Journal of proteome research

ARTICLE

pubs.acs.org/jpr

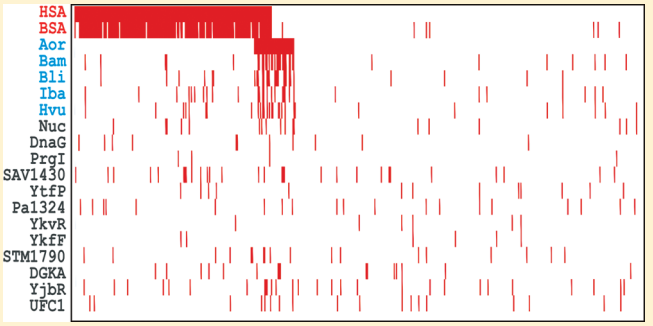
Correlation between Protein Function and Ligand Binding Profiles

Matthew D. Shortridge, Michael Bokemper, Jennifer C. Copeland, Jaime L. Stark, and Robert Powers*

Department of Chemistry, University of Nebraska-Lincoln, Lincoln, Nebraska 68588-0304, United States

Supporting Information

ABSTRACT: We report that proteins with the same function bind the same set of small molecules from a standardized chemical library. This observation led to a quantifiable and rapidly adaptable method for protein functional analysis using experimentally derived ligand binding profiles. Ligand binding is measured using a high-throughput NMR ligand affinity screen with a structurally diverse chemical library. The method was demonstrated using a set of 19 proteins with a range of functions. A statistically significant similarity in ligand binding profiles was only observed between the two functionally identical albumins and between the five functionally similar amylases. This new approach is independent of sequence, structure, or evolutionary information and, therefore, extends our ability to analyze and functionally annotate novel genes.



KEYWORDS: protein function, ligand binding, NMR ligand affinity screen, functional genomics, functional annotation

INTRODUCTION

The recent explosion in sequenced genomes has revealed a vast number of proteins that lack a functional annotation.[1] Many of these unannotated proteins may play an important role in human disease and, correspondingly, are critical for developing new therapeutics. Protein sequence and structure similarity methods are currently the most robust and widely used tools to annotate a protein of unknown function.[2] Nevertheless, these methods are limited in scope, prone to errors, and based on a small set of experimentally characterized proteins.[3] Only 40−60% of sequences suggest a potential functional assignment. Moreover, error rates of <30% occur even with conservative sequence identities of >60%. The accuracy of functional annotations decreases substantially in the twilight zone of 20−35% sequence identity.

Recent attempts to extend functional prediction beyond global sequence and structure similarity have led to the development of active-site similarity search methods.[4−7] These methods try to identify protein surface structures that interact with biologically important ligands since active-sites that share a similarity in sequence, structure and ligand binding are predicted to be functionally related. This is based on the fundamental principal that a protein's active-site has been optimized by nature to interact with a unique and specific set of targets, where this information can be leveraged to understand function. Consequently, protein surfaces have been shown to be exquisitely selective and to only bind ligands at very specific functionally relevant locations.[8−11] This understanding is also essential to drug discovery, where extensive resources are allocated by the pharmaceutical industry to identify high-affinity and selective compounds that target a specific therapeutically relevant protein.[12,13] The use of ligands as functional probes is the basis of our FAST-NMR methodology[4,14] that has been successfully applied to explore the function of *Staphylococcus aureus* protein SAV1430,[4] *Pseudomonas aeruginosa* protein PA1324,[15] *Pyrococcus horikoshii* OT3 protein PH1320,[14] human protein Q13206,[14] *Bacillus subtilis* protein YndB,[16] and *Salmonella typhimurium* PrgI protein.[17] Similar successes have been reported using ligand binding to infer function in virtual screens.[18,19] While promising, current active-site similarity techniques still rely on high-resolution protein structures to identify and measure functional similarity.[20] The availability of structures for the entire proteome remains a significant bottleneck for the high-throughput functional annotation of hypothetical proteins.

We report herein a new method to infer protein function that is independent of sequence and structural information. Our method uses a similarity in ligand binding profiles to annotate a protein of unknown function. This is similar in concept to the mapping of pharmacological space or the use of structure−activity relationships (SAR) for target selection and chemical lead identification in drug discovery.[21−24] A ligand binding profile is defined as a set of ligands that bind a protein from a high-throughput ligand affinity screen. Ligand binding is monitored using our 1D ^{1}H NMR line-broadening screen.[25] In essence, the chemical and structural diversity of a compound library provides an experimental means of mapping the physiochemical properties of a protein's active-site based on the compounds that do or do not bind the protein. Functional annotation is inferred by clustering unknown proteins with previously annotated proteins that share similar ligand binding profiles from the same chemical library. A modification of the E-value routinely used in sequence

Received: January 7, 2011
Published: March 03, 2011

homology is used to quantify ligand binding profile similarities. The methodology is demonstrated using 19 proteins with a range of function defined by Gene Ontology (GO) terms.[26]

