center conformations were chosen that represented the respective pocket conformations during the simulation. Docking was performed into each of the 12 representative structures (six of them with pocket 1 as docking center and six with pocket 2 as docking center). The dataset used in the virtual screen was the full NCI database. The rationale for screening such a large dataset was to comprehensively test the most accurate representation of the chemical space available through the NCI. Schroedinger's Virtual Screening Workflow was used to enable a hierarchical screening of the vast number of compounds. About 1% of all compounds were screened with the most accurate Glide XP scoring function. These docking results were ranked according to the predicted docking score. The best docking score for each compound was identified across the six input receptor structures for each pocket individually. For each of the two pockets, the top 40 compounds from a merged list of top scoring compounds were ordered for testing.

The identified HIGD1B pocket was targeted in a virtual screen. 86 equally spaced simulation structures (~1.16 ns apart in simulation time) were chosen to represent the receptor conformations over the course of the simulation. Docking was performed into each of the 86 structures using the NCI diversity set III in the virtual screen. With HIGD1B we tested the idea of screening a larger conformational variety albeit with a much smaller dataset compared to what was used for HIGD1A. To this end, AutoDock Vina was used for the virtual screen. All 1,013 compounds were docked into all 86 receptor structures, resulting in a total of 87,118 docking simulations. The docking results were ranked according to the predicted docking score. The best docking score for each compound was identified across all 86 input receptor structures. Finally, the top 80 compounds from the merged list of top scoring compounds were ordered for testing.

Supplemental Tables 1 and 2 summarize the results of HIGD1A and HIGD1B virtual screenings and provide NCI database IDs and docking scores for the top scoring compounds for each pocket in both proteins. While generally a wide structural variety of compounds can be seen, one motif seems to be recurring in HIGD1A top scoring compounds. Many compounds contain a tetraol or higher order polyol group. This specific geometry seems to fit the HIGD1A pocket shapes particularly well. Since the diversity set was used for HIGD1B screening no clear common structural motif between the top scoring compounds can be identified.

3.4. NMR screening approach

NMR is an unique method, which provides atomic-level information on protein–ligand interactions in a native-like environment. Advanced methods of NMR screening allow researchers to identify interacting compounds (hits) in a high-throughput manner (up to thousands of compounds per day) and, using protein-based detection, to obtain the detailed information on protein's binding pockets simultaneously [39]. The protein-based NMR detection of interaction with a compound follows the perturbation of the protein chemical shifts, in particular, ^{1}H and ^{15}N chemical shifts using ^{1}H–^{15}N-HSQC "fingerprint" spectra. Since we had a backbone assignment for both HIGD1A and HIGD1B [16], we decided to use the protein chemical shift perturbation as an indicator of interactions.

The NMR approaches for screening compound libraries usually use pooling strategies in order to reduce the number of experiments and, correspondingly, necessary amounts of an isotope-labeled protein. The adaptive pooling strategies divide the pool of compounds into groups of 3–30 molecules and screening these mixtures separately. The positive groups are deconvoluted subsequently in order to identify the active compounds. The adaptive pooling strategies reduce the number of experiments if the fraction of positive responses is below 29% [40]. The non-adaptive pooling approaches imply a strategic pooling, which gives to every compound an unique pattern of presence in the mixtures, thus making unnecessary the deconvolution of the mixtures [41,42]. Due to a relatively small size of libraries and higher probability of positive hits, the non-adaptive pooling strategies usually do not have advantages over adaptive schemes in NMR screens [40].

The deconvolution of groups with the positive response is required because standard group testing algorithm presumes a binary response (positive or negative) in the tests, and that is usually the case for many HTP drug screens. In turn, NMR screens can provide more complex response such as, for example, relaxation-induced line width broadening of resonances, attributed to a particular compound, in ligand-based screens [40] or chemical shifts perturbation for the different resonances in 2D NMR spectra in response to interactions with different compounds in protein-based screens. We took into account this feature of the NMR screen and designed a hybrid group-testing strategy. Our strategy allows direct identification of hits out of total N compounds using $\log_2 N$ samples in a case of small number of expected hits (below 2%). Otherwise, if the number of the positive responses is too high to resolve the active compounds, the strategy can easily adapt the pooling scheme for deconvolution of the intermediate results.

