Published in final edited form as:

Bioorg Med Chem. 2016 October 1; 24(19): 4750–4758. doi:10.1016/j.bmc.2016.08.019.

Novel small molecule binders of human N-glycanase 1, a key player in the endoplasmic reticulum associated degradation pathway

Bharath Srinivasan^{a,b}, Hongyi Zhou^a, Sreyoshi Mitra^a, and Jeffrey Skolnick^{a,e}

^aCenter for the Study of Systems Biology, School of Biology, Georgia Institute of Technology, 950, Atlantic Drive, Atlanta, Georgia 30332, United States

Abstract

Peptide:N-glycanase (NGLY1) is an enzyme responsible for cleaving oligosaccharide moieties from misfolded glycoproteins to enable their proper degradation. Deletion and truncation mutations in this gene are responsible for an inherited disorder of the endoplasmic reticulum-associated degradation pathway. However, the literature is unclear whether the disorder is a result of mutations leading to loss-of-function, loss of substrate specificity, loss of protein stability or a combination of these factors. In this communication, without burdening ourselves with the mechanistic underpinning of disease causation because of mutations on the NGLY1 protein, we demonstrate the successful application of virtual ligand screening (VLS) combined with experimental high-throughput validation to the discovery of novel small-molecules that show binding to the transglutaminase domain of NGLY1. Attempts at recombinant expression and purification of six different constructs led to successful expression of five, with three constructs purified to homogeneity. Most mutant variants failed to purify possibly because of misfolding and the resultant exposure of surface hydrophobicity that led to protein aggregation. For the purified constructs, our threading/structure-based VLS algorithm, FINDSITE^{comb}, was employed to predict ligands that may bind to the protein. Then, the predictions were assessed by high-throughput differential scanning fluorimetry. This led to the identification of nine different ligands that bind to the protein of interest and provide clues to the nature of pharmacophore that facilitates binding. This is the first study that has identified novel ligands that bind to the NGLY1 protein as a possible starting point in the discovery of ligands with potential therapeutic applications in the treatment of the disorder caused by NGLY1 mutants.

^eCorresponding author, skolnick@gatech.edu, Tel: (404) 407-8975, Fax: (404) 385-7478.

^bbharath.srinivasan@biology.gatech.edu

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Author Contributions

BS conceived of the study, participated in its design, carried out the experiments, analyzed and interpreted the results, and drafted the manuscript. HZ carried out the VLS predictions. SM carried out the experiments, analyzed and interpreted the results. JS conceived of the study, participated in its design and coordination, provided appropriate resources, helped analyze the data, and was involved in drafting and critically reviewing the manuscript. All authors read and approved the final manuscript.

Supplementary Material: Supplementary data is included with this manuscript and contains.

Graphical abstract

Keywords

drug discovery; N-glycanase 1; virtual ligand screening; Thermal shift assay; ERAD

1. Introduction

Protein glycosylation is an important posttranslational modification, PTM, whereby sugar moieties are added onto specific amino acids in proteins [1]. Glycosylation is a critical process responsible for a wide range of biological processes including, but not limited to, intercellular communication, cell adherence to extracellular matrix and recognition [2]. Glycosidic bonds between proteins and sugars can be broadly classified into N-, O-, C-linked glycosylation, with the former two being the most common [1]. Glycosylation is critical for the proper functioning and regulation of the biosynthetic-secretory pathway in the endoplasmic reticulum (ER) and Golgi apparatus, with a lot of proteins typically expressed in a cell undergoing this modification [3, 4]. Glycosylation is widely found in secreted proteins, surface receptors and ligands, organelle-resident proteins and also in proteins that are trafficked from the Golgi through the cytoplasm in post-Golgi carriers [3].

In the endoplasmic reticulum, glycosylation is used to monitor the status of protein folding, acting as a quality control mechanism to ensure that only properly folded proteins are trafficked to the Golgi [3, 5, 6]. Upon detection of misfolded proteins, proper recycling requires enzymes that can segregate the sugars from the proteins for the lysosomal and proteasomal degradation of the respective portions [7]. The enzyme that plays the most important role in such scenarios are peptide:N-glycanases (for N-linked glycans) [8]. Other enzymes that play critical roles are ENGase and β -galactosidase [9].

NGLY1 is an N-glycanase gene that encodes an enzyme that catalyzes the hydrolysis of an N(4)-(acetyl-beta-D-glucosaminyl) asparagine residue to N-acetyl-beta-D-glucosaminylamine and a peptide containing an aspartate residue [7]. The encoded enzyme, a cytoplasmic protein, plays an important role in the proteasome-mediated degradation of misfolded glycoproteins synthesized in the ER and subsequently translocated to the

cytoplasm. By performing the initial cleavage of bulky N-glycan chains on misfolded glycoproteins, NGLY1 makes further proteolysis possible [7]. Glycoproteins thus trimmed can enter the cylinder of the 20S proteasome to be acted upon by proteases [9]. UniProt entry reports multiple transcript variants encoding different isoforms for peptide:N-glycanase (Q96IV0). The protein has a three domain architecture with the amino terminal domain implicated as having interactions with p97 and HR23B, the carboxy terminal domain implicated in interacting with mannose that results in sugar substrate recognition and the central transglutaminase domain implicated in catalysis [7].

Improper degradation of misfolded glycoproteins leads to disease [10]. Several different variants of NGLY1 have been implicated, with the most predominant being R402X, R402del and R542X (human transcript 1 numbering) [11]. These variations lead to a novel autosomal recessive disorder of the endoplasmic reticulum associated degradation (ERAD) pathway manifest in neurological dysfunction, abnormal tear production, and liver disease [11].

We have employed our threading/structure-based FINDSITE^{comb} algorithm [12] to predict ligands that might bind to NGLY1. FINDSITE-based approaches have the advantage that they are as fast as traditional ligand-based approaches and yet overcome the limitations of traditional ligand- or structure-based approaches [12]. Furthermore, this study is a successful example for the effective integration of computational methods and high-throughput experimental screening to arrive at small-molecules that show binding to protein targets of immediate therapeutic importance. Given the critical role that glycosylation plays in several important biological processes, it might be desirable to regulate, rather than abolish, the role some of the enzymes play in glycosylation pathways [13]. With that intent, this work reports several binders of human NGLY1 as possible small-molecules that might stabilize/modulate the protein's function.