EXPERIMENTAL SECTION

Materials

The human serum albumin (HSA) (essentially fatty acid free, ≥ 96% pure), bovine serum albumin (BSA) (minimum 98% agarose gel electrophoresis, lyophilized), α-amylase from *Bacillus lincheniformis* (Bli) (500−1500 units/mg protein, 93−100% (SDS page)), α-amylase from *Aspergillus oryzae* (Aor) (powder, ∼30 units/mg), α-amylase from *Bacillus amyloliquefaciens* (Bam) (liquid, ≥250 units/g protein), β-amylase from barley (Hvu) (type II−B 20−80 units/mg protein), and β-amylase from sweet potato (Iba) (Type I−B, ammonium sulfate suspension, ≥750 units/mg protein) protein samples were all purchased from Sigma (St. Louis, MO). The *S. typhimurium* PrgI protein samples and assigned $^{1}H-^{15}N$ HSQC spectrum were generously provided by Dr. Roberto DeGuzman (University of Kansas). *Staphylococcus aureus* primase C-Terminal domain (CTD) protein sample was purchased from Nature Technologies Corporation (Lincoln, NE). *H. sapiens* diacylglycerol kinase alpha (DGKA), *P. aeruginosa* unannotated protein PA1324, *S. aureus* unannotated protein SAV1430, *S. typhimurium* unannotated protein STM1790, *H. sapiens* ubiquitin-fold modifier-conjugating enzyme 1 (UFC1), *E. coli* unannotated protein YjbR, *E. coli* unannotated protein YkfF, *B. subtilis* unannotated protein YkvR and *E. coli* unannotated protein YtfP protein samples were provided by Dr. Gaetano Montelione, Director of the Northeast Structural Genomics Consortium (NESG, www.nesg.org). The *S. aureus* nuclease was overexpressed in house from a cell stock of *E. coli* Bl21 DE3 codon+ (Stratagene) containing the pET28-(a)+plasmid with the *dnuc* gene provided by Dr. Greg Somerville (University of Nebraska-Lincoln) grown in LB broth and purified using a Talon cobalt affinity resin (Clontech). The deuterium oxide (99.9 atom % D) and the dimethyl sulfoxide-d_6 (99.9 atom % D) were purchased from Aldrich (Milwaukee, WI) The 3-(trimethylsilyl)propionic acid-2,2,3,3-d_4 (TMSP-d_4) was purchased from Cambridge Isotope (Andover, MA). The Bis-Tris-d_{19} (98 atom % D) was purchased from Isotec (Milwaukee, WI). The compound library was previously complied as described elsewhere.[27]

NMR Data Collection and Sample Preparation

All NMR data was collected on a Bruker 500 MHz Avance spectrometer (Billerica, MA) equipped with a triple resonance, *Z*-axis gradient cryoprobe and using a Bruker BACS-120 sample changer and IconNMR software for automated data collection. The screening data for this study was compiled over a 5 year time span in which two different 1D ^{1}H solvent suppression pulse sequences were used for the measurement of ligand 1D ^{1}H NMR line broadening. Data for the HSA, BSA, *S. aureus* primase CTD, PrgI, PA1324, and SAV1430 were collected as previously described.[4,15,17,25] Data for DGKA, STM1790, UFC1, YjbR, YkfF, YkvR and YtfP, the 5 amylases and *S. aureus* nuclease proteins was collected at 298 K using 64 transients with a spectrum width of 6009 Hz with 8 K data points and a 1.0 s relaxation delay using the excitation sculpting[28] method for solvent suppression of the residual H_2O resonance signal. The samples for the HSA, BSA, *S. aureus* primase CTD, PrgI, PA1324, and SAV1430 NMR screens were prepared as previously described.[4,15,17] *S. aureus* nuclease, DGKA, STM1790, UFC1, YjbR, YkfF, YkvR, YtfP, and the 5 amylases were screened at 5 μM protein concentration and 100 μM ligand concentration in a screening buffer of 2% DMSO-d_6, 20 mM Bis-Tris-d_{19} pH 7.0 (uncorrected), 11.1 μM TMSP-d4 in "100%" D_2O.

Chemical Library

All NMR ligand affinity assays were completed by screening each protein individually with a library of 437 biologically active compounds (http://bionmr-c1.unl.edu/ligands).[27] The library contains amino-acids, carbohydrates, cofactors, fatty-acids, hormones, inhibitors, known drugs, metabolites, neurotransmitters, nucleotides, and substrates. The compound library is divided into 116 mixtures with 3−4 ligands per mixture and is described in detail elsewhere. In order to assess the structural diversity of the library, 1300 molecular descriptors were calculated for each compound using the online software eDragon (VCClabs, http://www.vcclab.org/lab/edragon/).[29] MM2 minimized 3D MOL2 files were generated using ChemBio 3D Ultra 12.0 (CambridgeSoft, Cambridge, MA), converted to SMILES using OpenBabel (http://openbabel.org) and then uploaded to the eDragon Web site. The molecular descriptors calculated for each structure were incorporated into a single Excel spreadsheet and imported into SIMCA (UMETRICS, Kinnelon, NJ). Each molecular descriptor was treated as a separate bin or data point for each structure. A 3D PCA scores plot was generated using the calculated molecular descriptors for the structures in the library.

False positive and false negative rates were simulated to determine if the screening library of 437 compounds is of sufficient size to make meaningful comparisons between proteins of unknown function. An in-house program was written that randomly generates a ligand binding profile using a Gaussian distribution about two means: (i) average hit rate of 32 ± 44 bound ligands, or (ii) a lower hit rate of 16 ± 6. Either 1×10^6 random pairs of ligand binding profiles were generated or a single randomly generated ligand binding profile was compared against a random set of 1×10^6 ligand binding profiles. The simulations were done in triplicate and the library sizes used in the simulations corresponded to 437, 1000, 2000, 5000, and 10 000 compounds. An E-value of $\leq 1 \times 10^{-9}$ was used to define a similar ligand binding profile. A histogram of the Log(E-values) were plotted and fitted using EasyFit V5.4 (MathWave Technologies).

To estimate a false negative rate, an error was introduced to randomly generated pairs of identical ligand binding profiles. Each ligand binding profile has false binders added or true binders removed at a percentage of the rate that a true binder was added to the original ligand binding profile (based on the original number of predicted binders (m and n) chosen from the Gaussian distribution):

$$m_e = m_o \pm em_o \text{ and } n_e = n_o \pm en_o \quad (1)$$

where e is the error rate (10−50%), m_e and n_e ($m_e \neq n_e$) are the new number of bound ligands after the error rate is applied, and m_o and n_o ($m_o = n_o$) are the original number of bound ligands predicted from the Gaussian distribution.