Similar to the non-adaptive algorithms, our hybrid strategy defines an unique code of presence or absence of the compound in each of the samples. In contrast to the non-adaptive algorithms, we do not restrict the codes to make them independent in terms of Boolean algebra, i.e. a code may be a Boolean sum of two or more other codes. As a result, we have a number of samples within the theoretical upper bound for the adaptive algorithms ($d \log_2 N$, where d is the number of positive responses and N is the total number of compounds in the test) and still can determine the active compounds in many cases without deconvolution. The negative effect of our simplification appears when NMR spectra of different samples show similar response (the same affected cross peaks) and this response may be an impact of several compounds with the dependent (in terms of Boolean algebra) codes. For example, for the codes 1000 (A), 0001 (B), 1100 (C), and 1101 (D), the identical positive responses, detected for samples 1, 2, and 4, may indicate the single effect of a compound with code

D (1101), or combined effect of two, three or all four compounds in following combinations: BC, CD, ABC, ABD, or ABCD. In such cases additional experiments are necessary to deconvolute the compounds. Often, when the number of hit compounds is not expected to be very high, the deconvolution step can be combined with the "one-by-one" experiment for binding confirmation of the hit compounds.

According to the standard practice in description of group test problems, let us present the codes as columns in a matrix $M(s, c)$ with elements $m_{i,j} = \{0, 1\}$. Here s is the number of samples and c is the number of compounds (and codes). In each code c_i, $m_{i,j} = 1$ if this compound is present in sample j, otherwise $m_{i,j} = 0$. In non-adaptive algorithms in order to be able to resolve d positive responses the matrix should be d-disjunct, which means that any column is not a part of the Boolean sum of any other d columns. This requirement significantly increases the number of samples (length of the codes) for the tests with high positive response rates [40,42,43]. In order to simplify the pooling approach we avoided this restriction, leaving only the simple uniqueness of the codes for each compound.

The lists of compounds, selected by computational docking for both HI proteins, consist of 80 molecules (Supplemental Tables 3 and 4). We defined 7 samples for each list and distributed the codes in the list in such a way, that compounds with lower (better) score were present in the samples less frequently (see Section 2 and Supplemental Tables 3 and 4). The analysis of the codes shows, that if we consider all responses in 2D NMR spectra as binary events (yes or no, without differentiation between cross peaks and, therefore, between different residues), a response in a single sample gives us exactly one code (one hit compound), response in two samples give 3 codes (3 possible hit compounds), in 3 samples give 7 codes, and in 4, 5, and 6 samples give 14, 15, 25–28, and 47–50 codes, respectively. The codes for possible hit compounds are not independent: as we show in the example above, only specific combinations of four codes (A, B, C, and D) would correspond to the observed response in sample 1, 2, and 4. So, the deconvolution of these codes may require fewer trials, than an one-by-one approach, but since separate confirmation of the compound activity is necessary, in a case of small number of possible hit compounds it is always easier to test them "one-by-one".

3.5. NMR screening results

The CF-produced HIGD1A and HIGD1B were split to 8 equivalent samples for each protein. One sample was used as a control, without the compounds, and 7 samples were used to solubilize the compound mixtures as described in Section 2. The simple check of "fingerprint" $^{1}H–^{15}N$-HSQC-TROSY (TROSY) spectra shows a consistency of the samples. The ^{1}H and ^{15}N chemical shifts of backbone signals were retrieved from TROSY spectra and analyzed. The WSC values calculated for each cross peak in the TROSY spectra (see Section 2) allows us to identify the possible hit compounds as well as affected protein regions (Fig. 4, Supplemental Fig. 3). Persistent in all samples WCSs above a threshold value of 0.02 indicate sensitivity of the N-terminal residues and the inter-helical loop residues in both proteins to changes of the LMPG micelle properties (micelle shape, detergent density, interface properties, etc.). Since the LMPG micelles are very sensitive to the environment [44], these changes could be induced by non-specific interaction of several different compounds with the micelle.

HIGD1A shows specific, sample-dependent response in WCSs above the threshold in samples 2, 3 and 6 (Fig. 4A). In turn, only sample 7 shows a specific response of HIGD1B (Fig. 4B).