2. Materials and methods

2.1. Reagents

All reagents and chemicals, unless mentioned otherwise, were procured from Sigma-Aldrich (St. Louis, MO). HEPES, pH 7.3, buffer and dimethyl sulfoxide (DMSO) were obtained from Fischer Bioreagents. Extrinsic fluorophore dye, Sypro-orange, was from Invitrogen (Carlsbad, CA). 96-well PCR-plates and plate seals were from Eppendorf (NY, USA). The artificial gene synthesis exercise was carried out by Genscript (Piscataway, NJ). The small molecule library containing oncology drug set III (97 compounds), mechanistic set II (816 compounds) diversity set III (synthetic) (1597 compounds) and natural product set (118 compounds) were obtained from the open chemical repository of the Developmental Therapeutics Program (DTP) of the National Cancer Institute (NCI), National Institutes of Health (NIH) (<http://dtp.cancer.gov>).

2.2. Cloning, Expression and purification of NGLY1 constructs

2.2.1. Design and synthesis of the genes—Six different gene constructs were designed to increase the chances of protein expression and subsequent purification (Fig 1). The source sequence for human NGly1 gene was transcript 1 (NM_018297) (See

Supplementary Material). The six constructs were full-length NGLY1 (FLNGLY1), the R402 truncation variant of full length NGLY1 (R402XFLNGLY1), the catalytic domain of NGLY1 (CDNGLY1), the R402 truncation variant of the catalytic domain of NGLY1 (R402XCDNGLY1), the R542 truncation variant of full length NGLY1 (R542XFLNGLY1) and the R402 deletion variant of full length NGLY1 (R402del FLNGLY1). Since all six gene constructs are segments of eukaryotic genes, there are several factors that could potentially hamper their expression in a prokaryotic expression system like *E. coli*. The principal factor that determines expression failure in heterologous system is the differential codon usage patterns for the two different organisms (Human and *E. coli* in the present case). This is mainly because of the differential tRNA gene pools that account for codon bias. This limitation can be circumvented by either expressing the gene of interest in *E. coli* with plasmid encoded tRNAs for the deficient tRNAs or, a more common approach is to resynthesize the gene of interest taking into account the codon bias while keeping the amino acid sequence invariant. The latter approach has the advantage of not having to burden the expression system with multiple plasmids and their maintenance. Thus, we have designed and synthesized a codon-optimized version of each gene, the sequence of which is specified in SI material.

2.2.2. Cloning of the genes in pET30—The synthesized gene constructs were cloned into the pET30a vector with a carboxy-terminus hexa-histidine tag, a T7 promoter to drive the transcription and kanamycin resistance for selection. The genes were cloned into the NdeI and HindIII sites of the Multiple cloning site (MCS) on the vector. The pET series of vectors have a bacterial origin of replication and can be expressed in the BL21(DE3) *Escherichia coli* strain with a prophage (engineered phage) that has T7 RNA polymerase controlled by Lac regulatory construct. Further, these strains are deficient in Lon and Omp T proteases that could potentially cleave the recombinantly expressed proteins.

All the synthesized and cloned genes were assessed by sequencing and insert release to assess whether the inserted fragments are of the correct length (Fig S1).

2.3. Expression optimization

Expression optimization exercises were undertaken for all the six constructs (FLNGLY1, R402XFLNGLY1, CDNGLY, R402XCDNGLY, R542XFLNGLY1 and R402del FLNGLY1). The concentration of IPTG and the temperature of induction were varied to optimize the right conditions under which to express the protein.

2.4. Protein Purification

The expression plasmid was transformed into *E. coli* BL21 (DE3) for expression. A preinoculum was grown overnight in Luria-Bertani broth. 10 % was inoculated into terrific broth with kanamycin (20 µg/ ml) as the selection marker and incubated at 37° C till the optical density was 0.6 and induced with IPTG at a final concentration of 0.3 mM that was standardized from the previous step. Induction was carried out at 17° C for 15–18 hours. The cells were pelleted down at 2600 × g for 30 minutes and were resuspended in 30 ml of lysis buffer (50 mM Tris HCl, pH 8.0, 100 mM NaCl, 2 mM DTT and 10 % glycerol). The cells were lysed using sonicator (40% amplitude, 3 seconds on, 5 seconds off for 35–40 cycles)

and the obtained cell lysate was centrifuged at $12900 \times g$ for 35 min. The supernatant was kept for binding with buffer-equilibrated nickel nitrilotriacetic acid beads (Ni-NTA His-Bind® Resin, Qiagen, USA) for 3 hrs. The beads were washed with 30 ml of wash buffer (lysis buffer containing 15 mM and 20 mM imidazole, respectively). The bound enzyme was eluted with 5 ml of elution buffer (lysis buffer containing 250 mM imidazole). After the elution of the protein from Ni-NTA beads, 1 mM EDTA was added to protect the protein from oxidation by trace nickel. The protein was dialyzed extensively to get rid of the excess salt. It was then concentrated to the required volume using Macrosep centrifugal devices (Pall Co., USA) with 10 KDa cutoff membrane and stored at -20°C . Then, SDS-PAGE analysis was performed using standard protocols [14]. Protein concentrations were determined by the method of Bradford [15] with bovine serum albumin (BSA) as a protein standard. Recombinant hexa-histidine tagged full-length NGLY1 and R402del NGLY1 showed a single band on the SDS-polyacrylamide gel corresponding to molecular masses of ~75 kDa and ~48 KDa, respectively (Fig. S2).

2.5. Structure modeling for full-length NGLY1

Full length NGLY1 sequence was parsed into three domains (1–119, 158–464 & 465–654) by the structure prediction module TASSER^{vmt}-lite in FINDSITE^{comb}, which were then modeled independently. We also modeled the full length NGLY1 sequence without domain parsing using TASSER^{vmt}-lite [16]. Since this full length model only provides a reliably predicted central, catalytic domain of NGLY1, we superimposed the independently modeled domains on to the full length model to create a better full length prediction.