Binding Assay

Ligand binding was manually identified from a decrease in the free ligand 1D ^{1}H NMR signal upon the addition of protein. This decrease is determined by visually comparing ligand peak intensities to the TMSP-d_4 methyl resonance (0.00 ppm) from

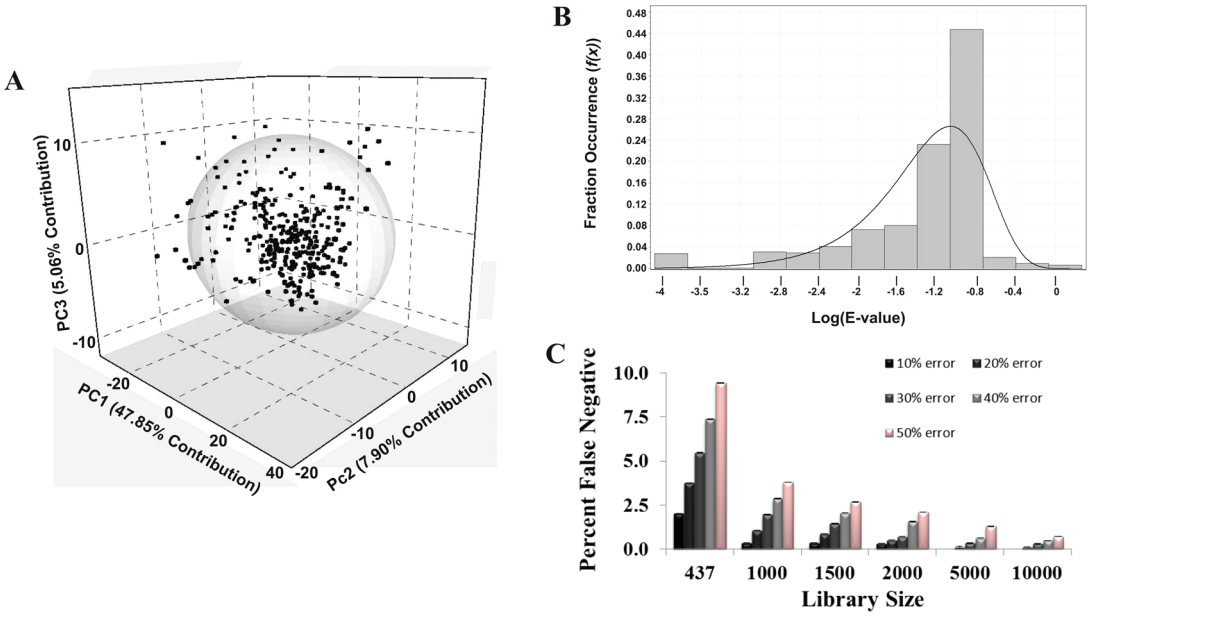


Figure 1. (A) Three dimensional PCA scores plot where each point represents one compound from the functional chemical library. The placement of each point in the PCA scores plot is indicative of the unique structural identity for each compound. The contribution of each principal component is labeled on the axis. The sphere represents the 95% confidence limit. (B) Histogram distribution of E-values calculated from a simulation of ligand binding profiles. A random ligand-binding profile was compared against a random set of 1×10^6 ligand binding profile using a library of 437 compounds. The solid line corresponds to the best fit curve from the Weibull Distribution (Extreme Value Type III Distribution) model. (C) Plot of the percentage of false negatives as a function of error rate (10−50%) and library size (437−10 000) from a simulation of ligand-binding profiles.

the 11.1 μM TMSP-d_4 internal standard. Any ligand with a visually observable decrease in peak height from the addition of a protein is considered to be a binder. A detailed analysis of the relationship between K_D and NMR line-broadening has been previously discussed in detail.[30] From this analysis, a conservative estimate of our limit of detection can be made, which corresponds to ligands with a K_D of >100−300 μM. Of course, this limit is dependent on the molecular weight of the protein, where sensitivity increases with MW. Thus, our ligand binding assay will be dominated by biologically relevant protein−ligand interactions, where nonspecific or irrelevant interactions start to dominate as the K_D increases beyond 300 μM.[31] Conversely, tight-binders ($K_D \leq$ nM) that are governed by slow-off rates may simply result in a decrease in peak intensity proportional to the limiting protein concentration. A 5% change in peak intensity may be difficult to decipher and correspond to a false positive. Nevertheless, encountering tight binders in ligand binding assay is generally a rare event. Binders from our chemical library are typically structural homologues to the natural ligand. Also, these tight binders would be expected to be uniformly missed for functionally similar proteins. The methods for data processing and identifying binding ligands have been previously discussed in detail.[25,27,30] Overall, for our library of 437 compounds, the 1D ^{1}H NMR line-broadening screen requires approximately a day to complete both the data acquisition and the data analysis.

Ligand Binding Profiles

A similarity in ligand binding profiles was measured between each pair of proteins using eq 1. Overlapping binding ligands (S) for every protein in a pairwise manner were identified by comparing a list of all binding ligands and counting the number of overlapping ligands. Each pairwise E-value was calculated using a library size of 437 compounds ($p' = 1/437 = 0.00229$). An Excel spreadsheet program was written to match overlapping ligands and measure E-values.

Functional Similarity Measurement

The Uniprot accession number was obtained for each protein in the study. The list of Uniprot accession numbers was uploaded to the semantic similarity tool FunSimMat (http://funsimmat.bioinf.mpi-inf.mpg.de/). All reported functional similarities are expressed as a *funsim* score measured as previously described.[32]

RESULTS AND DISCUSSION

Structural Diversity of the Screening Library

Our chemical library for NMR ligand affinity screening was designed to maximize functional diversity.[27] In addition to practical considerations such as solubility, stability and cost, compounds were added to our library based on a known biological activity involving a distinct protein or protein class. Compounds correspond to known drugs, inhibitors, substrates or cofactors. Not surprisingly, the compounds are also consistent with typical "drug-like" characteristics and with fragment libraries.[33,34] These characteristics include good aqueous solubility, low molecular-weights, and low number of rings, heteroatoms, and hydrogen-bond donors and acceptors. Diversity in biological activity was also anticipated to result in a correlated diversity in chemical structure. To validate the structural diversity of our functional chemical library, ~1300 different molecular descriptors were calculated for each compound.[29] A principal component analysis (PCA) of the set of molecular descriptors indicates a uniform coverage of structural space. A 3D PCA scores plot is shown in Figure 1A. The structures are distributed throughout the structural space defined by the molecular descriptors. Conversely, if there was an overabundance of any structural class, distinct clustering patterns would be apparent in

the 3D PCA scores plot. Clearly, our chemical library is an acceptable set of molecular probes to evaluate a diversity of protein function.