HIGD1A responses in samples 2, 3 and 6 restrict the pool of possible hits to seven compounds. Their NCI library IDs are 8127 (short list ID a02), 252035 (a03), 408122 (a06), 368270 (a11), 71286 (a34), 83960 (a41), and 1972 (b08). More detailed analysis of the response allows us to reorganize the shortlist. Indeed, the WCS responses can be easily divided into two groups. The first group contains cross peaks for residues K49, T54, S57, L60, and A66 (group I),

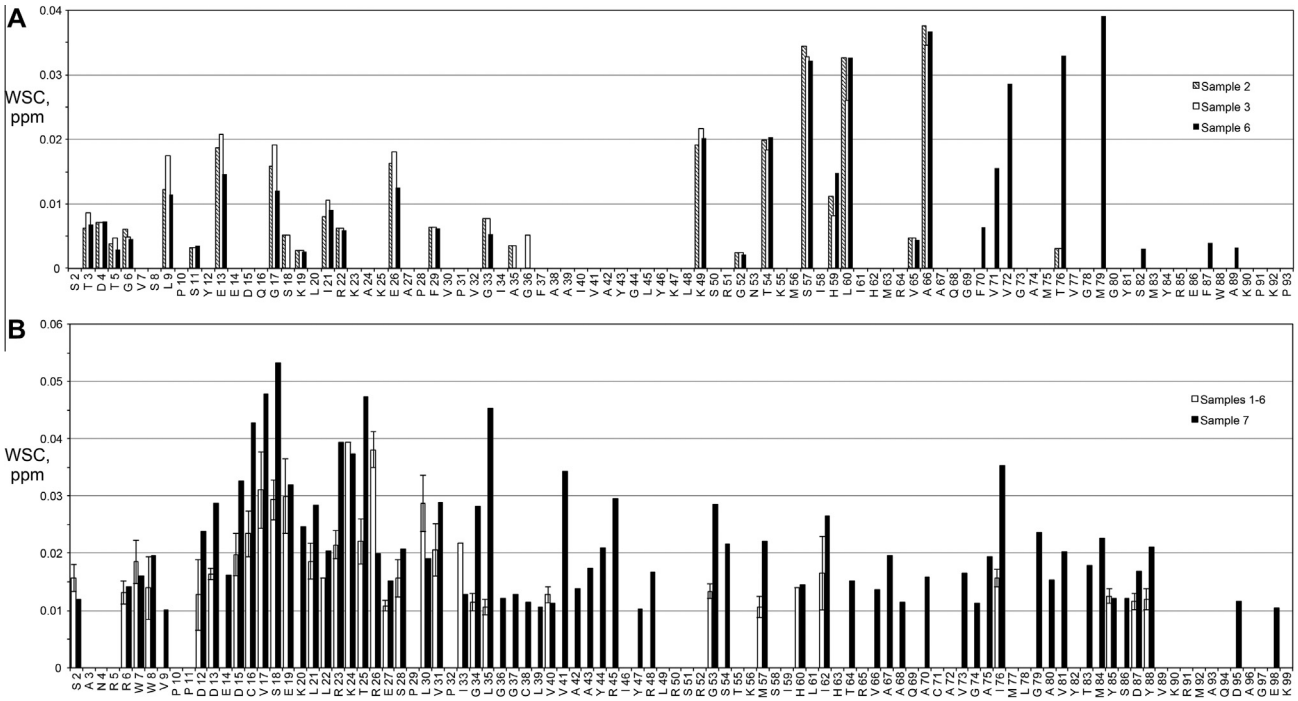

Fig. 4. Weighted chemical shifts difference (WCS) for samples of HIGD1A (A) and HIGD1B (B) with ligand mixtures. WCS were calculated using ^{1}H and ^{15}N chemical shifts measured in $^{1}H–^{15}N$-TROSY-HSQC spectra of the sample with and without the ligands. (A) WCS for samples 2 (white bars), 3 (dashed bars), and 6 (black bars) with the compounds mixed according to the pooling matrix for HIGD1A. (B) WCS for sample 7 (black-filled bars) in comparison with average WCS value for samples 1–6 (empty bars). Standard deviations for average WCS values are shown.