2.6. FINDSITE^{comb} for virtual ligand screening

FINDSITE^{comb} was employed to carry out virtual ligand screening on both full-length NGLY1, the central transglutaminase domain of NGLY1 and the various truncation and deletion constructs. The FINDSITE^{comb} methodology for ligand virtual screening is described in detail in Ref [12], with its experimental validation found in Ref [17]. FINDSITE^{comb} is a composite approach consisting of the improved FINDSITE-based approach FINDSITE^{filt} [12, 18] and the extended FINDSITE-based approach FINDSITE^X [19]. The basic assumption of FINDSITE^{comb} is that evolutionarily related proteins have a high chance of binding similar ligands. For a given protein target with unknown ligands, we predict its potential ligands by finding ligands that bind evolutionary related proteins.

An overview of FINDSITE^{comb} is shown in Fig 2. For a given protein with unknown binding ligands and unknown structure, FINDSITE^{comb} first builds a structure model using a state-of-the-art threading method [20, 21] and one of the best protein structure modeling approaches, TASSER^{VMT} [22]. The structure model is subsequently employed to search a library of proteins with experimentally bound ligands (PDB holo structures [23]) in the FINDSITE^{filt} component [12] as well as a library of modeled structures of proteins having experimentally determined binders but no experimental structures of the complex (DrugBank [24] & ChEMBL [25]) in the FINDSITE^X component [19]. The ligands of the top 100 ranked pockets from the PDB, and of the top ranked proteins from DrugBank & ChEMBL (called “template ligands”) are used as “seeds” to search against a compound library. The compound library ligands are independently ranked by the three components

(FINDSITE^{filt} using PDB, FINDSITE^X using DrugBank & FINDSITE^X using ChEMBL) according to their similarity to the respective template ligands. The best of the three rankings gives the final prediction, and is the maximal score of a compound in the ranking of FINDSITE^{filt}, FINDSITE^X using DrugBank and FINDSITE^X using ChEMBL. In practice, the top 1% of ranked ligands of the compound library is considered for further experimental testing.

FINDSITE^{comb} has been benchmarked on the DUD set [26] and is shown to be significantly better than the structure-based docking methods that require experimental structures [27, 28]. It is far faster and is not limited by the availability of experimental structures. FINDSITE^{comb} has also been extensively experimentally verified [29].

In this study, a compound library with molecules from the National Cancer Institute (NCI) and ZINC8 [30] (culled to TC<0.7) as background was used. The NCI molecules (see (http://dtp.nci.nih.gov/branches/dscb/repo_open.html)) library consists of 1597 molecules from the Diversity Set III, 97 from the Approved Oncology Drugs Set IV, and 118 from the Natural Product Set II (total 1812 NCI molecules). Together with the ZINC8 background, FINDSITE^{comb} screened a total of 69,683 molecules. NCI molecules ranked within the top 1% (i.e. higher than 700th) for each target were subsequently considered for experimental validation. The central catalytic domain (NGLY1_D:158-464AA) of NGLY1 was also screened.

2.7. Acquisition and quantification of thermal shift assays

Experimental validation of the VLS hits was carried out using a high-throughput thermal shift assay methodology following established guidelines [17, 31–33]. Briefly, protein samples were dispensed into 96-well PCR plates with sypro-orange, SO, as the extrinsic fluorophore dye and thermal melting was carried out by heating the samples from 25° C–74° C using a 1° C/min heating ramp in a RealPlex quantitative PCR instrument (Eppendorf, NY, USA). SO was kept at 5X concentration and 5 μM of protein was used. The 20 μl reaction mix contained 100 mM HEPES pH 7.3, 150 mM NaCl and the protein. Appropriate dye and protein controls were included in each plate as an internal reference. Each melting curve was assigned a quality score (Q), which is the ratio of the melting-associated increase in fluorescence (F_{melt}) to the total fluorescence range (F_{total}). Q = 1 is a high-quality curve, while Q = 0 indicates no thermal transition [34].

2.7.1. Data analysis—The validity of FINDSITE^{comb}'s top 75 predictions on CDNGLY1 was assessed by the thermal melt assay methodology. The curves were fit to Boltzmann's equation (Eq. 1) to obtain the melting temperature, T_m , from the observed fluorescence intensity, I (Fig 5A–C)

$$I = I_{min} + \frac{[I_{max} - I_{min}]}{1 + e^{\left(\frac{T_m - T}{a}\right)}} \quad (1)$$

where I_{min} and I_{max} are the minimum and maximum intensities; T is the temperature and a denotes the slope of the curve at the unfolding transition midpoint temperature, T_m . Further,

the first derivative of each melting isotherm was derived and fit to a Gaussian whose mean gave an accurate estimate of the T_m (Fig 5B–D). The fluorescence intensity was used to compute the ratio of fraction unfolded (f_u) and fraction folded ($1-f_u$). Approximate thermodynamic parameters were estimated by van't Hoff [35] and Gibbs-Helmholtz analyses [36]. Further, rough estimates of ligand-binding affinity at T_m were computed by employing Equation 2 [37], with slight modifications.

$$K_L(T_m) = \frac{\exp\left\{\frac{-\Delta H}{R}\left(\frac{1}{T_m} - \frac{1}{T_0}\right)\right\}}{[L]}, \quad (2)$$

where K_L is the ligand association constant, $[L]$ is the free ligand concentration at T_m ($[L_{Tm}] \sim [L]_{total}$, when $[L]_{total} \gg$ the total concentration of protein), T_0 is the melting temperature of the protein in the absence of ligand, ΔH is the van't Hoff enthalpy of binding at temperature T and R is the gas constant. K_D is the inverse of $K_L(T_m)$.