Calculation of Ligand Binding Profile Similarities

Measuring a significant similarity between two ligand binding profiles requires the development or adaptation of a robust scoring function. Current similarity scoring methods used for sequence analysis, such as the E-value developed by Karlin and Altschul,[35] are also well-suited for measuring a similarity between ligand binding profiles.

$$E = Kmne^{-\lambda S} \tag{2}$$

Here, the E-value is only dependent on the total number of compounds that bind each protein (m and n) and the total number of compounds that bind both proteins (S). Additionally, the probability of finding a significant similarity is proportional to the probability search space (K) and scoring function (λ).

$$K = \frac{(q - p')^2}{q} \text{ and } \lambda = \ln \frac{q}{p'} \tag{3}$$

Unlike sequence similarity, a similarity between ligand binding can be thought of as a binary system (binding vs nonbinding) therefore the probabilities p' and q simply becomes the probability of finding a hit within a library:

$$p' = \frac{1}{\text{library size}} \tag{4}$$

and the probability of finding a ligand that binds both proteins:

$$q = \frac{S}{m \times n} \tag{5}$$

The standard E-value also provides a robust measure of the probability that the ligand binding similarity is not due to chance using the standard P-value.

$$P = 1 - e^{-E} \tag{6}$$

As expected, the ligand binding profile E-value rapidly becomes insignificant ($P > 0.0001$) as the probability of finding a ligand that binds both proteins (q) decreases. Binding profiles that have a $P < 0.0001$ are significant at the 99.99% confidence interval ($E = 10^{-5}$). Thus, our method is only dependent on comparing the total number of binding events (m or n) and the set of overlapping binding ligands (S) between two proteins.

Sufficient Size of a Screening Library

Obtaining a balance between library depth and breadth is very challenging and has been a focus of compound library design for over a decade—without a clear consensus conclusion.[36] Clearly, the size of the library would be expected to impact the number of observed binders (m and n) and the corresponding similarity in ligand binding profiles (S and E-value). Fundamentally, determining the optimal size of the chemical library is an open-ended, and at some level, a very difficult question to adequately answer. It is always plausible for a protein to be screened that results in a complete absence of binders regardless of the size or composition of the chemical library. If the protein is a true unknown, how is it possible to ascertain *a priori* that the library composition is adequate? The only recourse is to explore the probability of identifying binders within a given set of reasonable assumptions and given experimental hit rates.

On average, 32 ± 44 ligands were observed to bind a protein target in our NMR ligand affinity screen. Our simulations indicate that even with a modest library size of 437 compounds, the probability of randomly finding two similar (E-value $\leq 1 \times 10^{-9}$) ligand binding profiles was shown to be effectively zero. This is not too surprising considering that in theory there are 2^{437} (3.5×10^{131}) different binding profiles, where the product (1.3×10^{263}) leads to an effectively miniscule probability of finding two similar ligand binding profiles. Of course, only a small subset of these potential ligand binding profiles are possible given 32 ± 44 bound ligands, but this still represents a very large number of dissimilar pairs of ligand binding profiles. A randomly selected ligand binding profile using a Gaussian distribution of bound ligands with a smaller mean (larger potential false positive rate) of 16 ± 6 was compared against a random set of 1×10^6 ligand binding profiles using the same Gaussian distribution to select binders. A histogram of the Log(E-values) is shown in Figure 1B and best fitted with the Weibull Distribution (Extreme Value Type III Distribution), which indicates the calculated E-values are significant.[37] Consequently, the comparison did not yield any significant similarities and the most common occurrence was an overlap (S) of zero (S ranged from 0 to 7).

While the false positive rate is effectively zero, a false negative rate was measurable and, as expected, decreased for increasing library size. Again, a total of 1×10^6 pairs of *identical* ligand binding profiles ($m = n = S$) was randomly generated using a Gaussian distribution with a mean of 16 ± 6 bound ligands (m and n). An error rate ranging from 10%-50% was introduced into each ligand binding profile, independently changing the two identical ligand binding profiles. The simulations were repeated for library sizes that ranged from 437 to 10 000 compounds. The percentage of false negatives (E-value of $>1 \times 10^{-9}$) found in each simulation are plotted as a function of library size in Figure 1C. The false negative rate increases proportional to the error rate and decreases proportional to the library size. For our library of 437 compounds, the percentage of false negatives is ~9% with a 50% error rate (see eq 1). Conversely, only a ~2% false negative rate is observed for a library of 2000 compounds at the maximum error rate of 50%. The false negative rate is below 1% for a library of 10 000 compounds. Correspondingly, ligand binding profile similarities are relatively tolerant to erroneous binders. This is consistent with the lack of any false negatives in the 19 screens reported herein. Thus, the simulations indicate that even a modest library of 437 compounds provides a relatively robust and reliable measure of functional similarity, but a slight increase in the library size may improve the methods accuracy. Of course, increasing the library size also increases assay time, but a library of 1500−2000 compounds is still practical since the assay time is only estimated to increase to ~1.5−2 days.

The library size also defines the minimal number of binders (m and n) and overlapping binders (S) required for obtaining a significant E-value of 1×10^{-9}. For a modest library of 437 compounds, the minimal number of binders and overlapping binders is 5 compounds. The number drops to 4 compounds for a library size of 1000−2000 compounds and to 2 compounds for a library size of 5000−10 000. Considering the average number of binders is 32 ± 44, these are effectively inconsequential improvements for a substantial increase in screening time. Alternatively, false negatives in the binding assay (missed tight binders) may be potentially detrimental to proteins that bind a very limited number of ligands (<5). In principal, a single false negative may be the difference between a significant or insignificant E-value.