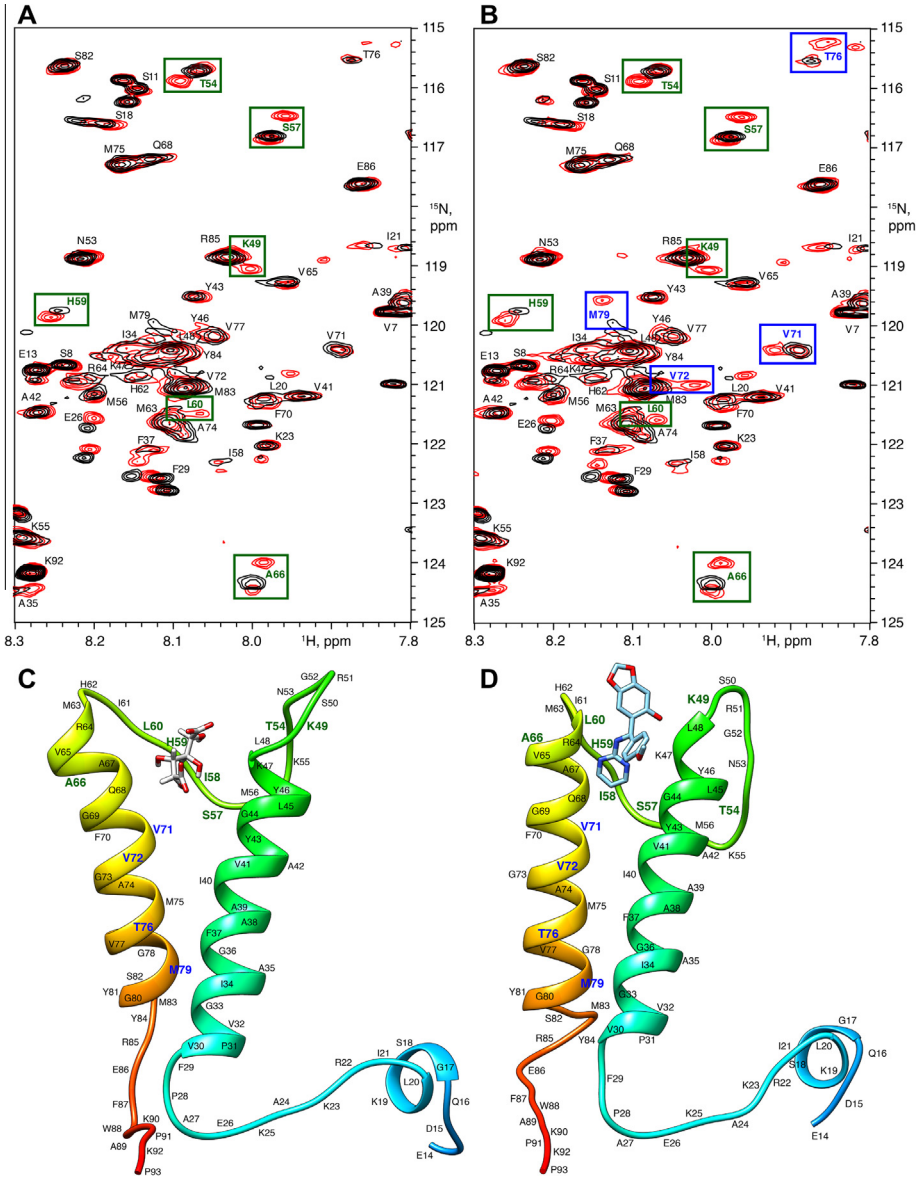

Fig. 5. Hit compounds interaction with HIGD1A confirmed by NMR spectroscopy. (A and B) Superimposed ^{1}H–^{15}N-TROSY-HSQC spectra (^{1}H: 7.8–8.3 ppm, ^{15}N: 115–125 ppm) of HIGD1A samples with (A: sample 2, B: sample 6) and without hit compounds. The spectrum of the sample without ligands is shown in black. The cross peaks, affected by the compounds are colored and marked with rectangles. Spatial structures of HIGD1A with bound hit compounds 408122 (C) and 368270 (D). The binding conformation of compounds 408122 and 368270 were calculated by Glide program [26]. The representative structures of HIGD1A were selected from a 100 ns MD simulation of the backbone NMR structure (2LOM) in DMPC lipid bilayer. The residues in close proximity to the ligand are labeled. Panels C and D of the figure were prepared using Chimera program [38].