3. Results

3.1. Cloning, expression and purification of different constructs of NGLY1

Figure 1 shows the various constructs that were designed to increase the chances of success to get soluble pure protein. Subsequent to gene optimization, gene synthesis was carried out by Genscript (Piscataway, NJ). The codon-optimized, artificially synthesized genes were cloned into pET30a vector and expressed in BL21(DE3) cells. The clones were confirmed by insert release (Fig S1) and sequencing. The various constructs (full-length NGLY1, the R402 truncation variant of NGLY1, the R402del variant of NGLY1, the R542 truncation variant of NGLY1, the catalytic domain of NGLY1 and the R402 truncation variant of the catalytic domain of NGLY1) were expressed. However, only FLNGLY1, the central catalytic domain and R402X FLNGLY1 were purified to homogeneity (Fig S2). The R402delFLNGLY1 and R542XFLNGLY1 constructs, though expressed, resulted in pull down of a lot of contaminating proteins. This could be because possible protein misfolding exposes surface hydrophobicity. However, R402XCDNGLY1 showed no expression whatsoever as assessed by SDS-PAGE gel analysis.

To gain further insights into possible structural changes effected by the truncation of proteins, structure models for representative constructs FLNGLY1 and R402X FLNGLY1 were predicted with high confidence. The predicted TM-scores of the structure predictions for PUB domain was 0.702344, for transglutaminase domain it was 0.665067 and for PAW domain it was 0.747682. It should be noted here that the values of TM-score range from 0 to 1 with 1 meaning perfect model quality. All TM-scores are highly significant. The full length and truncated models are shown in Fig 3. We calculated the accessible surface area (ASA) of the full length NGLY1 model and truncated model (1-402) to compute possible differences in surface hydrophobicity. The ASA for the full length and truncated model was 8875.9 Å² and 8994.3 Å², respectively. The difference in ASA is 118.4 Å² corresponding to the hydrophobic surface area that is exposed. It could be reasonably speculated that this increased hydrophobicity results in undesirable interactions with proteins in the *E. coli*

cytosol, thereby obviating attempts at purifying the aberrant constructs. Lack of ability to purify the disease-causing variants in recombinant protein expression systems is in itself an indication of one possible mechanism for disease causation. It could be hypothesized that, among other possibilities, the truncated variants precipitate and pull down other proteins to form aggregates. However, further careful investigations are required for substantiating this plausible hypothesis.

3.2. Thermal stability assessment for wild-type FLNGLY1 and R402XFLNGLY1

Having successfully purified recombinant FLNGLY1 and R402XFLNGLY1, the proteins were assessed for their thermal stability. This is to ensure their utility for high-throughput experimental differential scanning fluorimetry [17, 31, 33]. Fig 4A–B shows the thermal figure, the stability curves for the wild-type FLNGLY1 and the R402XFLNGLY1. As is seen in the figure, thermal melt curves of FLNGLY1 are multiphasic possibly corresponding to the differential stability of individual domains making up the FLNGLY1. The midpoint of the first transition was around 38 °C, while that of the second transition was 52 °C. Human NGLY1 is a protein with three distinct domains, viz., the peptide:N-glycanase ubiquitin associated domain (also called the protein-protein interaction domain) (PUB), the “Present in PNGase and other worm proteins” domain (PAW) and the transglutaminase domain (TG) [9]. It is very difficult to deconvolute the composite thermal melt curve into individual components corresponding to the three domains.

Thermal melt curves of R402X, (which essentially led to the truncation of the carboxy terminus PAW domain responsible for the recognition of carbohydrate substrate), though marginally better than the wild-type FLNGLY1, gave a broad plateau as seen in the first derivative of the thermal melting curves. This indicates that the N-terminal PUB domain and the central TG domains may still have an order of magnitude difference in their thermal stabilities. Moreover, it should be noted that the first transition in the thermal melt curve dramatically shifted from ~38 °C to ~48 °C indicating better thermal stability for the truncated version as compared to the wild-type. This observation was counterintuitive, since it was expected that the truncated variant would be less stable than the wild-type full-length protein.

To further assess whether this multiphasic thermal melt profile is conducive for carrying out high-throughput experimental thermal melting curves to identify small-molecule binders, we carried out the experiment in the presence of a small molecule, Actinomycin D. Biochemical intuition and presence of several carboxyl oxygens and amino groups, which can serve as hydrogen bonding acceptors and donors, respectively, led to the assumption that actinomycin D might bind to the protein. As is seen from Fig 4 C–D, it is clear that the small-molecule shows binding to the full-length protein. An exact estimate for T_m was hard to obtain because of the inability to deconvolute the spectra into its individual components. However, approximate manual estimates yielded a T_m value of ~ 6 °C indicating reasonable binding. It has to be pointed out here that there is no literature for small-molecule binders for NGLY1.

Lack of unambiguous thermal melt isotherms was discouraging, especially with respect to the utility of the protein for high-throughput thermal melt experiments to discover small-

molecule binders for this protein. To circumvent this problem, DSF experiments were carried out with CDNGLY1.

3.3. Virtual screening for central catalytic domain (amino acid residues 158–464)

FINDSITE^{comb} was employed to carry out virtual ligand screening as delineated in the “Materials and Methods” section (Fig 2). Top first templates for template ligands by FINDSITE^{filt} (that uses experimental holo template proteins and their associated ligands) [12, 18] and FINDSITE^X (that uses predicted holo template protein structures based on known binding ligands) [19] were: 1X3WA from PDB [23] (30% sequence identity to target); *Cytochrome c oxidase subunit 7C, mitochondrial* (11% sequence identity to target from the DrugBank [24]); *Endoribonuclease dicer* (8% sequence identity to target from ChEMBL database), respectively. The set of 76 NCI molecules (http://dtp.nci.nih.gov/branches/dscb/repo_open.html) within the top 1% (700) of the VLS predictions are listed in Table S1. The small-molecules obtained through VLS predictions can be clustered into 22 clusters with a Tanimoto Coefficient [39, 40] cutoff of 0.7. The cluster centroids are also listed in Table 1. A cursory glance at the kinds of ligands picked up by the VLS algorithm shows that the most populated cluster contains 19 small molecules that are cholate-like. The second most populated cluster has 14 small nucleoside analogue like molecules. It should be noted that the former group of compounds contains cyclic scaffolds with hydroxylation while the latter group contain pentose sugar moieties.