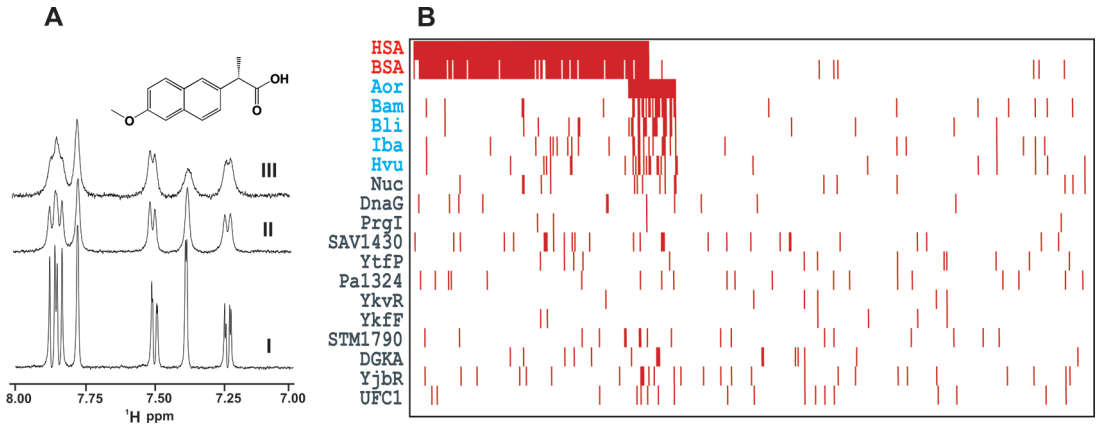


Figure 2. (A) Ligand binding is identified by a decrease in ligand peak intensity upon addition of a target protein. The 1D ^{1}H NMR spectrum of the nonsteroidal anti-inflammatory drug naproxen (I) is shown to broaden in the presence of *H. sapiens* serum albumin (HSA) (II) and *B. taurus* serum albumin (BSA) (III) indicating a positive binding event. The NMR line broadening experiments used 100 μM ligand and 5 μM protein as described in the methods section. (B) Heat map summarizing the NMR ligand affinity screens for 19 proteins, where the albumins are colored red, the amylases cyan and the remainder of the proteins gray. A binding ligand is indicated by a red line. The 437 ligands were sorted to maximize the clustering of binding ligands for the albumins and amylases.

Of course, the number of binders is expected to scale with the library size, assuming a relatively constant hit rate.[38] Correspondingly, increasing the library size to 1500−2000 compounds is expected to make proteins that bind only four or less ligands a relatively rare event.

Correlating Protein Function with Ligand Binding Profiles

To experimentally support the ligand binding profile hypothesis, 19 proteins were screened by NMR using our chemical library of biologically active compounds.[27] Binding events were identified as previously described by measuring a decrease in ligand ^{1}H NMR peak intensities in the presence of a protein (Figure 2a).[4,25] Thus, the ligand binding profile is simply a binary list that indicates which compounds out of the library of 437 compounds were shown to bind the protein. The complete summary of results from the NMR ligand affinity screen for the 19 proteins can be found in Table 1S (Supporting Information).

For the 19 proteins screened in the NMR ligand affinity assay, 13 proteins have a previously annotated function based on GO terms and 6 proteins have an unknown function. The 19 proteins were chosen to contain two sets of functionally similar proteins mixed with a third set of functionally diverse proteins. The two sets of functionally related proteins are 2 serum albumins and 5 amylases. The serum albumins and amylases were chosen because the proteins have a function related to ligand binding and were readily available from commercial sources. The additional 12 proteins are from NESG or other ongoing functional annotation projects involving our FAST-NMR methodology.[4,14] The primary intent of these additional proteins is to provide a "functional background" to test the ability of the ligand binding profile to distinguish the serum albumins and amylases from each other and from the remaining proteins. Will the addition of the 12 functionally diverse proteins cause erroneous similarities to the albumins or amylases that is not correlated with function?

A FunSimMat functional similarity score was calculated for each pair of proteins within the set of 19 proteins.[32] FunSimMat uses GO terms to generate a semantic similarity score that ranges from 0 for no functional similarity to 1 for identical functions. An average FunSimMat similarity score of 0.98 and 0.67 ± 0.04 was calculated between the albumins and amylases, respectively. The remaining 12 proteins exhibited no functional relationship to any other protein in the screening set, yielding an average FunSimMat similarity score of 0.1 ± 0.1. The complete list of FunSimMat similarity scores can be found in Table 2S (Supporting Information). A weak functional similarity was observed between the two albumins and the human protein ubiquitin-fold modifier-conjugating enzyme 1 (UFC1, Uniprot: Q9Y3C8). However, this similarity is limited to one overlapping and generic "protein binding" GO number (GO:0005515).

An all-vs-all pairwise comparison of the 19 ligand binding profiles gave a total of 171 ligand binding profile comparisons with only 11 comparisons giving a significant similarity score ($P <$ 0.0001). The comparisons with the highest similarity scores corresponded to the set of albumins (E-value 1×10^{-58}) and

Table 1. Functionally Similar Proteins Yield Significantly Similar Ligand Binding Profiles[a]

Comparison	*m*/*n*	*S*	E-value	Funsim score
HSA-BSA	178/171	162	2.16×10^{-58}	0.98
Bam-Aor	35/36	22	6.38×10^{-19}	0.68
Bam -Hvu	35/29	14	1.17×10^{-10}	0.63
Bli- Aor	28/36	18	1.19×10^{-15}	0.68
Bli - Bam	28/35	16	1.42×10^{-14}	0.68
Bli - Hvu	28/29	9	3.86×10^{-06}	0.63
Hvu - Aor	29/36	13	2.98×10^{-08}	0.64
Iba- Aor	29/36	12	2.98×10^{-08}	0.67
Iba - Bam	29/35	15	7.56×10^{-12}	0.63
Iba - Bli	29/28	11	2.43×10^{-08}	0.63
Iba - Hvu	29/29	12	2.98×10^{-08}	0.71

[a] Number of hits per protein (*m* and *n*), overlapping ligands (*S*), E-values and functional similarity scores (FunSim) are reported for significantly (99.99% confidence interval) similar ligand binding profiles from a comparison of 19 proteins, including a set of serum albumins from *H. sapiens* (HSA) and *B. taurus* (BSA) and amylases (Aor, Bam, Bli, Hvu, and Iba) gave significant similarity. The set of amylases was composed of 3 α-amylases from *A. oryzae* (Aor), *B. amyloliquefaciens* (Bam), *and B. licheniformis* (Bli) and 2 β-amylases *H. vulgare* (Hvu) and *I. batatas* (Iba). A complete list of binding profiles is reported in Supplementary Table 1 (Supporting Information).