which are affected in all three samples 2, 3, and 6 (Fig. 4A). The second group contains residues V71, V72, T76, and M79 (group II), which are affected only in sample 6. Basing on the definition of binding pockets 1 and 2 for HIGD1A in computational docking (see above and Fig. 3A), we presumed that these separately affected groups of residues belong to the different pockets. We re-grouped the shortlist and separated the combinations of possible hits, which could be responsible for effect only in samples 2 and 3. These combinations are a02 and a03 together, a11 alone, or any combination of a11 with a02 and/or a03. All of these compounds were predicted to bind to the pocket 1. In turn, in order to detect responses in samples 2, 3, and 6 at least one of the following combinations must contain all active compounds: b08, a06a11, a02a41, a03a34, a02a03a06, or any merger of these 5 combinations with the possible hits. Among these compounds only b08 was predicted to be active on the pocket 2.

The one-by-one binding tests show that compounds a11 and a06 (NCI library IDs 408122 and, 368270 respectively) affect chemical shifts in TROSY spectra of HIGD1A (Fig. 5), while other five compounds have no effect. Compound a11 affects chemical shifts of residues from the group I in the inter-helical loop and compound a06 affects chemical shifts of the residues in both groups I and II. Since the test with single compound b08 shows no response, no one of the hit compounds, pre-selected by computational docking to the pocket 2, were confirmed experimentally. On the other hand, responses in chemical shifts of atoms, located in the second TM helix of HIGD1A, to the compound a06 reflect changes in the chemical environment of the amide groups of V71, V72, T76, and M79 from the second TM helix. At the moment we don't have clear explanation for this observation and hypothesize that the interaction of the bulky aromatic compound 368270 with the inter-helical loop (Fig. 5D) may affect the packing of the helices inside the micelle.

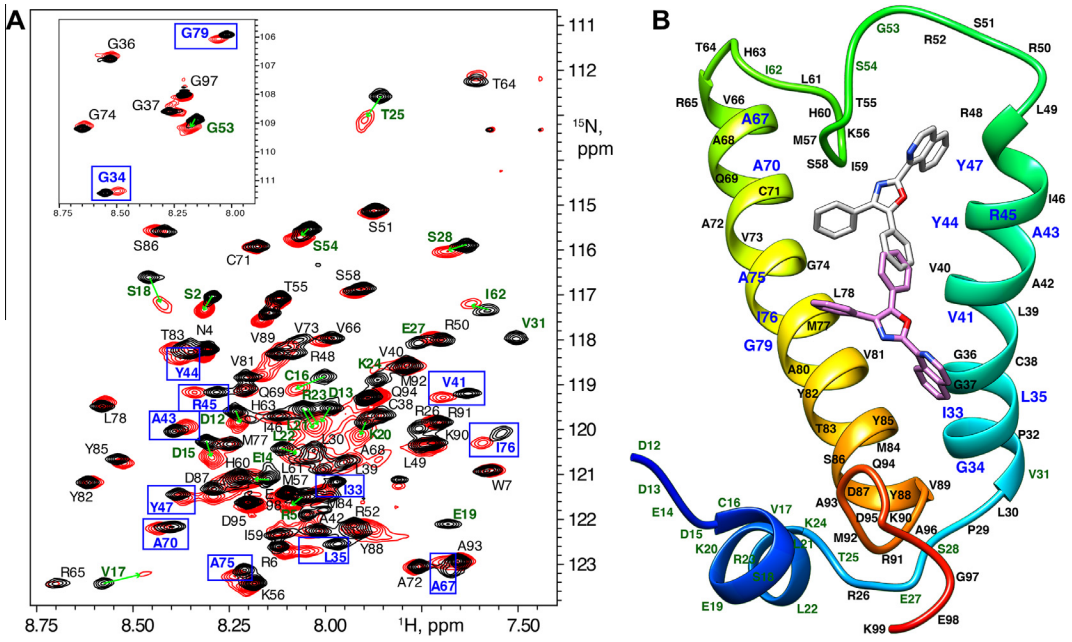

Fig. 6. Hit compound interaction with HIGD1B confirmed by NMR spectroscopy. (A) Superimposed ^{1}H–^{15}N-TROSY-HSQC spectra (^{1}H: 7.40–8.75 ppm, ^{15}N: 110.7–124.0 ppm) of HIGD1B samples with hit compound 25457 (red lines) and without any compound (black lines). The cross peaks, affected by the ligands are marked with arrows (N-terminal amphiphilic helix) and rectangles (TM helices). (B) Spatial structure of HIGD1B with compound 25457 bound in two poses. The binding conformations of 25457 were calculated by Vina-Autodock program [33]. The structure of HIGD1B was selected from a 100 ns MD simulation of the HIGD1B backbone NMR structure (2LON) in DMPC lipid bilayer. The residues in close proximity to the ligand are labeled. Panel B of the figure was prepared using Chimera program [38].