3.4. Differential scanning fluorimetry, DSF, based assessment of ligand binding

Before carrying out high-throughput experimental screening, the thermal melt curve for the transglutaminase domain of NGLY1 alone was assessed. As seen in Fig 5, the protein yielded a single unfolding transition in its thermal melt profile with a T_m value of ~38 °C. A total of 76 molecules were screened and yielded 29 interpretable curves. Eight out of the 29 molecules that yielded interpretable thermal melt curves gave positive shifts indicative of a ~27.6 % success rate. Fig 5A shows the thermal melt assay curves for seven novel small molecules, and Fig 5B shows the first derivative of the obtained curves. Actinomycin D (NSC3053) is also included, and as expected, stabilized the CDNGLY1 protein. However, it was not the best binder, which was NSC40275. The T_m values ranged from 0.8 °C to 4.7 °C indicative of reasonable μ M binding affinities (Table 2 and Fig 5A–B). The small-molecule, NSC30275 gave an atypical thermal shift curve (Fig 5C). However, repetition of the thermal shift assay experiment gave a consistent shift. Further, inspection of the first derivative for the thermal melt curve too didn't indicate any anomalous behavior (Fig 5D). To eliminate the possibility of any undesirable interaction of the extrinsic fluorophore dye Sypro-orange, SO, with the small molecules giving rise to aberrant signal, appropriate controls were carried out. The obtained success rate is comparable to that obtained in our previous work on several different therapeutically important proteins [17].

The resulting eight novel hits were castanospermine (NSC614552), antibiotic A-31438 (NSC209870), batyl alcohol (NSC284200), Aleuritic acid (NSC7668), N-ethylanthranilic acid (NSC16162), 3-anilino-2-naphthoic acid (NSC40275), Methyl 2-(benzoylamino)-4,6-o-benzylidene-2-deoxyhexopyranosid-3-ulose (NSC99925), all with reasonable T_m values (Fig 6). The best hit, NSC40275, gave a T_m of 4.66 °C. NSC40275 belongs to the cluster

containing N-ethylanthranilic acid (NSC16162, Rank 364, with a mTC score 0.679591). A detailed analysis on the nature of these molecules and their properties, especially with respect to their binding of NGLY1 is elaborated upon in the discussion.

4. Discussion

Rare diseases are a group of afflictions with mostly genetic underpinnings that affect a tiny fraction of the population. Often, big pharmaceutical companies do not focus on discovering therapies for such ailments given the minimal returns on investment for the research and development costs associated with drug development. Hence, it is contingent upon the research community to help develop therapies for such rare diseases to improve the quality of life for these patients. For most rare diseases, there is minimal information on the underlying biochemistry, and hence, no therapy or treatments exist. NGLY1 deficiency is one such condition that requires considerable research efforts to understand the underlying biochemical basis for a therapeutic solution. A key challenge is to identify small molecule binders of NGLY1 that may be effective at modulating the biological process or disease state by either stabilizing, activating or inhibiting NGLY1.

Experimental high-throughput screening, HTS, of small molecules has been widely applied to meet several objectives in basic and clinical research. Small molecules and their interaction with proteins implicated in disease causation have proven to be indispensable tools in evaluating the improvement or modification of defective functions [41]. Such molecules represent pharmacological leads that can eventually be developed for treating diseases. Innovations in the development of miniaturized, cost-effective, robust and reproducible assay formats is a crucial first step in the identification of potential 'hits' to be optimized as downstream drug candidates. However, the rate limiting steps in such experimental HTS screens has been the absence of reliable virtual ligand screening algorithms that can narrow down the chemical space to make it accessible for cost effective experimental high-throughput screening. In the present study, we have successfully demonstrated the application of our novel virtual ligand-screening algorithm FINDSITE^{comb} to a rare genetic disorder of glycoprotein processing. This study is a proof of concept that, apart from generic application in finding novel protein-ligand interactions [17], demonstrates that VLS methodology can be applied to real-life situations to deduce small-molecules that can lead to possible therapeutic applications. The methodology found eight novel binders with μM affinities, representing a success rate of 28%. In the following paragraphs, we present a brief discussion on the classes of molecules obtained highlighting the non-trivial nature of the obtained interacting small-molecules, thus emphasizing the qualitatively better performance of our VLS over the existing prior-art.

Castanospermine (NSC614552) is a tetrahydroxylated alkaloid isolated from plant sources. It is a potent inhibitor of select glucosidases and possesses antiviral activity [42–46]. It was highly surprising that castanospermine binding was observed for the central transglutaminase domain of NGLY1. It should be noted that previous literature implicates the carboxy terminus PAW domain in binding and recognizing the carbohydrate substrate [47]. However, our results seem to suggest that the central domain of NGLY1 might have a

role in not only carrying out the hydrolysis of the N-linked glycan but also in recognizing the sugar moiety (however weak) by stabilizing the geometry for optimal catalysis.

Batyl alcohol (NSC284200) and aleuritic acid (NSC7668) are important components of medical formulations and cosmetics. However, to the best of our knowledge, no reports exist that show specific interaction of these small-molecules with enzymes. Binding and stabilization of CDNGLY1 by both aleuritic acid and batyl alcohol, which share substantial structural similarity, is non-intuitive and requires sophisticated protein fold and small-molecule similarity metrics. Similarly, 3-anilino-2-naphthoic acid (NSC40275) and N-ethylanthranilic acid (NSC16162) are used in coloring applications. A close inspection of their structures show marked similarities especially vis-à-vis the ortho substitution of either an aryl or alkyl amino group with respect to the carboxylate group on a benzene ring structure. This observation, combined with the juxtaposition of amino and carbonyl groups in methyl-2-(benzoylamino)-4,6-o-benzylidene-2-deoxyhexopyranosid-3-ulose (NSC99925) and dactinomycin (NSC3053) is too compelling to be mere coincidence. We posit that the spatial arrangement of NH and CO groups anchored onto a ringed system might form the pharmacophore enabling their binding to the protein. Further, 2,3,4,5-tetrakis (hydroxymethyl) cyclopentanone (NSC9064) and the antibiotic A-31438 (3-O-alpha-mycarosylerythronolide B; NSC209870) are polyhydroxy compounds mimicking sugar groups. This reinforces our earlier observation that the central transglutaminase domain of NGLY1 is also, possibly, involved in sugar recognition.