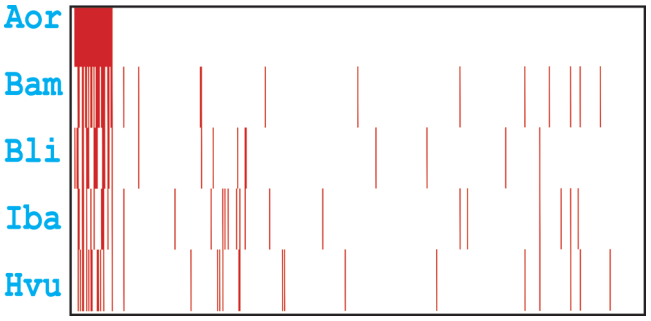


Figure 3. Expanded view of the heat map shown in Figure 2 highlighting the similarity in ligand binding profiles for the amylases. The 437 ligands were sorted to maximize the clustering of binding ligands for the amylases.

the set of amylases (median E-value 3×10^{-13}). Conversely, the median E-value for the remaining ligand binding comparisons was 0.1. For comparison, a median E-value of 3×10^{-26} was obtained when all the ligand binding profiles are compared to themselves. These results clearly indicate that ligand binding profile similarities are strongly correlated with functional similarity. All the protein pairs with a significant ligand binding similarity score along with the corresponding FunSimMat functional similarity score are listed in Table 1. The complete list of ligand binding similarity scores can be found in Table 3S (Supporting Information). It is also important to note that the absolute magnitude of a ligand binding profile E-value is directly dependent on the total number of ligands shown to bind a protein. This is equivalent to sequence homology where the E-value scales by the length of the protein sequences.

The overall similarity in the ligand binding profiles is also easily visualized in a heat map (Figure 2b). A ligand identified by NMR to bind a protein is simply indicated by a red line in the heat map. The ligands are sorted first by their ability to bind human serum albumin (HSA) and then by their binding to *A. oryzae* α-amylases (Aor). An expansion of the heat map focused on the amylases and sorted by ligand binding to Aor is also shown in Figure 3. The heat map clearly shows overlapping clusters of ligands between the albumins and amylases. The remainder of proteins exhibits no similarity in the ligand binding profile based on the obvious random scatter in the heat map.

There was also a minimal similarity in ligand binding between *S. aureus* nuclease and the α-amylases from *A. oryzae* and *B. amyloliquefaciens* (median E-value 4×10^{-5}). It is plausible that this minimal similarity is simply due to a serendipitous overlap in nonspecific ligand binding between the three proteins. However, the similarity in the ligand binding profiles was limited to the nucleosides in the library. Additionally, the remaining 3 amylases did not bind these ligands or exhibit a significant ligand binding similarity to nuclease. The observed ligand binding similarity between the nuclease and two of the α-amylases is potentially due to trace amounts of a nuclease that may be present in the *A. oryzae* and *B. amyloliquefaciens* α-amylases samples. This is a likely occurrence since the samples were purchased as crude mixtures, where size-exclusion chromatography only yielded a modest improvement in purity. This illustrates an important consideration in the general application of ligand binding profiles. False positives in the ligand affinity assay due to impurities, nonspecific binding, or experimental concerns (precipitation, aggregation, etc.) may lead to an inaccurate functional assignment. Proper care in the execution and analysis of ligand binding profiles should minimize these concerns.

As shown in Table 1, HSA and BSA had a large number of binding ligands (178 and 171, respectively) compared to the overall size of the library. The relative hit rate for these two proteins was 40.7 and 39.1% respectively. With a large hit rate, false similarities may arise if a second protein serendipitously binds to a small subset of compounds that were shown to bind HSA or BSA. However, the ligand binding similarity score (see eq 2) effectively eliminates this concern by scaling the score based on both the total number of compounds found to bind each protein and by the number of overlapping binding ligands. As an example, the *S. typhimurium* type III secretion system protein PrgI bound to a total of five compounds, where each compound was also shown to bind HSA and BSA. The corresponding E-values for the ligand binding profile comparisons between PrgI and HSA (7×10^{-2}) and BSA (6×10^{-2}) were not significant at a $P = 0.0001$.

Ligand binding profiles are independent of sequence and structural information and thus provide an experimentally based approach to predict protein function in a relatively robust and high-throughput fashion. The results reported herein demonstrate a clear correlation between ligand binding similarity scores and FunSimMat functional similarity scores. Specifically, only the set of albumins and amylases gave significant ligand binding similarity scores. Unfortunately, the ligand binding profiles were unable to differentiate between the two α and β amylase families. A further refinement of the functional annotation would require a second screening step using a focused library to differentiate these functional classes. In the case of the amylases, this would involve screening the proteins with a carbohydrate library, where a subset of the compounds would selectively bind to the α- or β-amylase proteins. Alternatively, a larger chemical library with an increase in the number of compounds per representative class, such as additional carbohydrates, would be expected to enhance the functional resolution of the technique.