HIGD1B response in a single spectrum allows identification of the single hit compound (NCI library ID 25457) using the pooling matrix. The affected cross peaks, according to HIGD1B assignment [16], belong to the following residues: L35, V41, Y44, R45 (all from the first TM helix), I76, G79, V81, M84, and Y88 (second TM helix, Fig. 6). Analysis of the possible binding pocket for the compound 25457 shows two preferable poses, first one under the inter-helical loop, where it may affect amides of residues V41, Y44, Y47, and A70, and another pose in the middle of the TM region, close to I33, V41, and V81 (Fig. 6B). The detected effect on amide chemical shifts of L35, R45, I76, and G79 can be attributed to ligand-induced changes in bundle packing and, subsequently, to changes in local environment, such as interaction with detergent aliphatic chains within a micelle.

4. Discussion

The paper describes the CNDY strategy for search of potential ligands for hIMPs. The backbone structures of HIGD1A and HIGD1B, determined by accelerated NMR methods [16], were refined in a lipid bilayer. Representative structures from 100 ns MD simulations in the bilayer were used for computational docking of the compounds from the NCI Open Chemical Repository. Predicted computational docking hit compounds (80 top-scored for each protein) were tested experimentally by protein-based NMR screening using combinatorial mixtures of the ligands. Two compounds were found interacting with HIGD1A and one compound was active on HIGD1B (Supplemental Table 5). Based on the number of hits it is not possible to identify one of the virtual screening strategies to be superior.

The described strategy for the fast screening of potential hit compounds for IMPs identified and experimentally confirmed three hits starting from the library of several hundred thousand compounds and the backbone-only, low-to-medium resolution NMR structures of the IMPs. Despite the high throughput, the approach allows us to assign the protein responses to different compounds and outline the binding pockets.

Computational identification of a compound binding within the TM region of IMP is an incredibly challenging task. Standard computational docking approaches are tailored for the search of a hydrophobic ligand, which fits into a hydrophobic pocket on soluble proteins or protein domains. These approaches also work fine with domains, distant from the membrane, such as tails and inter-helical loops. The domains of MPs, exposed from the membrane, may have hydrophobic cavities far removed from the solvent, resembling those of folded globular proteins. In turn, TM helical bundles have extended hydrophobic surfaces, which extensively interact with aliphatic chains of lipids, and a less hydrophobic helix–helix interface. Since standard docking programs do not model the lipid environment, the hydrophobic surface of the TM bundle became “naked” and exposed to hydrophobic ligands. As a result the computational docking to the intra-membrane pocket may produce a large number of false-positive hits, which in the experiment could not compete with the lipids for binding to the protein surface. In light of the demonstrated success in identifying three compounds interacting with the HI proteins, there remains potential for further improvement. For instance, we ask if the computational ligand search for TM helical bundles can be modified (1) to account for the lipid environment of the hydrophobic surface of a TM helical bundle and, concurrently, (2) to search for a binding in the hydrophilic cavities inside the bundles, which are formed by interacting TM helices and are shielded from the detergent. These improvements combined with the use of genomic information will greatly facilitate the drug discovery process in the near future.

Acknowledgments

This work has been partly supported by the National Institute of Health Grants GM098630 and GM095623 and by IFEZ (Joint Center for Biosciences). Work at UCSD is supported in part by NSF, NIH, HHMI, NBCR, and the NSF Supercomputer Centers.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.bbrc.2014.01.179.