Having presented the above analysis, we recognize that the thermal stability imparted by the small-molecule binding is poor and requires considerable efforts to improve the initial hits. However, the compounds represent promising hits that are non-intuitive and non-trivial. In summary, this paper demonstrates the necessity for reliable and novel approaches to carry out virtual ligand screening with particular emphasis on molecular targets involved in/ implicated with the causation of rare disorders and provides a proof-of-concept for NGLY1.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would also like to thank the Developmental Therapeutics Program of the National Cancer Institute for providing the small molecules used in this study.

Funding

This work was supported in part by the Grace-Wilsey Foundation and by the National Institutes of Health, Division of General Medical Sciences grant No. 1R35GM118039.

Abbreviations

NGLY1	N-glycanase
ER	Endoplasmic reticulum
ERAD	Endoplasmic reticulum associated degradation

SDS-PAGE sodium dodecyl sulfate polyacrylamide gel electrophoresis

TSA thermal shift assay

DSF differential scanning fluorimetry

References

- Ohtsubo K, Marth JD. Glycosylation in cellular mechanisms of health and disease. *Cell*. 2006; 126:855–867. [PubMed: 16959566]
- Dwek RA. Biological importance of glycosylation. *Developments in biological standardization*. 1998; 96:43–47. [PubMed: 9890515]
- Roth J. Protein N-glycosylation along the secretory pathway: relationship to organelle topography and function, protein quality control, and cell interactions. *Chemical reviews*. 2002; 102:285–303. [PubMed: 11841244]
- Apweiler R, Hermjakob H, Sharon N. On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochimica et biophysica acta*. 1999; 1473:4–8. [PubMed: 10580125]
- Kleizen B, Braakman I. Protein folding and quality control in the endoplasmic reticulum. *Current opinion in cell biology*. 2004; 16:343–349. [PubMed: 15261665]
- Helenius A, Aebi M. Roles of N-linked glycans in the endoplasmic reticulum. *Annual review of biochemistry*. 2004; 73:1019–1049.
- Hirayama H, Hosomi A, Suzuki T. Physiological and molecular functions of the cytosolic peptide:N-glycanase. *Seminars in cell & developmental biology*. 2015; 41:110–120. [PubMed: 25475175]
- Spiro RG. Protein glycosylation: nature, distribution, enzymatic formation, and disease implications of glycopeptide bonds. *Glycobiology*. 2002; 12:43R–56R.
- Suzuki T. The cytoplasmic peptide:N-glycanase (Ngly1)-basic science encounters a human genetic disorder. *Journal of biochemistry*. 2015; 157:23–34. [PubMed: 25398991]
- Valastyan JS, Lindquist S. Mechanisms of protein-folding diseases at a glance. *Disease models & mechanisms*. 2014; 7:9–14. [PubMed: 24396149]
- Enns GM, Shashi V, Bainbridge M, Gambello MJ, Zahir FR, Bast T, Crimian R, Schoch K, Platt J, Cox R, Bernstein JA, Scavina M, Walter RS, Bibb A, Jones M, Hegde M, Graham BH, Need AC, Oviedo A, Schaaf CP, Boyle S, Butte AJ, Chen R, Chen R, Clark MJ, Haraksingh R, F.C. Consortium. Cowan TM, He P, Langlois S, Zoghbi HY, Snyder M, Gibbs RA, Freeze HH, Goldstein DB. Mutations in NGLY1 cause an inherited disorder of the endoplasmic reticulum-associated degradation pathway. *Genetics in medicine: official journal of the American College of Medical Genetics*. 2014; 16:751–758. [PubMed: 24651605]
- Zhou H, Skolnick J. FINDSITE^{comb}: A Threading/Structure-Based, Proteomic-Scale Virtual Ligand Screening Approach. *Journal of Chemical Information and Modeling*. 2013; 53:230–240. [PubMed: 23240691]
- Dwek RA, Butters TD, Platt FM, Zitzmann N. Targeting glycosylation as a therapeutic approach. *Nature reviews. Drug discovery*. 2002; 1:65–75. [PubMed: 12119611]
- Laemmli UK. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature*. 1970; 227:680–685. [PubMed: 5432063]
- Bradford MM. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Analytical biochemistry*. 1976; 72:248–254. [PubMed: 942051]
- Zhou H, Skolnick J. Template-based protein structure modeling using TASSER^{VM}. *Proteins*. 2011; 80:352–361. [PubMed: 22105797]
- Srinivasan B, Zhou H, Kubanek J, Skolnick J. Experimental validation of FINDSITE^{comb} virtual ligand screening results for eight proteins yields novel nanomolar and micromolar binders. *Journal of Cheminformatics*. 2014; 6:16. [PubMed: 24936211]