While our methodology has been shown to be effective with the proteins examined, limitations may be encountered with other classes of proteins. An NMR ligand affinity screen using intrinsically disordered proteins would be unproductive unless ligand binding induced a folded state or a binding partner that stabilized a folded state was present. Of course, the presence of a binding partner would complicate the data analysis; does the ligand bind the complex or binding partner instead of the targeted protein? Membrane proteins would be equally challenging, requiring methods to prepare adequate quantities of the protein for the NMR screen while requiring lipid bicelles, micelles or detergents to stabilize the protein. A similar data analysis problem would arise. Do the ligands interact with the lipid bicelles, micelles or detergents instead of or in addition to the protein target? Furthermore, does the NMR sample preparation procedure affect the solubility or aggregation state of compounds in the library? Finally, proteins that bind a very limited number of compounds from our library (<5) would result in a ligand binding profile that would only yield insignificant E-values ($>1 \times 10^{-9}$). Despite these potential limitations and challenges, ligand binding profiles are expected to be broadly applicable to the majority of the proteome.

CONCLUSION

The success of whole-genome sequencing has generated an enormous data set of functionally uncharacterized proteins.

Sequence and structure homology are routinely used to leverage functional annotations, but >30% of the proteome lack a sequence or structure similarity to proteins of known function. Alternatively, detailed experimental analysis may require upward of a decade of effort to characterize a single protein. Instead, we describe the use of high-throughput NMR ligand affinity screens to infer a biological function through a similarity in ligand binding profiles. A diverse chemical library is used to map the physiochemical properties of a protein's active-site, where the identity of the ligands that bind a protein provides information about the biological activity of the protein. A modification to the E-value developed by Karlin and Altschul allows for a similarity between ligand-binding profiles to be measured, where an E-value $\leq 1 \times 10^{-5}$ suggests functional similarity. We demonstrated that the preponderance of binding ligands identified from 19 NMR ligand affinity screens were uniquely associated with each functional class and were shown to correlate with the protein's function based on GO terms (Figure 2b and Table 1).

■ ASSOCIATED CONTENT

Supporting Information

Three supplementary tables that include the complete summary of the NMR ligand affinity screen for all 19 proteins, all pairwise functional similarity scores, and all pairwise ligand binding profile similarity scores. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Department of Chemistry, 722 Hamilton Hall, University of Nebraska-Lincoln, Lincoln, NE 68588-0304. Tel: (402) 472-3039. Fax: (402) 472-9402. E-mail: rpowers3@unl.edu.

■ ACKNOWLEDGMENT

We would like to thank Dr. Roberto De Guzman from the University of Kansas for supplying the PrgI sample, and Dr. Gaetano Montelione, Director of the Northeast Structural Genomics Center (NESG, www.nesg.org) for supplying the DGKA, PA1324, SAV1430, STM1790, UFC1, YjbR, YkfF, YkvR and YtfP protein samples for the NMR ligand affinity screen. We would like to thank Dr. Greg Somerville from the University of Nebraska-Lincoln for providing the *S. aureus* nuclease expression vector and Drs. Byron Taylor and Rathnam Chaguturu from the University of Kansas for providing numerous high-throughput screening data sets. This work was supported by the National Institute of Allergy and Infectious Diseases Nebraska [R21AI081154], and by the Tobacco Settlement Biomedical Research Development Fund. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Allergy and Infectious Diseases. Research was performed in facilities renovated with support from the National Institutes of Health [RR015468-01].

■ REFERENCES

(1) Janitz, M. Assigning functions to genes--the main challenge of the post-genomics era. *Rev. Physiol. Biochem. Pharmacol.* **2007**, *159*, 115–29.

(2) Rentzsch, R.; Orengo, C. A. Protein function prediction--the power of multiplicity. *Trends Biotechnol.* **2009**, *27* (4), 210–9.

(3) Rost, B.; Liu, J.; Nair, R.; Wrzeszczynski, K. O.; Ofran, Y. Automatic prediction of protein function. *Cell. Mol. Life Sci.* **2003**, *60* (12), 2637–50.

(4) Mercier, K. A.; Baran, M.; Ramanathan, V.; Revesz, P.; Xiao, R.; Montelione, G. T.; Powers, R. FAST-NMR: functional annotation screening technology using NMR spectroscopy. *J. Am. Chem. Soc.* **2006**, *128* (47), 15292–9.

(5) Kinnings, S. L.; Jackson, R. M. Binding site similarity analysis for the functional classification of the protein kinase family. *J. Chem. Inf. Model.* **2009**, *49* (2), 318–29.

(6) Ferre, F.; Ausiello, G.; Zanzoni, A.; Helmer-Citterich, M. Functional annotation by identification of local surface similarities: A novel tool for structural genomics. *BMC Bioinf.* **2005**, *6*, 194.

(7) Skolnick, J.; Brylinski, M. FINDSITE: a combined evolution/structure-based approach to protein function prediction. *Briefings Bioinf.* **2009**, *10* (4), 378–91.

(8) Hajduk, P. J.; Huth, J. R.; Fesik, S. W. Druggability Indices for Protein Targets Derived from NMR-Based Screening Data. *J. Med. Chem.* **2005**, *48* (7), 2518–25.

(9) English, A. C.; Groom, C. R.; Hubbard, R. E. Experimental and computational mapping of the binding surface of a crystalline protein. *Protein Eng.* **2001**, *14* (1), 47–59.

(10) Allen, K. N.; Bellamacina, C. R.; Ding, X.; Jeffery, C. J.; Mattos, C.; Petsko, G. A.; Ringe, D. An Experimental Approach to Mapping the Binding Surfaces of Crystalline Proteins. *J. Phys. Chem.* **1996**, *100* (7), 2605–11.

(11) Liepinsh, E.; Otting, G. Organic solvents identify specific ligand-binding sites on protein surfaces. *Nat. Biotechnol.* **1997**, *15* (3), 264–268.

(12) Kubinyi, H. Structure-based design of enzyme inhibitors and receptor ligands. *Curr. Opin. Drug Discovery Dev.* **1998**, *1* (1), 4–15.

(13) Gubernator, K.; Boehm, H. J. Examples of active areas of structure based-design. *Methods Princ. Med. Chem.* **1998**, *6*, 15.