References

[1] J. Drews, Drug discovery: a historical perspective, Science 287 (2000) 1960–1964.
[2] A. Shekhtman, D.S. Burz (Eds.), Protein NMR Techniques, Humana Press, 2012.
[3] J. Steyaert, B.K. Kobilka, Nanobody stabilization of G protein-coupled receptor conformational states, Curr. Opin. Struct. Biol. 21 (2011) 567–572.
[4] G. Sciara, F. Mancia, Highlights from recently determined structures of membrane proteins: a focus on channels and transporters, Curr. Opin. Struct. Biol. 22 (2012) 476–481.
[5] I. Maslennikov, S. Choe, Advances in NMR structures of integral membrane proteins, Curr. Opin. Struct. Biol. 23 (2013) 555–562.
[6] T. Qureshi, N.K. Goto, Contemporary Methods in Structure Determination of Membrane Proteins by Solution NMR, in: G. Zhu (Ed.), NMR Proteins Small Biomol, Springer, Berlin Heidelberg, 2012, pp. 123–185.
[7] S. Ou-Yang, J. Lu, X. Kong, Z. Liang, C. Luo, H. Jiang, Computational drug discovery, Acta Pharmacol. Sin. 33 (2012) 1131–1140.
[8] J.D. Durrant, J.A. McCammon, Molecular dynamics simulations and drug discovery, BMC Biol. 9 (2011) 71.
[9] W. Sinko, S. Lindert, J.A. McCammon, Accounting for receptor flexibility and enhanced sampling methods in computer-aided drug design, Chem. Biol. Drug Des. 81 (2013) 41–49.
[10] J.-H. Lin, A.L. Perryman, J.R. Schames, J.A. McCammon, The relaxed complex method: accommodating receptor flexibility for drug design with an improved scoring scheme, Biopolymers 68 (2003) 47–62.
[11] R.E. Amaro, R. Baron, J.A. McCammon, An improved relaxed complex scheme for receptor flexibility in computer-aided drug design, J. Comput. Aided Mol. Des. 22 (2008) 693–705.
[12] S.B. Shuker, P.J. Hajduk, R.P. Meadows, S.W. Fesik, Discovering high-affinity ligands for proteins: SAR by NMR, Science 274 (1996) 1531–1534.
[13] A.L. Skinner, J.S. Laurence, High-field solution NMR spectroscopy as a tool for assessing protein interactions with small molecule ligands, J. Pharm. Sci. 97 (2008) 4670–4695.
[14] E.R. Zartler, M.J. Shapiro, Protein NMR-based screening in drug discovery, Curr. Pharm. Des. 12 (2006) 3963–3972.
[15] M. Pellecchia, I. Bertini, D. Cowburn, C. Dalvit, E. Giralt, W. Jahnke, et al., Perspectives on NMR in drug discovery: a technique comes of age, Nat. Rev. Drug Discov. 7 (2008) 738–745.
[16] C. Klammt, I. Maslennikov, M. Bayrhuber, C. Eichmann, N. Vajpai, E.J. Chiu, et al., Facile backbone structure determination of human membrane proteins by NMR spectroscopy, Nat. Methods 9 (2012) 834–839.
[17] N. Denko, Cornelia Schindler, Albert Koong, K. Laderoute, C. Green, A. Giaccia, Epigenetic Regulation of Gene Expression in Cervical Cancer Cells by the Tumor Microenvironment, Clin. Cancer Res. 6 (2000) 480–487.
[18] M. Vukotic, Silke Oeljeklaus, Sebastian Wiese, F.N. Vögtle, C. Meisinger, H.E. Meyer, et al., Rcf1 mediates cytochrome oxidase assembly and respirasome formation, revealing heterogeneity of the enzyme complex, Cell Metab. 15 (2012) 336–347.
[19] S. Jo, T. Kim, W. Im, Automated builder and database of protein/membrane complexes for molecular dynamics simulations, PLoS One 2 (2007) e880.
[20] S. Jo, J.B. Lim, J.B. Klauda, W. Im, CHARMM-GUI membrane builder for mixed bilayers and its application to yeast membranes, Biophys. J. 97 (2009) 50–58.
[21] A.D. MacKerell, N. Banavali, N. Foloppe, Development and current status of the CHARMM force field for nucleic acids, Biopolymers 56 (2000) 257–265.
[22] J.C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, et al., Scalable molecular dynamics with NAMD, J. Comput. Chem. 26 (2005) 1781–1802.
[23] J.-P. Ryckaert, G. Ciccotti, H.J.C. Berendsen, Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes, J. Comput. Phys. 23 (1977) 327–341.