18. Brylinski M, Skolnick J. FINDSITE: A threading-based method for ligand-binding site prediction and functional annotation. *Proc Natl Acad Science*. 2008; 105:129–134.
19. Zhou H, Skolnick J. FINDSITE^X: A Structure-Based, Small Molecule Virtual Screening Approach with Application to All Identified Human GPCRs. *Molecular Pharmaceutics*. 2012; 9:1775–1784. [PubMed: 22574683]
20. Zhou H, Zhou Y. Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. *Proteins*. 2005:321–328. [PubMed: 15523666]
21. Skolnick J, Kihara D, Zhang Y. Development and large scale benchmark testing of the PROSPECTOR_3 threading algorithm. *Proteins*. 2004; 56:502–518. [PubMed: 15229883]
22. Zhang Y, Skolnick J. Automated structure prediction of weakly homologous proteins on genomic scale. *Proc Natl Acad Sci (USA)*. 2004; 101:7594–7599. [PubMed: 15126668]
23. Bernstein FC, Koetzle TF, Williams GJB, Meyer EF, B MD Jr, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. The Protein Data Bank: A Computer-based Archival File for Macromolecular Structures. *J Mol Biol*. 1977; 112:535–542. [PubMed: 875032]
24. Wishart D, Knox C, Guo A, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucl Acid Res*. 2006; 34:D668–672.
25. Gaulton A, Bellis L, Bento A, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington J. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucl Acid Res*. 2012; 40:D1100–1107.
26. Mysinger MM, Carchia M, Irwin JJ, Shoichet BK. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J Med Chem*. 2012; 55:6582–6594. [PubMed: 22716043]
27. Ewing TJA, Makino S, Skillman AG, Kuntz ID. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J Comput-Aided Molec Design*. 2001; 15:411–428.
28. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. *Journal of Computational Chemistry*. 2010; 31:455–461. [PubMed: 19499576]
29. Srinivasan B, Zhou H, Kubanek J, Skolnick J. Experimental validation of FINDSITEcomb virtual ligand screening results for eight proteins yields novel nanomolar and picomolar binders. *Journal of Cheminformatics*. 2014; 6:16–29. [PubMed: 24936211]
30. Irwin JJ, Shoichet BK. ZINC – A Free Database of Commercially Available Compounds for Virtual Screening. *J Chem Inf Model*. 2005; 45:177–182. [PubMed: 15667143]
31. Roy A, Srinivasan B, Skolnick J. PoLi: A Virtual Screening Pipeline Based on Template Pocket and Ligand Similarity. *J Chem Inf Model*. 2015; 55:1757–1770. [PubMed: 26225536]
32. Srinivasan B, Skolnick J. Insights into the slow-onset tight-binding inhibition of Escherichia coli dihydrofolate reductase: detailed mechanistic characterization of pyrrolo [3,2-f] quinazoline-1,3-diamine and its derivatives as novel tight-binding inhibitors. *The FEBS journal*. 2015; 282:1922–1938. [PubMed: 25703118]
33. Srinivasan B, Tonddast-Navaei S, Skolnick J. Ligand binding studies, preliminary structure-activity relationship and detailed mechanistic characterization of 1-phenyl-6,6-dimethyl-1,3,5-triazine-2,4-diamine derivatives as inhibitors of Escherichia coli dihydrofolate reductase. *European journal of medicinal chemistry*. 2015; 103:600–614. [PubMed: 26414808]
34. Crowther GJ, He P, Rodenbough PP, Thomas AP, Kovzun KV, Leibly DJ, Bhandari J, Castaneda LJ, Hol WG, Gelb MH, Napuli AJ, Van Voorhis WC. Use of thermal melt curves to assess the quality of enzyme preparations. *Analytical biochemistry*. 2010; 399:268–275. [PubMed: 20018159]
35. John DM, Weeks KM. van't Hoff enthalpies without baselines. *Protein science: a publication of the Protein Society*. 2000; 9:1416–1419. [PubMed: 10933511]
36. LiCata VJ, Liu CC. Analysis of free energy versus temperature curves in protein folding and macromolecular interactions. *Methods in enzymology*. 2011; 488:219–238. [PubMed: 21195230]

37. Lo MC, Aulabaugh A, Jin G, Cowling R, Bard J, Malamas M, Ellestad G. Evaluation of fluorescence-based thermal shift assays for hit identification in drug discovery. *Analytical biochemistry*. 2004; 332:153–159. [PubMed: 15301960]
38. Caspi R, Altman T, Dreher K, Fulcher CA, Subhraveti P, Keseler IM, Kothari A, Krummenacker M, Latendresse M, Mueller LA, Ong Q, Paley S, Pujar A, Shearer AG, Travers M, Weerasinghe D, Zhang P, Karp PD. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucl Acid Res*. 2012; 40:D742–D753.
39. Ballester PJ, Mitchell JBO. A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. *Bioinformatics*. 2010; 26:1169–1175. [PubMed: 20236947]
40. Tanimoto TT. An elementary mathematical theory of classification and prediction. IBM Interanal Report. 1958
41. Gomes CM. Protein misfolding in disease and small molecule therapies. *Current topics in medicinal chemistry*. 2012; 12:2460–2469. [PubMed: 23339300]
42. Walker BD, Kowalski M, Goh WC, Kozarsky K, Krieger M, Rosen C, Rohrschneider L, Haseltine WA, Sodroski J. Inhibition of human immunodeficiency virus syncytium formation and virus replication by castanospermine. *Proceedings of the National Academy of Sciences of the United States of America*. 1987; 84:8120–8124. [PubMed: 2825177]
43. Pan YT, Hori H, Saul R, Sanford BA, Molyneux RJ, Elbein AD. Castanospermine inhibits the processing of the oligosaccharide portion of the influenza viral hemagglutinin. *Biochemistry*. 1983; 22:3975–3984. [PubMed: 6615812]
44. Saul R, Chambers JP, Molyneux RJ, Elbein AD. Castanospermine, a tetrahydroxylated alkaloid that inhibits beta-glucosidase and beta-glucocerebrosidase. *Archives of biochemistry and biophysics*. 1983; 221:593–597. [PubMed: 6404222]
45. Saul R, Ghidoni JJ, Molyneux RJ, Elbein AD. Castanospermine inhibits alpha-glucosidase activities and alters glycogen distribution in animals. *Proceedings of the National Academy of Sciences of the United States of America*. 1985; 82:93–97. [PubMed: 3881759]
46. Whitby K, Pierson TC, Geiss B, Lane K, Engle M, Zhou Y, Doms RW, Diamond MS. Castanospermine, a potent inhibitor of dengue virus infection in vitro and in vivo. *Journal of virology*. 2005; 79:8698–8706. [PubMed: 15994763]
47. Zhou X, Zhao G, Truglio JJ, Wang L, Li G, Lennarz WJ, Schindelin H. Structural and biochemical studies of the C-terminal domain of mouse peptide-N-glycanase identify it as a mannose-binding module. *Proceedings of the National Academy of Sciences of the United States of America*. 2006; 103:17214–17219. [PubMed: 17088551]

Figure 1.