(14) Powers, R.; Copeland, J.; Mercier, K. Application of FAST-NMR in Drug Discovery. *Drug Discovery Today* **2008**, *13* (3−4), 172–9.

(15) Mercier, K. A.; Cort, J. R.; Kennedy, M. A.; Lockert, E. E.; Ni, S.; Shortridge, M. D.; Powers, R. Structure and function of Pseudomonas aeruginosa protein PA1324 (21−170). *Protein Sci.* **2009**, *18* (3), 606–18.

(16) Stark, J. L.; Mercier, K. A.; Mueller, G. A.; Acton, T. B.; Xiao, R.; Montelione, G. T.; Powers, R. Solution structure and function of YndB, an AHSA1 protein from Bacillus subtilis. *Proteins Struct. Funct. Bioinf.* **2010**, *78* (16), 3328–40.

(17) Shortridge, M. D.; Powers, R. Structural and functional similarity between the bacterial type III secretion system needle protein PrgI and the eukaryotic apoptosis Bcl-2 proteins. *PLoS One* **2009**, *4* (10), e7442.

(18) Keiser, M. J.; Roth, B. L.; Armburster, B. N.; Ernsberger, P.; Irwin, J. J.; Shoichet, B. K. Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* **2007**, *25* (2), 197–206.

(19) Xie, L.; Li, J.; Xie, L.; Bourne, P. E. Drug discovery using chemical systems biology: identification of the protein-ligand binding network to explain the side effects of CETP inhibitors. *PLoS Comput. Biol.* **2009**, *5* (5), e1000387.

(20) Dessailly, B. H.; Lensink, M. F.; Orengo, C. A.; Wodak, S. J. LigASite--a database of biologically relevant binding sites in proteins with known apo-structures. *Nucleic Acids Res.* **2008**, No. Database issue, D667–73.

(21) Paolini, G. V.; Shapland, R. H. B.; van, H. W. P.; Mason, J. S.; Hopkins, A. L. Global mapping of pharmacological space. *Nat. Biotechnol.* **2006**, *24* (7), 805–15.

(22) Vieth, M.; Sutherland, J. J.; Robertson, D. H.; Campbell, R. M. Kinomics: Characterizing the therapeutically validated kinase space. *Drug Discovery Today* **2005**, *10* (12), 839–46.

(23) Bamborough, P.; Drewry, D.; Harper, G.; Smith, G. K.; Schneider, K. Assessment of Chemical Coverage of Kinome Space and Its Implications for Kinase Drug Discovery. *J. Med. Chem.* **2008**, *51* (24), 7898–914.

(24) Weinstein, J. N.; Myers, T. G.; O'Connor, P. M.; Friend, S. H.; Fornace, A. J., Jr.; Kohn, K. W.; Fojo, T.; Bates, S. E.; Rubinstein, L. V.;

Anderson, N. L.; Buolamwini, J. K.; van, O. W. W.; Monks, A. P.; Scudiero, D. A.; Sausville, E. A.; Zaharevitz, D. W.; Bunow, B.; Viswanadhan, V. N.; Johnson, G. S.; Wittes, R. E.; Paull, K. D. An information-intensive approach to the molecular pharmacology of cancer. *Science* **1997**, *275* (5298), 343–9.

(25) Mercier, K. A.; Shortridge, M. D.; Powers, R. A multi-step NMR screen for the identification and evaluation of chemical leads for drug discovery. *Comb. Chem. High Throughput Screen* **2009**, *12* (3), 285–95.

(26) Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; Harris, M. A.; Hill, D. P.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S.; Matese, J. C.; Richardson, J. E.; Ringwald, M.; Rubin, G. M.; Sherlock, G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **2000**, *25* (1), 25–9.

(27) Mercier, K. A.; Germer, K.; Powers, R. Design and characterization of a functional library for NMR screening against novel protein targets. *Comb. Chem. High Throughput Screen* **2006**, *9* (7), 515–34.

(28) Hwang, T.-L.; Shaka, A. J. Water suppression that works. Excitation sculpting using arbitrary waveforms and pulsed field gradients. *J. Magn. Reson., Ser. A* **1995**, *112* (2), 275–9.

(29) Tetko, I. V.; Gasteiger, J.; Todeschini, R.; Mauri, A.; Livingstone, D.; Ertl, P.; Palyulin, V. A.; Radchenko, E. V.; Zefirov, N. S.; Makarenko, A. S.; Tanchuk, V. Y.; Prokopenko, V. V. Virtual computational chemistry laboratory--design and description. *J. Comput. Aided Mol. Des.* **2005**, *19* (6), 453–63.

(30) Shortridge, M. D.; Hage, D. S.; Harbison, G. S.; Powers, R. Estimating protein-ligand binding affinity using high-throughput screening by NMR. *J. Comb. Chem.* **2008**, *10* (6), 948–58.

(31) Breuker, K. The study of protein-ligand interactions by mass spectrometry-a personal view. *Int. J. Mass Spectrom.* **2004**, *239* (1), 33–41.

(32) Schlicker, A.; Albrecht, M. FunSimMat: a comprehensive functional similarity database. *Nucleic Acids Res.* **2008**, No. Database issue, D434–9.

(33) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23* (1−3), 3–25.

(34) Hajduk, P. J.; Greer, J. A decade of fragment-based drug design: strategic advances and lessons learned. *Nat. Rev. Drug Discovery* **2007**, *6* (3), 211–9.

(35) Karlin, S.; Altschul, S. F. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. U.S.A.* **1990**, *87* (6), 2264–8.

(36) Dolle, R. E. Historical overview of chemical library design. *Methods Mol. Biol.* **2010**, *685* (1), 3–25.

(37) Baldi, P.; Nasr, R. When is Chemical Similarity Significant? The Statistical Distribution of Chemical Similarity Scores and Its Extreme Values. *J. Chem. Inf. Model.* **2010**, *50* (7), 1205–22.

(38) Chen, I. J.; Hubbard, R. E. Lessons for fragment library design: analysis of output from multiple screening campaigns. *J. Comput.-Aided Mol. Des.* **2009**, *23* (8), 603–20.