[24] M. Christen, Philippe H. Hünenberger, D. Bakowies, R. Baron, R. Bürgi, D.P. Geerke, et al., The GROMOS software for biomolecular simulation: GROMOS05, J. Comput. Chem. 26 (2005) 1719–1751.
[25] LigPrep, version 2.6, Schrödinger, LLC, New York, 2013.
[26] R.A. Friesner, Jay L. Banks, Robert B. Murphy, T.A. Halgren, J.J. Klicic, D.T. Mainz, et al., Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy, J. Med. Chem. 47 (2004) 1739–1749.
[27] T.A. Halgren, R.B. Murphy, R.A. Friesner, H.S. Beard, L.L. Frye, W.T. Pollard, et al., Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening, J. Med. Chem. 47 (2004) 1750–1759.
[28] R.A. Friesner, R.B. Murphy, Matthew P. Repasky, L.L. Frye, J.R. Greenwood, T.A. Halgren, et al., Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein−ligand complexes, J. Med. Chem. 49 (2006) 6177–6196.
[29] QikProp, version 3.5, Schrödinger, LLC, New York, 2012.
[30] C.A. Lipinski, F. Lombardo, B.W. Dominy, P.J. Feeney, Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, Adv. Drug Deliv. Rev. 46 (2001) 3–26.
[31] J.L. Banks, H.S. Beard, Y. Cao, A.E. Cho, W. Damm, R. Farid, et al., Integrated Modeling Program, Applied Chemical Theory (IMPACT), J. Comput. Chem. 26 (2005) 1752–1780.
[32] J.C. Shelley, A. Cholleti, L.L. Frye, J.R. Greenwood, M.R. Timlin, M. Uchimaya, Epik: a software program for pK(a) prediction and protonation state generation for drug-like molecules, J. Comput. Aided Mol. Des. 21 (2007) 681–691.
[33] O. Trott, A.J. Olson, AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading, J. Comput. Chem. 31 (2010) 455–461.
[34] C. Klammt, Frank Löhr, Birgit Schäfer, W. Haase, V. Dötsch, H. Rüterjans, et al., High level cell-free expression and specific labeling of integral membrane proteins, Eur. J. Biochem. 271 (2004) 568–580.
[35] I. Maslennikov, C. Klammt, E. Hwang, G. Kefala, M. Okamura, L. Esquivies, et al., Membrane domain structures of three classes of histidine kinase receptors by cell-free expression and rapid NMR analysis, Proc. Natl. Acad. Sci. U S A 107 (2010) 10902–10907.
[36] R. Keller, The Computer Aided Resonance Assignment Tutorial, Cantina Verlag, Goldau, Switzerland, 2004.
[37] J. Cavanagh, W.J. Fairbrother, A.G. Palmer III, M. Rance, N.J. Skelton, Chapter 9 – Larger proteins and molecular interactions, in: J. Cavanagh, W.J. Fairbrother, A.G. Palmer, M. Rance, N.J. Skelton (Eds.), Protein NMR Spectrosc, second ed., Academic Press, Burlington, 2007, pp. 725–780.
[38] E.F. Pettersen, T.D. Goddard, C.C. Huang, G.S. Couch, D.M. Greenblatt, E.C. Meng, et al., UCSF Chimera – A visualization system for exploratory research and analysis, J. Comput. Chem. 25 (2004) 1605–1612.
[39] P.J. Hajduk, Tobias Gerfin, Jean-Marc Boehlen, M. Häberli, D. Marek, S.W. Fesik, High-throughput nuclear magnetic resonance-based screening, J. Med. Chem. 42 (1999) 2315–2317.
[40] K.A. Mercier, R. Powers, Determining the optimal size of small molecule mixtures for high throughput NMR screening, J. Biomol. NMR 31 (2005) 243–258.
[41] J.J. Devlin, Amy Liang, Lan Trinh, M.A. Polokoff, D. Senator, W. Zheng, et al., High capacity screening of pooled compounds: identification of the active compound without re-assay of pool members, Drug Dev. Res. 37 (1996) 80–85.
[42] R.M. Kainkaryam, P.J. Woolf, Pooling in high-throughput drug screening, Curr. Opin. Drug Discov. Dev. 12 (2009) 339–350.
[43] Ding-Zhu Du, Frank K Hwang, Combinatorial Group Testing and its Applications, second ed., World Scientific, 1999.
[44] H.-X. Zhou, T.A. Cross, Influences of membrane mimetic environments on membrane protein structures, Annu. Rev. Biophys. 42 (2013) 361–392.