Schematic representation of the various constructs of NGLY1. The first sequence shows the reference sequence for transcript 1 of human full-length NGLY1 with numbers depicting amino acids. The domain boundary demarcations are as reported in Suzuki, 2015[9]. The additional residues in other constructs are coming either from the hexa-histidine tags or from restriction sites (see Supporting Information for the SDS-PAGE image and sequences).

Figure 2. Flowchart of the FINDSITE^{comb} virtual ligand screening method and the high-throughput experimental validation protocol 4e [17].

Figure 3.
Modelling of FLNGLY1 and R402X FLNGLY1.

Figure 4.

Thermal unfolding curves of FLNGLY1 and R402XFLNGLY1 **A)** Primary unfolding curves for the wild-type protein and the truncation mutation. **B)** Gaussian fit of the first-derivative for curves in (A). **C)** Primary unfolding curves for wild-type FLNGLY1 in the presence of the small-molecule, NSC3053. **D)** Gaussian fit of first-derivative for curves in (C). On the plots A and C, the y-axis represents the normalized fluorescence and the x-axis represents the temperature in $^\circ\text{C}$ while in curves B and D, the y axis represents the first derivative $dF/d(T)$ and the x-axis represents the temperature in $^\circ\text{C}$. The top panel of the figure shows the domain organization of the wild-type FLNGLY1 protein. The experimental data points were fit to the respective equations using the non-linear curve-fitting algorithm of GraphPad Prism v 6.0e.

Figure 5.

Thermal unfolding curves of CDNGLY1 **A)** Primary unfolding curves for CDNGLY1 in the presence of small-molecules that were screened from FINDSITE^{comb} VLS predictions and yielded positive shifts. **B)** Gaussian fit of first-derivative for curves in (A). **C)** Atypical primary unfolding curves for CDNGLY1 in the presence of the small-molecule, NSC40275, that was obtained from the screened FINDSITE^{comb} VLS predictions and yielded positive shifts. **D)** Gaussian fit of first-derivative for curves in (C). The experimental data points were fit to the respective equations using the non-linear curve-fitting algorithm of GraphPad Prism v 6.0e.

Figure 6.

Structure of small molecules showing binding to CDNGLY1 as assessed by thermal shift assay methodology. The SDF files for the small molecules were downloaded from Pubchem (<http://pubchem.ncbi.nlm.nih.gov>) and the figure was generated using ChemBioDraw 14.0.

Table 1

Classification of the small-molecule hits obtained through VLS predictions that form 22 clusters with a Tanimoto Coefficient (TC) cutoff of 0.7.

fdRank	mTC score	ID (NSC/CAS)	Cluster size	Name
8	0.89668	1614	19	BHDTACA ¹
134	0.723355	127947	1	Ethyl 2-cyano-3-imino-3-(2-oxooctahydro-1-benzofuran-3-yl)propanoate
154	0.717943	280594	1	ADMBDRPDP ²
160	0.716085	20192	7	Streptovaricin complex, fraction c
164	0.714849	69-74-9	14	Cytarabine
190	0.709878	2561	2	Benzyl .beta.-d-arabinopyranoside
221	0.701268	47619	1	1,3-Di-tert-butyl-2-thiourea
282	0.691339	5451-09-2	1	Aminolevulinic
364	0.679591	16162	2	N-Ethylanthranilic acid
400	0.675179	317003	1	TDEAHMM ³
416	0.673912	227309	1	Chlorobiocin
457	0.668769	81660	1	Reserpinolic acid
487	0.665655	14974	5	Glaucarubol
496	0.664494	67546	1	(+)-(S)-Boldine
533	0.66046	99925	2	Khelline-quinone
548	0.659466	15133	5	Glaucarubol
579	0.65685	153365	1	Xylopyranoside, methyl 3-deoxy-3-(3-methyl-3-nitrosoureido), .beta.-D-
595	0.655579	6731	7	Dicyclohexyl phthalate
615	0.653608	15571	1	Antineoplastic-15571
627	0.652109	407628	1	WLN: T C666 A GVN LM&TTJ
676	0.647523	14540	1	WLN: T56 BM DN FNVNVJ C3 F1 H1
681	0.647059	47617	1	1,3-Di-tert-butyl-2-thiourea

¹ 3.beta.-hydroxy-7,11-dioxo-4,4,14-trimethyl-5.alpha.-cholan-24-oic acid;

² 3-Amino-1,5-dihydro-5-methyl-1-.beta.-D-ribofuranosyl-1,4,5,6,8-pentaazaacenaphthylene 5'-(dihydrogen phosphate);

³ 9H-Thioxanthen-9-one, 1-[[2-(dimethylamino)ethyl]amino]-7-hydroxy-4-methyl-, monohydriodide

Table 2

Summary of virtual ligand screening, thermal shift assay and binding parameters for the hits obtained on *CDNGLY1*.

Identity	Rank	Q [#]	T _m (° C)	T _m (° C)	K _D (μM) ^b
Protein	NA	1.0	37.23	NA	NA
NSC3053	NA	0.1 ^c	40.18	2.95	88.5
NSC40275	451	0.7	41.89	4.66	32.9
NSC614552	534	1.2	40.07	2.84	94.3
NSC209870	21	0.9	39.55	2.32	127.7
NSC7668	11	0.7	39.3	2.07	150.9
NSC99925	533	1.1	39.14	1.91	162.3
NSC284200	57	1.2	38.61	1.38	221.5
NSC9064	15	1.1	38.24	1.01	275.3
NSC16162	364	0.4	38.01	0.78	315.3

[#] quality score (Q) is the ratio of the melting-associated increase in fluorescence (F_{melt}) and total range in fluorescence (F_{total}). A Q value of 1 represents a high-quality curve, while a value of 0 shows an absence of melting as described earlier [34].

^b K_D is the dissociation constant computed from the magnitude thermal shifts obtained relative to the protein alone.

^c In spite of the low Q value for NSC3053, the absolute fluorescence value is 59000, way above the baseline of negative fluorescent value of the dye sypro orange with increasing temperature.